Methods for Estimating Hidden Structure and Network Transitions in Genomics

A DISSERTATION PRESENTED
BY
DANIEL JEN SCHLAUCH
TO
THE DEPARTMENT OF BIOSTATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOSTATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS

APRIL 2017

©2017 – Daniel Jen Schlauch
all rights reserved.

Thesis advisor: Professor John Quackenbush Daniel Jen Schlauch

Methods for Estimating Hidden Structure and Network Transitions in Genomics

Abstract

The explosion of data arising from advances in high throughput sequencing has allowed scientists to study genomics in far greater detail. However, this high resolution picture of cells often makes it difficult to see the higher level functions, features and structure in the biology that lead to phenotypic outcomes. With these advancements come new challenges in genomic analysis. The biological complexity of sample is not fully captured by a single snap shot of the sample's genome or gene expression profile. It would be simple if all phenotypes could be explained with a simple additive model consisting of the measured gene expression or variants in an organism. However, it has become increasingly clear that the drivers of fundamental biological processes involve complex systems of interactions between numerous genomic components. These interactions are often described using network models, such as gene regulatory models. Understanding how these networks should be constructed is an active field involving many types of interpretations of edges and methods for inference. However, relatively few methods exist for identifying network changes between phenotypic states or experimental conditions. In chapter 2, we address this problem by proposing a method for

estimating transcription factors that characterize changes in these networks.

In network inference we are seeking to find genes with related expression patterns. However, unknown and unmeasured structure is known to exists in genomic studies and has the potential to bias associations by confounding the relationship between features and phenotypes. Substantial work has already been published which attempts to identify and remove the impact of this unwanted variation, but subtle effects will continue to remain. Left uncorrected, this hidden structure may cause spurious correlations in genome wide association studies and gene expression analyses. As the field advances, new applications and new technologies call for methods which improve on existing tools and utilize all available information to provide better estimates. Two of the chapters presented here deal with problems of identifying and removing unwanted structure in data. Chapter 4 addresses a previously undocumented problem in the topic of gene coexpression by identifying and controlling for batch effect at the covariance level. In other words, we address confounding where coexpression is induced in a subset of samples by those samples' membership in a batch. We propose a method to correct the coexpression matrix that recognizes the modular nature of gene expression using a regression model for the eigenvectors of the expression correlation. Chapter 3 presents a method for identifying heterogeneity in genome studies aimed at high-throughput DNA-sequencing. This approach exploits the increased informativeness of low frequency variants to provide a higher resolution picture of population structure by more precisely measuring genetic similarity.

Thesis advisor: Professor John Quackenbush Daniel Jen Schlauch

Contents

Acknowledgments

I WISH to thank my dissertation advisor, Professor John Quackenbush, whose leadership, support, wisdom and perspective sets the gold standard for any student fortunate enough to mentored by him. He was always available to patiently guide me through my research and continuously replenished my passion for science along the way with his own excitement. He inspired me in the fight against cancer (literally and figuratively) and served as a role model for perseverance, ambition and attitude in the long road to the PhD.

I am grateful to my committee members, Kimberly Glass and Christoph Lange, whose patience and expertise were invaluable in my progression through the program. Both spent countless hours with me discussing, planning, guiding and editing. These chapters could not have been written without them.

I am incredibly grateful for the network of students, faculty and staff in the Biostatistics department who all contributed positively to the creation of a community that fosters growth and challenges us all to succeed. Additionally, I want to thank the postdocs and grad students of the Quackenbush lab – Joe Paulson, John Platig, Marieke Kuijjer, Joe Barry, Maude Fagny, Camila Lopes-Ramos, Megha Padi, Joey Chen, and Heather Selby for making the group a warm, welcoming, exciting, and creative environment that I looked forward to working in every day.

I want to thank my amazing family and awesome friends, in particular my Dad, my brother Mike, my sister Amy for their love and support through these last five years. My wife, Jen, who stood with me throughout the years and is my motivation to persevere. Also, my dog Tails and my cat Willow, who provided critical emotional support. And finally, my Mom, who continues to remind me everyday to strive to be the best version of myself that I can be.

Author List

The following authors contributed to Chapter 1:

Daniel Schlauch[1,2,3], Kimberly Glass[3,4], Craig P. Hersh[3,4,5], Edwin K. Silverman[3,4,5], and John Quackenbush[1,2,4]

The following authors contributed to Chapter 2:

Daniel Schlauch[1,2,3], Heide Fier[2,6], and Christoph Lange[2,7]

The following authors contributed to Chapter 3:

Daniel Schlauch[1,2,3], Joseph N. Paulson[1,2], Kimberly Glass[3,4], John Quackenbush[1,2,4]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, [2]Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA 02115, [3]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA 02115, [4]Department of Medicine, Harvard Medical School, Boston, MA 02115, [5]Pulmonary and Critical Care Division, Brigham and Women's Hospital and Harvard Medical School, Boston, USA, [6]Institute of Genomic Mathematics, University of Bonn, Bonn, Germany,[7]Channing Division of Network Medicine, Brigham and Women's Hospital, 181 Longwood Ave; CL-3; Boston, MA 02115

Listing of figures

Listing of tables

1

Introduction

The development of high-throughput technologies over the last two decades has brought significant promise towards understanding molecular biology and has shed light on the genomic involvement in the progression of human disease. Microarray and sequencing technologies have allowed us to interrogate many different types of biological problems at the molecular level, including the study of the genome, the transcriptome, the epigenome and other 'omics and dramatically reduced cost. For example, RNA-Seq is commonly used to measure the abundance of RNA (often selected

for mRNA) in a biological sample with the hope that relative quantities of RNA mapping to particular genes will help explain proteomic, cellular and phenotypic differences we observe at a higher level. But, as this ability to collect data has increased, so too have the statistical, biological and computational challenges that accompany questions about how healthy disease transitions to disease. It is clear that most phenotypic differentiation is not attributable to single units (genes or variants, for example) responsible for high level function. Instead, we observe that cell states are more adequately described with models that include numerous interacting features.

We are interested in hidden structures within the data that can tell us more about biological systems than the sum of individual components of the data. However, this underlying framework may represent many types of hidden systems, depending on the particular area of study and data. In some cases, this structure is important to understanding the mechanisms which drive disease, such as when an important biological process is disrupted in disease cells. But other times that structure arises from other variables that may confound analyses, such as the presence of ancestral heterogeneity in genetic studies. It is critical to find and address these artifacts where possible, as failing to do so leads to inflated rates of false positives.

Two such examples of unwanted structure in the data are seen with batch effect in gene expression and population stratification in statistical genetics. In each case, heterogeneity of the samples creates unwanted variation in the data. These effects have been widely studied and are known to produce spurious results in analyses such as

Genome Wide Association Studies (GWAS) and differential gene expression analyses. Many methods have been proposed to address these issues. Some of these approaches require the advanced knowledge of the sample variables which cause the underlying structure (such as batch effect), but others require the estimation of the sources of variation in the data. If the source of unwanted variation is unknown, it is common to estimate it by first estimating a similarity matrix across samples. This matrix can be used with an eigendecomposition or with a linear mixed model to control the type I error. It's clear that the efficacy of these methods is directly impacted by our ability to infer genetic similarity. With the increased use of DNA sequencing relative to DNA microarrays, we have increased resolution to infer genetic similarity. Furthermore, it has been shown that the lower frequency variants, visible only for sequencing studies, are the most informative of ancestry, owing to their more recent average emergence in human evolution. Chapter 3 in this dissertation proposed a new method in this field which exploits this new information to generate a higher resolution picture of ancestry. This tool has wide utility in identifying subtle population structure and cryptic relatedness in studies, particularly those involving purportedly homogeneous populations, but in fact exhibit subtle structure.

In the case of gene expression studies, we also see the presence of unwanted structure in the form of batch effect. This problem is typically motivated by differential gene expression analyses where we are most interested in determining genes or gene sets which show relative changes in mRNA abundances across phenotypic groups.

Methods such as ComBat and Surrogate Variable Analysis have been demonstrated to be very effective in this context and are widely used. More recently, scientists have become interested in developing gene networks which focus on gene coexpression rather than gene expression. In these analyses, it is not the relative abundance of a gene that we are concerned with - it is the manner in which that gene's expression pattern matches others that provide clues to its biological function. Significant work has been undertaken describing cellular states with gene networks. Simply put, researchers are interested in uncovering the manner in which genes are functionally connected. There are many ways to describe gene relationships, but often, if these models describe direct regulatory function we refer to them as Gene Regulatory Networks (GRN) and if they more generally imply "guilt-by-association" coexpression in an undirected graph, we call them Gene Coexpression Networks (GCN). This has lead to a growth in our understanding of biological function along with an appreciation for the complexity of molecular pathways and the biological processes that accompany them. However, the batch correction tools which are in wide use today make corrections at the individual gene level, which does not necessarily adjust for confounding by coexpression. Chapter 4 describes this problem and presents an approach that produces a model for the coexpression matrix that can be used to control for batch effect and other confounding covariates. This tool has applications for any of the numerous methods which utilize a coexpression matrix in the analysis, which is common in gene network inference.

Gene regulatory network inference is important for understanding how transcription factors influence downstream genes, but many studies involve cases and controls and are designed to uncover the molecular mechanisms which separate the two. These investigations are less interested in the topology overall networks, but rather choose to focus the interactions that differ between states. In this context, our goals can be divided into two parts, (1) the construction of gene regulatory networks and (2) the analysis of the structural changes between those networks. Due to the complexity of the underlying networks and the high dimensionality of typical datasets, these challenges remain open problems. In this dissertation we the explore novel methods developed for gaining insight into network transformations between cases and controls in a complex disease. Biological states are characterized by distinct patterns of gene expression that reflect each phenotype's active cellular processes. Driving these phenotypes are GRNs in which transcriptions factors control when and to what degree individual genes are expressed. Phenotypic transitions, such as those that occur when disease arises from healthy tissue, are associated with changes in these networks. In this context, we are less interested in inferring the general biological landscape, but are more interested in interrogating the transcription factor-gene relationships that change from one phenotype to another. While many methods exist for network inference, few approaches are designed for evaluating differential gene regulatory networks. A simple approach involves simply finding the difference between two inferred networks. In Chapter 2, we present a new approach to understanding these transi-

tions. MONSTER models phenotypic-specific regulatory networks and then estimates a "transition matrix" that converts one state to another. By examining the properties of the transition matrix, we can gain insight into regulatory changes associated with phenotypic state transition.

"No models are true. When you construct a model you leave out all the details which you, with the knowledge at your disposal, consider inessential. Models should not be true, but it is important that they are applicable. This also means that a model is never accepted finally, only on trial."

-Georg Rasch

2

Estimating Drivers of Cell State Transitions Using Gene Regulatory Network Models

Specific cellular states are often associated with distinct gene expression patterns. These states are plastic, changing during development, or in the transition from health to disease. One relatively simple extension of this concept is to recognize that

we can classify different cell-types by their active gene regulatory networks and that, consequently, transitions between cellular states can be modeled by changes in these underlying regulatory networks. Here we describe **MONSTER**, **MO**deling **N**etwork **S**tate **T**ransitions from **E**xpression and **R**egulatory data, a regression-based method for inferring transcription factor drivers of cell state conditions at the gene regulatory network level. As a demonstration, we apply MONSTER to four different studies of chronic obstructive pulmonary disease to identify transcription factors that alter the network structure as the cell state progresses toward the disease-state. Our results demonstrate that MONSTER can find strong regulatory signals that persist across studies and tissues of the same disease and that are not detectable using conventional analysis methods based on differential expression. An R package implementing MONSTER is available at github.com/QuackenbushLab/MONSTER.

2.1 Introduction

Cell state phenotypic transitions, such as those that occur during development, or as healthy tissue transforms into a disease phenotype, are fundamental processes that operate within biological systems. Understanding what drives these transitions, and modeling the processes, is one of the great open challenges in modern biology. One way to conceptualize the state transition problem is to imagine that each phenotype has its own characteristic gene regulatory network, and that there are a set of processes that are either activated or inactivated to transform the network in the initial

state into one that characterizes the final state. Identifying those changes could, in principle, help us to understand not only the processes that drive the state change, but also how one might intervene to either promote or inhibit such a transition.

Each distinct cell state consists of a set of characteristic processes, some of which are shared across many cell-states ("housekeeping" functions) and others which are unique to that particular state. These processes are controlled by gene regulatory networks in which transcription factors (and other regulators) moderate the transcription of individual genes whose expression levels, in turn, characterize the state. One can represent these regulatory processes as a directed network graph, in which transcription factors and genes are nodes in the network, and edges represent the regulatory interactions between transcription factors and their target genes. A compact representation of such a network, with interactions between m transcription factors and p target genes, is as a binary $p \times m$ "adjacency matrix". In this matrix, a value of 1 represents an active interaction between a transcription factor and a potential target, and 0 represents the lack of a regulatory interaction.

When considering networks, a cell state transition is one that transforms the initial state network to the final state network, adding and deleting edges as appropriate. Using the adjacency matrix formalism, one can think of this as a problem in linear algebra in which we attempt to find an $m \times m$ "transition matrix" $\mathbf{T}$, subject to a set of constraints, that approximates the conversion of the initial network's adjacency

matrix $\mathbf{A}$ into the final network's adjacency matrix $\mathbf{B}$, or

$$\mathbf{B} = \mathbf{AT} \tag{2.1}$$

In this model, the diagonal elements of $\mathbf{T}$ map network edges to themselves. The drivers of the transition are those off-diagonal elements that change the configuration of the network between states.

While this framework, as depicted in Figure 2.1, is intuitive, it is a bit simplistic in that we have cast the initial and final states as discrete. However, the model can be generalized by recognizing that any phenotype we analyze consists of a collection of individuals, all of whom have a slightly different manifestation of the state, and therefore a slightly different active gene regulatory network. Practically, what that means is that for each state, rather than having a network model with edges that are either "on" or "off," a phenotype should be represented by a network in which each edge has a weight that represents an estimation of its presence across the population. In other words, the initial and final state adjacency matrices are not comprised of 1's and 0's, but of continuous variables that estimate population-level regulatory network edge-weights. Consequently, the problem of calculating the transition matrix is generalized to solving $\mathbf{B} = \mathbf{AT} + \mathbf{E}$, where $\mathbf{E}$ is an $p \times m$ error matrix. In this expanded framework, modeling the cell state transition remains equivalent to estimating the appropriate transition matrix $\mathbf{T}$, and then identifying state transition drivers based on

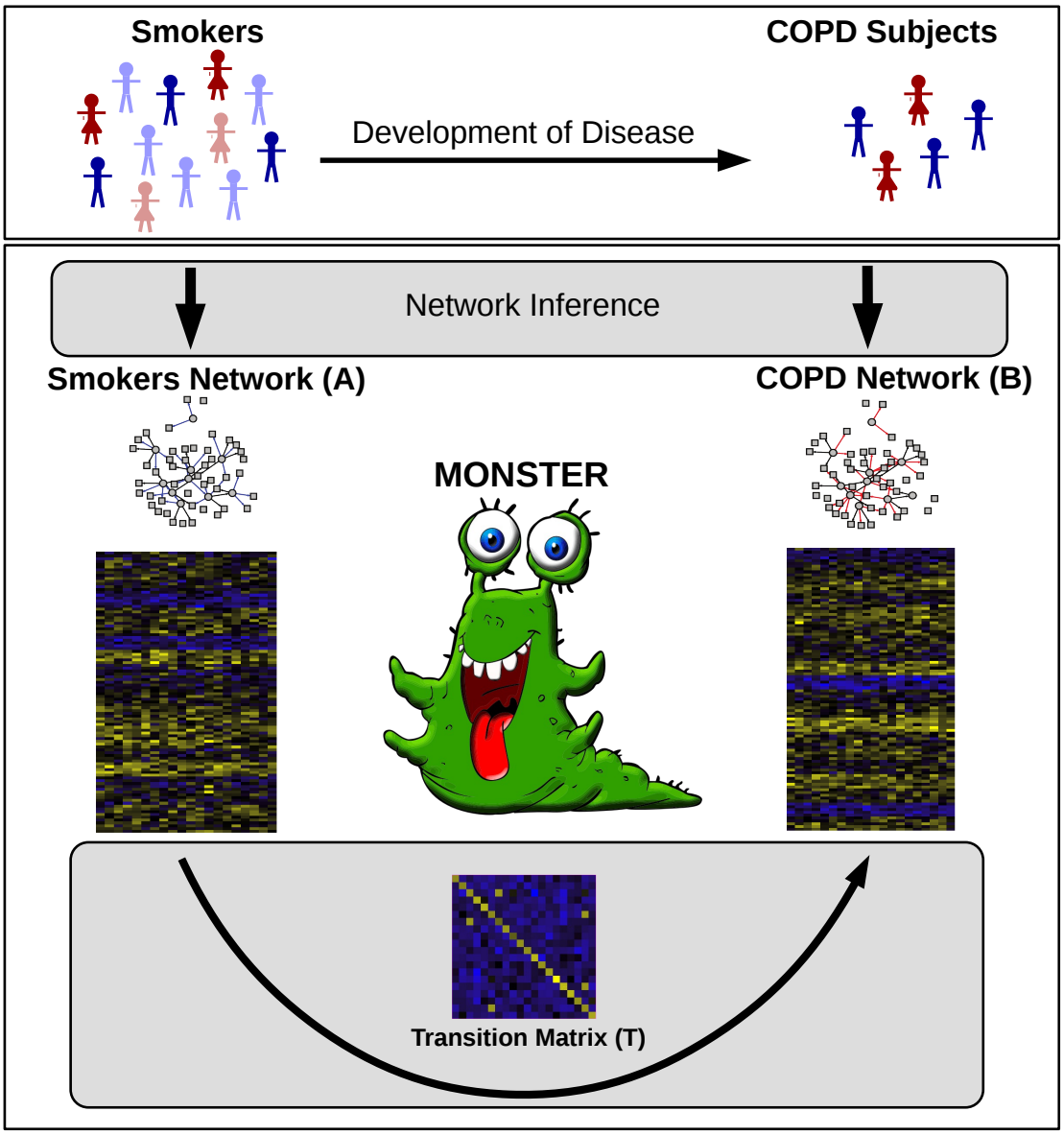


Figure 2.1: Overview of the MONSTER approach, as applied to the transition between smokers and those suffering from chronic obstructive pulmonary disease (COPD). MONSTER's approach seeks to find the $TF \times TF$ transition matrix that best characterizes the state change in network structure between the initial and final biological conditions. Subjects are first divided into two groups based on whether they have COPD or are smokers that have not yet developed clinical COPD. Network inference is then performed separately on each group, yielding a bipartite adjacency matrix connecting transcription factors to genes. Finally, a transition matrix is computed which characterizes the conversion from the consensus Smokers Network to the COPD Network.

features of that matrix.

2.2 Methods

Data for COPD Network Inference and Analysis

Sequence binding motifs

A regulatory network prior between transcription factors and target genes was created by using position weight matrices for 205 transcription factor motifs obtained from JASPAR 2014 (http://jaspar2014.genereg.net/),[74] and running Haystack[85] to scan the hg19 genome for occurrences of these motifs. Sequences were identified as hits for a transcription factor if they satisfied the significance threshold of $p < 10^{-5}$. We then used HOMER (http://homer.salk.edu/homer/ngs/index.html)[48] to identify transcription factor binding motifs that map to a window ranging from 750 base pairs downstream to 250 base pairs upstream of each gene's transcription start site under the assumption that transcription factors falling in this region may actively regulate expression of the gene.

ECLIPSE

Gene expression data from the ECLIPSE study (GSE54837)[102] was collected using blood samples from 226 subjects classified as non-smokers (6), smoker controls (84) or COPD (136). Blood samples from each individual were profiled using Affymetrix

Human Genome U133 Plus 2.0 microarrays. CEL data files from these assays were RMA-normalized[55] in R using the Bioconductor package 'affy'[41]. Array probes were collapsed to 19,765 Entrez-gene IDs using a custom CDF[23] and the 220 samples for COPD or smoker control subjects were retained for analysis. Finally, genes were associated with potential regulatory transcription factors using a motif scan (described above). 1,553 genes were not associated with any transcription factor and excluded from further analysis, leaving 17,342 genes that were used to construct network models.

COPDGene

Gene expression data from the COPDGene study (GSE42057)[4,95] was collected from blood samples obtained from 136 subjects classified as smoker controls (42) or COPD (94) and profiled on Affymetrix Human Genome U133 Plus 2.0 microarrays. Similar to the ECLIPSE data, CEL data files from these microarray assays were RMA-normalized using the 'affy' package and array probes were collapsed to Entrez-gene IDs using a custom CDF[23], yielding 18,960 genes. After removal of genes that did not match with our motif scan, the COPDGene data contained 17,253 genes.

LGRC

Gene expression data from 581 lung tissue samples in the LGRC (GSE47460)[43] was profiled using two array platforms: Agilent-014850 Whole Human Genome Microarray

4x44K G4112F and Agilent-028004 SurePrint G3 Human GE 8x60K arrays. LIMMA was used to background correct and normalize gene expression across samples within each of these two platforms. Genes that were represented by more than one probe were then removed and the expression data was merged between the two array platforms by matching probes that represented the same gene, leaving 17,573 genes. Next, batch effect due to the array platform was addressed by running ComBat[56]. Genes not present in our motif scan were then removed, yielding 14,721 genes. After normalization we filtered the samples included in the LGRC data-set by removing those that corresponded to subjects that (1) were not designated as either a COPD case or control (mostly subjects with Interstitial Lung Disease), (2) had a diagnosis of COPD, but spirometric measures in the normal range, (3) had been identified as non-Caucasian, (4) had been labeled as a former smoker, but had zero or unknown pack years, (5) had high pre-bronchodilator FEV1/FVC ratios, or (6) had been taken as a biological replicate of another sample which was included. After removal of those samples we were left with 164 COPD cases and 64 controls for which we had gene expression data.

LTCDNM

Gene expression data from the LTCDNM (GSE76925)[92] was collected using HumanHT-12 BeadChips. Quality control was performed using quantile, signal-to-noise, correlation matrix, MA, and principal component analysis (PCA) plots using R statistical

software (v 3.2.0) to identify outliers and samples with questionable or low-quality levels, distributions, or associations. This process yielded 151 samples for analysis, including 115 subjects classified as either diagnosed with COPD (87) or as a smoker control (28). After filtering for low variance and percentage of high detection p-values, 32,831 probes representing 20,794 genes were retained. The R package lumi[27] was then used for background correction, log2 transformation and quantile normalization. Finally, we collapsed probes to gene symbols based on maximum gene expression and removed genes that were not matched with our motif scan, yielding 14,273 genes.

TFs included in analysis

For each study, we identified transcription factors for which we had gene expression data, removing those transcription factors that lacked expression values. This mapping and filtering left 164 transcription factors in ECLIPSE and COPDGene, 148 in LGRC, and 145 in LTCDNM. MONSTER was run separately on each of these studies. Comparisons of differential transcription factor involvement across studies were performed using the 143 transcription factors that were common to all four studies.

MONSTER: MOdeling Network State Transitions from Expression and Regulatory data

The MONSTER algorithm models the regulatory transition between two cellular states in three steps: (1) Inferring state-specific gene regulatory networks, (2) mod-

eling the state transition matrix, and (3) computing the transcription factor involvement.

Inferring state-specific gene regulatory networks:

Before estimating the transition matrix, $\mathbf{T}$, we must first estimate a gene regulatory starting point for each state. While there have been many methods developed to infer such networks[50,45,46,28,13,77,97], we have found the bipartite framework used in PANDA[44] to have features that are particularly amenable to interpretation in the context of state transitions. PANDA begins by using genome-wide transcription factor binding data to postulate a network "prior", and then uses message-passing to integrate multiple data sources, including state-specific gene co-expression data.

Motivated by PANDA, we developed a highly computationally efficient, classification-based network inference method that uses common patterns between transcription factor targets and gene co-expression to estimate edges and to generate a bipartite gene regulatory network connecting transcription factors to their target genes.

This approach is based on the simple concept that genes affected by a common transcription factor are likely to exhibit correlated patterns of expression. To begin, we combine gene co-expression information with information about transcription factor targeting derived from sources such as ChIP-Seq or sets of known sequence binding motifs found in the vicinity of genes. we then calculate the direct evidence for a regulatory interaction between a transcription factor and gene, which we define as

the squared partial correlation between a given transcription factor's gene expression, g_i, and the gene's expression, g_j, conditional on all other transcription factors' gene expression:

$$\hat{d}_{i,j} = cor\left(g_i, g_j | \left\{g_k : k \neq i, k \in \mathbf{TF_j}\right\}\right)^2,$$

where g_i is the gene which encodes the transcription factor TF_i, g_j is any other gene in the genome, and $\mathbf{TF_j}$ is the set of gene indices corresponding to known transcription factors with binding site in the promoter region of g_j. The correlation is conditioned on the expression of all other potential regulators of g_j based on the transcription factor motifs associated with g_j.

Next, we fit a logistic regression model which estimates the probability of each gene, indexed j, being a motif target of a transcription factor, indexed i, based on the expression pattern across the n samples across p genes in each phenotypic class:

$$logit(P\left[\mathbf{M}_{i,j} = 1\right]) = \beta_{0,i} + \beta_{1,i} g_j^{(1)} + \cdots + \beta_{N,i} g_j^{(N)}$$

$$\hat{\theta}_{i,j} = \frac{e^{\hat{\beta}_{0,i} + \hat{\beta}_{1,i} g_j^{(1)} + \cdots + \hat{\beta}_{N,i} g_j^{(N)}}}{1 + e^{\hat{\beta}_{0,i} + \hat{\beta}_{1,i} g_j^{(1)} + \cdots + \hat{\beta}_{N,i} g_j^{(N)}}}$$

where the response $\mathbf{M}$ is a binary $p \times m$ matrix indicating the presence of a sequence motif for the i^{th} transcription factor in the vicinity of each of the j^{th} gene. And where $g_j^{(k)}$ represents the gene expression measured for sample k at gene j. Thus, the fitted probability $\hat{\theta}_{i,j}$ represents our estimated indirect evidence. Combining the scores for

the direct evidence, $\hat{d}_{i,j}$, and indirect evidence, $\hat{\theta}_{i,j}$, via weighted sum between each transcription factor-gene pair yields estimated edge-weights for the gene regulatory network (see below).

Applying this approach to gene expression data from two distinct phenotypes results in two $p \times m$ gene regulatory adjacency matrices, one for each phenotype. These matrices represent estimates of the targeting patterns of the m transcription factors onto the p genes. This network inference algorithm finds validated regulatory interactions in *Escherichia coli* and Yeast (*Saccharomyces cerevisiae*) data sets (see below).

Modeling the state transition matrix:

Once we have gene regulatory network estimates for each phenotype, we can formulate the problem of estimating the transition matrix in a regression framework in which we solve for the $m \times m$ matrix that best describes the transformation between phenotypes (2.1). More specifically, MONSTER predicts the change in edge-weights for a transcription factor, indexed i, in a network based on all of the edge-weights in the baseline phenotype network.

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

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

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

The regression is performed m times corresponding to each of the transcription factors in the data. In this sense, columns in the transition matrix can be loosely interpreted as the optimal linear combination of columns in the initial state adjacency matrix which predict the column in the final state adjacency matrix. (see below).

This framework allows for the natural extension of constraints such as $L1$ and/or $L2$ regularization. For the analysis we present in this manuscript, we use the normal equations and do not impose a penalty on the regression coefficients.

Computation of MONSTER's transition matrix

The hypothesis behind MONSTER is that different phenotypes are characterized by distinct regulatory networks and that transitions between networks are associated with large-scale changes in the regulatory structure of the network. Essentially, transcription factors gain or lose targets and in doing so, alter the structure of the

network from one phenotypic state to another. The task of identifying meaningful network transitions then becomes an evaluation of the relative refinement of edge weights.

Our analysis of validation data sets (shown below) indicates that the reconstructed networks are strongly driven by the structure of the motif prior, with small changes defining differences between phenotypes. Hence, in comparing networks between phenotypes, the problem becomes one of of understanding changes in edges that have relatively low signal and high noise. In other words, state transitions are characterized by a large number of individually unreliable edge weights.

Consider two adjacency matrices, $\mathbf{A}$ and $\mathbf{B}$, that represent two gene regulatory networks estimated from a case-control study. Each matrix has dimensions $(p \times m)$ representing the set of p genes targeted by m transcription factors. We seek a matrix, $\mathbf{T}$, such that

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

where $\mathbf{E}$ is our error matrix, which we want to minimize. Intuitively, we may frame this as a set of m independent regression problems, where m is the number of transcription factors and also the column rank of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{T}$, and $\mathbf{E}$. For a column in $\mathbf{B}$, $\mathbf{b}_i$, we note that a corresponding column in $\mathbf{T}$, τ_i, represents the ordinary least squares solution to

$$E\left[\mathbf{b}_i\right] = \tau_{i1}\mathbf{a}_{1i} + \tau_{i2}\mathbf{a}_{2i} + \cdots + \tau_{im}\mathbf{a}_{mi}$$

or alternatively expressed

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

where $E[\epsilon_{ij}] = 0$. This can be solved with normal equations,

$$\begin{aligned} \tau_i &= \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}_i \\ \mathbf{T} &= [\tau_1, \tau_2, \ldots, \tau_m] \end{aligned}$$

which produces the least squares estimate. In other words, the loss function $L(\mathbf{T}) = \sum_{gene=1}^{N} ||\mathbf{B}_{gene} - \mathbf{A}_{gene}\mathbf{T}||^2$ is minimized.

It is easy to see how this allows for a straightforward extension via the inclusion of a penalty term. For example, an L_1 regularization[108] can be used to create an identity penalty model matrix for each column regression such that only the k^{th} diagonal element is 0 and all other diagonals are 1. This gives unpenalized priority for the k^{th} regression coefficient in the k^{th} regression model:

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

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

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

Although not used in the analysis presented in the main text, an implementation of this extension is available in the R package MONSTER.

Computing the transcription factor involvement:

For a transition between two nearly identical states, we expect that the transition matrix would approximate the identity matrix. However, as initial and final states diverge, there should be increasing differences in their corresponding gene regulatory networks and, consequently, the transition matrix will also increasingly diverge from the identity matrix. In this model, the transcription factors that most significantly alter their regulatory targets will have the greatest "off-diagonal mass" in the transition matrix, meaning that they will have very different targets between states and so are likely to be involved in the state transition process. We define the "differential transcription factor involvement" (dTFI) as the magnitude of the off-diagonal mass associated with each transcription factor, or,

$$\hat{dTFI}_j = \frac{\sum_{i=1}^{m} I\left(i \neq j\right)\hat{\tau}_{i,j}^2}{\sum_{i=1}^{m}\hat{\tau}_{i,j}^2} \tag{2.2}$$

where, $\hat{\tau_{i,j}}$ is the value in of the element i^{th} row and j^{th} column in the transition matrix, corresponding to the i^{th} and j^{th} transcription factors . To estimate the significance of this statistic, we randomly permute sample labels $n = 400$ times across phenotypes.

Analyzing the Transition Matrix

The derivation described above illustrates a key feature of the MONSTER method. Specifically, that the transition matrix ($\mathbf{T}$) reduces the case-control network transformation from a set of $2 \times p \times m$ estimates to a set of $m \times m$ estimates that are more easily interpreted. We can think of a column, τ_i, on the matrix $\mathbf{T}$ as containing the linear combination of regulatory targets of TF_i in $\mathbf{A}$ that best approximates the regulatory targets of TF_i in $\mathbf{B}$. As one would expect, a large proportion of the matrix "mass" would be on the diagonal for those transcription factors which do not change regulatory behavior between case and control. It is therefore of interest to evaluate values off of the diagonal as indicative of a network transition.

