

Why I've been so busy

A DISSERTATION PRESENTED

BY

DANIEL 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

MAY 2017

©2017 – DANIEL SCHLAUCH
ALL RIGHTS RESERVED.

Why I've been so busy

ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

Quisque facilisis erat a dui. Nam malesuada ornare dolor. Cras gravida, diam sit amet rhoncus ornare, erat elit consectetuer erat, id egestas pede nibh eget odio. Proin tincidunt, velit vel porta elementum, magna diam molestie sapien, non aliquet massa pede eu diam. Aliquam iaculis. Fusce et ipsum et nulla tristique facilisis. Donec eget sem sit amet ligula viverra gravida. Etiam vehicula urna vel turpis. Suspendisse sagittis ante a urna. Morbi a est quis orci consequat rutrum. Nullam egestas feugiat felis. Integer adipiscing semper ligula. Nunc molestie, nisl sit amet cursus convallis, sapien lectus pretium metus, vitae pretium enim wisi id lectus. Donec vestibulum. Etiam vel nibh. Nulla facilisi. Mauris pharetra. Donec augue. Fusce ultrices, neque id dignissim ultrices, tellus mauris dictum elit, vel lacinia enim metus eu nunc.

Contents

o INTRODUCTION	I
1 ESTIMATING DRIVERS OF CELL STATE TRANSITIONS USING GENE REGULATORY NETWORK MODELS	3
2 IDENTIFICATION OF GENETIC OUTLIERS DUE TO SUB-STRUCTURE AND CRYPTIC RELATIONSHIPS	45
2.1 Introduction	46
2.2 Methods	49
2.3 Results	67
2.4 Discussion	76
3 CONSECTETUER ADIPISCING ELIT	88
4 CONCLUSION	89
APPENDIX A SOME EXTRA STUFF	91
REFERENCES	101

DEDICATED TO THE ENDLESS PURSUIT OF KNOWLEDGE ABOUT NOTHING...

Acknowledgments

THANKS, to my dog, cat, and wife.

0

Introduction

From Author summary on MONSTER paper...

Biological states are characterized by distinct patterns of gene expression that reflect each phenotype's active cellular processes. Driving these phenotypes are gene regulatory networks in which transcription 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. MONSTER is a new approach to understanding these transitions. 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. We demonstrate the power of MONSTER by applying it to data from four independent studies of chronic obstructive pulmonary disease and find a robust set of transcription factors that help explain the development of the disease.

*Nulla facilisi. In vel sem. Morbi id urna in diam dignis-
sim feugiat. Proin molestie tortor eu velit. Aliquam erat
volutpat. Nullam ultrices, diam tempus vulputate egestas,
eros pede varius leo.*

Quoteauthor Lastname

1

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, MOdeling Network State Transitions from Expression and Regulatory 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.

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” T , subject to a set of constraints, that approximates the conversion of the initial network’s adjacency matrix A into the final network’s adjacency matrix B , or

$$B = AT \quad (1.1)$$

In this model, the diagonal elements of 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 1.1, 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 $B = AT + E$, where 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 T , and then identifying state transition drivers based on features of that matrix.

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/>),⁴⁸ and running Haystack⁵⁴ 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>)³² 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)⁶⁸ 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-normalized³⁶ in R using the Bioconductor package 'affy'²⁶. Array probes were collapsed to 19,765 Entrez-gene IDs using a custom CDF¹⁵ 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)^{4,64} 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 CDF¹⁵, 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)¹ 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 ComBat³⁷. 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)⁶ was collected using HumanHT-12 Bead-Chips. 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 lumi¹⁸ 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.

METHODS

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

INFERRING STATE-SPECIFIC GENE REGULATORY NETWORKS:

Before estimating the transition matrix, T , we must first estimate a gene regulatory starting point for each state. While there have been many methods developed to infer such networks^{34,29,30,19,8,50,66}, we have found the bipartite framework used in PANDA²⁸ 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:

$$\hat{d}_{i,j} = \text{cor}(g_i, g_j | \{g_k : k \neq i, k \in \text{TF}_j\})^2,$$

where g_i is the gene which encodes the transcription factor TF_i , g_j is any other gene in the genome, and 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:

$$\text{logit}(P[M_{i,j} = 1]) = \beta_{0,i} + \beta_{1,g_j^{(1)}} + \cdots + \beta_{N,g_j^{(N)}}$$

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

where the response 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 $\hat{\vartheta}_{i,j}$ represents our estimated indirect evidence. Combining the scores for the direct evidence, $\hat{d}_{i,j}$, and indirect evidence, $\hat{\vartheta}_{i,j}$, 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 (1.1). 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 B and 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,