There are many biological processes involved in gene regulation that may differ between phenotypic states, including RNA degradation, post-translational modification, protein-level interactions and epigenetic alterations. These all have the ability to impact transcription factor targeting without impacting the expression level of the transcription factor itself. Because our hypothesis is that changes in phenotype are associated with changes in regulatory networks, we want to identify those transcription

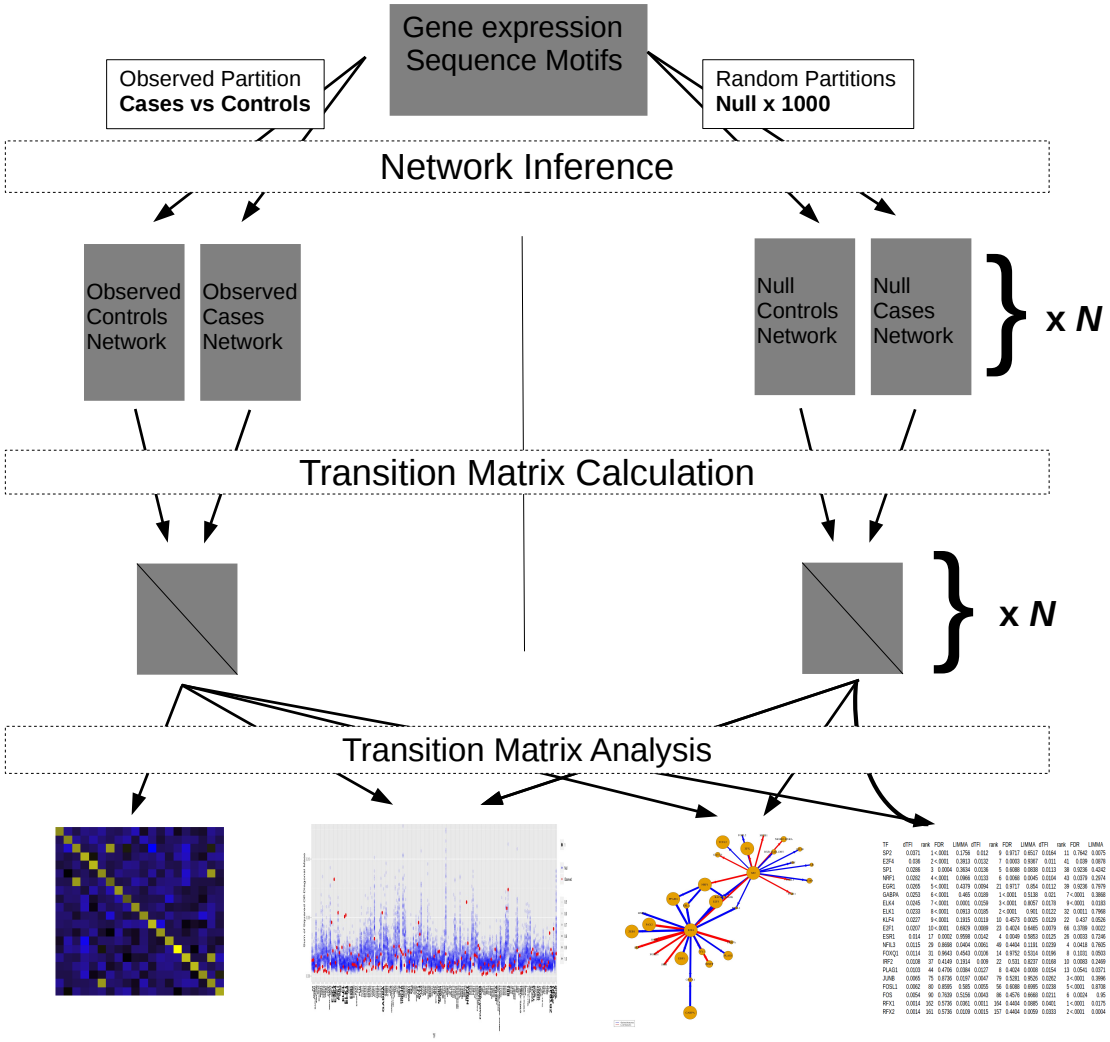


Figure 2.2: Overview of MONSTER analysis workflow. (1) Network inference is computed separately to subsets of the gene expression data including the case group, the control group and N permutations of the case and control labels. (2) The transition matrix is estimated between the cases and controls and each of the pairs of permuted "case" and "control" groups. (3) The transition matrix computed between the case and control group is interpreted within the context of the N matrices estimated for the permuted groups.

factors that have undergone significant overall changes in behavior between states. As a measure to quantify such changes, we define the differential Transcription Factor Involvement (dTFI),

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

The dTFI can be loosely interpreted as the proportion of transcription factor targeting that is gained from or lost to other available transcription factors as the state changes. It is a statistic on the interval $[0, 1]$ that can be used to identify transitions which are systematic, informative, and non-arbitrary in nature. In other words, the dTFI can capture edge weight signal for which there is an attributable regulatory pattern based on the inferred networks.

The distribution of the dTFI statistic under the null has a mean and standard deviation that depends to a large extent on the motif-based network prior structure. In particular, we find that both mean and standard deviation of the dTFI are higher for transcription factors that have fewer prior regulatory targets. From a statistical perspective, transcription factors with relatively more targets are able to generate more stable targeted expression patterns, which leads to more consistent estimates in "agreement". From a biological perspective, increased motif presence may indicate that transcription factors are more likely to be involved in "housekeeping" or tissue specific processes that are unlikely to change between cases and controls.

We address the dependence of the null distribution of the dTFI on the motif struc-

ture using the following resampling procedure (Figure 2.2):

0. Gene regulatory networks are reconstructed based on a prior regulatory structure and gene expression from case and control samples and the transition matrix and the dTFI values for each transcription factor are computed.
1. Gene expression samples are randomly assigned as case and control forming null-case and null-control groups with sizes reflecting the true case and control groups.
2. Gene regulatory networks are reconstructed for the null-case and null-control groups with the same prior regulatory structure.
3. The transition matrix algorithm is applied to the two null networks.
4. The dTFI is calculated for each transcription factor based on the computed null transition matrix.
5. Steps 1-4 are repeated n times.

For the analysis presented in the main text, we set $n = 400$. This procedure allows us to estimate a background distribution of dTFI values based on the underlying motif prior network structure and therefore test the significance of observed dTFI values between cases and controls.

2.3 Results

MONSTER finds significantly differentially involved transcription factors in COPD with strong concordance in independent data sets

As a demonstration of the power of MONSTER to identify driving factors in disease, we applied the method to case-control gene expression data sets from four independent Chronic Obstructive Pulmonary Disease (COPD) cohorts: Evaluation of COPD

Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE)[102,109] (2.3), COPDGene[95,4,84], Lung Genomics Research Consortium (LGRC)[43] and Lung Tissue from Channing Division of Network Medicine (LT-CDNM)[92]. The tissues assayed in ECLIPSE and COPDGene were whole blood and peripheral blood mononuclear cells (PBMCs), respectively, while homogenized lung tissue was sampled for LGRC and LT-CDNM.

As a baseline comparison metric, we evaluated the efficacy of applying commonly used network inference methods on these case-control studies. In analyzing phenotypic changes, networks are generally compared directly, with changes in the presence or weight of edges between key genes being of primary interest. It is therefore reasonable to assume that any reliable network results generated from a comparison of disease to controls will be reproducible in independent studies. We investigated whether this is the case for our four COPD data sets using three widely used network inference methods - Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE)[73], Context Likelihood of Relatedness (CLR)[31], and Weighted Gene Correlation Network Analysis (WGCNA)[115] - computing the difference in edge weights between cases and controls for each of the four studies. We found no meaningful correlation ($R^2 < .01$) of edge weight difference across any of the studies regardless of network inference method or tissue type (Figure 2.7). Edge weight differences, even when very large in one study, did not reproduce in other studies. This suggests that a simple direct comparison of edges between inferred networks is insufficient for extract-

ing reproducible drivers of network state transitions. This finding may be unsurprising given the difficulty in inferring individual edges in the presence of heterogeneous phenotypic states, technical and biological noise with a limited number of samples.

The lack of replication in edge-weight differences between independent data sets representing similar study designs indicates that we need to rethink how we evaluate network state transitions. MONSTER provides a unique approach for making that comparison. In each of the four COPD data sets, we used MONSTER to calculate the differential transcription factor involvement ($dTFI$, Equation 2.2) for each transcription factor and used permutation analysis to estimate their significance (Figure 2.3, Figures 2.8-2.9). We observed strongly significant ($p < 1e-15$) correlation in dTFI values for each pairwise combination of studies. In addition, out of the top 10 most differentially involved transcription factors in the ECLIPSE and COPDGene studies, we found 7 to be in common. Furthermore, three of these seven transcription factors (GABPA, ELK4, ELK1) also appeared as significant in the LGRC results with FDR<0.01 and each of the top five ECLIPSE results were among the top seven in the LT-CDNM results (Table 2.2, Figure 2.10). This agreement is quite striking considering that the there was almost no correlation in the edge-weight differences across these same studies when we tested the other methods. But it is exactly what we should expect—that the same method applied to independent studies of the same phenotypes should produce largely consistent results.

Many of the top dTFI transcription factors, especially those identified by MON-

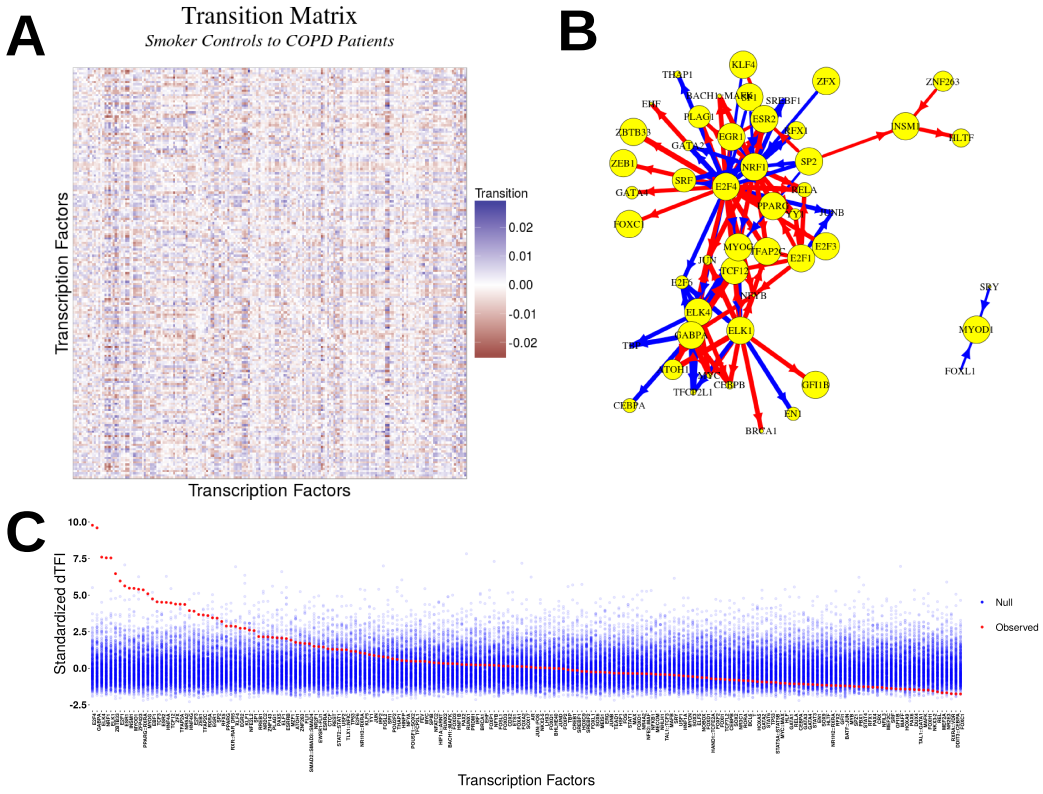


Figure 2.3: MONSTER analysis results in the ECLIPSE study. A Heatmap depicting the transition matrix calculated for smoker controls "transitioning" to COPD by applying MONSTER to ECLIPSE gene expression data. For the purposes of visualization, the magnitude of the diagonal is set to zero. **B** A network visualization of the 100 largest transitions identified based on the transition matrix in (A). Arrows indicate a change in edges from a transcription factor in the Smoker-Control network to resemble those of a transcription factor in the COPD network. Edge thickness represents the magnitude of the transition and node (TFs) sizes represent the dTFI for that TF. Blue edges represent a gain of targeting features and red represents the loss. **C** The dTFI score from MONSTER (red) and the background null distribution of dTFI values (blue) as estimated by 400 random sample permutations of the data.

STER across all four studies, are biologically plausible candidates to be involved in the etiology of COPD (Table 2.2, Figures 2.8,2.9,2.10). For example, E2F4 is a transcriptional repressor important in airway development[24] and studies have begun to demonstrate the relevance of developmental pathways in COPD pathogenesis[9].

Some of the greatest effect sizes across all four studies were found for SP1 and SP2. An additional member of the SP transcription factor family, SP3, has been shown to regulate HHIP, a known COPD susceptibility gene[117]. Both SP1 and SP2 form complexes with the E2F family[96,58] and may play a key role in the alteration of E2F4 targeting behavior. Furthermore, E2F4 has been found to form a complex with EGR-1 (a highly significant transcription factor in ECLIPSE and LT-CDNM) in response smoke exposure, which may lead to autophagy, apoptosis and subsequently to development of emphysema[14].

Mitochondrial mechanisms have also been associated with COPD progression[19]. Two of most highly significant transcription factors based on dTFI in ECLIPSE were NRF1 and GABPA (FDR<.001). Indeed, these TFs had highly significant dTFI (FDR<0.1) in all four studies. NRF1 regulates the expression of nuclear encoded mitochondrial proteins[47]. GABPA, also known as human nuclear respiratory factor-2 subunit alpha, may have a similar role in nuclear control of mitochondrial gene expression. Furthermore, GABPA interacts with SP1[40] providing evidence of a potentially shared regulatory mechanism with E2F4.

Overall, we found a strong correlation across studies in transcription factors iden-

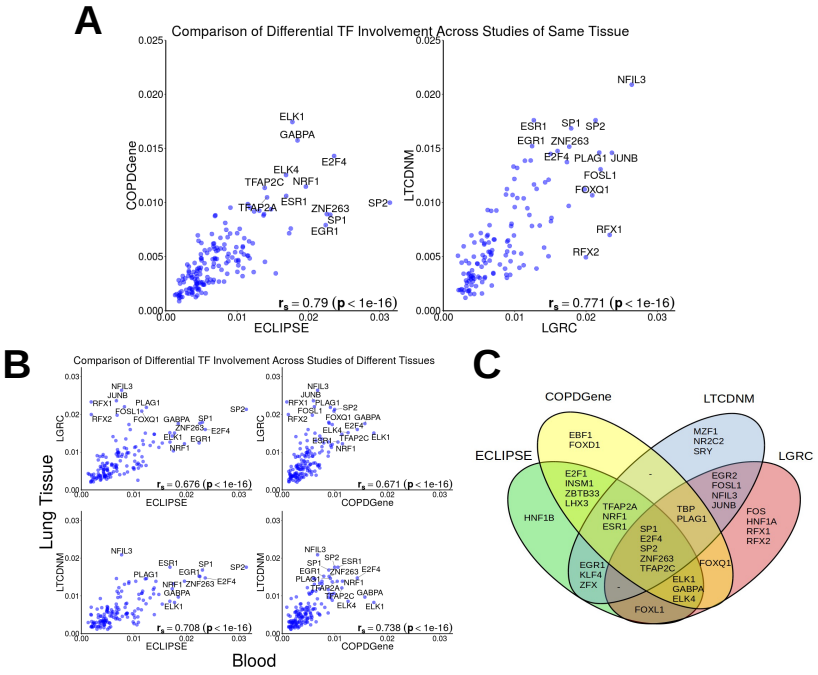


Figure 2.4: Strong reproducibility in top differential transcription factor involvement found in case-control COPD studies. ECLIPSE and COPDGene profiled gene expression in whole-blood and PBMC while the gene expression data in LGRC and LT-CDNM were assayed in lung tissue. **A** Results for studies with gene expression data obtained from the same-tissue. Both the blood based (left) and lung tissue studies (right) demonstrate very high Spearman correlation of differential involvement. **B** Despite using data from different sources we found agreement between studies of different tissues. **C** Venn diagram depicting the top 20 transcription factors found in each study. The union of all top 20 lists contains 36 transcription factors.

tified as significantly differentially involved (Figure 3A-3B). It is reassuring that we find the strongest agreement when comparing studies that assayed similar tissues. However the fact that we see similar dTFI signal across studies involving different tissue types is also notable as it suggests that the transition from smoker control to disease phenotype affects multiple tissues and supports the growing evidence for a role in immune response in COPD pathogenesis.

Gene regulatory networks, and results derived from their comparison, are notoriously difficult to replicate across studies[103]. The four studies we used each has unique aspects, including the choice of microarray platform, study demographics, location, time, and tissue. Nevertheless, MONSTER identified similar sets of transcription factors associated with the transition between cases and controls. This consistency in biologically-relevant transcription factors, associated with the transition from the control phenotype to disease, in four independent studies suggests that MONSTER can provide not only robust network models, but also can identify reliable differences between networks.

Despite the overall consistency, some transcription factors had variable $dTFI$ across studies. For example, using the LGRC dataset, we discovered a highly significant ($FDR < .0001$) differential targeting pattern involving the transcription factors RFX1 and RFX2 (Table 2.2). However, these same TFs were not identified as potential drivers of the control to COPD transition in either the ECLIPSE or COPDGene study. This difference is likely due the differences in tissue type as the RFX family

transcription factors are known to regulate ciliogenesis[17]. Cilia are critical for clearing mucous from the airways of healthy individuals, but disruption can lead to infection and potentially to chronic airflow obstruction[49,52,30].

The hypothesis behind MONSTER is that each phenotype has a unique gene regulatory network and that a change in phenotypic state is reflected in changes in transcription factor targeting. That hypothesis translates to an expectation that transcription factors driving change in phenotype will have the greatest $dTFI$ scores. One might expect that these "driving transcription" factors would also be differentially expressed. We compared $dTFI$ to differential expression (ECLIPSE Figure 4, other studies shown in Figure 2.11) and found that many of the transcription factors with high dTFI values were not differentially expressed. This suggests that there are other mechanisms, such as epigenetic modification of the genome or protein modifications, that alter the structure of the regulatory network by changing which genes are targeted by key transcription factors.

MONSTER recovers network edges in *in silico*, *Escherichia coli* and Yeast (*Saccharomyces cerevisiae*)

For its initial step, MONSTER uses gene expression together with a prior network structure to infer regulatory network edges. For method testing and validation of MONSTER's network estimates we used four data sets of increasing biological complexity: (1) *in silico*, (2) *Escherichia coli*, and (3) *Saccharomyces cerevisiae* (yeast)

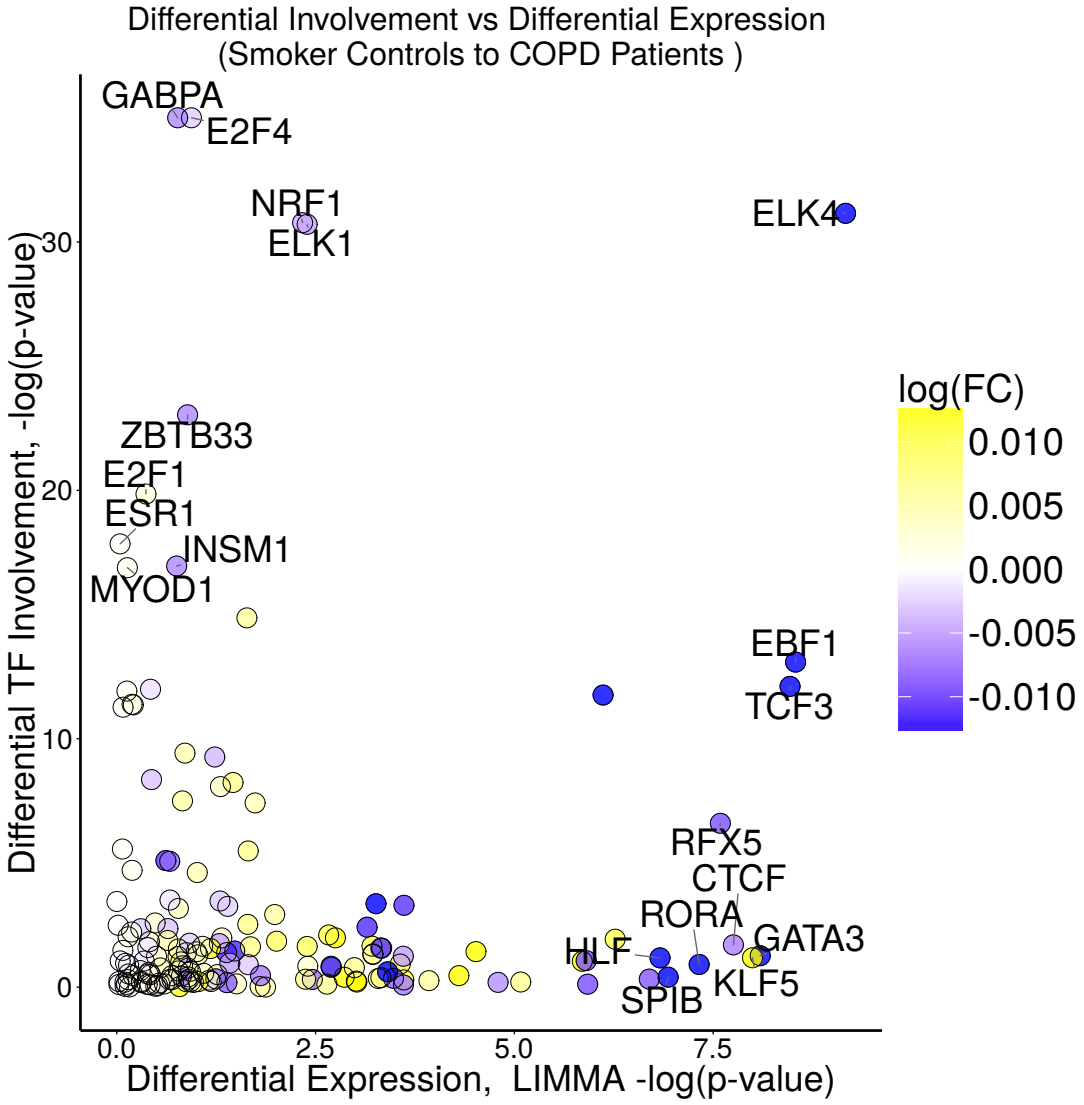


Figure 2.5: Differentially involved transcription factors are not necessarily differentially expressed. A plot of the differential expression versus the differential involvement for transcription factors based on our analysis of the ECLIPSE data. MONSTER commonly finds transcription factors which are differentially involved but are expressed at similar levels across cases and controls. Importantly, these transcription factors would not have been identified using conventional differential expression methods. This demonstrates the unique potential MONSTER has for discovery beyond standard gene expression analysis.

expression data together with simulated motif priors derived from reference networks and (4) yeast expression data together with a biological motif prior generated independently of the reference. For data set (4), we used the yeast motif prior, 106 gene expression samples from transcription factor knockout or overexpression conditions, and ChIP gold standard described in Glass *et. al.*[44]. Data for the first three sources was obtained from the 2012 DREAM5 challenge data set[72]. This challenge asked contestants to infer gene networks from expression data alone, using a reference standard for evaluation. For the purposes of validating MONSTER, we instead started with the reference network and randomly perturbed TF-gene pairs to create the type I and type II error rates consistent with biological yeast motif prior used in the fourth data set. Specifically, if an edge appeared in the reference network, that edge appeared in the simulated motif data with probability 0.3; if an edge was absent from the reference network, that edge appeared in the simulated motif data with probability 0.1. These probabilities result in an area under the Receiver-Operator Characteristic curve (AUC-ROC) of approximately 0.7 for prediction of the reference edges by the simulated edges.

For each of the data sets, we evaluated the accuracy of MONSTER's network inference method using AUC-ROC. For the DREAM5 data sets we applied MONSTER to the expression data together with the simulated priors and used the original reference networks as our gold-standards. For the fourth data set we applied MONSTER to the expression and motif data, and used the ChIP-chip data as our gold-standard.

AUC-ROC for edge weight differences vs Transition Matrix using various NI methods

NI Method	Network AUC	edge weight differences	MONSTER
Pearson	.704	.512 (p=.61)	.688 (p<.0001)
TOM	.703	.51 (p=.62)	.689 (p<.0001)
ARACNE	.515	.523 (p=.58)	.566 (p=.09)
CLR	.694	.57 (p=.19)	.814 (p<.0001)

Table 2.1: Comparison of edge weight difference to Transition Matrix in simulated case-control gene expression. Several network inference methods were run on our *in silico* case-control data. The overall network area under the curve of the receiver-operator characteristic (AUC-ROC) was performed for each method averaged across cases and controls. The naive transcription factor-transcription factor transitions were calculated as the difference in transcription factor-transcription factor edge weight between cases and controls. The transition matrix transcription factor-transcription factor transitions used the absolute transition matrix values.

We found that in all four of these data sets, the accuracy of the estimated edges from MONSTER's network inference was superior to the accuracy of the input motif prior data (Figure 2.6).

MONSTER accurately predicts transcription factor transitions in *in silico* gene expression data

We next used simulated data to evaluate MONSTER's transition matrix. To begin, we randomly generated a "true" control adjacency matrix, $\mathbf{M}_0$, which contained information for all possible edges between $m = 100$ transcription factors and $p = 10,000$ genes with "edge weights" sampled from a standard uniform distribution. We then defined a state transition matrix, $\mathbf{T}$, with diagonal elements set equal to one and $1,000$ random off-diagonal elements (representing random pairs of transcription factors) set

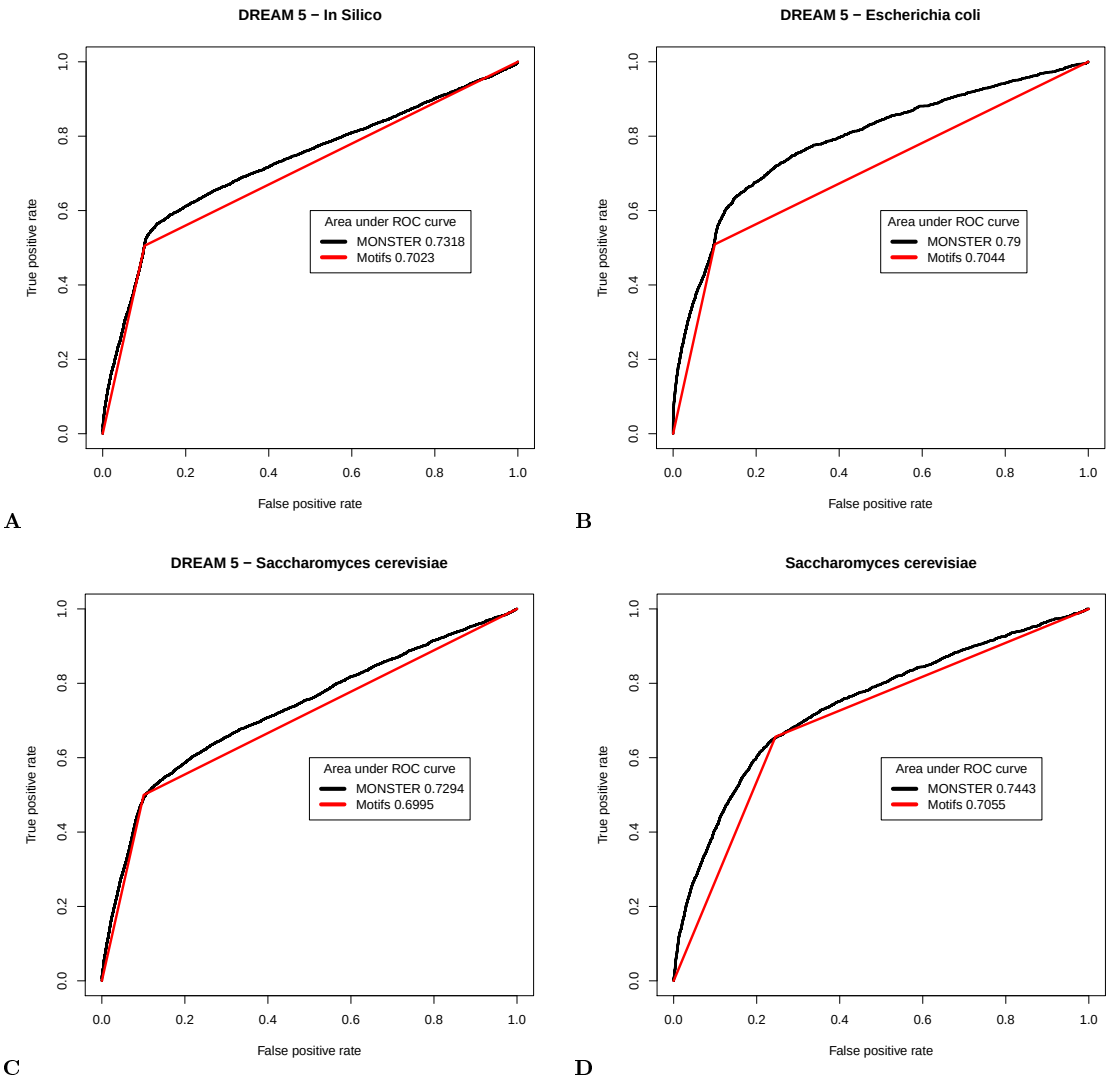


Figure 2.6: Receiver-Operator Characteristic curves for three DREAM 5 data sets (A) *in silico*, (B) *Escherichia coli*, (C) *Saccharomyces cerevisiae*, and an (D) additional *Saccharomyces cerevisiae* data set as described in Glass *et. al.*[44]. The prior network for each of the DREAM5 data set analyses was derived from the validation standard, with error introduced (both type I and type II) bringing the area under the ROC curve to ≈ 0.70. In the other *Saccharomyces cerevisiae* data set analysis, sequence motifs were used as the prior and a ChIP-chip derived network was used as the validation standard. In each of these tests, we observed a measurable improvement in performance of MONSTER's network inference method over the prior.

equal to values sampled from a uniform random distribution between -1.0 and 1.0. These off-diagonal elements (transcription factor pairs) ultimately represent the transitions that we seek to recover and their corresponding values represent the magnitude of the regulatory transition. Finally, based on $\mathbf{M}_0$ and $\mathbf{T}$ we defined the "true" cases network as $\mathbf{M}_1 = \mathbf{T}\mathbf{M}_0$.

Next, we generated two *in silico* gene expression datasets, one each for the case and control networks. To do this, we sampled 500 times from each of two multivariate Gaussian distributions with the variance-covariance matrix, Σ, defined as $\mathbf{M}_0\mathbf{M}_0'$ and $\mathbf{M}_1\mathbf{M}_1'$ for controls and cases, respectively. We note that we scaled the magnitude of the diagonal elements of Σ by 4 to simulate noise in the *in silico* data. This value was chosen such that the networks predicted using the *in silico* gene expression data had an AUC-ROC of approximately .70 when evaluated using the "true" networks (see below).

We next used this simulated data to reconstruct networks using several commonly used network inference methods, including the Pearson correlation (used in WGCNA) [60] [62], Topological Overlap Measure (TOM) [94], Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE) [73], and Context Likelihood of Relatedness (CLR) [31]. The implementation of each method was from the R package nettools [35].

We next constructed a gold-standard for our network transitions, defined as $\mathbf{T}_{GS} = ceil(|\mathbf{M}|)$. For each of the five network inference methods, we then evaluated the accuracy of two potential approaches for identifying network alterations. First, we simply

subtracted edge weights between the inferred cases network and the inferred controls network and selected those edges that extended between the 100 TFs in our model (excluding those genes that were not TFs). Second, we used MONSTER to predict the transition needed to map the control network to the case network. The results are summarized in Table 2.1. For each of the network inference methods tested, we found that the transition matrix showed substantial improvement over the edge weight difference method in identifying transitions between transcription factors. In all cases, the edge weight difference (column 3) was not statistically significant for predicting transitions, but when the transition matrix was used (column 4) a strong predictive signal appeared.

MONSTER finds significant protein-protein interactions

There are numerous biological regulatory mechanisms that may play a role in transitions between phenotypic states. Of particular interest to us are those that are not readily detectable via conventional methods for the analysis of gene expression data. For example, gene regulation involves complex processes in which transcription factors, either singly or in multiprotein complexes, bind to DNA in the region of a gene to activate or repress the transcriptional process. Such multi-protein interactions create combinatorial complexity that can explain much of the variation in organism complexity which is unexplained by gene expression alone[67].

As reported in the main text, we ran MONSTER on data from 84 smoker controls

and 136 COPD subjects in the ECLIPSE study. To test whether MONSTER could reliably detect protein-protein interactions between regulatory transcription factors, we evaluated whether our estimated transitions between case and control COPD networks in this analysis recapitulated known protein-protein interactions, as reported in Ravasi *et. al.*[93] and processed in Glass *et. al.*[46]. This dataset contained 223 interactions between the transcription factors we used as input of our model; of these, 39 were self-interacting and were removed. We attempted to predict the remaining 184 interactions between transcription factors using MONSTER.

We used the absolute value of the transition matrix and tested whether that value predicted protein-protein interactions based on the area under the ROC curve. To assess the significance of AUC-ROC, we also applied this evaluation to the 400 "random" transition matrices generated based on the randomized phenotypic labels. MONSTER achieved an AUC-ROC score of .548, suggesting predictive power to identify known PPI between transcription factors. While weak, this result exceeded all randomized phenotype results and was significant at $p < .0025$. This indicates that MONSTER is able to extract a small but significant protein interaction signal from highly obfuscated data.

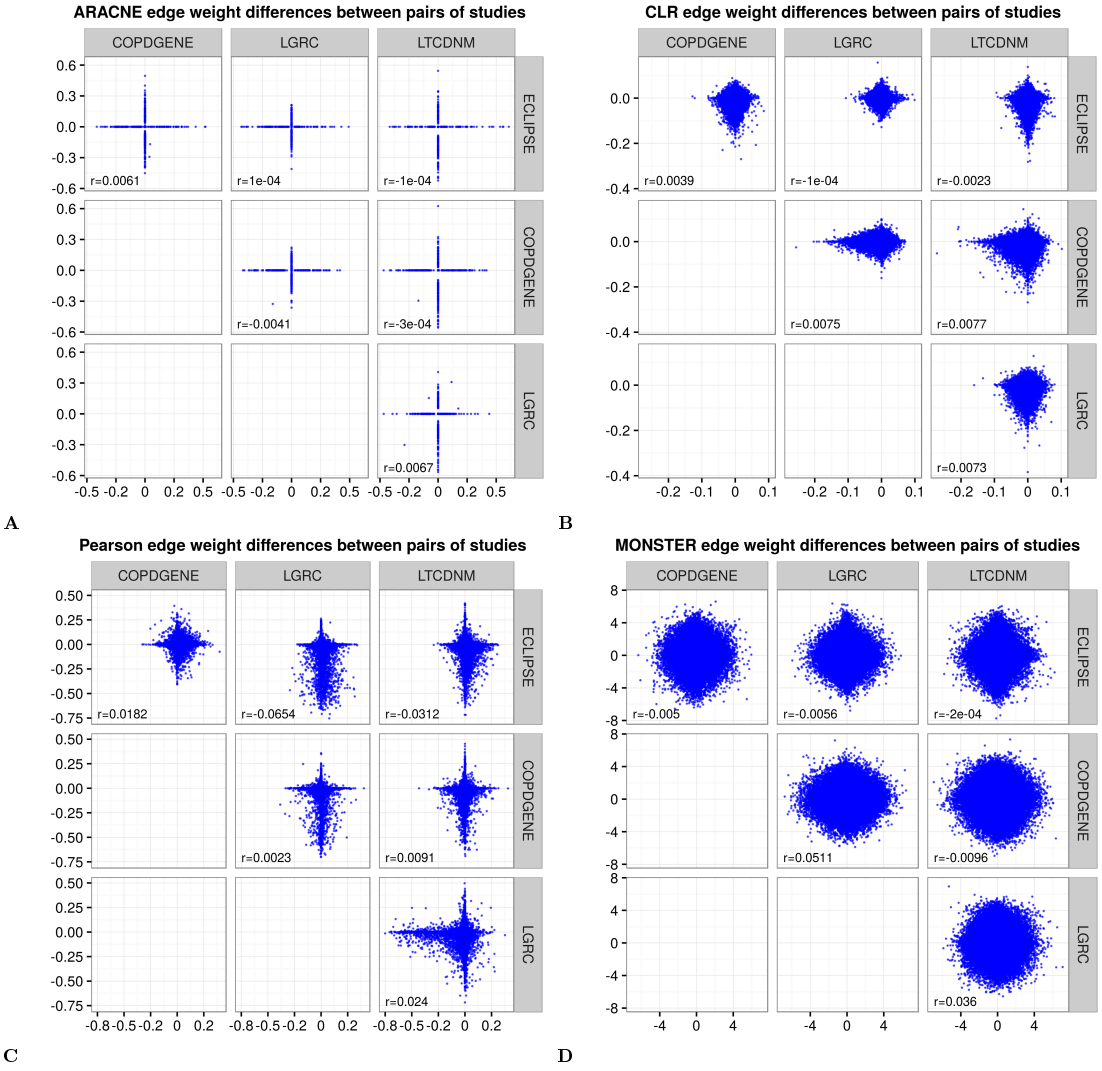


Figure 2.7: Edge weight differences between cases and controls do not correlate across studies. Using MONSTER and three other commonly used methods, we performed network inference separately on cases and controls in four COPD data sets. Here, the case-control difference is compared for each method in each data set. Most methods had very poor overall concordance in the edge weight differences they estimated. The methods tested were **A** Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE), **B** Context Likelihood of Relatedness (CLR), **C** Pearson correlation networks, such as in Weighted Gene Correlation Network Analysis (WGCNA), and **D** MONSTER. No detectable agreement between studies exist were found, regardless of network inference method or tissue type.

Irreproducibility of network inference methods in estimating transcription factor - gene edge-weights in COPD

Conceptually, MONSTER is comprised of two elements. The first infers gene regulatory networks from transcriptional data while the second uses the networks inferred for two different phenotypes to calculate the transition matrix between states. Instead of using the second part of the MONSTER approach to understand the transition between one state and another, one could imagine instead substracting the edge-weights predicted for two networks and using those differences to define a transition between two phenotypic states. To test whether this is a reasonable approach we examined the reproducibility of edge weight differences between case and control networks estimated for four COPD datasets using MONSTER's network reconstruction approach as well as three other widely used network inference methods: Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE), Context Likelihood of Relatedness (CLR), and the standard Pearson correlation used in such methods as Weighted Gene Correlation Network Analysis (WGCNA).

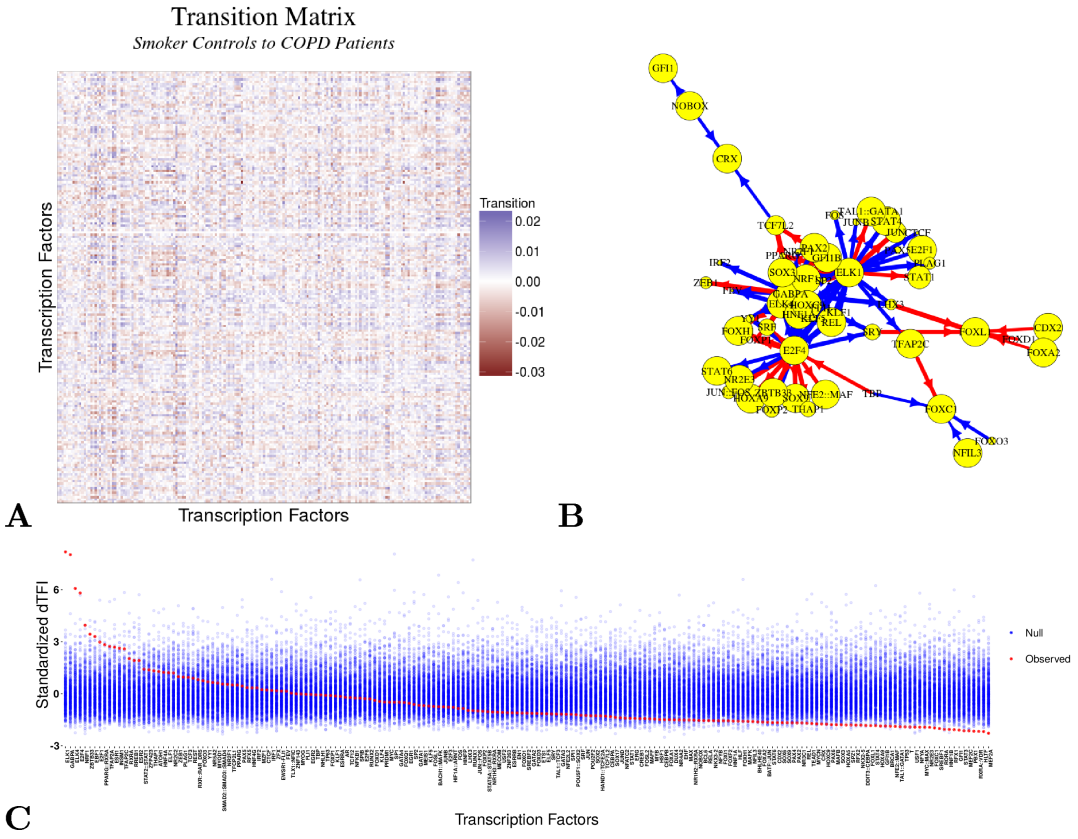
We used each of the four methods to separately estimate networks for cases and controls in each of the COPD studies. We then calculated the difference between case and control edges (differential edge weights) in each study for each method. We reasoned that if edge-differences were reflective of biologically meaningful associations, these should be present in each study and should appear as a correlated set of differ-

ential edge weights.

We plotted the differential edge weights for each pairwise combination of studies (Figure 2.7) and found that the differential edges found by ARACNE, CLR, WGCNA and MONSTER were almost entirely study specific, meaning that edges are found in one study comparing smoker controls to COPD patients are not found in a second study comparing the same phenotypes. Clearly, evaluation of individual edge-weight differences is not a reproducible approach for comparing inferred networks and stands in stark contrast to the highly reproducible set of differentially-involved set of transcription factors that we were able to identify across all four studies (as presented in the main text).

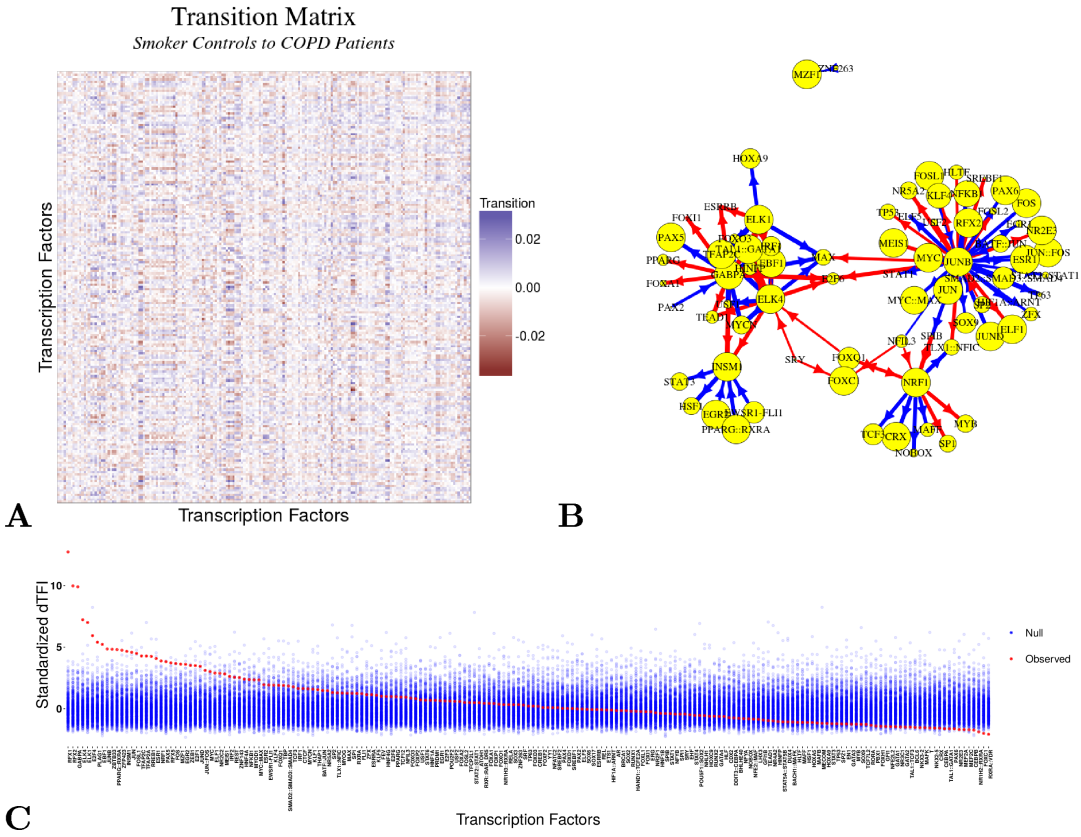
2.4 Discussion

One of the fundamental problems is biology is modeling the transition between biological states such as that which occurs during development or as a healthy tissue transforms into a disease state. As our ability to generate large-scale, integrative multi-omic data sets has grown, there has been an increased interest in using those data to infer gene regulatory networks to model fundamental biological processes. There have been many network inference methods published, each of which uses a different approach to estimating the "strength" of interactions between genes (or between transcription factors and their targets). But all suffer from the same fundamental limitation: every method relies on estimating weights that represent the likelihood of an



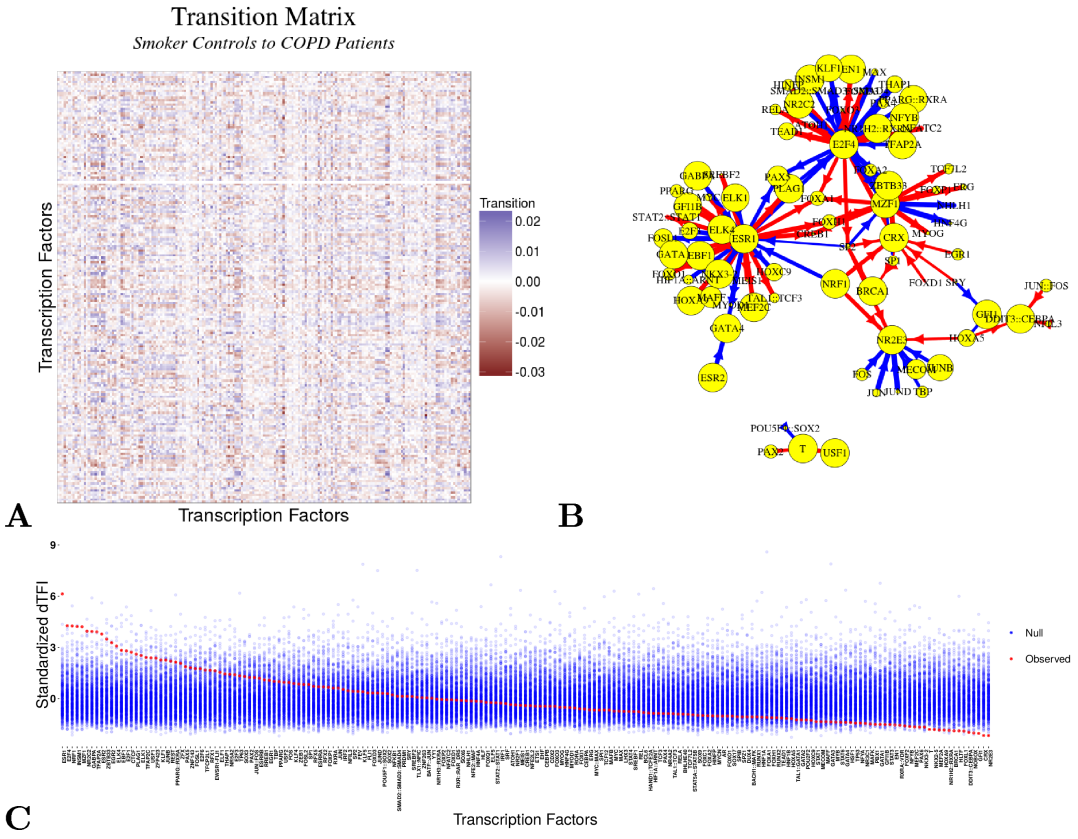
COPDGENE

Figure 2.8: MONSTER analysis results for COPDGENE study. A Heatmap depicting the transition matrix calculated from smoker controls to COPD cases by applying MONSTER to the COPDGene study. For the purposes of visualization, the magnitude of the diagonal is set to zero. **B** A network visualization of the strongest 100 transitions identified based on the transition matrix shown in **A**. Arrows indicate a change in edges from a transcription factor in the Control network to resemble those of a transcription factor in the COPD network. Edges are sized according to the magnitude of the transition and nodes (transcription factors) are sized by the dTFI for that transcription factor. The gain of targeting features is indicated by the color blue while the loss of features is indicated by red. **C** The dTFI score from MONSTER (red) and the background null distribution of dTFI values (blue) as estimated by 400 random sample permutations of the data.



LGRC

Figure 2.9: MONSTER analysis results for LGRC study. A Heatmap depicting the transition matrix calculated from smoker controls to COPD cases by applying MONSTER to the LGRC study. For the purposes of visualization, the magnitude of the diagonal is set to zero. **B** A network visualization of the strongest 100 transitions identified based on the transition matrix shown in **A**. Arrows indicate a change in edges from a transcription factor in the Control network to resemble those of a transcription factor in the COPD network. Edges are sized according to the magnitude of the transition and nodes (transcription factors) are sized by the dTFI for that transcription factor. The gain of targeting features is indicated by the color blue while the loss of features is indicated by red. **C** The dTFI score from MONSTER (red) and the background null distribution of dTFI values (blue) as estimated by 400 random sample permutations of the data.



LTCDNM

Figure 2.10: MONSTER analysis results for LTCDNM study. A Heatmap depicting the transition matrix calculated from smoker controls to COPD cases by applying MONSTER to the LTCDNM study. For the purposes of visualization, the magnitude of the diagonal is set to zero. **B** A network visualization of the strongest 100 transitions identified based on the transition matrix shown in **A**. Arrows indicate a change in edges from a transcription factor in the Control network to resemble those of a transcription factor in the COPD network. Edges are sized according to the magnitude of the transition and nodes (transcription factors) are sized by the dTFI for that transcription factor. The gain of targeting features is indicated by the color blue while the loss of features is indicated by red. **C** The dTFI score from MONSTER (red) and the background null distribution of dTFI values (blue) as estimated by 400 random sample permutations of the data.

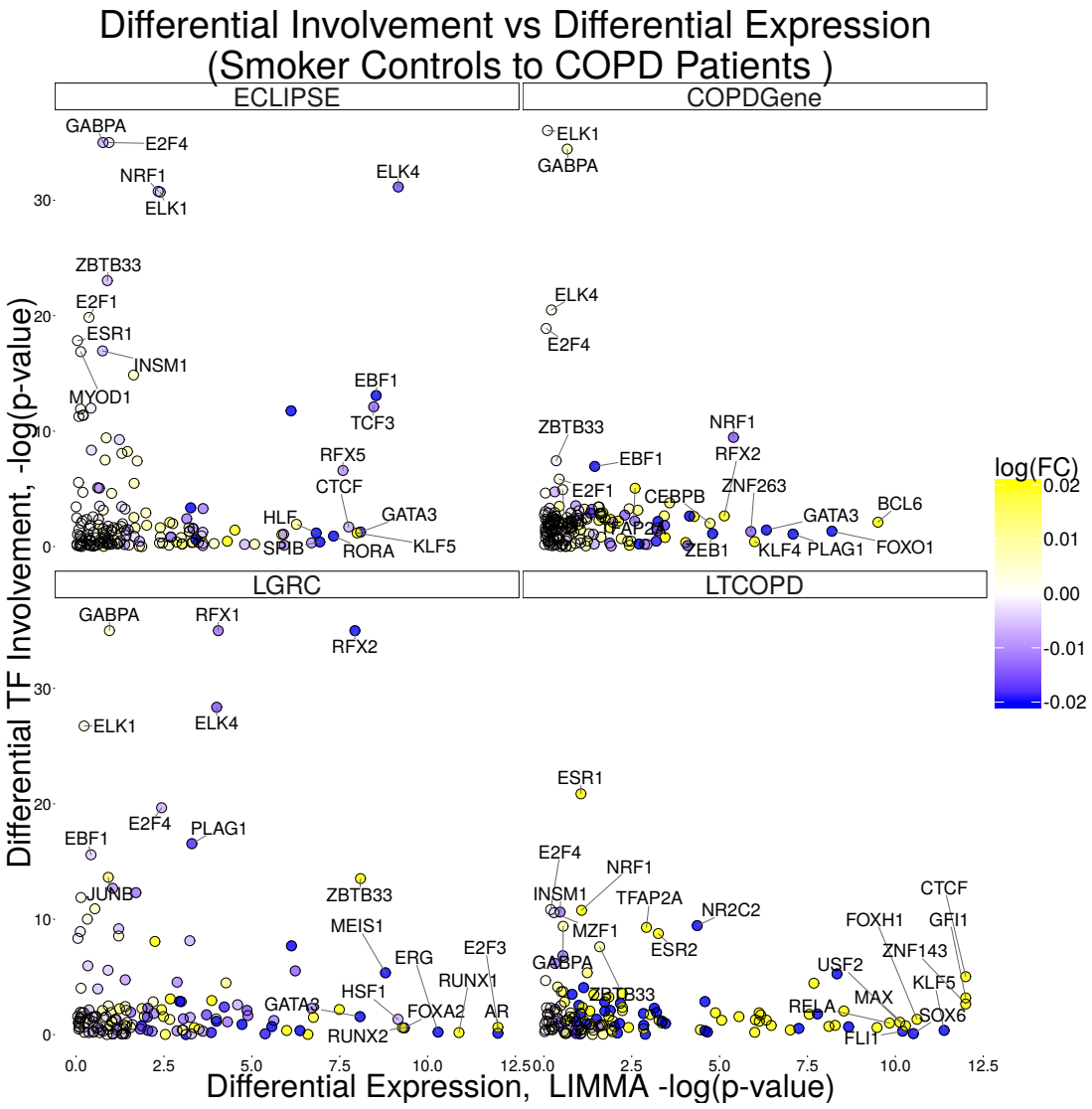


Figure 2.11: Differentially transcription factor involvement vs differential gene expression in four studies of COPD. Plots of the differential expression of transcription factors based on LIMMA, and their different involvement (dTF1) based on MONSTER. We observe much higher consistency between the transcription factors highlighted using MONSTER compared to LIMMA. In addition, we note that MONSTER commonly finds transcription factors which are differentially involved but are expressed at similar levels across cases and controls. This demonstrates the unique potential MONSTER has for discovery beyond standard gene expression analysis.

A Top significantly differentially involved transcription factors

	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

B Differential gene expression for significantly involved transcription factors.

	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

Table 2.2: Top Transcription Factor Hits. A Combined list of transcription factors which were among the top 10 hits (out of 166 available transcription factors) in any of the 4 studies, ordered by the dTFI in the ECLIPSE study. For each study, columns indicate the transcription factor's (1) differential transcription factor Involvement, (2) dTFI Rank within list of transcription factors, (3) and Significance of dTFI by false discovery rate. **B** The same list of top transcription factors evaluated for differential gene expression analysis using LIMMA. A substantial number of differentially involved transcription factors do not exhibit gene expression differentiation, highlighting the ability of MONSTER to identify key features distinguishing phenotypes which are not detectable via gene expression analysis.

interaction between two genes to identify "real" (high confidence) edges. In comparing phenotypes, most methods then subtract discretized edges in one phenotype from those in the other to search for differences.

MONSTER represents a new way of looking at phenotypic transitions, but one that captures many aspects of what we should expect. First, we have to recognize that there is no single network that represents a phenotype, but that each phenotype is represented by a family of networks that all vary slightly from each other, yet which have essential features that are consistent with the phenotype. What this means is that each regulatory edge in a network representation has to be represented by continuous, rather than discrete, variables. This captures the fact that regulatory interactions are stronger in certain individuals and weaker in others, or present in some and absent in others, but that, on average, they represent a distribution.

Second, when we consider a change in phenotype, that will be reflected in altered patterns of gene expression, and ultimately in the networks that represent the phenotype. In a transition, some individuals will experience a greater change while others will experience a smaller change. But overall, regulatory patterns in the network will shift as the phenotype changes.

Third, the change in the gene regulatory network structure between phenotypes will be driven by changes in the connectivity of the regulators—the transcription factors that alter when, how, and how strongly genes are expressed. A natural hypothesis in this model is that the transition between phenotype is likely associated with the

transcription factors that experience the greatest change in their regulatory patterns between states, and that the activation or inactivation of their target genes, and the functions carried out by those genes, likely reflect the phenotypic differences between states.

MONSTER captures these features, creating initial and final state network representations and estimating the change in transcription factor regulatory patterns by estimating a transition matrix. For each transcription factor, the "off diagonal mass" calculated as the differential transcription factor involvement (dTFI), identifies those transcription factors that are ultimately likely to drive the phenotypic state transition.

In applying MONSTER to four independent COPD gene expression data sets surveying both COPD and smoker controls, a highly consistent picture of the transcription factors associated with disease development emerges. This consistency is, to some, surprising as gene expression data is notoriously noisy, with each study finding sets of differentially expressed genes that often are not concordant. By focusing on transcriptional regulators, MONSTER seems to be able to separate a cleaner signal from the noise and one that makes some biological sense. Indeed, when one looks at the transcription factors found by MONSTER as associated with the transition, all are biologically plausible candidates which provide new and important opportunities for future molecular studies of COPD pathogenesis. It is also noteworthy that many of these transcription factors could not have been found through a simple differential

expression analysis as their transcriptional levels do not change significantly between disease and control populations. Rather, it is the regulatory patterns of these transcription factors, possibly driven by epigenetic or other changes, that shifts with the phenotype.

Acknowledgements

The project described was supported by Award Number P01 HL105339, R01 HL111759, R01 HL089897, R01 HL089856 and K25HL133599 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

Author Contributions

DS, KG, and JQ designed research; DS performed method development and application; DS, KG, CPH, EKS and JQ interpreted results; DS, KG, CPH, EKS and JQ wrote the paper. The authors declare no conflict of interest.

Most people stop looking when they find the proverbial needle in the haystack. I would continue looking to see if there were other needles.

-Albert Einstein

3

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

In order to minimize the effects of genetic confounding on the analysis of high-throughput genetic association studies, e.g. (whole-genome) sequencing (WGS) studies, genome-wide association studies (GWAS), etc., we propose a general framework to assess and to test formally for genetic heterogeneity among study subjects. As the approach fully utilizes the recent ancestor information captured by rare variants, it is

especially powerful in WGS studies. Even for relatively moderate sample sizes, the proposed testing framework is able to identify study subjects that are genetically too similar, e.g cryptic relationships, or that are genetically too different, e.g. population substructure. The approach is computationally fast, enabling the application to whole-genome sequencing data, and straightforward to implement.

Results: Simulation studies illustrate the overall performance of our approach. In an application to the 1000 Genomes Project, we outline an analysis/cleaning pipeline that utilizes our approach to formally assess whether study subjects are related and whether population substructure is present. In the analysis of the 1000 Genomes Project data, our approach revealed subjects that are most likely related, but had previously passed standard qc-filters.

Availability: An implementation of our method, Similarity Test for Estimating Genetic Outliers (STEGO), is available in the R package stego from Github at https://github.com/dschlauch/stego.

3.1 Introduction

The fundamental assumption in standard genetic association analysis is that the study subjects are independent and that, at each locus, the allele frequency is identical across study subjects ([16,91,112]). In the presence of population heterogeneity, e.g. population substructure or cryptic relatedness, these assumptions are violated. It can introduce confounding into the analysis and lead to biased results, e.g. false positive

findings ([86,57,90,110]). Given the generality of the problem, it has been the focus of methodology research for a long time. An early approach, genomic control, was developed for candidate gene and later for genome-wide association studies (GWAS) ([26,3]), adjusting the association test statistics at the loci of interest by an inflation factor that is estimated at a set of known null-loci. With the arrival of GWAS data, it became possible to estimate the genetic dependence between study subjects and the overall genetic variation for each study subject by computing the empirical genetic variance-covariance matrix between study subjects at a whole genome level. The genetic variance-covariance matrix can then be utilized in two ways to minimize the effects of population substructure on the association analysis.

The first method is to compute an eigendecomposition of the matrix and to include the eigenvectors that explain the most variation as covariates in the association analysis ([86,88]). An alternative approach is to incorporate the estimated dependence structure of the study subjects directly into a generalized linear model and account directly for the dependence at the model-level ([70,69,116]). Both approaches have proven to work well in numerous applications. While the first approach is computationally fast and easy to implement, the direct modeling of the dependence structure between study subjects can be more efficient ([75]).

However, both approaches benefit if, prior to the analysis, study subjects whose genetic profile is very different from the other study subjects, e.g. "genetic outliers", are removed from the data set. A common practice is currently to examine the eigen-

value plots visually and to identify outliers by personal judgment on how far study subjects are from the "clouds" of study subjects. As typically up to 10 eigenvectors have to be considered, this process of identifying outliers can become a complicated and subjective procedure. Alternatively, a software tool SMARTPCA ([83]), provides a more quantitative utility for removal of outliers by iteratively recomputing PCs in the genetic data. The method assumes a set of unrelated individuals and uses the covariance-based genetic similarity matrix to identify these individuals.

Many methods exist for inferring relatedness which make the strong assumption of population homogeneity ([91,112,57,16]). These methods have been shown to be biased in the context of population heterogeneity ([71]). More recently, methods have been developed which attempt to estimate relatedness with population structure ([107,71]). These developments improve the ability to detect existing pedigrees, which can aid in the removal of individuals who violate homogeneity. However, there is currently no quantitative measure of homogeneity which can be used to test a dataset prior to the application of GWAS.

In this communication, we propose a formal statistical test that assesses whether two study subjects come from the same population and whether they are unrelated. The test statistic is based on an adaptation of the Jaccard Index which utilizes the idea that variants are differentially informative of relatedness based on their allele frequency. Recent work has shown that the Jaccard Index alone can be used to reveal finer scale population structure compared with existing methods such as EIGEN-

STRAT ([89]). Furthermore, the distribution of our statistic can be derived under the null-hypothesis which makes it computationally fast, enabling the application to whole-genome sequencing data. Our measure has clearly defined properties which can be used to test for homogeneity in a population and in particular identify individuals who are likely be related in a study population. Applications to the 1,000 Genome Project suggests that our approach is better suited to detect sub-populations than genetic variance-covariance approach. This is most likely attributable to the emphasis of our approach on small allele frequencies.

3.2 Methods

Exploiting the information in rare variants (RVs), such as one with minor allele frequency (MAF) $< 1\%$, is fundamental to our method, as our approach utilizes the features of RVs that they are typically more recent than common variants and that many of them are population/family specific. Since allele frequencies can differentially confound association studies ([75]), we developed a method that utilizes the differential informativeness of variants by allele frequency to obtain a high resolution picture of population structure and protect the association study against bias due to genetic confounding. Our approach uses an intuitive, computationally straightforward approach towards identifying similarity between two study subjects which is also directly linked to the kinship coefficient.

3.2.1 Similarity measure among haploid genomes

Consider a matrix of n individuals ($2n$ haploid genomes), with N independent variants described by the genotype matrix $\mathbf{G}_{2n\times N}$. $\mathbf{G}$ is a binary matrix with value 1 indicating the presence of the minor allele and 0 indicating the major allele. We define the similarity index between two haploid genomes, $s_{i,j}$

$$s_{i,j} = \frac{\sum_{k=1}^{N} w_k \mathbf{G}_{i,k}\mathbf{G}_{j,k}}{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]} \tag{3.1}$$

where

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

and $I[\cdot]$ is an indicator function, evaluating to 1 if true and 0 if false.

We can consider the weight variable, w_k, to be the inverse of the probability that two alleles selected randomly without replacement both belong to the set of minor alleles. Intuitively, for $\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1$, w_k is monotonically decreasing as the minor allele count increases (See below).

In the absence of population structure, i.e. homogeneous population we have

$$E\left(s_{i,j}\right) = 1 \tag{3.2}$$

It therefore follows from the Central Limit Theorem that in the absence of population structure, cryptic relatedness and dependence between loci (such as linkage disequilibrium) the distribution of the similarity index, $s_{i,j}$ is Gaussian.

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

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

$$\hat{\sigma^2_{i,j}} = \hat{Var}\left(s_{i,j}\right) = \frac{\sum_{k=1}^{N}\left(w_k - 1\right)}{\left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^2} \tag{3.3}$$

The similarity index $s_{i,j}$ provides an easily interpreted statistical test for evaluating possible relatedness between individuals in a purportedly homogeneous dataset of unrelated individuals. Note that this formulation (3.3) is independent of the samples i, j and depends only on the allele counts for each variant across the study group. (See below).

3.2.2 Similarity measure among diploid genomes

This approach is easily generalized to the diploid scenario. A diploid similarity score, $s_{diploid}$, is obtained by averaging each of the four pairwise haploid $s_{haploid}$ scores between each person's two haploid genotypes. For n individuals, $2n$ genotypes per locus, the similarity between individuals i and j is defined as

$$s_{i,j}^{(diploid)} = \frac{\sum_{k=1}^{N}\sum_{l=1}^{2}\sum_{m=1}^{2}\left[w_k \mathbf{G}_{i_l,k}\mathbf{G}_{j_m,k}\right]/4}{\sum_{k=1}^{N} I\left[\left(\sum_{l=1}^{n}\left[\mathbf{G}_{l_1,k}+\mathbf{G}_{l_2,k}\right]\right)>1\right]}$$

where $\mathbf{G}_{i_2,k}$ refers to the 2^{nd} genotype of individual i at locus k.

Here it becomes clear that the method can be applied to phased and unphased data alike. For an unphased data matrix $\mathbf{H}_{n\times N}$, where $\mathbf{H}$ contains the number of minor alleles, $\{0,1,2\}$, for a subject at a particular variant.

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

This formulation will have the same mean

$$E\left[s_{i,j}^{(diploid)}\right] = 1$$

and assuming independence of each individual's haploid genomes, such as in the absence of inbreeding,

$$\hat{Var}\left(s_{i,j}^{(diploid)}\right) = \frac{\hat{Var}\left(s_{i,j}^{(haploid)}\right)}{4} = \hat{\sigma^2}_{i,j}$$

Which yields the asymptotic result

$$s_{i,j} \sim N\left(1, \hat{\sigma^2}_{i,j}\right) \tag{3.4}$$

3.2.3 Tests of Heterogeneity

We can test the null hypothesis that population structure does not exist and all subjects are unrelated, with respect to the alternative that at least one pair of individuals is related.

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

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

Under the null hypothesis, we have clearly defined the distribution of test statistics. However, violations of homogeneity may come in many forms and thus there is no most powerful test which can be applied to detect all possible alternatives. Below, we examine two separate rejection criteria designed for two types of heterogeneity-population structure and cryptic relatedness.

Given the complex nature with which population structure may manifest itself, we recommend the conservative Kolmogorov-Smirnov test for detection of population structure.

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

where F_s is the empirical distribution function for s, and $\Phi(x)$ is the cumulative distribution function defined in (3.4). Under the null, K follows the Kolmogorov distribution ([111]), so we define K_α as the value such that $P(K > K_\alpha | H_0) = \alpha$, and reject homogeneity in the data if $K > K_\alpha$.

While this approach is effective for a large number of small effects, as would be expected with subtle population structure, it will not be particularly effective at detecting a small number of large effects, as expected with cryptic relatedness. For this scenario we recommend a test with more power at the far right tail of the distribution.

In a homogeneous dataset lacking relatedness, we consider each of the $\binom{n}{2}$ comparisons to be independent. To achieve a family-wise error rate α, we use the Šidák procedure ([101]) or the approximately equivalent Bonferroni procedure. We reject the null at the α level when we obtain similarity scores in the rejection region

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

3.2.4 Estimating cryptic relatedness

The measure described here is particularly powerful for measuring relatedness. Intuitively, we can imagine two subjects which have a kinship coefficient, ϕ, indicating a probability of a randomly chosen allele in each person being identical by descent (IBD). For an allele which belongs to the one person, the probability of it belonging to a related person with kinship coefficient ϕ is $\phi + (1 - \phi) \times p$, where p is the allele frequency in the population. We can clearly see that for rare alleles, such that p is small compared to ϕ, there will be a much larger relative difference in the probability

of shared alleles among related individuals ($\phi > 0$) compared to unrelated individuals ($\phi = 0$). Given that STEGO weights more highly these rarer alleles, there is increased sensitivity to detection of relatedness.

Consider a coefficient of kinship between two individuals i, j, $\phi_{i,j} > 0$ with no other population structure present in the data. For an individual variant, k, with sufficient allele frequency, the expected contribution to the statistic for an allele from each individual, s_{i_1,j_1} is

$$E\left(s_{i_1,j_1,k}|\phi_{i,j}\right) = 1 + \phi_{i,j}\left[p_k \frac{\binom{2n}{2}}{\binom{p_k(2n-2)+2}{2}} - 1\right]$$

and the expectation for the similarity score between those haploid genomes is

$$E\left(s_{i_1,j_1}|\phi_{i,j}\right) =$$

$$\frac{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\left[1 + \phi_{i,j}\left[p_k \frac{2n(2n-1)}{\left(p_k(2n-2)+2\right)\left(p_k(2n-2)+1\right)} - 1\right]\right]}{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]} \tag{3.5}$$

It can be seen that in the presence of cryptic relatedness, $\phi_{i,j} > 0$,

$$E\left(s_{i_1,j_1}|\phi_{i,j} > 0\right) > 1$$

With $\sum_{i=1}^{2n} \mathbf{G}_{i,k}$ as the maximum likelihood estimator for $p_k n$, by the invariance prin-

ciple, w_k is a consistent estimator for $\frac{\binom{2n}{2}}{\binom{p_k(2n-2)+2}{2}}$.

This yields a maximum likelihood estimate of this kinship defined as

$$\hat{\phi}_{i,j} = \frac{s_{i,j} - 1}{\left[\frac{\sum_{k=1}^{N} \hat{p_k} w_k}{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]} - 1\right]} \tag{3.6}$$

with

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

For example, in an otherwise homogeneous study group of unrelated individuals a pair of cousins ($\phi = .0625$), with $MAF \sim Uniform\,(.02, .1)$ we can directly calculate the expectation of their similarity statistic, $s_{i,j}$

$$E\left(s_{i,j} | \phi = .0625, \text{No other structure}\right) \approx 2.19$$

3.2.5 Statistical power to detect outliers

The properties of this similarity measure lend themselves toward straightforward power calculations. It is often of interest to consider some coefficient of relatedness, γ, that is acceptable for a study. Setting a $\phi \geq \gamma$ allows for the calculation of the probability of obtaining a pair of samples inside the rejection region given two unacceptably

closely related individuals.

$$P\left(Reject\, H_0 | \phi_{i,j} = \gamma\right) = \alpha + (1-\alpha)\left(1 - \Phi\left(\frac{\mu_{i,j} - 1}{\sqrt{\hat{\sigma^2}_{i,j}}}\right)\right) \tag{3.7}$$

Where $\Phi(x)$ is the cumulative distribution function for a standard normal random variable. Also note that this power is computed under the assumption of homogeneity among all unrelated individuals, which will yield a conservative estimate of the probability of rejection. The presence of unknown population structure will necessarily increase the power of the test.

It is of interest in any study seeking to quantitatively demonstrate the homogeneity of participants to produce this statistic which can demonstrate that heterogeneity would have been observed with some probability, given the presence of some specified degree of relatedness, γ.

3.2.6 Example of $\hat{s}_{i,j}$ computation

Consider a binary matrix of 20 haploid genomes from 10 samples (labeled a-t).

The matrix $\boldsymbol{G}$ is encoded such that 1 indicates the presence of the minor allele and 0 indicates the presence of the major allele for a particular variant (column) and haploid sample (row). For visual purposes, we limit the calculation to 4 variants and we show $\boldsymbol{G}^T$ here:

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
variant 1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
variant 2	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
variant 3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
variant 4	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

For the purposes of this explanatory example, note that columns g and n share the low frequency allele (1), but differ along the common variants (2-4). Intuitively, we want to consider the relative informativeness of the low frequency variant compared to the common variants.

The minor allele frequencies for variants 1,2,3,4 are 0.1, 0.5, 0.5, and 0.5, respectively.

For each locus, we compute w_k, $k \in \{1, 2, 3, 4\}$

$$w_k = \frac{\binom{20}{2}}{\binom{\sum_{l=1}^{20} \boldsymbol{G}_{l,k}}{2}}$$

Such that $w_1 = 190$ and $w_2 = w_3 = w_4 = 4.22$

Using equation (1), we compute the relatedness matrix for across these variants:

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
a	-	4.2	8.4	4.2	8.4	4.2	8.4	4.2	8.4	4.2	4.2	0	4.2	0	4.2	0	4.2	0	4.2	0
b	4.2	-	4.2	4.2	4.2	4.2	4.2	4.2	4.2	4.2	0	0	0	0	0	0	0	0	0	0
c	8.4	4.2	-	4.2	8.4	4.2	8.4	4.2	8.4	4.2	4.2	0	4.2	0	4.2	0	4.2	0	4.2	0
d	4.2	4.2	4.2	-	4.2	4.2	4.2	4.2	4.2	4.2	0	0	0	0	0	0	0	0	0	0
e	8.4	4.2	8.4	4.2	-	4.2	8.4	4.2	8.4	4.2	4.2	0	4.2	0	4.2	0	4.2	0	4.2	0
f	4.2	4.2	4.2	4.2	4.2	-	4.2	4.2	4.2	4.2	0	0	0	0	0	0	0	0	0	0
g	8.4	4.2	8.4	4.2	8.4	4.2	-	4.2	8.4	4.2	4.2	0	4.2	**190**	4.2	0	4.2	0	4.2	0
h	4.2	4.2	4.2	4.2	4.2	4.2	4.2	-	4.2	4.2	0	0	0	0	0	0	0	0	0	0
i	8.4	4.2	8.4	4.2	8.4	4.2	8.4	4.2	-	4.2	4.2	0	4.2	0	4.2	0	4.2	0	4.2	0
j	4.2	4.2	4.2	4.2	4.2	4.2	4.2	4.2	4.2	-	0	0	0	0	0	0	0	0	0	0
k	4.2	0	4.2	0	4.2	0	4.2	0	4.2	0	-	4.2	8.4	4.2	8.4	4.2	8.4	4.2	8.4	4.2
l	0	0	0	0	0	0	0	0	0	0	4.2	-	4.2	4.2	4.2	4.2	4.2	4.2	4.2	4.2
m	4.2	0	4.2	0	4.2	0	4.2	0	4.2	0	8.4	4.2	-	4.2	8.4	4.2	8.4	4.2	8.4	4.2
n	0	0	0	0	0	0	**190**	0	0	0	4.2	4.2	4.2	-	4.2	4.2	4.2	4.2	4.2	4.2
o	4.2	0	4.2	0	4.2	0	4.2	0	4.2	0	8.4	4.2	8.4	4.2	-	4.2	8.4	4.2	8.4	4.2
p	0	0	0	0	0	0	0	0	0	0	4.2	4.2	4.2	4.2	4.2	-	4.2	4.2	4.2	4.2
q	4.2	0	4.2	0	4.2	0	4.2	0	4.2	0	8.4	4.2	8.4	4.2	8.4	4.2	-	4.2	8.4	4.2
r	0	0	0	0	0	0	0	0	0	0	4.2	4.2	4.2	4.2	4.2	4.2	4.2	-	4.2	4.2
s	4.2	0	4.2	0	4.2	0	4.2	0	4.2	0	8.4	4.2	8.4	4.2	8.4	4.2	8.4	4.2	-	4.2
t	0	0	0	0	0	0	0	0	0	0	4.2	4.2	4.2	4.2	4.2	4.2	4.2	4.2	4.2	-

3.2.7 Expectation of $s_{i,j}$

The expectation of s, under the null of population homogeneity is defined as

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

By this definition and that of w_k, variants which contain one or fewer minor alleles contribute a zero to the summation in both the numerator and denominator and can be ignored. Let N^* be the number of variants indexed k such that $\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1$, and let k^* index those variants. We condition on the minor allele count for each variant, $a_{k^*} = \sum_{l=1}^{2n} \mathbf{G}_{l,k^*}$, where $n \geq a > 1$,

$$
\begin{aligned}
E\left[s_{i,j}\right] &= E\left[\frac{\sum_{k^*=1}^{N^*} w_{k^*}\mathbf{G}_{i,k^*}\mathbf{G}_{j,k^*}}{N^*}\right] \\
E\left[s_{i,j}\right]N^* &= \sum_{k^*=1}^{N^*} \frac{\binom{2n}{2}}{\binom{a}{2}} E\left[\mathbf{G}_{i,k^*}\mathbf{G}_{j,k^*}\right] \\
&= \sum_{k^*=1}^{N^*} \frac{\binom{2n}{2}}{\binom{a}{2}} E\left[\mathbf{G}_{i,k^*}\right] E\left[\mathbf{G}_{j,k^*}|\mathbf{G}_{i,k^*}=1\right] \\
&= \sum_{k^*=1}^{N^*} \frac{\binom{2n}{2}}{\binom{a}{2}} \left(\frac{a}{2n}\right)\left(\frac{a-1}{2n-1}\right) \\
&= \sum_{k^*=1}^{N^*} \frac{\frac{2n!}{2!(2n-2)!}}{\frac{a!}{2!(a-2)!}} \left(\frac{a}{2n}\right)\left(\frac{a-1}{2n-1}\right) \\
&= \sum_{k^*=1}^{N^*} \frac{2n\left(2n-1\right)}{a\left(a-1\right)} \left(\frac{a}{2n}\right)\left(\frac{a-1}{2n-1}\right) \\
&= \sum_{k^*=1}^{N^*} 1 \\
&= N^* \\
E\left[s_{i,j}\right] &= 1
\end{aligned}
$$

Intuitively, we can consider the weight factor, w_k, to be the inverse of the probability of selecting two minor alleles at random without replacement. Therefore, the expectation for the numerator for each variant is one, and consequently, $s_{i,j}$, the mean over all variants with multiple minor alleles is also one.

3.2.8 Variance of $s_{i,j}$

The variance of $s_{i,j}$ can be estimated by

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

This formulation is independent of the samples i, j and depends only on the allele counts for each variant across the study group.

$$\begin{aligned}
Var\left(s_{i,j}\right) &= Var\left(\frac{\sum_{k=1}^{N} w_k \mathbf{G}_{i,k}\mathbf{G}_{j,k}}{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]}\right) \\
&= \left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^{-2} Var\left(\sum_{k=1}^{N} w_k \mathbf{G}_{i,k}\mathbf{G}_{j,k}\right) \\
&= \left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^{-2} \sum_{k=1}^{N} Var\left(w_k \mathbf{G}_{i,k}\mathbf{G}_{j,k}\right) && \text{Independence} \\
&= \left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^{-2} \sum_{k=1}^{N} w_k^2 Var\left(\mathbf{G}_{i,k}\mathbf{G}_{j,k}\right) \\
&= \left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^{-2} \sum_{k=1}^{N} w_k^2 P\left[\mathbf{G}_{i,k}\mathbf{G}_{j,k} = 1\right]\left(1 - P\left[\mathbf{G}_{i,k}\mathbf{G}_{j,k} = 1\right]\right) && \text{Var of Bernouli} \\
&= \left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^{-2} \sum_{k=1}^{N} w_k^2 \frac{1}{w_k}\left(1 - \frac{1}{w_k}\right) \\
&= \frac{\sum_{k=1}^{N}\left(w_k - 1\right)}{\left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^2}
\end{aligned}$$

3.2.9 Linkage Disequilibrium Pruning

Phase 3 of the 1000 Genomes Project contains 2504 individuals with a combined total of over 80 million variants. Assumptions of STEGO include the independence of variants, which may be violated in the presence of Linkage Disequilibrium (LD). Our method focuses variants with low minor allele frequency, which are less susceptible to high R^2 between loci. However, to help reduce the impact of correlated variants, we filtered the data such that the impact of LD was limited. Prior to analysis, we divided the data into blocks of 800 consecutive variants and selected only one locus from each block. The selected variant within each block was chosen based on the smallest minor allele frequency observed which was larger than our cutoff of 1%. We chose this cutoff as a balance between our interest in focusing on lower frequency alleles and the recognition that QC concerns may become an increasingly valid concern at the lowest allele frequencies (see subsection "Ancestry informativeness by allele frequency" below). We recommend the use of a cutoff which balances the value of rare variants with the confidence in the technology used to obtain the data. This filtering yielded approximately 100,000 variants for each of the 26 populations in the TGP.

3.2.10 STEGO algorithm computation time

Of interest is the computation time of STEGO in comparison to other similarity metrics. Principal Components Analysis can be performed in multiple ways, including

a decomposition of the correlation or covariance matrix and a singular value decomposition, implemented in base R as the functions **prcomp()** and **princomp()**, respectively. We compared our method in terms of computation time of generating a correlation matrix. We simulated a study of $p = 100,000$ phased variants across N individuals and ran an R implementation of STEGO against the base implementation of correlation, **cor()** and the two Principal Components analysis methods with default parameters (Figure 3.11). An R script implementing this simulation is available at https://github.com/dschlauch/Genetic-Outliers/blob/master/timingComparison.R Using a computer with Intel(R) Core(TM) i7-3630QM CPU @ 2.40GHz, and Microsoft R Open 3.2.5 linked with multi-threaded BLAS/LAPACK libraries, we found a that our method ran substantially faster than correlation (**cor**) and PCA (**princomp**) in R.

STEGO's weights are independently computed for each of the p variants based solely on the minor allele count and thus are computed in linear time, $\mathcal{O}(p)$. The computationally intensive step is implemented via matrix multiplication of the weighted genotypes matrix, which with naive implementation has complexity $\mathcal{O}(pN^2)$. Both singular value decomposition and variance-covariance matrix computations have complexity $\mathcal{O}(pN^2)$[53], and therefore each of the three methods compared have the same asymptotic computational complexity.

3.2.11 Ancestry informativeness by allele frequency

An important motivating principle of our method is the assertion that rare variants are useful for identifying fine-scale population structure. The reasoning is that rare variants are less stable than common variants and can more easily become fixed at 0%. It is reasonable to suspect that rare variants are more likely to have arisen recently in the ancestral history of a population and may therefore be informative in separating recently related populations.

To confirm this concept in the 1000 Genomes Project, we compared the relative abilities of MAF bins to separate populations. We used the Jaccard Index to compute a similarity score between all pairs of individuals and computed the ratio of within-population to between-population Jaccard Index means. Figure **??** shows the comparison of the Yoruban population (YRI) with all others and demonstrates the improved performance of rare variants compared with more common variants. It is notable, however, that the improvement clearly ceases at the lowest MAF bin (0%-0.4%), suggesting a lack of reliability for the rarest variants as a result of the imperfect nature of sequencing. It is for this reason that we recommend a minimum MAF for analyses which considers features of the analysis such as sequencing depth.

3.2.12 Simulations demonstrate sensitivity STEGO to subtle population stratification

We compared the GSM derived from STEGO against a GSM obtained via normalized variance-covariance, such as that used in EIGENSTRAT's implementation of PCA[86]. We first generated an ancestral allele frequency distribution for 20k variants, each as an observation from an $exponential\,(\lambda = .05)$ distribution. Next, two descendant allele frequencies were obtained by applying a small deviation from the ancestral allele frequency of $Uniform\,(-0.003, 0.003)$. The purpose was to create a very subtle genetic drift loosely representing a relatively short timeframe of breeding isolation for each population. We chose the size of small allele frequency deviation as one which pushes the limits of standard approaches to detect. The 1000 simulated individuals were sampled independently from the descendant allele frequencies (500 from each population), and GSMs were generated as described above.

The results from this straightforward simulation support our intuition regarding rare variants (Figure 3.12). Owing to the fact that co-occurrence of rare variants is more informative of shared population membership than co-occurrence of common variants, we see that STEGO outperforms standard PCA in this scenario. In this plot, using the top two eigenvectors shown, the ratio of within-population variance to total variance is .81 for STEGO. For variance-covariance, this ratio is .99, as this standard method was ineffective at picking up virtually any signal. At this level of subtle pop-

ulation stratification and using only 20k variants, we do not observe any meaningful separation between the two groups using PCA. Conversely, STEGO clearly demonstrates a tendency to distinguish the groups along the first principal component. The source code for this simulation is available at

https://github.com/dschlauch/Genetic-Outliers/blob/master/simulated_stego_varcov.R.

3.2.13 Simulations demonstrate power to detect heterogeneity

We ran STEGO on simulated genotypes derived from a homogeneous dataset containing varying degrees of relatedness. A homogenized version of a real dataset was generated by randomly resampling each variant across all samples. This eliminates correlations between individuals and variants, preserving only the allele frequency distribution. To test the power of our method to identify relatedness we generated an additional sample, S_{N+1} which was related to an arbitrarily chosen individual, S_N, in the homogenized dataset. The genotype for S_{N+1} was generated by assigning one of their values for each allele to be the same as one of the alleles of S_N with probability 4ϕ and assigning the other to be a randomly chosen allele across all samples. With probability $1 - 4\phi$, both haplotypes for S_{N+1} were selected randomly from the homogenized data.

For variant i, allele j, the genotype at $S_{N+1,i,j}$ is given as

$$S_{N+1,i,j} = \begin{cases} S_{N,i,1} & \text{with probability } \phi + \frac{1-2\phi}{2N} \\ S_{N,i,2} & \text{with probability } \phi + \frac{1-2\phi}{2N} \\ S_{1,i,1} & \text{with probability } \frac{1-2\phi}{2N} \\ \vdots & \vdots \\ S_{N-1,i,2} & \text{with probability } \frac{1-2\phi}{2N} \end{cases}$$

For each coefficient of kinship we simulated 1,000 studies containing 301 individuals across 100,000 variants in the above manner to evaluate the power of STEGO. Each simulated study contained only a single related pair with relatedness, ϕ, among an otherwise homogeneous dataset. We demonstrate that under the null hypothesis, $H_0 : \phi = 0$, the family-wise type I error rate, $\alpha = .05$ is preserved. We then compared the proportion of simulated studies which were found to have significantly related pairs to the analytically derived probability of type II error. Figure 3.13 demonstrates that our findings that computed Type II error aligns to the formula in Equation 7. Further investigation into the computational complexity of our method show that our method does not sacrifice speed compared to standard PCA (see subsection "Simulations demonstrate sensitivity STEGO to subtle population stratification" below).

3.3 Results

3.3.1 Identification of relatedness and structure in 1000GP data

We applied our method to data from the 1000 Genomes Project (TGP) ([21,22]), an international consortium which has sequenced individuals from 26 distinct populations sampled from around the globe.

These populations were not identified by the TGP to have cryptic relatedness or had known cryptic relatedness removed ([81]). However, subsequent analyses have discovered numerous inferred relationships closer than first cousins ([42,1,34]).

Phase 3 of the 1000 Genomes Project contains 2504 individuals with a combined total of over 80 million variants. To test STEGO, we selected a subset of approximately 100,000 variants across each of the 26 populations which limited the impact of linkage disequilibrium ([87]) and increased the independence of consecutive measurements (See below). These 100,000 variants were each chosen from 100,000 separate blocks based on low minor allele frequency and a qc-control cutoff of 1% (See subsection "Linkage Disequilibrium Pruning" below, Figure 3.1 was then run on each of these populations separately to test for heterogeneity and relatedness within population groups (Figure 3.1, 3.3A).

Our investigation revealed a great deal of variation in the presence of cryptic relatedness and population structure across the 26 populations of the study. Under the assumptions that each study contained a homogeneous population of unrelated indi-

viduals, only a handful of groups contained neither large outliers nor heavily inflated numbers of significant results.

We defined the presence of population structure as applying to those populations which deviated from the normal distribution defined under the null model. From Equation (3.3), we have the expected distribution under H_0 which we tested for in each of the populations using a standard Kolmogorov-Smirnov test. Using a significance cutoff of $\alpha = .01$, 15 of the 26 populations were found to have violated population homogeneity.

In addition to investigating population structure, we examined the presence of cryptic relatedness in the study. We defined relatedness as those individual pairs which exceed the cutoff for a family-wise error rate of $\alpha = .01$ and were estimated to have a coefficient of relatedness $\hat{\phi} > \frac{1}{32}$, which approximately corresponds to half first cousins. By this measure, cryptic relatedness was observed in all but six of the 26 populations using this method. Eleven pairs of first order (parent-offspring or full sibling) relationships were detected among individuals within the same population group, $\left(.2 < \hat{\phi_{i,j}} < .3\right)$, a set of pairings which corresponds identically with the conclusions of Gazal et al([42]).

Inference on our kinship estimate is made under the assumption of homogeneity of the background study population. Identified significant relatedness may be due to the fact that the variance of the similarity score is inflated in the presence of population structure. So it is incomplete to identify cryptic relatedness in this manner in popula-

tions which contain identified structure. However, in populations which do not exhibit detectable structure, we still find many instances of related individuals in this study. For example, two individuals from the ACB population (African Caribbeans in Barbados) produced a $s_{i,j}$ score of 2.6 $\left(p < 10^{-30}\right)$, whereas no other pairing exceeded the family-wise cutoff of 1.3 (Figure 3.1). Using the formula above, the estimated coefficient of kinship is $\hat{\phi} = .27$, suggesting that those individuals are first degree relatives. Additionally, two pairs of individuals in the STU population- (HG03899/HG03733 and HG03754/HG03750) were both estimated to have a kinship coefficient $\hat{\phi} \approx .25$, similarly indicating a relatedness of the first degree.

Interestingly, not all related pairs belonged to the same population groups. We additionally discovered a pair of individuals, HG03998 from the STU population and HG03873 from the ITU population, which exhibited strikingly high relatedness. The plot below (Figure 3.2) was generated by placing HG03998 into the ITU population and running STEGO on that population. An individual who belongs to a separate population from all others in a dataset would be expected to produce similarity scores less than 1. However, the similarity between HG03998 and HG03873 was found to be $s = 3.9$, significant at $p < 10^{-30}$ with an estimated relatedness $\hat{\phi} > .25$, suggesting that these individuals are first order relatives despite belonging to different population groups. Both populations were sampled from locations in the United Kingdom, further supporting the evidence that these individuals are related.

With strong evidence of relatedness in the data, we sought to test our method on

pruned data which contained no known related pairs. Gazal et al propose a subset of the TGP which removes 243 individuals such that no two individuals are as related as cousins or closer. These 243 samples include all those with cryptic relatedness inferred by the FSUITE and RELPAIR ([8,29]) methods. This results in a reduced set of 2261 individuals which are assumed to be no more closely related than half first cousins ($\phi = .0312$)([42]). We applied this filter and re-analyzed each of the 26 populations again to test for heterogeneity and cryptic relatedness.(Figure 3.3B, Table 3.1, Figure 3.6)

Eleven populations which had been identified as violating homogeneity ($\alpha = .01$) in the full TGP dataset were no longer identified as violating homogeneity after removal of suspected related pairs. However, four populations, including each of the admixed American groups, continued to violate homogeneity even after the attempts to limit the impact of related individuals. The three most significant populations are all part of the Ad Mixed American super population and represent "new world" groups which have undergone extensive admixture in recent centuries. - CLM (Colombians from Medellin, Colombia) $\left(p = 7 \times 10^{-8}\right)$, PUR (Puerto Ricans from Puerto Rico) $\left(p = 3 \times 10^{-31}\right)$, and PEL (Peruvians from Lima, Peru) $\left(p = 2 \times 10^{-27}\right)$. It is therefore reassuring that these groups of individuals would exhibit the greatest amount of structure among the populations surveyed.

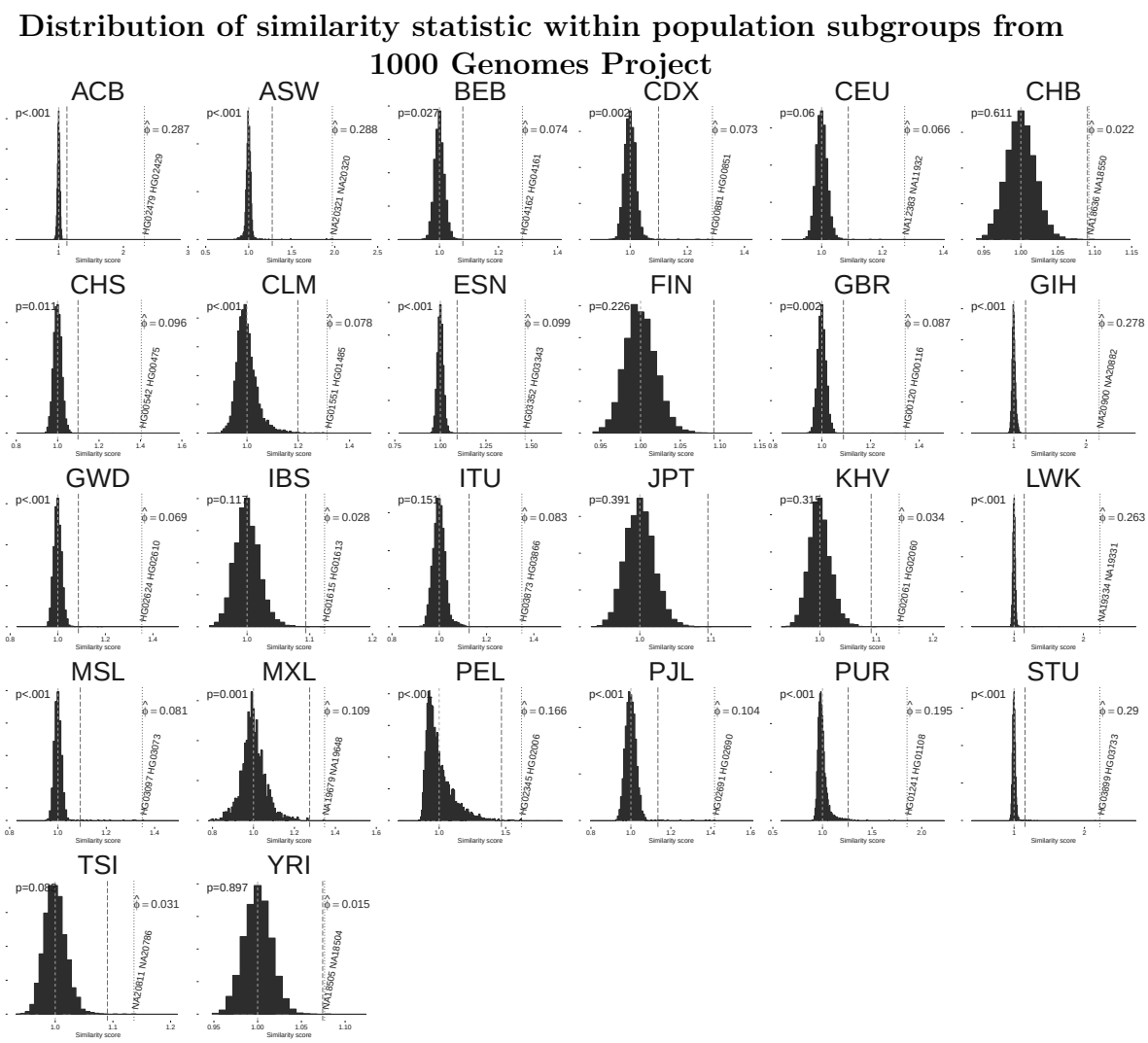


Figure 3.1: Distribution of similarity coefficients for each of the 26 populations in the 1000 Genomes Project. Homogeneous populations lacking cryptic relatedness should be expected to exhibit distributions centered around 1 with no outliers. The red dotted vertical line on each plot indicates the family-wise$\alpha = .01$ level cutoff for $\binom{n}{2}$ comparisons. The most significant related pair is labeled for each population with the estimated kinship for that pairing indicated in blue. The p-value for the KS test for homogeneity is reported for each population. Many of the population groups do demonstrate the null behavior (e.g. JPT, KHV, FIN), however, a number of populations show the presence of extreme outliers (e.g. STU, PUR) or systematic right skew (e.g. MXL, PEL)

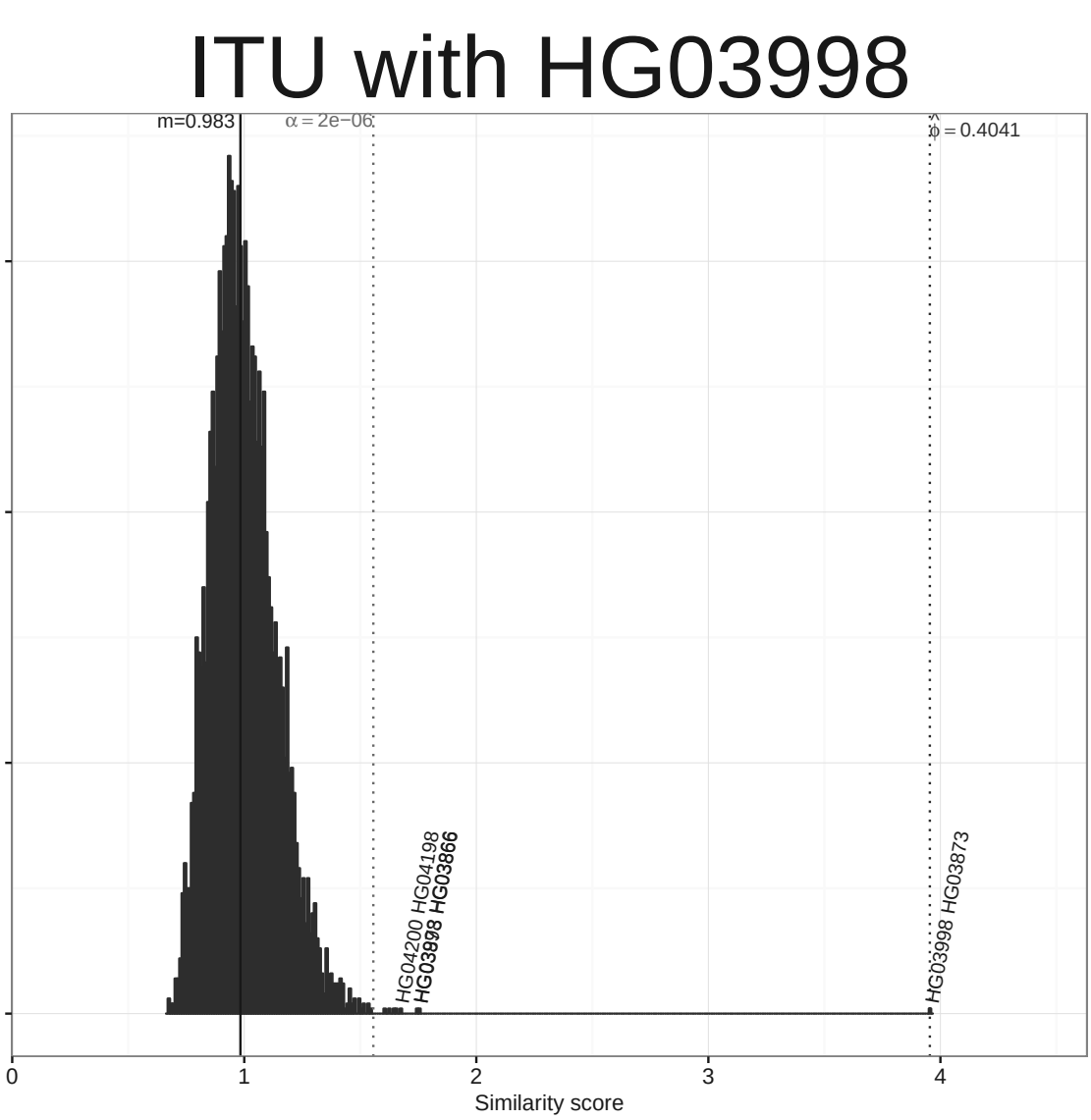


Figure 3.2: Distribution of all pairwise s statistics for population Indian Telugu from the UK (ITU) with individual HG03998 included. HG03998 is now believed to be related to HG03873, despite being labeled in the Sri Lankan Tamil from the UK (STU) population. The family-wise $\alpha = .01$ cutoff is indicated by the dotted red vertical line and the s statistic for HG03998 and HG03873 is seen as an extreme outlier at 3.97.

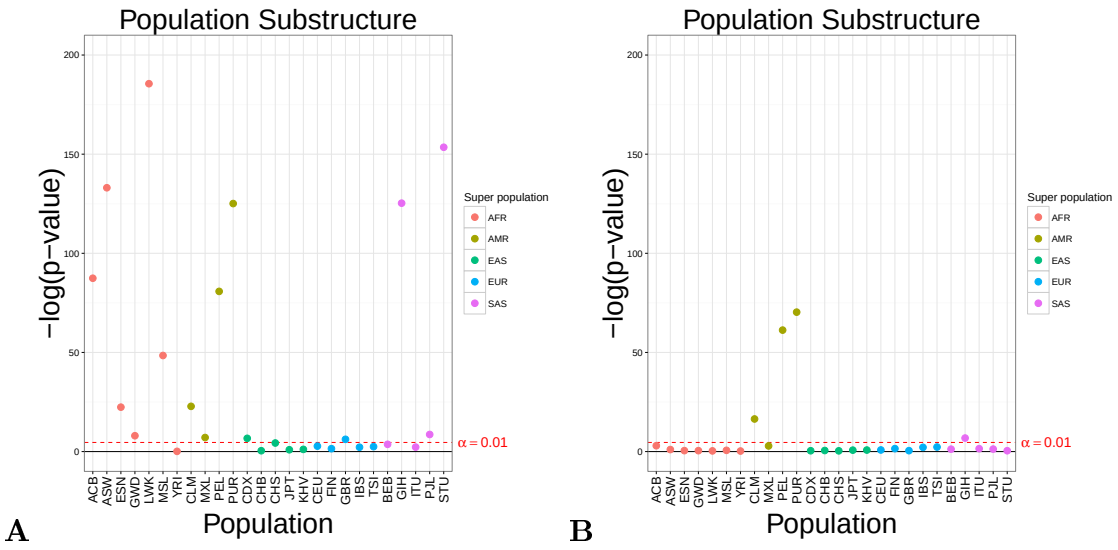


Figure 3.3: Significance of population heterogeneity in 26 populations of the TGP. Detection of population structure was found at $p < .01$ in 15 of the 26 populations using the full dataset (A). Upon removal of suspected related individuals, four populations (CLM, PEL, PUR and GIH) violated homogeneity in the relatedness-removed populations (B).

3.3.2 Population differentiation in 1000 Genomes Project

There are many methods for detecting population structure. Most commonly, Principal Components Analysis ([86,88]) is used, identifying the vectors of largest variation which ideally corresponds to the population structure. Commonly, this procedure first involves the computation of a genetic similarity matrix (GSM) via the correlation between all samples, which is followed by an eigendecomposition of that matrix. There are a number of limitations to this straightforward approach, one of which is that the calculation of a variance-covariance matrix equally weights the impact of all loci, failing to fully utilize the fact that each variant's allele frequency is informative of the value of each variant. Recently, the use of the Jaccard Index has been used to esti-

mate genetic similarity ([89]). This approach provides a higher resolution picture of the genetic landscape by exploiting the co-occurrence of rare-variants in sequencing data. STEGO directly utilizes this the differential value of alleles based on minor allele frequency by weighting variants by how unlikely such a co-occurrence would have been in a homogeneous population.

We evaluated the effectiveness of our similarity measure to differentiate populations in the TGP in both global and localized contexts. For the global scenario we used data from all 26 populations in a single analysis. In the localized scenarios, we ran 57 separate analyses corresponding to all possible pairs of populations within each of the five superpopulations. In each analysis, STEGO and covariance (PCA) were used to compute the GSMs containing all pairwise similarity scores. An eigendecomposition of each GSM was performed and each individual in the study was plotted against the top two eigenvectors for each method.

In comparing our results with those of PCA on the global scale, we achieve highly similar results depicting the two dimensional linear migrations of ancient human history. This is notable because despite a focus on separating recently related populations, STEGO remains effective at partitioning samples of more distant common ancestry as well (Figure 3.7).

Despite no loss of performance on the global scale, STEGO outperforms standard PCA when the task involves classifying individuals of recent ancestry. Focusing only on populations belonging to the same continental superpopulation, every possible pair

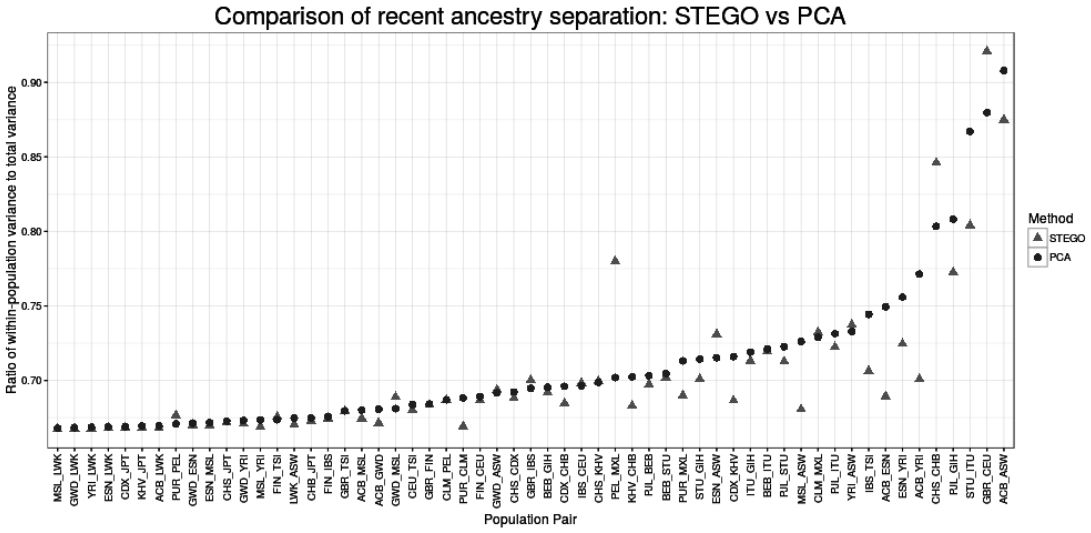


Figure 3.4: The eigendecomposition of the STEGO matrix separates individuals of populations belonging to the same continent with greater efficiency than PCA. In 43 of 57 possible within-continent population pairs, STEGO had superior separation of populations. Separation was measured as the ratio of within population variance to total variance along the first 3 eigenvectors for STEGO and PCA.

of populations were merged following the removal of suspected related pairs. This yielded 57 sets of unrelated samples in which a subtle binary population stratification existed. STEGO and standard PCA were then run on each merged dataset and the two methods were compared by computing the ratio of mean within-population variance to total variance across the first three principal components.

The results show that STEGO outperforms PCA by this measure in 41 of the 57 possible comparisons (binomial test $p < .001$) (Figure 3.4). We chose a pair of closely related populations from the 1000 Genomes Project in order to illustrate this performance. The populations Sri Lankan Tamil (STU) and Indian Telugu (ITU) have relatively small geographical separation and recent common ancestry relative to other

populations in the TGP. We demonstrate the clearer separation in comparing our method with that of standard Principal Components Analysis (Figure 3.5). We additionally explored a case in which the ratio of mean within-population variance to total variance was greater for STEGO compared to PCA, a group containing Utah Residents with Northern and Western Ancestry (CEU) and British in England and Scotland (GBR). Despite a clear trend of superior performance with STEGO, CEU-GBR is a notable exception. Closer inspection reveals that the first eigenvector from STEGO clearly isolates 11 samples exclusively from the GBR population. To our knowledge, these individuals have not previously been identified as a distinct subset of the CEU population. It is not readily apparent what features of the data are being captured here or the relative value of those features (this may be a result of population structure, relatedness, batch effect, etc.). But it is notable that all 11 samples came from the same population in the TGP. It is reasonable to infer that this subset of samples contains a disproportionate number of co-occurrences of low frequency variants, which were not detected by PCA (Figure 3.8).

The reasoning behind the superior performance in fine scale population stratification is due to the focus on rarer alleles. Rare alleles tend to be less stable over generations and become fixed at 0% with high probability. Therefore, rare alleles that are observed are more likely to have arisen recently (see subsection "Ancestry informativeness by allele frequency"). It stands to reason that these alleles would therefore be the most informative of recently related populations. By appropriately recognizing the in-

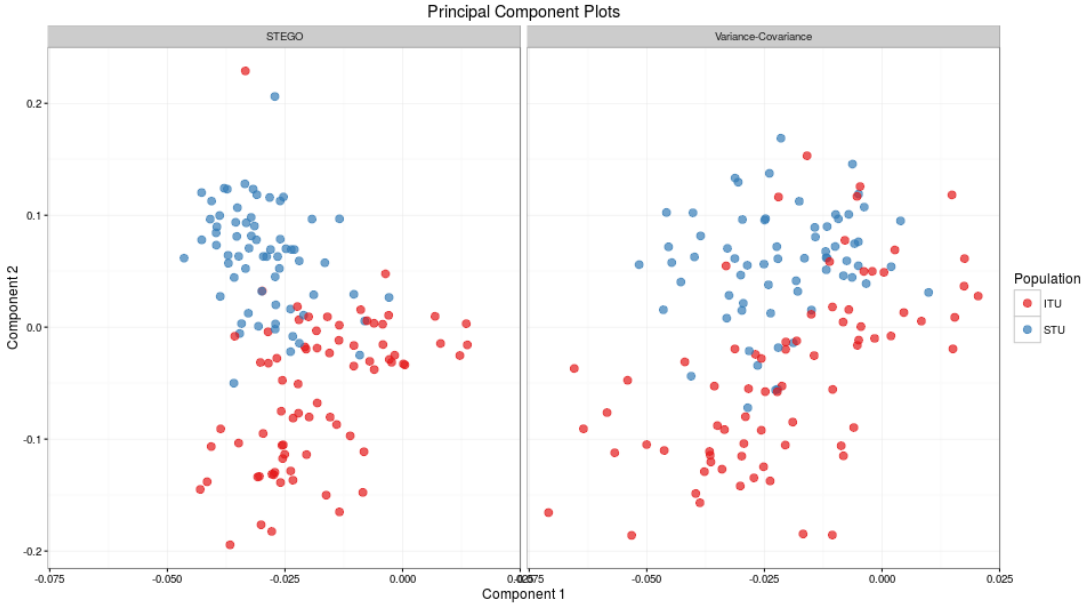


Figure 3.5: Example: ITU vs STU. Two populations of Southern Asian origin, Indian Telugu from the UK (ITU) and Sri Lankan Tamil from the UK (STU). A genetic similarity matrix was computed using STEGO and standard correlation on the same set of variants. An eigendecomposition of each matrix was then performed. These plots show the set of unrelated individuals projected on to the first two eigenvectors. We see clearer clustering by population (colored) in our method (Left) compared to standard PCA (right). This performance boost is attributed to the value added by preferentially considering genetic agreement in less frequent alleles.

creased information contained in the co-occurrences of less frequent alleles, we achieve superior separation of recently related populations.

3.4 Discussion

The ability to identify genetic outliers has well-established utility in genome-wide association studies. Many existing methods for identification of genetic associations are predicated on the assumptions that population homogeneity holds in the study. Checking for violations of these assumptions typically involves a qualitative assess-

ment without any specific concern for effect size and power. STEGO provides an analytical approach for quantitatively assessing homogeneity and a formal test for the identification of cryptic relatedness and population stratification.

Moreover, our approach involves the estimation of a GSM which, due to its preferential weighting towards rare variants, provides higher resolution for distinguishing populations which have recently diverged. As sequencing costs have plummeted and our ability to measure rare variants has increased, there will be increased demand for tools which make use of the differential informativeness of variants according to frequency. Recent work ([12]) has already demonstrated the use of pre-calculated SNP weights to infer the ancestry of samples of unknown origin, and STEGO's GSM in combination with large scale sequencing projects, such as the TGP, promises to further improve the resolution of this approach.

Several limitations exist with our approach. First, the method assumes that the variants are independent. We satisfy this assumption by performing LD sampling, but in doing so limit the number of informative markers to less than 100k, potentially omitting much of our data and reducing our power to detect heterogeneity. The choice of LD sampling method will necessarily impact the performance of the method, but due to the nature of STEGO focusing on rare variants as opposed to SNPs, the impact of LD will be limited. Additionally, with respect to the detection of population structure, we cannot design a uniformly most powerful test for structure due to the complex nature in which structure can exist. Future work will include quantifica-

tion of the specific gains achieved in controlling type I error and power in the context of rare variant association studies. Higher resolution population structure is always preferred, though the exact gains achieved in GWAS remain to be quantified.

In spite of these limitations, STEGO provides a formal, interpretable tool which is directly linked to the kinship coefficient. It provides a formal statistical test for population substructure, identifying study subjects which are related and subjects which are genetic outliers in their assigned population.

Acknowledgements

This work was supported by the National Institutes of Health [Grant numbers 1P01HL105339, T32HL007427] and Cure Alzheimer. The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

Distribution of similarity statistic within population subgroups from 1000 Genomes Project after removal of related individuals

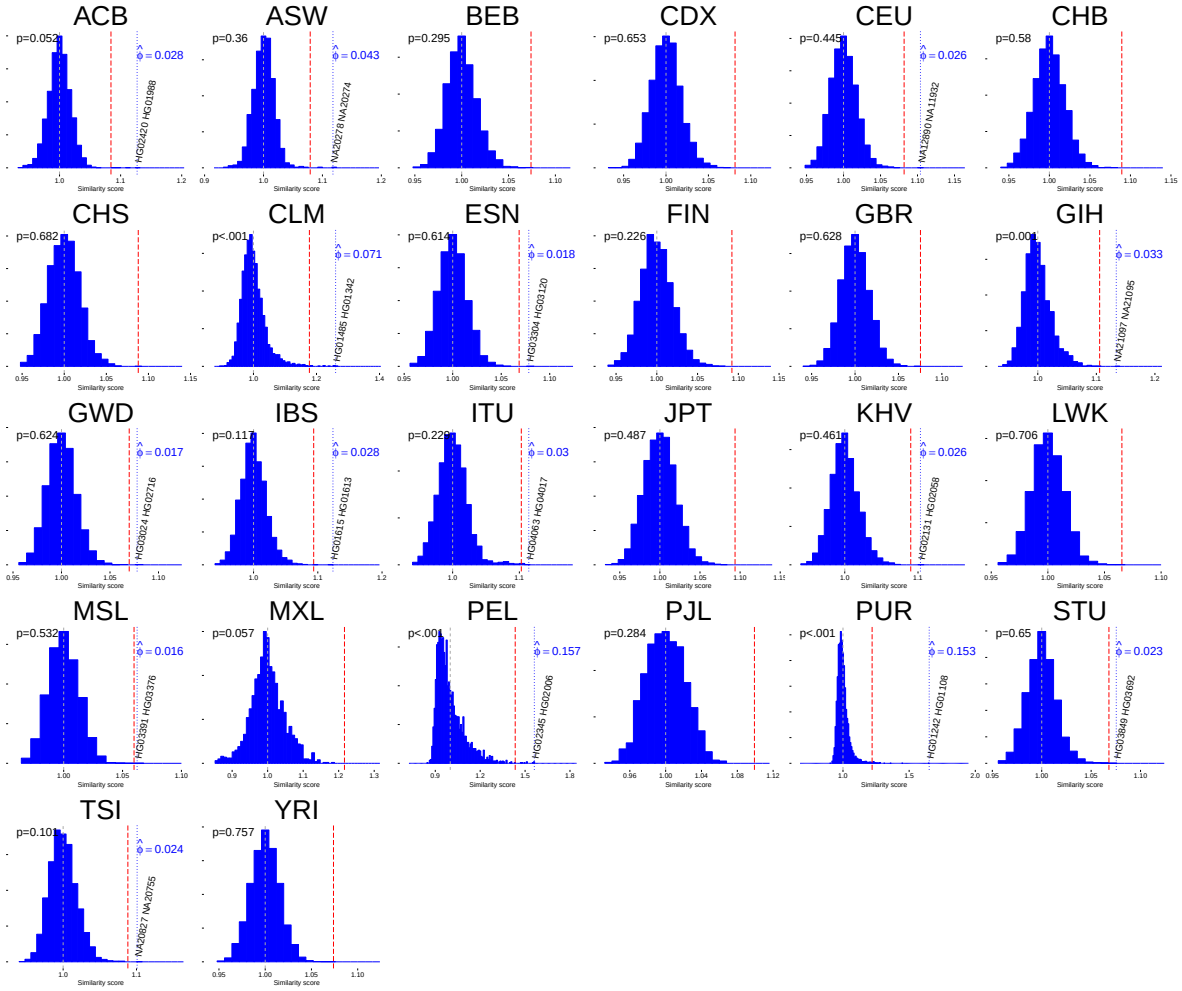


Figure 3.6: Distribution of similarity coefficients for each of the 26 populations in the 1000 Genomes Project after the removal of suspected related individuals. Homogeneous populations lacking cryptic relatedness should be expected to exhibit distributions centered around 1 with no outliers. A heterogeneous population is expected to exhibit a normal distribution centered around 1. Non-normal distributions such as right-skewed (e.g. PUR, PEL, CLM) or bimodal are indicative of population structure. The red dotted vertical line on each plot indicates the family-wise$\alpha = .01$ level cutoff for $\binom{n}{2}$ comparisons. The most significant related pair is labeled for each population with the estimated kinship for that pairing indicated in blue. The p-value for the KS test for homogeneity is reported for each population. Outliers in the absence of non-normally distributed statistics are an indication of relatedness among pairs of individuals.

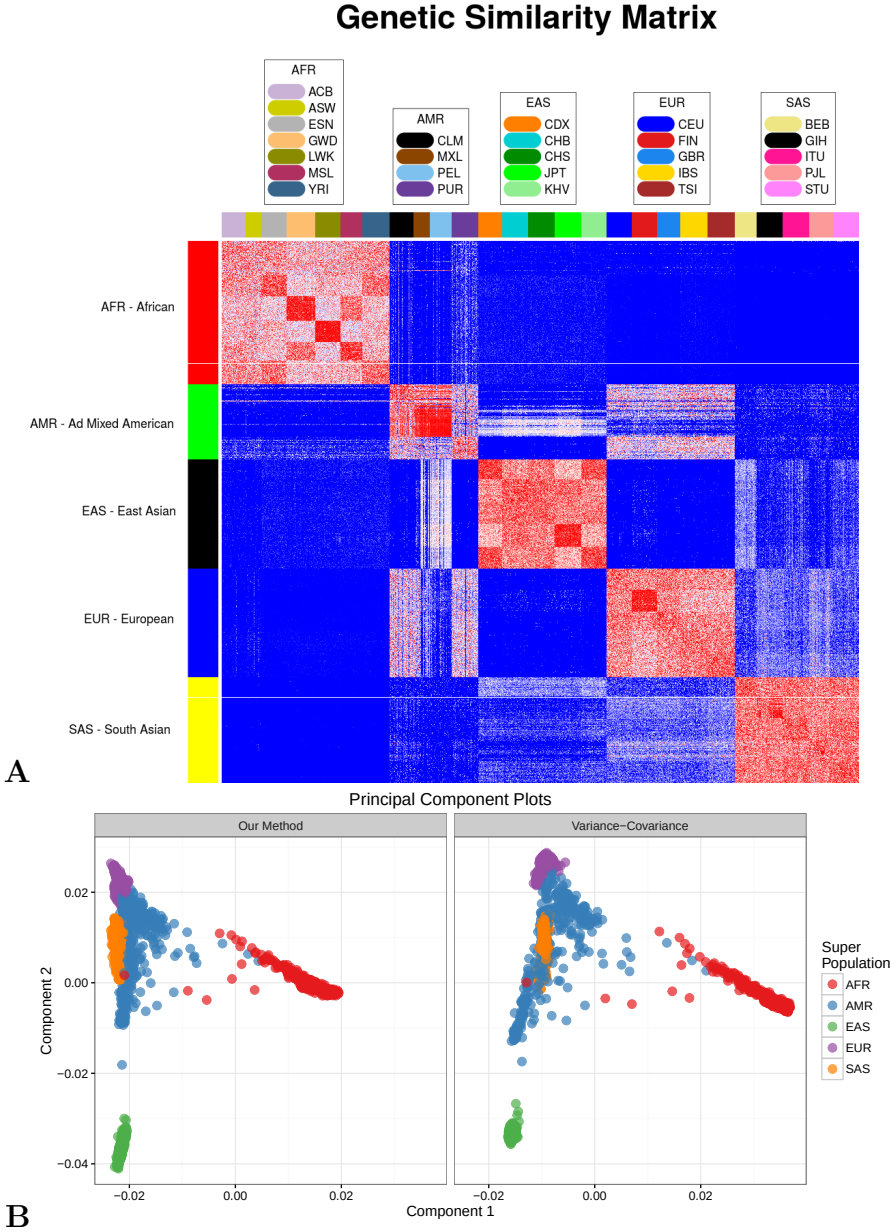


Figure 3.7: Population structure in 2504 samples from 1000 Genomes Project. (**A**) Heatmap of the GSM generated by STEGO using 80,000 LD-sampled variants. The vertical colorbar indicates membership in one of the five superpopulations, while the horizontal colorbar indicates membership in one of the 26 populations. (**B**) Projecting each individual onto the top two eigenvectors resulted in a similar 2-dimensional distribution of global ancestry. Both STEGO and PCA show similar projections which elucidate the migratory patterns of early humans.

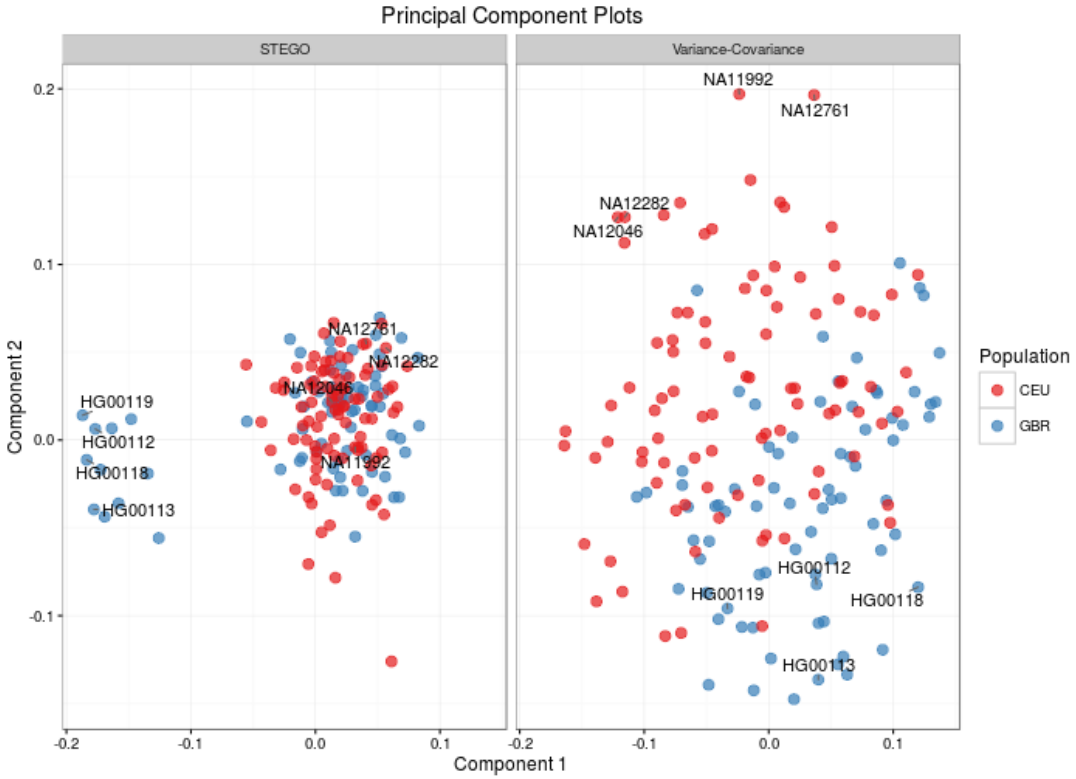


Figure 3.8: Despite a clear trend of superior performance with STEGO, notable exceptions occur. For example, by this measure, the populations GBR and CEU were more clearly divided by PCA (Right) than by STEGO (Left). Closer inspection revealed that the first eigenvector from STEGO isolates 11 samples exclusively from the GBR population. It is not readily apparent what features of the data are being captured here or the relative value of those features (this may be a result of population structure, relatedness, batch effect, etc.). But it is notable that all 11 samples came from the same population in the 1000 Genomes Project. It is reasonable to infer that this subset of samples is scientifically relevant. It most likely contains a disproportionate number of co-occurences of rare variants, which were not observed separately by PCA.

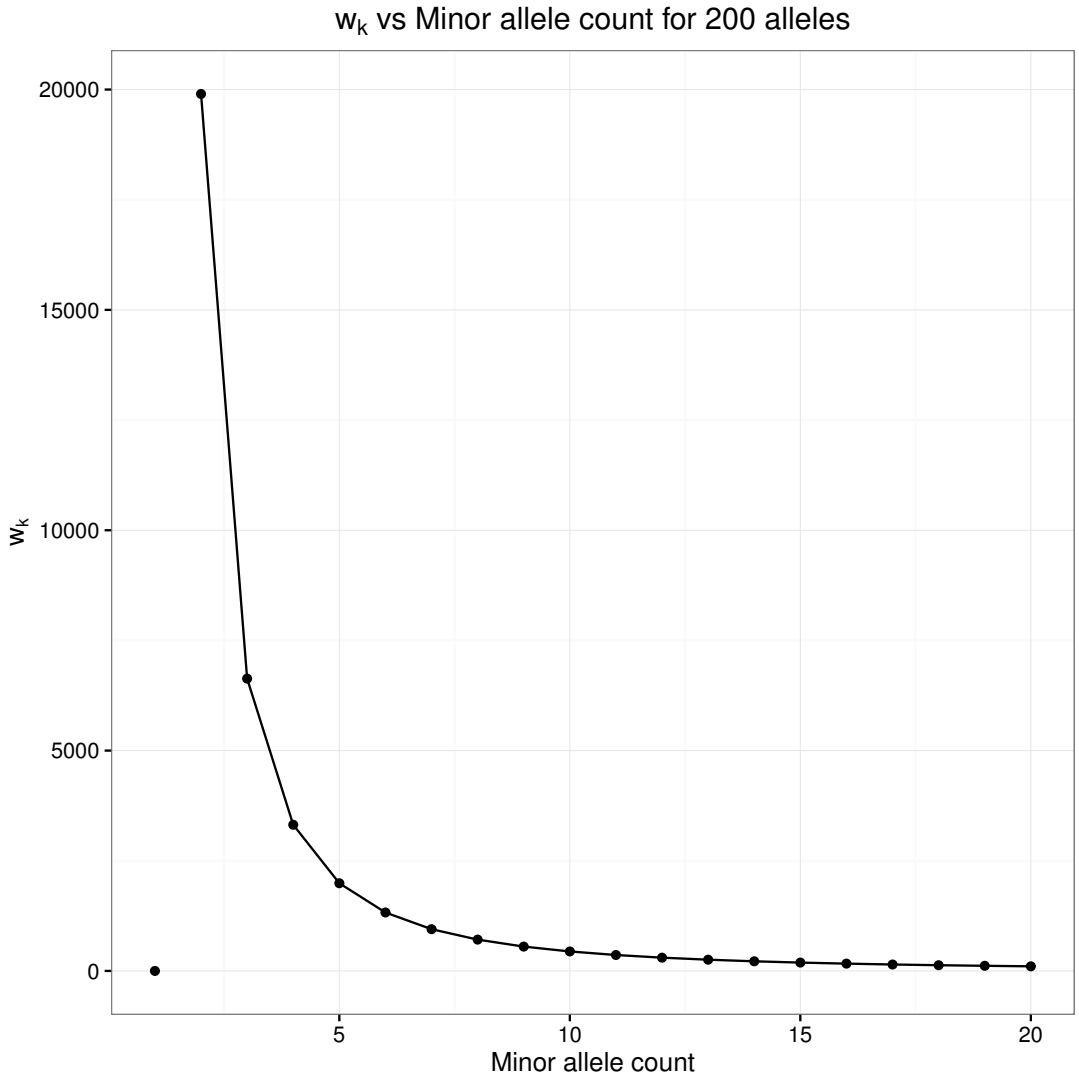


Figure 3.9: The weight factor, w_k, is shown here as a function of the minor allele count in an example of 100 individuals (200 alleles) for each locus, k. w_k is monotonically decreasing for minor allele counts greater than 1, lending greater power to lower frequency variants. Typically, there will be a minimum minor allele count such that the largest values for $w_k$are never obtained in practice

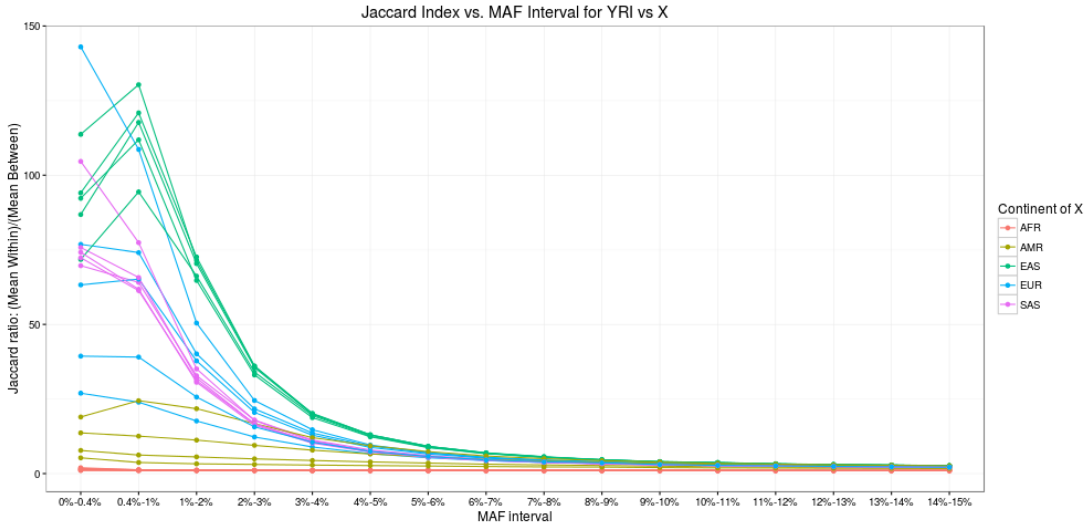


Figure 3.10: Lower frequency alleles are more informative of ancestry. For the Yoruban population (YRI), this plot compares the average unweighted Jaccard Index between individuals within group the to individuals in all other populations of the 1000 Genomes Project. When filtering by each minor allele frequency, we observe that low frequency alleles create the strongest separation between populations. This trend holds true for all but the lowest interval (0-0.4% MAF), likely owing to a tradeoff between rare variant informativeness and quality control reliability.

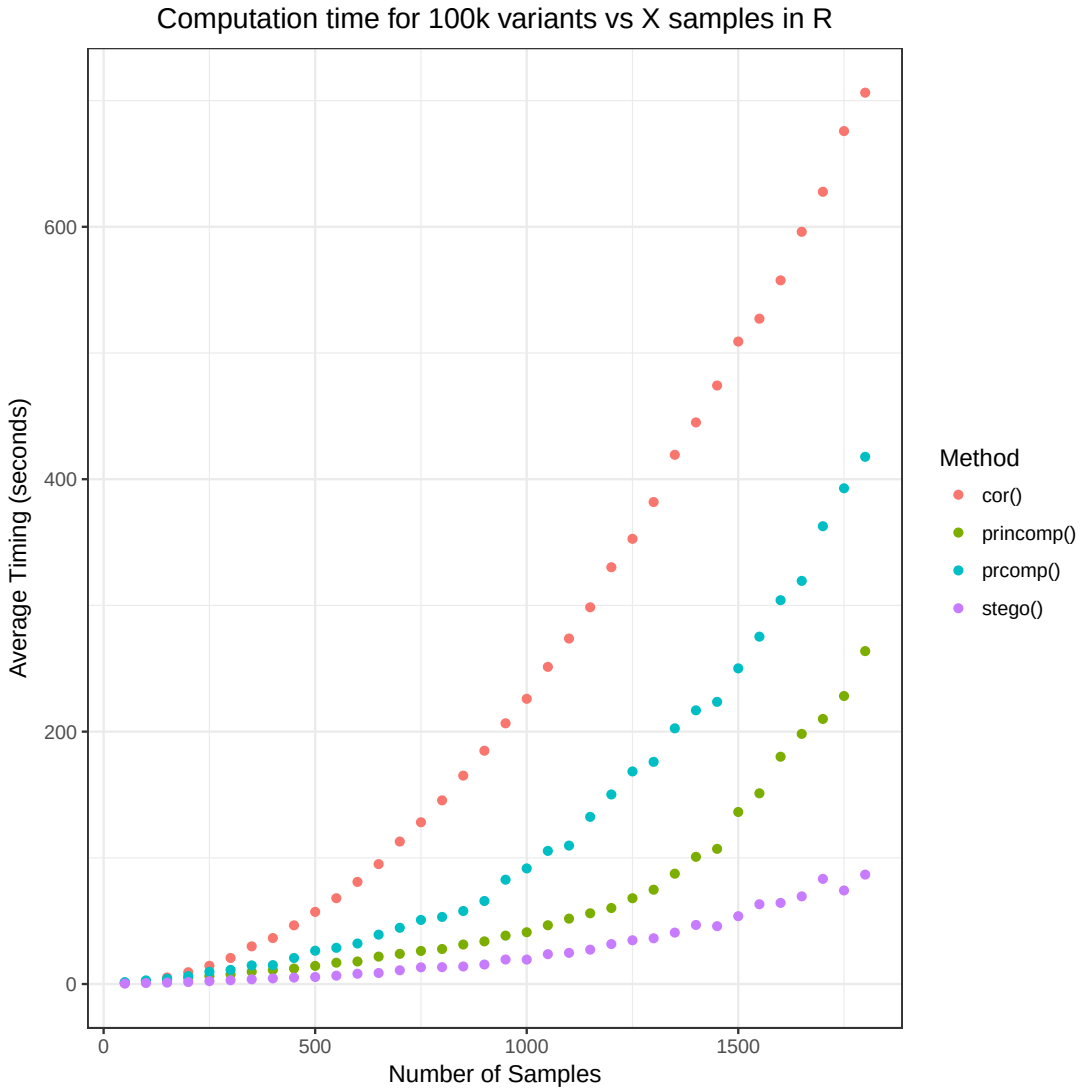


Figure 3.11: Average running time of STEGO is compared to the default implementations of cor() and princomp() in R. For each sample size on the x-axis, 100,000 variants were randomly generated across the samples. R functions for STEGO, correlation, and two implementations of PCA (**prcomp** and **princomp**) were run 10 times on each simulated dataset. Each of these methods has asymptotic computational complexity of $\mathcal{O}(pN^2)$, and we observe a consistent speed improvement of approximately 3x for STEGO compared to **princomp()**. This improvement scales linearly with increased number of variants, which is most appealing for large whole genome sequencing studies involving thousands of subjects and millions of variants.

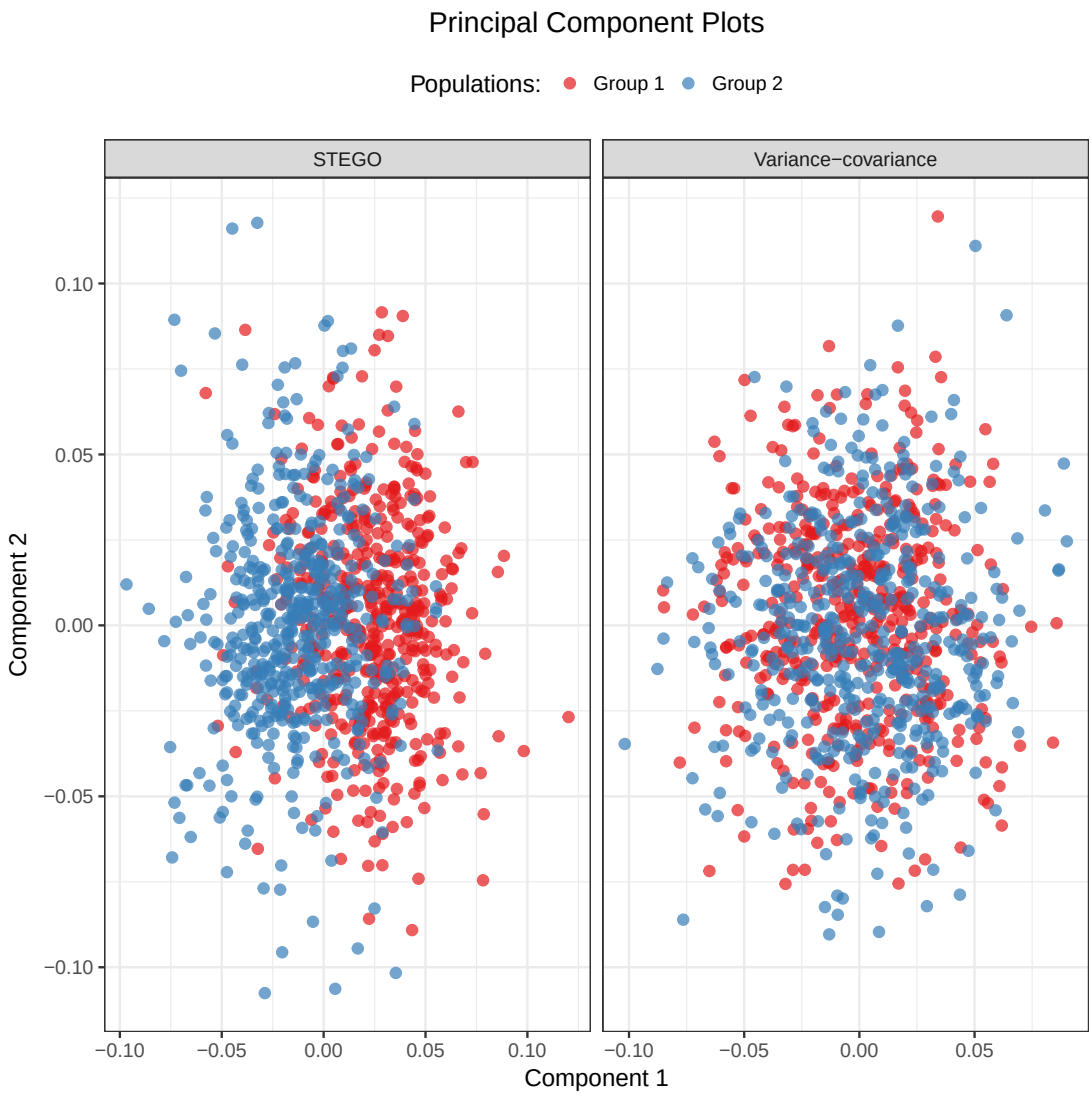


Figure 3.12: Principal Component plots for two methods for generating the genetic similarity matrix. On the left, the GSM is generated via STEGO and on the right the GSM uses the normalized covariance matrix. The STEGO method makes more efficient use of the more ancestry informative rare variants, providing a higher resolution separation of our two closely related populations. Across the first two components the ratio of within-population variance to total variance for STEGO vs variance-covariance is .81 and .99, respectively.

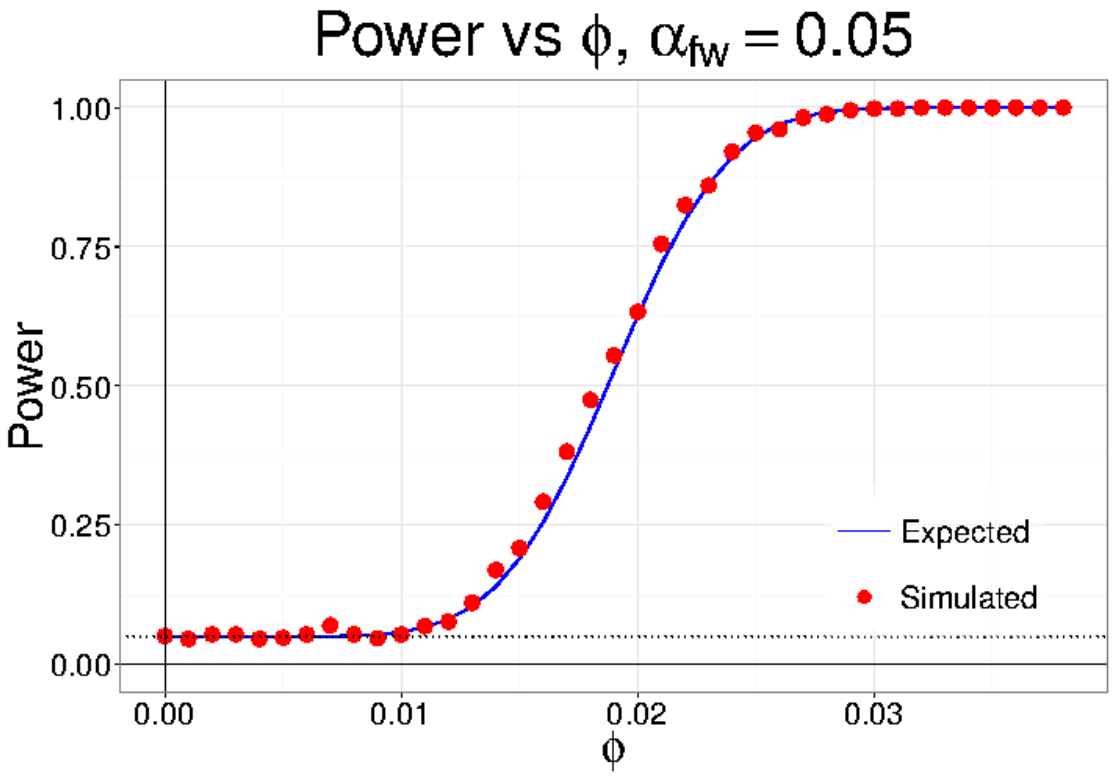


Figure 3.13: The probability of rejecting the null hypothesis given a simulated set of 301 homogeneous individuals containing a single related pair with coefficient of kinship, ϕ. The simulated power curve aligns with the analytically derived expectation demonstrating the clearly defined power of the method.

Population	Super Population	Structure	Cryptic Relatedness
ACB	AFR - African	NO	NO
ASW		NO	**YES**
ESN		NO	NO
GWD		NO	NO
LWK		NO	NO
MSL		NO	NO
YRI		NO	YES
CLM	AMR - Ad Mixed American	**YES**	**YES**
MXL		NO	NO
PEL		**YES**	**YES**
PUR		**YES**	**YES**
CDX	EAS - East Asian	NO	NO
CHB		NO	NO
CHS		NO	NO
JPT		NO	NO
KHV		NO	NO
CEU	EUR - European	NO	NO
FIN		NO	NO
GBR		NO	NO
IBS		NO	NO
TSI		NO	NO
BEB	SAS - South Asian	NO	NO
GIH		**YES**	NO
ITU		NO	NO
PJL		NO	NO
STU		NO	NO

Table 3.1: Presence of population structure and cryptic relatedness detected in each of the 26 populations in the 1000 Genomes Project. STEGO was run separately on each population group following the removal of suspected related individuals. Population structure was defined as a significant ($p < .01$) Kolmogorov-Smirnov statistic comparing the observed test statistic distribution to that expected under the assumption of homogeneity. Cryptic relatedness was defined as those populations containing at least one pair of individuals with estimated kinship $\hat{\phi} > \frac{1}{32}$ and statistically significant ($p < .01$) kinship after multiple testing correction.

Eliminate all other factors, and the one which remains must be the truth.

-Sherlock Holmes, *The Sign of Four*

4

Batch effect on covariance structure confounds gene coexpression

Systemic biases associated with gene expression experiments, batch effects, have been known to induce spurious associations and confound differential gene expression (DE) results. Commonly, to address the effects of batch on DE, methods have been developed to adjust expression values such that the mean and variance of each gene is conditionally independent of a set of batch covariates. To date, these methods

have not addressed the potential for differential coexpression (DC) across confounders. While this is of lesser concern in the context of DE, analyses that utilize a gene coexpression or correlation matrix will continue to see confounding due to batch effects even when applying standard batch correction techniques.

In this article, we demonstrate the persistence of confounding at the covariance level after standard batch correction using simulation and biological examples. We present an approach for computing a corrected gene coexpression matrix, called Coexpression Model Analysis (CMA), based on the estimation of a conditional covariance matrix. CMA estimates a reduced set of parameters that express the coexpression as a function of the sample covariates and can be used to control for continuous and categorical covariates. The method is computationally fast, and makes use of the inherently modular structure of features commonly found in genomic analyses.

4.1 Introduction

High-throughput data generation, including RNA-sequencing and microarrays, have revolutionized molecular biology. These technological advancements allow for the measurement of tens of thousands of gene expression patterns at once, giving us a window into the molecular activity of living cells. But as promising as these data generation methods are, the deluge of data that has arisen from these tools has revealed an extraordinarily level of complexity inherent to cells. Muddying the links between the information detected at the gene level to the higher level observations of phenotypes.

Plummeting costs have lead to increased accessibility of high-throughput genomic assays and with that we gain ability to investigate numerous hypotheses simultaneously. At the heart of most genomic studies is the analysis of the manner in which the biological variability of genomic features, such as RNA expression, differs in the context of phenotypes and/or other genomic features. We hope that understanding the joint distribution of gene expression, conditional on phenotypes, will lead to an understanding of the core biology. However, it can be difficult to distinguish which associations are driven by real biological mechanisms and which associations are observed because of confounding by undesirable batch effects or other extraneous experimental variables. It is critical to address this confounding in order to reduce the probability of false positive results.

In the context of gene expression studies, the measurement of biological sources of variation RNA abundance are typically of interest. Commonly, observed variation is the result of technical artifacts that may confound associations between experimental groups and gene expression[65,59].

Batch effects are known to come from many sources. Some sources are obvious, such as the array platform or the experimental reagents used, but others may be more unexpected. Timing, ozone[33], technician and lab humidity have all been identified as sources of unwanted variation and undoubtedly many other sources remain undiscovered[98]. In other words, variation attributed to batch is virtually unavoidable. Ideally, experimental design will allow for single-batch experiments to be performed, but

studies may be too large for this to be practical. Randomized batch assignment is recommended as an experimental design[20], but many investigations involve the use of publicly available data (e.g. Gene Expression Omnibus, Genomics Data Commons) for which no randomization is possible.

A common way to approach the batch correction problem is to consider the model $G_{ij} = \alpha_j + X\beta_j + B\gamma_{ij} + \delta_{ij}\epsilon_{ij}$, where G_{ij} is the gene expression of gene j for sample i, X is the design matrix, β_j is a vector of regression coefficients for gene j for the columns of X. The next two terms specify the additive and multiplicative impacts of batch. B is an matrix of indicators for each of the batches, and γ_j is a vector of additive batch effects on gene j. ϵ_{ij} is the error term and δ_{ij} is the multiplier of that error term. Controlling for batch necessarily involves estimating the impact of batch on the mean expression and the variance of that expression, specifically γ_{ij} and δ_{ij}, for each gene. It is generally not known what mechanism for batch effect is at fault for a particular study and consequently, it is unknown which set of genes and the magnitude of the effect on those genes. Therefore, without knowing which features are susceptible to batch effect, it is typical to estimate γ_{ij} and δ_{ij} for each separately gene in a study.

Despite widespread literature published regarding the identification and control of confounding due to batch effect[11,6,66,56,82], batch effect correction has focused on adjusting for the effects of batch on gene expression mean and variance at an individual level. For example, ComBat[56] uses an empirical bayes approach to estimate

the mean and variance parameters for each gene and then computes an adjusted gene expression that controls for these effects. Another approach, Surrogate Variable Analysis[66], uses a combination of measured covariates and singular value decomposition to identify unknown sources of variation. These variables are estimated and their effects regressed out of the gene expression matrix. These approaches amount to a location-scale adjustment that is critical for promoting the conditional independence of gene expression with batch. For the purposes of differential gene expression analysis, this approach is reasonably effective for both microarrays and RNA-seq data[20].

However, as our understanding of genomics grew, we recognized that finding differentially expressed genes do not give a complete picture of relationship between the transcriptome and phenotype. Cellular states involve the complex combination of numerous biological processes that are characterized by the behavior of large sets of interacting genes. This understanding has lead to increased interest measuring gene coexpression, the degree to which the expression between two gene is correlated, to gain an understanding of the network biology where simple differential gene expression falls short. Analogous to differential expression and complementary to network inference, we are interested in differential coexpression - the change in gene correlation across experimental conditions[38].

The difference between differential coexpression analysis and differential expression is that we focus on the pairwise joint distribution of genes as opposed to the marginal distribution of each gene. Essentially, we assume that genes that are functionally re-

lated will exhibit a correlated expression pattern across a set of experimental conditions or samples - a ”guilt by association” premise. Significant association between groups of genes may indicate a common functional interaction. With this in mind, a natural goal is the identification of those genes that are differentially correlated. Gene pairs or gene sets that gain or lose their common expression pattern across experimental conditions may implicate the biological pathways or functional mechanisms that drive a particular phenotypic change.

Many methods have been proposed in for differential coexpression analysis, most commonly in the context of gene network inference[54]. Often, these proposed algorithms start with the computation of correlation matrices from the gene expression data[106,61,63,44,105,15,104,113,2]. There is an assumption, often implicit, that the gene expression data lacks heterogeneity or has heterogeneity sufficiently corrected. The biases that persist in the coexpression matrix when homogeneity is violated are rarely discussed or considered in the literature. Each of the methods that use correlation matrices would benefit from estimated coexpression that has had the impact of batch effect reduced compared to a Pearson correlation alone.

In estimating coexpression matrices, standard batch correction is critical[39], but not sufficient. Location-scale confounding on gene expression will reduce power and bias results. This will inevitably lead to highly significant, but biologically meaningless associations between large volumes of genes. Though the common batch correction practices help mitigate this problem, they fail to remove the impact of the type of batch

effect that causes differential coexpression patterns among genes in the absence of differential expression. Current methods treat batch effect as acting on the marginal distribution of each gene and ignore the possibility of changes to joint distributions. While some impact on the joint distribution is addressed by removing the impact of differential means and variances across batches it is insufficient if the covariance itself is associated with batch.

It is easy to conceive of scenarios where this phenomenon plays out. For example, different experimental protocols across batches may induce a coexpression difference by preferentially sampling cells with certain active biological pathways for cell cycle or stress response. But even simpler, batch effect for coexpression may be introduced merely by differential biological variability. To illustrate this, recall that correlation is roughly interpreted as the square root of the proportion of total variability explained by true relationship between the genes (as opposed to other sources of error and variance). Then two genes which are functionally related will only be detected as such if there exists meaningful biological variability in both batches. For example, if the two genes are consistently in the same expression state across samples, their correlation will be near zero despite their interaction. Greater variability in one batch compared to another will lead to differences in ability to capture coexpression. Subtle differences in protocol that lead to differences in biological variability can not be removed with standard batch correction methods.

The demonstration in Figure 4.1 shows two examples of uncorrected batch effect

(left) impacting two genes in a study. In the top row, batch effect alters the means and variances of the two genes (location-scale model). In the bottom row, the means, variances *and* coexpression is impacted. Upon application of ComBat (Right) to the uncorrelated genes, the two genes become independent as desired. However, when applied to the conditionally coexpressed case (Bottom row) we continue to observe differential coexpression across batches.

Though methods exist to consider the correlation of genes in the presence of location-scale batch effect, no method exists that further allows for the coexpression itself to be a function of batch. Similarly, no method currently available returns a corrected coexpression matrix rather than a corrected expression matrix.

While the impact of ignoring pairwise interactions may be negligible for simple differential gene expression analyses, it is the differential coexpression patterns, rather than differential expression, that are widely considered in the field of network inference[25,11,37]. The impact of confounding due to differential coexpression in batches remains critically unexamined.

In order to solve this problem, we need to create a model that describes the coexpression matrix as a function of the experimental conditions and batches. Classical regression models predict the expectation of a response variable as a function of a set of predictors, but in coexpression analyses, we are interested in the covariance. Some work has recently been published on the subject of modeling the covariance matrix[51,118], but little has been studied in high dimensional or biological settings.

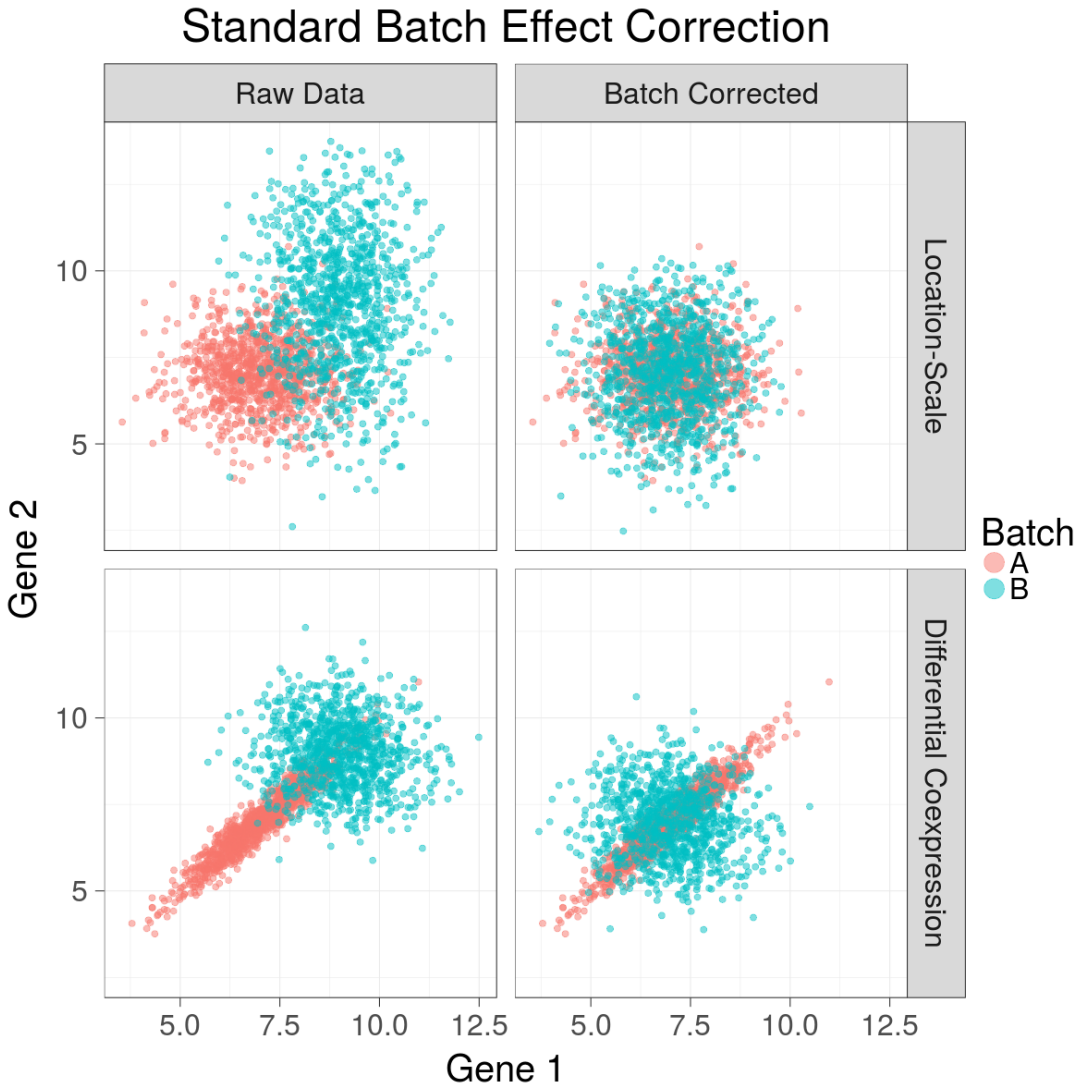


Figure 4.1: In this toy example, we demonstrate which artifacts standard batch correction is capable of correcting and which artifacts will remain. In A-D, we show plots of two example genes before (left) and after (right) correction, colored by their batch. In the top row (A,B), we show a comparison of two genes which are conditionally independent and demonstrate that location-scale batch correction appropriately removes the marginal dependence between the genes. In the bottom row (C,D), we show two genes that are conditionally coexpressed and illustrate that batch correction may help mitigate the measured coexpression, but the resulting coexpression is still a function of the batch membership. Importantly, when comparing coexpression matrices, differing batch proportions will bias the differential coexpression. In simulations we demonstrate that in the absence of batched differential coexpression, ComBat sufficiently controls the type I error. However, when coexpression differs by batch, our false positive rate increases above the expectation of the null model.

Estimating the coexpression matrix faces at least two major challenges. The first problem is that in the case of numerous batches or continuous covariates, it may not be possible to estimate a coexpression matrix using the sample covariance matrix form, $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(X_i-\bar{X}\right)^T$, where X_i is the set of all gene expression values for sample i. Another issue is that it requires the estimation of a very large number of parameters. Given p genes, there are $\binom{p}{2}$ pairwise correlations, and each of these must be a function of the number of covariates. For most high throughput gene expression studies where $N \ll p$, we want to limit this parameter space in some way. Previous work has shown the increased difficulty in reproducing coexpression across studies[99] likely owing to the high number of parameters to estimate in noisy data. Recent work has allowed for the imposition of sparsity on the gene covariance matrix[7] or precision matrix[36], but the complexity of biological systems make sparsity an unreasonable choice and computationally burdensome to implement.

In the method we describe here, CMA, we reduce the parameter space by exploiting the modular nature of gene expression, estimating only N variables for each covariate, with each weight corresponding to a eigenvector. This collects the information from many similarly expressed genes by effectively borrowing information from similarly patterned features. This allows us to estimate gene coexpression matrix as a function of sample covariates. Our method is presented in a regression framework that allows for the inclusion of continuous and categorical covariates into the adjustment model.

4.2 Methods

4.2.1 Approach

In this manuscript we present a method for estimating the coexpression matrix by modeling the matrix as a function of the largest components of variation. Critical to our approach is the idea that although there are $\binom{p}{2}$ pairwise gene-gene relationships, the true biology can be predominantly explained by a much smaller set of variance components. One way to identify these components is to compute the eigendecomposition of the gene correlation matrix. We can then write the coexpression matrix as a function of the experimental covariates and these eigenvectors. Solving this formulation by minimizing the squared error will yield a set of parameter estimates from which we can compute corrected coexpression estimates.

Consider a set of N samples with q covariates measuring gene expression across p genes. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})^T$ 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.

In multivariate regression form we can express this as

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

where β is a $q \times p$ matrix of coefficients.

Equivalently,

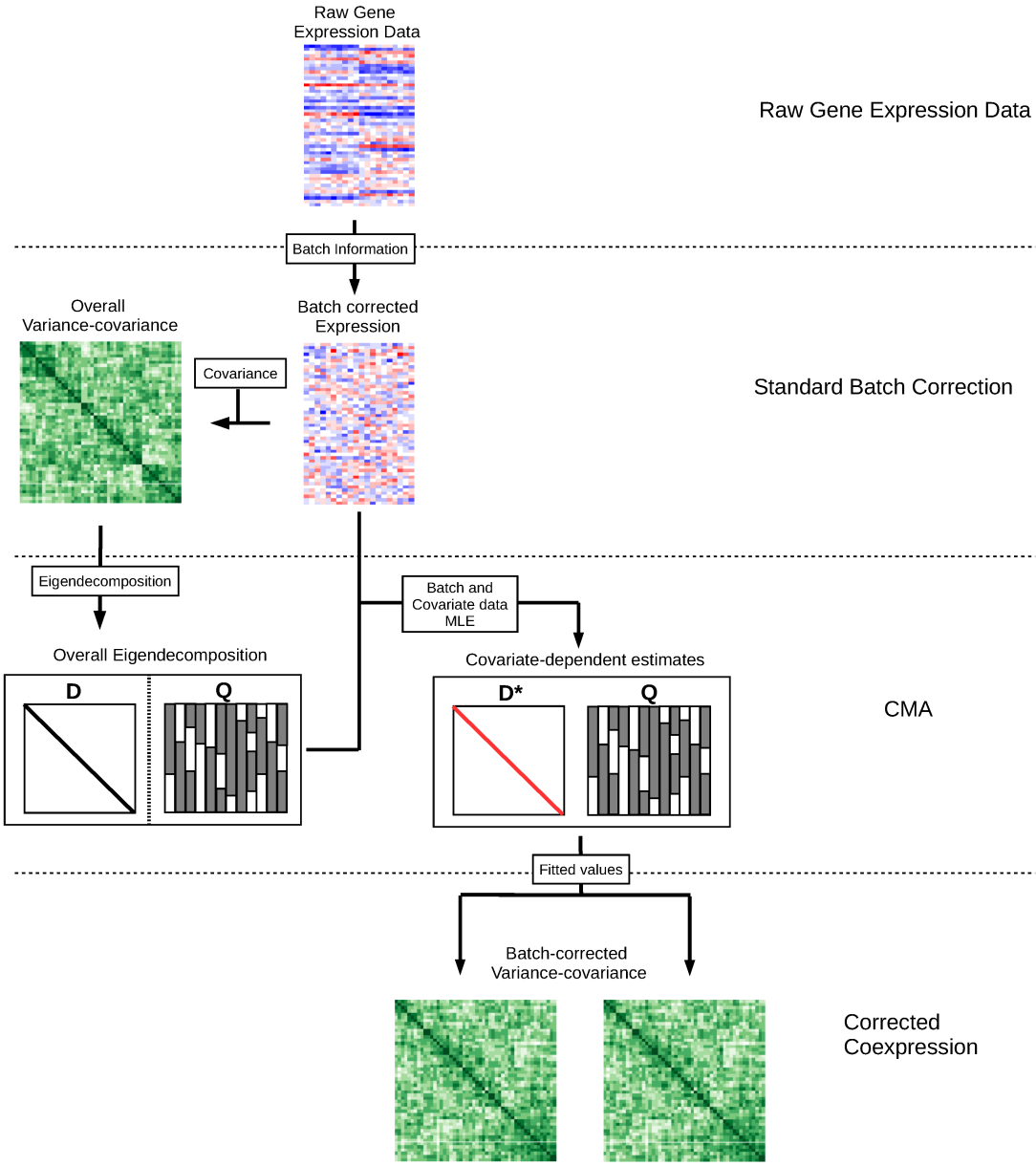


Figure 4.2: Workflow of CMA. CMA begins with a raw or normalized gene expression dataset. (1) Standard batch correction (ComBat) is applied to remove location-scale batch effect. (2) The overall coexpression matrix is calculated. (3) An eigendecomposition of the overall coexpression matrix is computed. The eigenvectors from this decomposition are then used to re-estimate "pseudo-eigenvalues" that minimize the coexpression error from the batch corrected expression data. (4) Fitted values obtained from this estimation, in combination with the eigenvector matrix, $\mathbf{Q}$, are used to estimated covariate-dependent coexpression matrices such as for batch corrected network inference or differential coexpression analysis.

$$\mathbf{G} = \beta^T \mathbf{X} + \mathbf{E}$$

where $\mathbf{G}$, $\mathbf{X}$ and $\mathbf{E}$ are each matrices with column i corresponding to $\mathbf{g}_i$, ϵ_i and $\mathbf{x}_i$, respectively.

Here, we make the usual multivariate assumption for $\mathbf{E}$ that the rows $\epsilon_i, \ldots, \epsilon_N$ are conditionally independent, and follow distribution, $MVN_p(\mathbf{0}_p, \Sigma_i)$. Notably in this paper, the covariance of ϵ_i differ according to i.

Estimating the covariance structure for a set of p genes typically involves computing the sample covariance matrix, S, with entries $s_{jk} = \frac{1}{N-1}\sum_{i=1}^{N}(G_{ij} - \bar{G_{\cdot j}})(G_{ik} - \bar{G_{\cdot k}})$. However, as is typical in high-throughput settings, $p \gg N$, producing an estimated covariance matrix $p \times p$ with column rank $\leq N$.

To address this "curse of dimensionality", numerous methods have been proposed. One might use a series of LASSO regressions to estimate parameters in the inverse covariance matrix[76], or perform penalized maximum likelihood estimation with the penalty on the inverse covariance matrix[5,114,36]. Each of these approaches imposes sparsity on the precision matrix, effectively assuming a large degree of conditional independence between genes. More recent work has explored imposing sparsity on the covariance matrix itself, rather than the precision matrix[7], which allows us to assume widespread marginal independence of genes.

The approach we take here involves estimating a covariance matrix Σ_i which de-

pends on the batch and experimental design features of sample i. An estimate of Σ_i that allows all elements of the matrix to vary freely can be obtained by separately estimating the covariance matrix for each unique row of $\mathbf{X}$. However, this approach in impractical for a large number of categorical covariates or any continuous covariates. Additionally, it neglects the information in other samples and other genes which can be used to gain a better estimate of the coexpression. Given that groups of genes often behave in distinct patterns, it is inefficient to estimate coexpression values for every pairwise combination of genes.

Instead, we approach the problem by making use of the fact that genes commonly behave in coexpressed modules, and that the dimensional space is effectively much smaller than p^2. To do this, we decompose the gene expression correlation matrix and find a set of eigenvectors which explain the variation. We then attempt to infer a diagonal matrix of "pseudo-eigenvalues", which minimize the square error. This procedure allows us to reduce the parameter space from p^2 to p or less while still considering the bulk of the variability in the data. Furthermore, in the application to the gene expression data, the column rank of the coexpression matrix will be $N-1$, and the number of non-zero eigenvalues will also be only $N-1$. Therefore, we need only estimate the parameters corresponding to eigenvectors with non-zero eigenvalues substantially reducing the parameter space from p to $N-1$.

Formally, for Σ_i we estimate $\Sigma_i = \mathbf{Q}\mathbf{\Lambda}_i\mathbf{Q}^T$, where $\mathbf{Q}$ is held constant as the set of eigenvectors from the full coexpression matrix. In this formulation,$\mathbf{\Lambda}_i$ is a diagonal

matrix with entries

$$\mathbf{\Lambda}_{i,kk} = \mathbf{x}_i \mathbf{\Psi}_{\cdot k} \tag{4.1}$$

where $\mathbf{x}_i$ is the predictors for sample i and$\mathbf{\Psi}$ is a $p \times q$ matrix of coefficients.

Because we don't estimate the pseudo-eigenvalues after $k = N - 1$, we set $\mathbf{\Psi}_{\cdot k} = \mathbf{0}_q$ for all $k \geq N$.

Intuitively, we can think of the parameter matrix $\mathbf{\Psi}$ as adjusting the eigenvalues as a function of the covariates to minimize the coexpression error. It is straightforward to show that in the case of a single batch and no experimental conditions, i.e. $\mathbf{x}_i = 1$ for all $i \in N$, then$\mathbf{\Psi}$ becomes identical to the vector of eigenvalues from the original covariance matrix.

4.2.2 Likelihood function

The likelihood function for a multivariate normal with mean μ and variance-covariance Σ is

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

The maximum likelihood estimation of μ is simply the vector $\bar{\mathbf{g}} = \frac{\sum_{i=1}^{N} \mathbf{g}_i}{N}$ and since μ is independent of Σ we can subtract off the row means, yielding $\mathbf{G}_i^* = \mathbf{G}_i - \bar{\mathbf{g}}$. And plugging in our index dependent covariance matrix from equation 4.1 we have

$$\mathcal{L}(\gamma) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{\frac{p}{2}} \left|\mathbf{Q}\text{diag}\left(\mathbf{x}_i\Psi\right)\mathbf{Q}^T\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\mathbf{G}_i^*\right)^T\left(\mathbf{Q}\text{diag}(\mathbf{x}_i\Psi)\mathbf{Q}^T\right)^{-1}\left(\mathbf{G}_i^*\right)}$$

where $\text{diag}\left(\mathbf{x}_i\Psi\right)$ is defined as a matrix with 0's in all off-diagonal entries and diagonal equal to $\mathbf{x}_i\Psi$.

4.2.3 Estimator

In estimating the parameters in the matrix Ψ, we may consider that each row, i, of Ψ corresponds to the vector of contributions from the i^{th} eigenvector of $\mathbf{Q}$. With $\mathbf{Q}_i$ specifying the i^{th} column of $\mathbf{Q}$ we have that $\mathbf{Q}_i^T\mathbf{Q}_j = 0$ for all $i \neq j$ and $\mathbf{Q}_i^T\mathbf{Q}_i = 1$ for all $i, j \in 1, 2, \ldots p$.

For some $h \in 1, 2, \ldots p$, we seek to find the estimates $\hat{\Psi}_h$ which minimize the squared error of the estimated correlation matrices defined as $\mathbf{G}_i\mathbf{G}_i^T$ for each sample $i \in 1, 2, \ldots, N$. By the Orthogonal Decomposition Theorem, the "error residuals" $\mathbf{Q}_h^T\left[\mathbf{G}_i\mathbf{G}_i^T - \mathbf{Q}_h\mathbf{X}_i\hat{\Psi}_h\mathbf{Q}_h^T\right]\mathbf{Q}_h$ will be minimized when they are orthogonal to the hyperplane spanned by $\mathbf{X}$. Therefore, we can set the product below (Equation 4.2) equal to the zero vector to solve for our estimator $\hat{\Psi}$.

$$\mathbf{0}_q = \sum_{i=1}^{N} \mathbf{X}_i^T \left[\mathbf{Q}_h^T \left[\mathbf{G}_i \mathbf{G}_i^T - \mathbf{Q}_h \mathbf{X}_i \hat{\Psi}_h \mathbf{Q}_h^T \right] \mathbf{Q}_h \right] \tag{4.2}$$

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

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

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

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

$$\hat{\Psi}_h = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \sum_{i=1}^{N} \left[\mathbf{X}_i^T \mathbf{Q}_h^T \mathbf{G}_i \mathbf{G}_i^T \mathbf{Q}_h \right] \tag{4.3}$$

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

Equation 4.3 provides an estimate for Ψ_h, a q-vector specifying the contribution of eigenvector h and the q covariates to the correlation structure in the N samples.

The estimate $\hat{\Psi}$ represents the least squares estimate for Ψ, which is equivalent to the maximum likelihood estimate under normal error. Given the generous assumption of a properly specified model, this estimate will be the most efficient estimator and will asymptotically converge to the true parameter Ψ.

This provides a closed form solution to our problem. Given that the computationally intensive steps involve matrix inversion, the computational complexity is $\mathcal{O}(n^3)$

or less, depending on the specific implementation. This allows for relatively fast computation of corrected coexpression matrix that is comparable to the simple Pearson correlation computation, which has similar complexity. Using a computer with Intel(R) Core(TM) i7-3630QM CPU @ 2.40GHz, and Microsoft R Open 3.2.5 linked with multi-threaded BLAS/LAPACK libraries, the R implementation of this method finished in 8.8 seconds on a dataset of 4000 genes, 400 samples and 2 covariates.

4.2.4 Corrected covariance matrix

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

$$\hat{\mathbf{S}} = \mathbf{Q} diag\left(\bar{\mathbf{x}}\hat{\Psi}\right)\mathbf{Q}^T \text{ or } \hat{\mathbf{S}} = \sum_{i=1}^{p} \bar{\mathbf{x}}\hat{\Psi}_i \mathbf{Q}_i \mathbf{Q}_i^T \tag{4.4}$$

where $\bar{\mathbf{x}}$ is a q-vector specifying the column means of $\bar{\mathbf{x}}$,

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{N} \mathbf{x}_i}{N}$$

The differential coexpression matrix between two conditions, defined in binary as

column 2 of $\mathbf{X}$, is computed

$$\hat{\mathbf{W}} = \mathbf{Q} diag\left(\mathbf{v}\hat{\Psi}\right)\mathbf{Q}^T \tag{4.5}$$

where $\mathbf{v} = (0, 1, 0, \ldots 0)_q$

4.3 Results

4.3.1 Simulations demonstrate residual coexpression batch effect after ComBat

To illustrate the presence of batch effect in purportedly corrected gene expression data, we performed an extremely simple simulation to capture the effect. We simply took a gene expression dataset of 100 samples and selected 1000 genes at random to be in batch 1 and labeled them *New Gene 1, …, New Gene 1000*. We sampled another set of 1000 genes and assigned them to batch 2, and added that data to *New Gene 1, …, New Gene 1000*. In essence, for each "simulated" gene, there were 200 total expression observations - 100 from one gene and 100 from a separate gene. Naturally, there was substantial association with batch across the dataset. This is seen in the highly significant differential expression across batches. We also compared the differential coexpression across batches by plotting the distribution of differential coexpression estimates between batch1 and batch 2. We compared this distribution to that of two randomly assigned groups to show that the absolute differential coexpression was

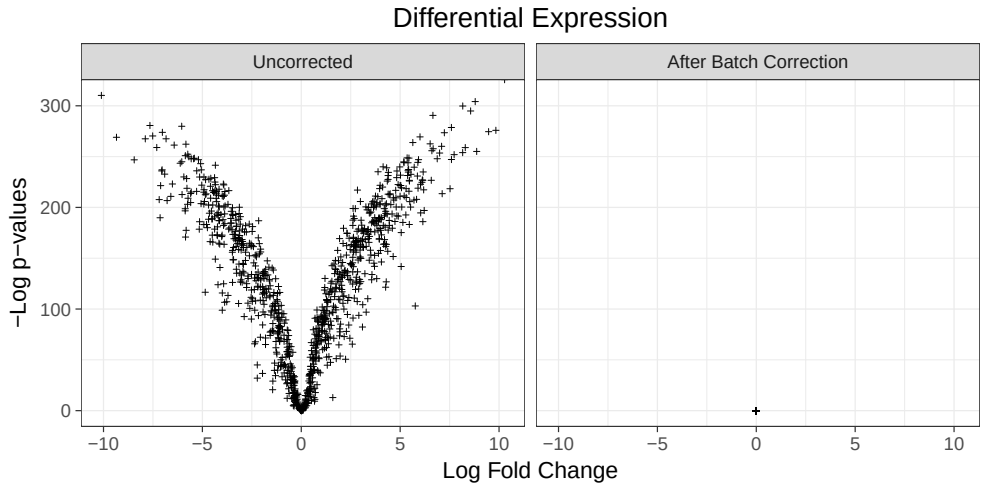
much greater across batches.

We then applied ComBat to the data, removing the effect of the batch assignment and performed the above assessments again. As expected, the differentially expressed genes virtually disappeared. Interestingly, the differential coexpression was only mildly reduced. Substantially more gene-pairs were highly coexpressed across batch-groups than across random-groups. This simple demonstration illustrates both the value of ComBat in addressing location-scale batch effect and the need for methods which address batch structure.

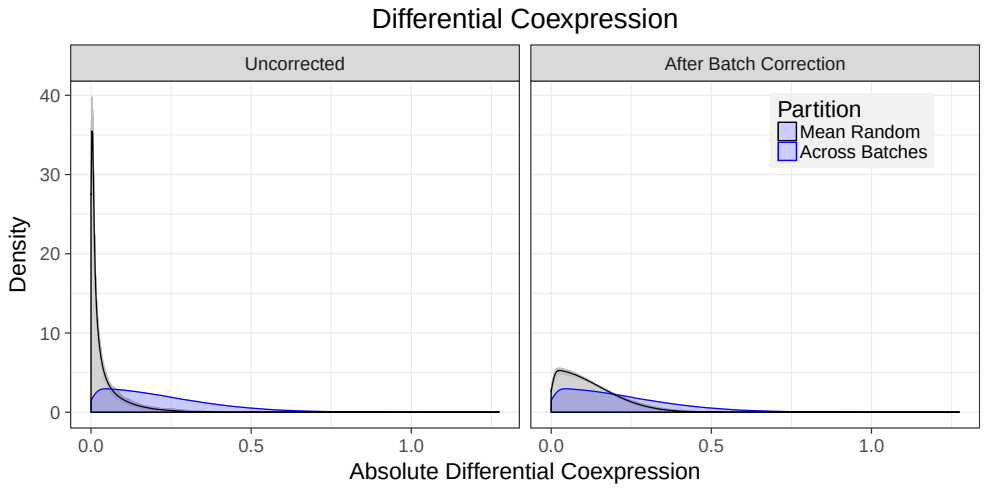
4.3.2 Improved coexpression estimates in *In Silico* analysis

We performed a simulation study to determine the relative performance of our method in identifying differential coexpression in the presence of coexpression batch effect. Gene expression for 400 samples were simulated across 4,000 genes. The simulation study contained a balance Cases/Control design with 200 samples per group. Similarly, "Batch A" and "Batch B" were each assigned 200 samples. To generate an unbalanced batch effect, 150/200 samples in Batch A were control group samples, whereas 150/200 samples in Batch B were cases.

Each gene was randomly assigned to one of 10 distinct modules, labeled A-J. Modules A,B were labeled as background modules with the coexpression pattern present in all samples, Module C was present in all Batch A samples, Module D was present in all Batch B samples, Module E was present in controls, Module F and G were



A



B

Figure 4.3: Simulations demonstrate batch correction on differential expression and coexpression. The impact of ComBat on differential expression relative to differential coexpression was evaluated using an illustrative example. A dataset was generated for 1000 genes using real data, but with batch effect induced via the replacement of half of each gene's expression values. This rewiring created substantial differential expression and coexpression, but only the differential expression was addressed completely with ComBat, highlighting the need for additional methods to address this problem.

present in cases. The coexpression pattern of all other modules were present in no samples. Within each module each gene was assigned a continuous value, γ_i, uniformly random between $-a$ and a. For case-control modules, a was chosen to be $\sqrt{0.1}$ and for all other modules a was set at $\sqrt{0.2}$. The true coexpression between any two genes was defined as $\rho_{i,j} = \gamma_i \gamma_j$. This yielded within module correlation values in the range $(-0.1, 0.1)$ for cases/controls and $(-0.2, 0.2)$ for batch and background modules. The average absolute correlation between two case-control coexpressed genes was $\rho = 0.025$ with $R^2 = 0.000625$. The average absolute correlation between two batch or background coexpressed genes was $\rho = 0.05$ with $R^2 = 0.0025$.

The simulated study was run by generating 400 samples from a multivariate normal distribution with 0-vector mean and covariance equal to the correlation matrix described above for each sample.

The eigenvectors obtained demonstrate the tendency to isolate distinct gene modules. Figure 4.4 shows this feature along with the pseudo-eigenvalue contribution of each covariate. It is important to note that the top eigenvectors do not necessarily identify genes of interest in the case-control context. The estimate $\hat{\Psi}$ is a $3 \times p$ matrix with the first 20 columns plotted in (Figure 4.4B). The i^{th} column and j^{th} row can be interpreted as the additional contribution of the i^{th} eigenvector for a 1 unit increase in the value of the j^{th} variable. This is analogous to standard regression, where we can identify the estimated mean differences associated with a change in a predictor. To identify differential coexpression for the j^{th} variable, such as case-control, control-

ling for batch we need only examine the values that deviate significantly from zero. The parameter corresponding to the case-control variable successfully finds the eigenvectors which best describe the genes differentially coexpressed across cases/controls.

We evaluated the ability of CMA to capture case-control differential coexpression relative to batch coexpression and background coexpression (Figure 4.5). For 4,000 genes there are 7,998,000 pairwise coexpression estimates of which 319,600 (4%) are considered case-control gene-pairs, and 159,600 (2%) are considered batch.

4.3.3 ComBat-corrected expression data still contains batch-associated coexpression in ENCODE

Above, we outline a theoretical basis for adjusting for differential coexpression by batch. In short, we demonstrate how this particular form of batch-effect could, in theory, lead to reduced power and biased results. However, it remains to be seen whether this purported phenomenon actually occurs in real gene expression datasets. One might hope that the impact of batch on gene expression data occurs on each gene independently, altering the distribution of expression within each batch. In that scenario, existing approaches would be sufficient for removal of batch effect and batch-associated differential coexpression would be virtually absent.

To demonstrate that differential coexpression by batch exists in the wild, we selected publicly available data from the ENCODE project. This dataset (GSE19480) contains 153 RNA-seq samples across 57820 ENSEMBL IDs from lymphoblastoid cell

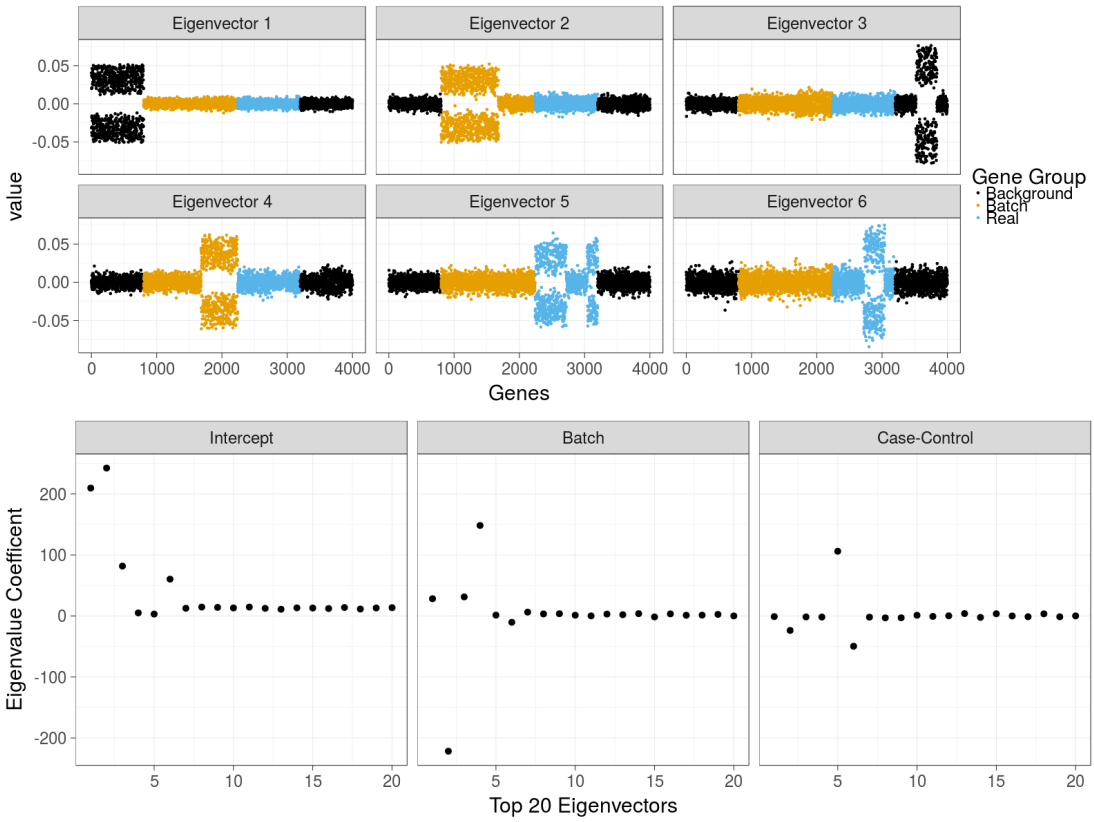


Figure 4.4: Eigenvector plots show separation of modular structure. CMA is designed to estimate sample specific coexpression as a function of the sample covariates and the overall coexpression eigenvectors. **(A)** Here we see the top six eigenvectors plotted for all 4000 genes. Each point is colored according to that gene's membership in a batch, case-control or background module. We see that the eigenvectors tend to separate along with coexpression modules. **(B)** Pseudo-eigenvalues for the top 20 eigenvectors corresponding to the three covariates (intercept, batch, case-control). Deviations from zero on the y-axis are indications of an unequal contribution of the corresponding eigenvector to the fitted coexpression estimate. Note that eigenvectors 5 and 6 have notable non-zero pseudo-eigenvalues corresponding to case-control parameter.

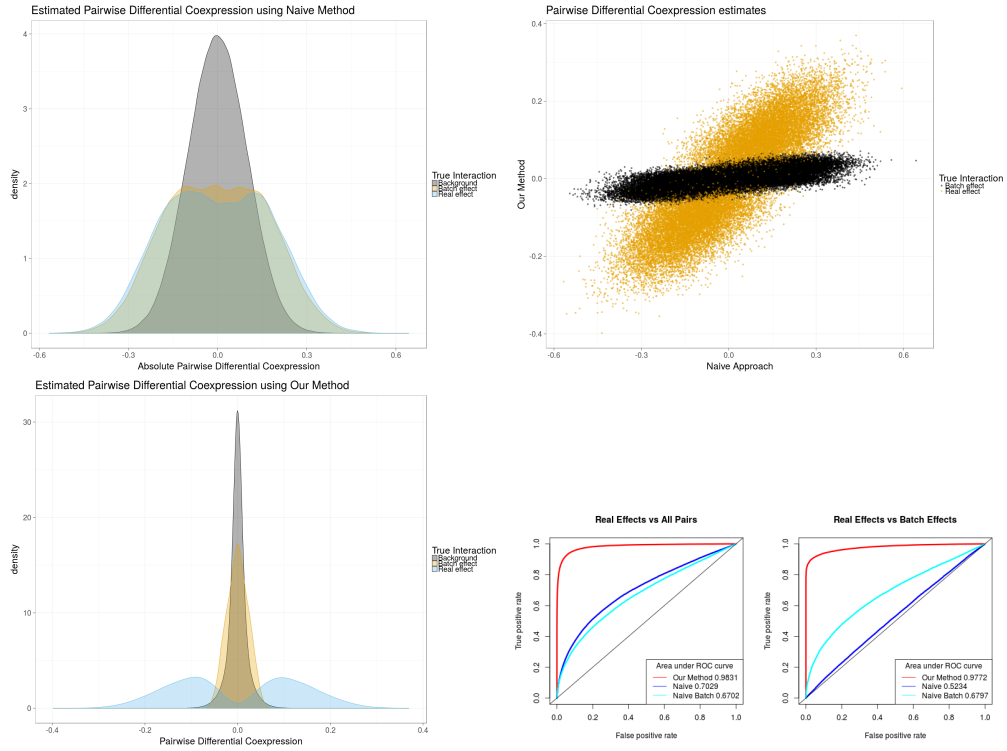


Figure 4.5: Comparison of methods for differential coexpression estimates in *In Silico* data. Fitted differential coexpression scores for standard differential coexpression (upper left) vs CMA (upper right) separated by true relationship. Pearson difference failed to generate much power to predict true coexpression compared to background and found batch effect at approximately the same rate as true effect. CMA found true effects at vastly superior rates compared to both background and batch effect. The predicted scores of non-background genes for Pearson difference (x-axis) vs CMA (y-axis) demonstrate improved ability to separate case-control effects (orange) from batch effects (black). ROC curves show the relative performance in identifying case-control genes compared to background genes (left) and batch genes (right).

lines obtained from Yoruban HapMap individuals. Reads were aligned using Bowtie[64], and counts were produced using *featureCounts* from the *Subread* program[68]. Replicates were removed, yielding samples from 63 individuals who were each sequenced at both the Yale University and Argonne National Laboratory (126 samples in total). Both centers used the Illumina Genome Analyzer II, which helps reduce, but is known to not eliminate, batch effect. These two centers represent the two batches to consider for correction. Since each batch contains RNA-seq experiments on the same group of 63 individuals, one would hope that in the absence of batch effect (or in the presence of satisfactory batch correction) that there would be minimal differential expression and coexpression between the batches.

We first ran LIMMA on the uncorrected data between Yale and Argonne and observed 495 significant (FDR<.01) genes (Figure 4.6A). We then applied ComBat to the data with Yale/Argonne as the batch. ComBat uses an empirical bayes approach to make location/scale adjustments for each gene, returning a gene expression matrix of corrected values. We then reran LIMMA using the same Center partition and observed 43 (FDR<.01) significant genes. Unsurprisingly, this procedure removed the differential expression across batches which is critical from the removal of confounding effects in differential expression.

Next we examined the distribution of differential coexpression between the two batches (Figure 4.6B). We also generated null distributions by randomly swapping the two centers for each of the 63 individuals 1000 times. Interestingly, despite the

absence of differential expression across batches, differential coexpression persists after batch correction.

4.3.4 CMA allows for separation of covariate specific modules with WGCNA in COPDGene study

Weighted Gene Coexpression Network Analysis (WGCNA) is one of the most popular network reconstruction methods in use today[61], with 1730 citations as of March 31, 2017. Its use continues to grow with 140 citations in the first 3 months of 2017. Like many other methods in the field, WGCNA begins with a standard Pearson correlation matrix of gene expression data. We were interested in whether the use of CMA could provide covariate-specific differential coexpression estimates that could integrate with WGCNA to find functionally relevant coexpression modules. While our method is motivated by the idea of removing batch, it is general enough to be applied to any confounding variable. In this application, we chose to treat three clinical covariates as confounders - sex, age and pack-years.

Gene expression data from the COPDGene study (GSE42057)[4,95] was collected from blood samples obtained from 136 subjects classified as smoker controls (42) or COPD (94) and profiled on Affymetrix Human Genome U133 Plus 2.0 microarrays. CEL data files from these microarray assays were RMA-normalized using the 'affy' package and array probes were collapsed to Entrez-gene IDs using a custom CDF[23], yielding 18,960 genes.

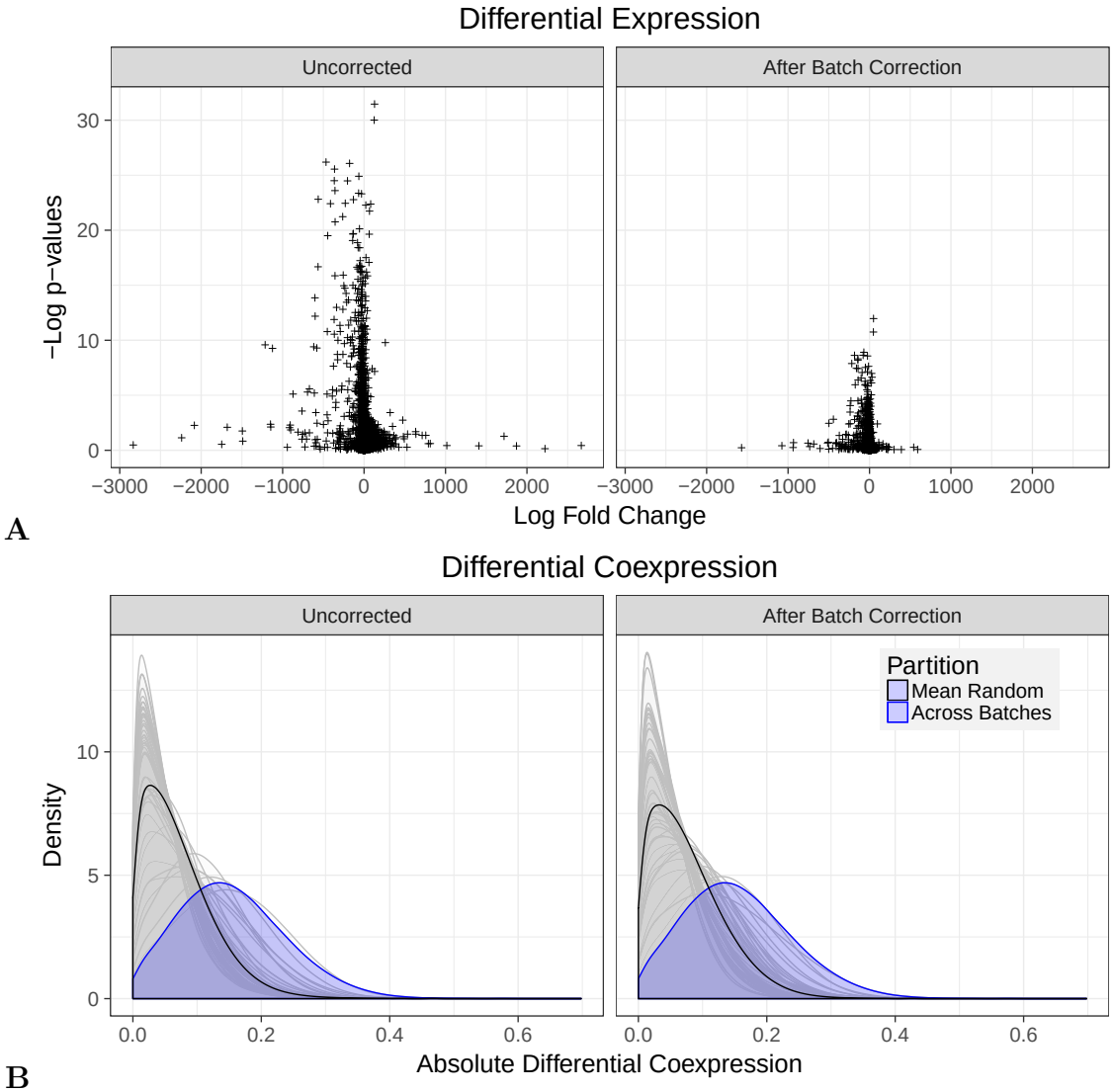


Figure 4.6: Differential expression and absolute differential coexpression in ENCODE data with batch correction. ComBat effectively mitigates the differential expression between samples run in two separate centers (**A**). However, with this same batch correction, differential coexpression continues to be strongly influenced by processing center. The lower plot (**B**) shows the distribution of differential coexpression when comparing groups that are randomly assigned (grey) compared to assignments based on batch. Despite the fact that ComBat helps mitigate differential expression between batches, these results show that batch-associated differential coexpression remains uncorrected.

Previous work applying WGCNA to this data has identified network modules associated with COPD diagnosis[79,80]. These studies involved applying topological overlap to an overall similarity score (e.g. Pearson, Euclidean, biweight midcorrelation) after standard batch correction (Surrogate Variable Analysis[66]). The similarity matrix typically does not consider sample covariates and consequently yields a collection of modules which are generally coexpressed, not necessarily differentially coexpressed. To identify which modules might be relevant to phenotypes of interest, an eigendecomposition by samples of each module is performed and the top eigenvector (eigengene) is regressed against the phenotypes and other covariates. This approach, while effective at identifying associated modules has limitations. The eigengene obtained through this method will capture the greatest axis of variation across the samples, not the greatest axis of covariation. By design, the eigengene will only be associated with a phenotype of interest if there is differential expression within the module across phenotypes. Given the wide availability of methods for differential expression analysis, the greatest value coming from the investigation of coexpression necessarily focuses on discovery of genes and modules which are not differentially expressed. In any scenario where we wish to consider differential coexpression as a potential driver of disease needs to consider these concepts.

Using a model similar to the one described by Morrow et al[80], we applied CMA to the COPDGene data and included Sex, Age at enrollment, and smoking history (measured in pack-years) as a covariates in the model.

The distribution for each of these covariates were uneven across cases and controls in this study, potentially leading to confounded results. Using Equation 4.5 we generated $p \times p$ matrix interpreted as the differential correlation for the case-control partition, holding the other variables constant. We applied a soft thresholding power of 6 and computed the topological overlap matrix, as described in[61]. Because we use a differential coexpression matrix instead as a similarity matrix, it is expected that the matrix will tend to be sparse compared to the overall coexpression. Unsurprisingly, this leads to a reduction in the strength of the modularity particularly in the background modules. The gives us the added benefit of being able to identify relatively few top modules and assume that the rest of the genes are not differentially coexpressed. For each of the covariates, including the case-control indicator, we examined the top module generated by analyzing it for functional enrichment using the R package GOstats (1.7.4)[32].

As is often a challenge in the field, there is no available benchmark for assessment of coexpression estimates. Instead, it is common to borrow information from the external sources, such as the Gene Ontology (GO) database, to evaluate a method's ability to infer known functional biology from the data.

The top differential coexpression module was found to be enriched for many biological processes involving development and morphogenesis, including top hits for anatomical structure development $\left(\text{FDR } = 2.6 \times 10^{-5}\right)$ and anatomical structure morphogenesis $\left(\text{FDR } = 1.6 \times 10^{-4}\right)$ (Table 4.1). The involvement of morphogenesis is identified

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 4.1: GO categories for differential coexpression in COPDGene identified with CMA found with FDR<0.1. In contrast with standard WGCNA, our method finds these 9 functional categories, which are independently established in the etiology of COPD.

in numerous studies of COPD and is previously been cited for its involvement in the progression of the disease[78,100]. In past studies, many variants associated with COPD have been found at chromosome 4q31, upstream of HHIP (hedgehog-interacting protein) gene[117]. Notably, the Hedgehog signaling pathway is important for the morphogenesis of the lung[18]. None of the top GO term hits, including these GO pathways, appeared in the enrichment analysis for the COPD-associated WGCNA modules for the original publication.

Even more striking about this differentially coexpressed module is that the top GO pathway - anatomical structure development (GO:0048856) $\left(\text{FDR } = 2.6 \times 10^{-5}\right)$ is the same top pathway identified in a separate study of African-Americans with COPD exacerbations[10] using DNA methylation data. The COPDGene gene expression dataset and the PA-SCOPE methylation dataset are two studies measuring the

same disease but with different populations of individuals, in a different location, using different technology to measure different biological features. It is therefore quite promising that the biological functions observed have strong overlap.

The sex covariate was not significantly (FDR<0.1) associated with any GO categories.

Discussion

This manuscript makes two important contributions to gene correlation networks. First, we identify the problem of confounding by differential coexpression, provide a theoretical basis for that artifact and demonstrate its presence in real data. Second, we propose a method for estimating coexpression matrices in the context of covariates which serve as coexpression confounders.

Incremental improvements in high-throughput data collection have dramatically increased the availability of large scale gene expression data. As we dive deeper into this data, we recognize that cellular states are rarely driven by the additive impacts of sets of suspect genes. Rather, it is the relationships, pairwise and higher, that these genes have with each other and their environment that leads to the phenotypes we seek to explain. Technological and methodological advancements in genomics allow us unprecedented ability to study these interactions. But with this new data come new statistical challenges that were not as impactful in differential expression analyses.

We argue that the batch correction methods that are designed for and are ubiq-

uitous in differential expression are important, but not sufficient, for removing unwanted variation from the data in gene coexpression. With respect to differential coexpression by batch, to our knowledge this is the first paper to address this problem.

Our proposed method uses a regression model for the coexpression matrix and reduces the parameter space by constraining the coexpression by the components of variation contained in the whole data. Future work may investigate a number of natural extensions of this approach. For example, we may wish to prespecify the $\mathbf{Q}$ matrix in some form other than the eigenvectors of the coexpression matrix. This may include a priori gene sets of known functional relevance or the eigenvectors of a separate training set.

Our results show successful estimation of coexpression when applied to a simulated dataset in the context of batch effect and identify coexpression modules in a dataset of gene expression from a COPD study that were not otherwise identified using standard WGCNA approach.

5

Conclusion

For the foreseeable future, we will continue to produce greater quantities of genomic data with improved precision at faster rates and cheaper costs. Scientists will have unprecedented access to the data that could lead to the understanding of human disease and point to potential targets for intervention. However, as French Mathematician Henri Poincare once said, "Science is built of facts the way a house is built of bricks. But an accumulation of facts is no more science than a pile of bricks is a house." It is already clear that many of the bottlenecks in the path to understanding lie not in our ability to generate but in the analysis of that data.

Perhaps the most significant bottleneck in genomic analysis comes from the recognition that biomolecular functions are extraordinarily complex. With respect to finding causative genomic features of diseases, many of the low hanging fruit have long been picked. For example, the relatively small set of heritable diseases which are adequately explained by the additive effects of small number causal variants have been mostly identified. Remaining are the vast range of complex and or rare diseases which are driven not by a single genetic risk, but by an intricate system with contributions from numerous genomic, transcriptomic, epigenomic, and environmental factors. Additionally, many of the characteristics of biological function that may be involved in disease are not even observable with single snapshots of a sample. The measurements of interest may be the way certain genomic features interact with one another, not their isolated abundances, which cannot be estimated with a single observations. Simple models may not be appropriate for these diseases and we must therefore consider interacting elements, such as we do in gene networks, to describe the molecular mechanisms which drive cellular states.

The growing field of personalized medicine calls for the tailoring of treatment regimens based on in part on the molecular signatures specific to an individual. For example, genomic biomarkers such as somatic or germline mutations in cancer can suggest a patient response to a drug. Cancer in particular would benefit network based characterizations, owing to the fact that its high degree of complexity and heterogeneity makes it more appropriately described in those terms. Biomarker discovery is pre-

dicted to be heavily dependent on the ability to infer networks and understand the mechanisms that drive the change from healthy to disease. Our work in this area, presented in chapter 2, is an important contribution to addressing this problem. In that chapter we outlined an approach to implicate certain transcription factors whose targeting pattern changes best explained the transition between cellular states. We operated under the recognition that the size and scale of the problem along with the degree of technical and biological noise in the data made it unreasonable to identify specific TF-gene interactions with a satisfactory rate of false positives. With improvements in data quality and quantity, future work will focus on specific regulatory events that are altered across experimental conditions. This may be accelerated via the integration of complementary data sources, such ChIP-Seq and methylation sequencing, where we may gain additional independent information on gene regulation.

As the effects that we search for become smaller and more complex, greater importance lies in the ability to remove the impact of sources subtle, complex bias such as those described in these chapters. Addressing these issues with proper quality control will be critical for preventing the reporting of spurious results. QC is not simply a process stemming from imperfect data generation, which can become obsolete as the technology improves. In some cases, the unwanted artifacts arise from real biology, such as in the case of population structure. Because of this we can't rely on superior technologies to address these problems. We showed examples of this concept in chapters 3 and 4 where important structural features of the data reveal themselves when

measured across samples.

The revolutionary advancements in data generation have touched all virtually fields of research. In addition to data creation, data access has improved as well. Whereas previously, clinical and molecular data was frequently housed in isolated data silos controlled by separate entities, we are now seeing greater collaboration and sharing. As Dr. John Quackenbush often puts it, "Every revolution in science has been driven by one and only one thing: access to data." The methods described in these chapters depend on the availability of new data and vice versa. This new data combined with new methods will be critical in bridging the gap from data to understanding.

References

[1] Al-Khudhair, A., Qiu, S., Wyse, M., Chowdhury, S., Cheng, X., Bekbolsynov, D., Saha-Mandal, A., Dutta, R., Fedorova, L., & Fedorov, A. (2015). Inference of distant genetic relations in humans using '1000 genomes'. *Genome biology and evolution*, 7(2), 481–492.

[2] Amar, D., Safer, H., & Shamir, R. (2013). Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput Biol*, 9(3), e1002955.

[3] Bacanu, S.-A., Devlin, B., & Roeder, K. (2002). Association studies for quantitative traits in structured populations. *Genetic epidemiology*, 22(1), 78–93.

[4] Bahr, T. M., Hughes, G. J., Armstrong, M., Reisdorph, R., Coldren, C. D., Edwards, M. G., Schnell, C., Kedl, R., LaFlamme, D. J., Reisdorph, N., et al. (2013). Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease. *American journal of respiratory cell and molecular biology*, 49(2), 316–323.

[5] Banerjee, O., Ghaoui, L. E., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar), 485–516.

[6] Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., & Marron, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1), 105–114.

[7] Bien, J., Tibshirani, R. J., et al. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4), 807.

[8] Boehnke, M. & Cox, N. J. (1997). Accurate inference of relationships in sib-pair linkage studies. *The American Journal of Human Genetics*, 61(2), 423–429.

[9] Boucherat, O., Morissette, M., Provencher, S., Bonnet, S., & F, M. (2016). Bridging lung development with chronic obstructive pulmonary disease. relevance of developmental pathways in chronic obstructive pulmonary disease

pathogenesis. *American Journal of Respiratory and Critical Care Medicine*, 193(4), 362–75.

[10] Busch, R., Qiu, W., Lasky-Su, J., Morrow, J., Criner, G., & DeMeo, D. (2016). Differential dna methylation marks and gene comethylation of copd in african-americans with copd exacerbations. *Respiratory Research*, 17(1), 143.

[11] Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., & Liu, C. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one*, 6(2), e17238.

[12] Chen, C.-Y., Pollack, S., Hunter, D. J., Hirschhorn, J. N., Kraft, P., & Price, A. L. (2013). Improved ancestry inference using weights from external reference panels. *Bioinformatics*, (pp. btt144).

[13] Chen, W. W., Schoeberl, B., Jasper, P. J., Niepel, M., Nielsen, U. B., Lauffenburger, D. A., & Sorger, P. K. (2009). Input–output behavior of erbb signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular systems biology*, 5(1).

[14] Chen, Z.-H., Kim, H. P., Sciurba, F. C., Lee, S.-J., Feghali-Bostwick, C., Stolz, D. B., Dhir, R., Landreneau, R. J., Schuchert, M. J., Yousem, S. A., et al. (2008). Egr-1 regulates autophagy in cigarette smoke-induced chronic obstructive pulmonary disease. *PloS one*, 3(10), e3316.

[15] Choi, J. K., Yu, U., Yoo, O. J., & Kim, S. (2005). Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, 21(24), 4348–4355.

[16] Choi, Y., Wijsman, E. M., & Weir, B. S. (2009). Case-control association testing in the presence of unknown relationships. *Genetic epidemiology*, 33(8), 668–678.

[17] Choksi, S. P., Lauter, G., Swoboda, P., & Roy, S. (2014). Switching on cilia: transcriptional networks regulating ciliogenesis. *Development*, 141(7), 1427–1441.

[18] Chuang, P.-T., Kawcak, T., & McMahon, A. P. (2003). Feedback control of mammalian hedgehog signaling by the hedgehog-binding protein, hip1, modulates fgf signaling during branching morphogenesis of the lung. *Genes & development*, 17(3), 342–347.

[19] Cloonan, S. M., Glass, K., Laucho-Contreras, M. E., Bhashyam, A. R., Cervo, M., Pabón, M. A., Konrad, C., Polverino, F., Siempos, I. I., Perez, E., et al. (2016). Mitochondrial iron chelation ameliorates cigarette smoke-induced bronchitis and emphysema in mice. *Nature medicine*.

[20] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., et al. (2016). A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1), 13.

[21] Consortium, . G. P. et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), 56–65.

[22] Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74.

[23] Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic acids research*, 33(20), e175–e175.

[24] Danielian, P. S., Kim, C. F. B., Caron, A. M., Vasile, E., Bronson, R. T., & Lees, J. A. (2007). E2f4 is required for normal development of the airway epithelium. *Developmental biology*, 305(2), 564–576.

[25] de la Fuente, A. (2010). From 'differential expression'to 'differential networking'–identification of dysfunctional regulatory networks in diseases. *Trends in genetics*, 26(7), 326–333.

[26] Devlin, B., Roeder, K., & Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theoretical population biology*, 60(3), 155–166.

[27] Du P, Kibbe WA, L. S. (2007). *Using lumi, a package processing Illumina Microarray*. Bioconductor R package.

[28] Eduati, F., De Las Rivas, J., Di Camillo, B., Toffolo, G., & Saez-Rodriguez, J. (2012). Integrating literature-constrained and data-driven inference of signalling networks. *Bioinformatics*, 28(18), 2311–2317.

[29] Epstein, M. P., Duren, W. L., & Boehnke, M. (2000). Improved inference of relationship for pairs of individuals. *The American Journal of Human Genetics*, 67(5), 1219–1231.

[30] Fahy, J. V. & Dickey, B. F. (2010). Airway mucus function and dysfunction. *New England Journal of Medicine*, 363(23), 2233–2247.

[31] Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., & Gardner, T. S. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1), e8.

[32] Falcon, S. & Gentleman, R. (2007). Using gostats to test gene lists for go term association. *Bioinformatics*, 23(2), 257–258.

[33] Fare, T. L., Coffey, E. M., Dai, H., He, Y. D., Kessler, D. A., Kilian, K. A., Koch, J. E., LeProust, E., Marton, M. J., Meyer, M. R., et al. (2003). Effects of atmospheric ozone on microarray data quality. *ANALYTICAL CHEMISTRY-WASHINGTON DC-*, 75(17), 4672–4675.

[34] Fedorova, L., Qiu, S., Dutta, R., & Fedorov, A. (2016). Atlas of cryptic genetic relatedness among 1000 human genomes. *Genome biology and evolution*, 8(3), 777–790.

[35] Filosi, M., Visintainer, R., & Riccadonna, S. (2014). *nettools: A Network Comparison Framework*. R package version 1.0.1.

[36] Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441.

[37] Fukushima, A. (2013). Diffcorr: an r package to analyze and visualize differential correlations in biological networks. *Gene*, 518(1), 209–214.

[38] Fuller, T. F., Ghazalpour, A., Aten, J. E., Drake, T. A., Lusis, A. J., & Horvath, S. (2007). Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome*, 18(6-7), 463–472.

[39] Furlotte, N. A., Kang, H. M., Ye, C., & Eskin, E. (2011). Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity. *Bioinformatics*, 27(13), i288–i294.

[40] Galvagni, F., Capo, S., & Oliviero, S. (2001). Sp1 and sp3 physically interact and co-operate with gabp for the activation of the utrophin promoter. *Journal of molecular biology*, 306(5), 985–996.

[41] Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy-analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3), 307–315.

[42] Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E., & Leutenegger, A.-L. (2015). High level of inbreeding in final phase of 1000 genomes project. *Scientific reports*, 5.

[43] genomics research consortium, L. (2015). Lung genomics research consortium (lgrc). Accessed: 2016-02-02.

[44] Glass, K., Huttenhower, C., Quackenbush, J., & Yuan, G.-C. (2013). Passing messages between biological networks to refine predicted interactions. *PloS one*, 8(5), e64832.

[45] Glass, K., Quackenbush, J., Silverman, E. K., Celli, B., Rennard, S. I., Yuan, G.-C., & DeMeo, D. L. (2014). Sexually-dimorphic targeting of functionally-related genes in copd. *BMC systems biology*, 8(1), 118.

[46] Glass, K., Quackenbush, J., Spentzos, D., Haibe-Kains, B., & Yuan, G.-C. (2015). A network model for angiogenesis in ovarian cancer. *BMC bioinformatics*, 16(1), 115.

[47] Gopalakrishnan, L. & Scarpulla, R. C. (1995). Structure, expression, and chromosomal assignment of the human gene encoding nuclear respiratory factor 1. *Journal of Biological Chemistry*, 270(30), 18019–18025.

[48] Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4), 576–589.

[49] Hessel, J., Heldrich, J., Fuller, J., Staudt, M. R., Radisch, S., Hollmann, C., Harvey, B.-G., Kaner, R. J., Salit, J., Yee-Levin, J., et al. (2014). Intraflagellar transport gene expression associated with short cilia in smoking and copd. *PloS one*, 9(1), e85453.

[50] Hill, S. M., Lu, Y., Molina, J., Heiser, L. M., Spellman, P. T., Speed, T. P., Gray, J. W., Mills, G. B., & Mukherjee, S. (2012). Bayesian inference of signaling network topology in a cancer cell line. *Bioinformatics*, 28(21), 2804–2810.

[51] Hoff, P. D. & Niu, X. (2012). A covariance regression model. *Statistica Sinica*, (pp. 729–753).

[52] Hogg, J. C. (2004). Pathophysiology of airflow limitation in chronic obstructive pulmonary disease. *The Lancet*, 364(9435), 709–721.

[53] Holmes, M., Gray, A., & Isbell, C. (2007). Fast svd for large-scale matrices. In *Workshop on Efficient Machine Learning at NIPS*, volume 58 (pp. 249–252).

[54] Hsu, C.-L., Juan, H.-F., & Huang, H.-C. (2015). Functional analysis and characterization of differential coexpression networks. *Scientific reports*, 5, 13295.

[55] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2), 249–264.

[56] Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1), 118–127.

[57] Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., Eskin, E., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4), 348–354.

[58] Karlseder, J., Rotheneder, H., & Wintersberger, E. (1996). Interaction of sp1 with the growth-and cell cycle-regulated transcription factor e2f. *Molecular and cellular biology*, 16(4), 1659–1667.

[59] Lander, E. S. (1999). Array of hope. *Nature genetics*, 21, 3–4.

[60] Langfelder, P. & Horvath, S. (2008a). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1), 1.

[61] Langfelder, P. & Horvath, S. (2008b). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, (1), 559.

[62] Langfelder, P. & Horvath, S. (2012a). Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software*, 46(11), 1–17.

[63] Langfelder, P. & Horvath, S. (2012b). Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software*, 46(11), 1–17.

[64] Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3), R25.

[65] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10), 733–739.

[66] Leek, J. T. & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9), e161.

[67] Levine, M. & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945), 147–151.

[68] Liao, Y., Smyth, G. K., & Shi, W. (2014). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7), 923–930.

[69] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10), 833–835.

[70] Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., & Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. *Nature methods*, 9(6), 525–526.

[71] Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22), 2867–2873.

[72] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., Stolovitzky, G., et al. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8), 796–804.

[73] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1), S7.

[74] Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C.-y., Chou, A., Ienasescu, H., et al. (2013). Jaspar 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research*, (pp. gkt997).

[75] Mathieson, I. & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, 44(3), 243–246.

[76] Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, (pp. 1436–1462).

[77] Molinelli, E. J., Korkut, A., Wang, W., Miller, M. L., Gauthier, N. P., Jing, X., Kaushik, P., He, Q., Mills, G., Solit, D. B., et al. (2013). Perturbation biology: inferring signaling networks in cellular systems. *PLoS Comput Biol*, 9(12), e1003290.

[78] Morrisey, E. E., Cardoso, W. V., Lane, R. H., Rabinovitch, M., Abman, S. H., Ai, X., Albertine, K. H., Bland, R. D., Chapman, H. A., Checkley, W., et al.

(2013). Molecular determinants of lung development. *Annals of the American Thoracic Society*, 10(2), S12–S16.

[79] Morrow, J. D., Qiu, W., Chhabra, D., Rennard, S. I., Belloni, P., Belousov, A., Pillai, S. G., & Hersh, C. P. (2015). Identifying a gene expression signature of frequent copd exacerbations in peripheral blood using network methods. *BMC medical genomics*, 8(1), 1.

[80] Morrow, J. D., Zhou, X., Lao, T., Jiang, Z., DeMeo, D. L., Cho, M. H., Qiu, W., Cloonan, S., Pinto-Plata, V., Celli, B., et al. (2017). Functional interactors of three genome-wide association study genes are differentially expressed in severe chronic obstructive pulmonary disease lung tissue. *Scientific Reports*, 7.

[81] Nemesh, J. & McCarroll, S. (2012). Addressing cryptic relatedness in candidate samples for 1kg. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/ supporting/cryptic_relation_analysis. Accessed: 2016-06-06.

[82] Nygaard, V., Rødland, E. A., & Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1), 29–39.

[83] Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS genet*, 2(12), e190.

[84] Pillai, S. G., Ge, D., Zhu, G., Kong, X., Shianna, K. V., Need, A. C., Feng, S., Hersh, C. P., Bakke, P., Gulsvik, A., et al. (2009). A genome-wide association study in chronic obstructive pulmonary disease (copd): identification of two major susceptibility loci. *PLoS Genet*, 5(3), e1000421.

[85] Pinello, L., Xu, J., Orkin, S. H., & Yuan, G.-C. (2014). Analysis of chromatin-state plasticity identifies cell-type–specific regulators of h3k27me3 patterns. *Proceedings of the National Academy of Sciences*, 111(3), E344–E353.

[86] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8), 904–909.

[87] Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D., et al. (2008). Long-range ld can confound genome scans in admixed populations. *The American Journal of Human Genetics*, 83(1), 132–135.

[88] Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7), 459–463.

[89] Prokopenko, D., Hecker, J., Silverman, E. K., Pagano, M., Nöthen, M. M., Dina, C., Lange, C., & Fier, H. L. (2016). Utilizing the jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 genomes project. *Bioinformatics*, 32(9), 1366–1372.

[90] Ptak, S. E. & Przeworski, M. (2002). Evidence for population growth in humans is confounded by fine-scale population structure. *Trends in Genetics*, 18(11), 559–563.

[91] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3), 559–575.

[92] Qiu, W., DeMeo, D. L., Houston, I., Pinto-Plata, V. M., Celli, B. R., Marchetti, N., Criner, G. J., Bueno, R., Morrow, G., Washko, K., et al. (2015). Network analysis of gene expression in severe copd lung tissue samples. In *A30. BIG DATA: HARVESTING FRUITS FROM COPD AND LUNG CANCER* (pp. A1253–A1253). Am Thoracic Soc.

[93] Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5), 744–752.

[94] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *science*, 297(5586), 1551–1555.

[95] Regan, E. A., Hokanson, J. E., Murphy, J. R., Make, B., Lynch, D. A., Beaty, T. H., Curran-Everett, D., Silverman, E. K., & Crapo, J. D. (2011). Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1), 32–43.

[96] Rotheneder, H., Geymayer, S., & Haidweger, E. (1999). Transcription factors of the sp1 family: interaction with e2f and regulation of the murine thymidine kinase promoter. *Journal of molecular biology*, 293(5), 1005–1015.

[97] Saez-Rodriguez, J., Alexopoulos, L. G., Zhang, M., Morris, M. K., Lauffenburger, D. A., & Sorger, P. K. (2011). Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer research*, 71(16), 5400–5411.

[98] Scherer, A. (2009). *Batch effects and noise in microarray experiments: sources and solutions*, volume 868. John Wiley & Sons.

[99] Schlauch, D., Glass, K., Hersh, C. P., Silverman, E. K., & Quackenbush, J. (2016). Estimating drivers of cell state transitions using gene regulatory network models. *bioRxiv*, (pp. 089003).

[100] Shi, W., Chen, F., & Cardoso, W. V. (2009). Mechanisms of lung development: contribution to adult lung disease and relevance to chronic obstructive pulmonary disease. *Proceedings of the American Thoracic Society*, 6(7), 558–563.

[101] Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318), 626–633.

[102] Singh, D., Fox, S. M., Tal-Singer, R., Bates, S., Riley, J. H., & Celli, B. (2014). Altered gene expression in blood and sputum in copd frequent exacerbators in the eclipse cohort. *PloS one*, 9(9), e107381.

[103] Sîrbu, A., Ruskin, H. J., & Crane, M. (2010). Comparison of evolutionary algorithms in gene regulatory network model inference. *BMC bioinformatics*, 11(1), 1.

[104] Siska, C. & Kechris, K. (2017). Differential correlation for sequencing data. *BMC Research Notes*, 10(1), 54.

[105] Southworth, L. K., Owen, A. B., & Kim, S. K. (2009). Aging mice show a decreasing correlation of gene expression within genetic modules. *PLoS Genet*, 5(12), e1000776.

[106] Tesson, B. M., Breitling, R., & Jansen, R. C. (2010). Diffcoex: a simple and sensitive method to find differentially coexpressed gene modules. *BMC bioinformatics*, 11(1), 497.

[107] Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., & Risch, N. (2012). Estimating kinship in admixed populations. *The American Journal of Human Genetics*, 91(1), 122–138.

[108] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 267–288).

[109] Vestbo, J., Anderson, W., Coxson, H. O., Crim, C., Dawber, F., Edwards, L., Hagan, G., Knobil, K., Lomas, D. A., MacNee, W., et al. (2008). Evaluation of copd longitudinally to identify predictive surrogate end-points (eclipse). *European Respiratory Journal*, 31(4), 869–873.

[110] Voight, B. F. & Pritchard, J. K. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genet*, 1(3), e32.

[111] Wang, J., Tsang, W. W., & Marsaglia, G. (2003). Evaluating kolmogorov's distribution. *Journal of Statistical Software*, 8(18).

[112] Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7), 565–569.

[113] Yu, H., Liu, B.-H., Ye, Z.-Q., Li, C., Li, Y.-X., & Li, Y.-Y. (2011). Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. *BMC bioinformatics*, 12(1), 315.

[114] Yuan, M. & Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, (pp. 19–35).

[115] Zhang, B. & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1).

[116] Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4), 355–360.

[117] Zhou, X., Baron, R. M., Hardin, M., Cho, M. H., Zielinski, J., Hawrylkiewicz, I., Sliwinski, P., Hersh, C. P., Mancini, J. D., Lu, K., et al. (2012). Identification of a chronic obstructive pulmonary disease genetic determinant that regulates hhip. *Human molecular genetics*, 21(6), 1325–1335.

[118] Zou, T., Lan, W., Wang, H., & Tsai, C.-L. (2016). Covariance regression analysis. *Journal of the American Statistical Association*, (just-accepted), 1–44.