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

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

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

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 L_1 and/or L_2 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 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, A and 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, T, such that

$$B = AT + E$$

where 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 A, B, T, and E. For a column in B, b_i , we note that a corresponding column in T, τ_i , represents the ordinary least squares solution to

$$E[b_i] = \tau_{i1}a_{1i} + \tau_{i2}a_{2i} + \cdots + \tau_{im}a_{mi}$$

or alternatively expressed

$$\begin{bmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{ip} \end{bmatrix} = \tau_{1,i} \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix} + \cdots + \tau_{p,i} \begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{pp} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{ip} \end{bmatrix}$$

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

$$\tau_i = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}_i$$

$$\mathbf{T} = [\tau_1, \tau_2, \dots, \tau_m]$$

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 regularization⁷ 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:

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

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

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

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,

$$dTFI_j = \frac{\sum_{i=1}^m I(i \neq j) \hat{\tau}_{i,j}^2}{\sum_{i=1}^m \hat{\tau}_{i,j}^2} \quad (1.2)$$

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 (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 T as containing the linear combination of regulatory targets of TF_i in A that best approximates the regulatory targets of TF_i in 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 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),

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

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 structure using the following resampling procedure (Figure 1.2):

- o. 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.
- i. 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.

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)^{68 72} (1,3), COPDGene^{64 4 53}, Lung Genomics Research Consortium (LGRC)¹ and Lung Tissue from Channing Division of Network Medicine (LT-CDNM)⁶¹. 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 investi-

gated 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)⁴⁷, Context Likelihood of Relatedness (CLR)²², and Weighted Gene Correlation Network Analysis (WGCNA)⁷⁶ - 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). 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 extracting 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 1.2) for each transcription factor and used permutation analysis to estimate their significance (Figure 1.3, Figures 1.8-1.9). 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 re-

sults were among the top seven in the LT-CDNM results (Additional Table 1, Additional Figure 3).

This agreement is quite striking considering that 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 MONSTER across all four studies, are biologically plausible candidates to be involved in the etiology of COPD (Additional Table 1, Additional Figures 1-3). For example, E2F4 is a transcriptional repressor important in airway development¹⁶ and studies have begun to demonstrate the relevance of developmental pathways in COPD pathogenesis⁶.

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 gene⁷⁸. Both SP1 and SP2 form complexes with the E2F family^{65,39} 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 to smoke exposure, which may lead to autophagy, apoptosis and subsequently to development of emphysema⁹.

Mitochondrial mechanisms have also been associated with COPD progression¹². 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 proteins³¹. 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 SP1²⁵ providing evidence of a potentially shared regulatory mechanism with E2F4.

Overall, we found a strong correlation across studies in transcription factors identified 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 studies⁶⁹. 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 (Additional Table 1). 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 to the differences in tissue type as the RFX family transcription factors are known to regulate ciliogenesis¹¹. Cilia are critical for clearing mucous from the airways of healthy individuals, but disruption can lead to infection and potentially to chronic airflow obstruction^{33,35,21}.

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 Additional Figure 4) 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) 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.*²⁸. Data for the first three sources was obtained from the 2012 DREAM5 challenge data set⁴⁶. 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. 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 1.6).

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 1.1: 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.

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, M_o , 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, T , with diagonal elements set equal to one and $1,000$ random off-diagonal elements (representing random pairs of transcription factors) set 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 M_o and T we defined the “true” cases network as $M_1 = TM_o$.

Next, we generated two *in silico* gene expression datasets, one each for the case and control net-

works. To do this, we sampled 500 times from each of two multivariate Gaussian distributions with the variance-covariance matrix, Σ , defined as $M_o M'_o$ and $M_i M'_i$ 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) ^{40 41}, Topological Overlap Measure (TOM) ⁶³, Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE) ⁴⁷, and Context Likelihood of Relatedness (CLR) ²². The implementation of each method was from the R package nettools ²⁴.

We next constructed a gold-standard for our network transitions, defined as $T_{GS} = \text{ceil}(|M|)$. 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 1.1. 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 alone⁴².

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.*⁶² and processed in Glass *et. al.*³⁰. 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.

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 subtracting 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 differential edge weights.

We plotted the differential edge weights for each pairwise combination of studies (Figure 1.7) 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).

DISCUSSION

One of the fundamental problems in 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 interaction between two genes to identify “real” (high confidence) edges. In comparing phenotypes, most methods then subtract dis-

A Top significantly differentially involved transcription factors

transcription factor	ECLIPSE			COPDGene			LGRC			LTCDNM		
	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.

transcription factor	ECLIPSE		COPDGene		LGRC		LTCDNM	
	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 1.2: 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.

cretized 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 HL05339, R01 HL11759, R01 HL089897, R01 HL089856 and K25 HL133599 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.

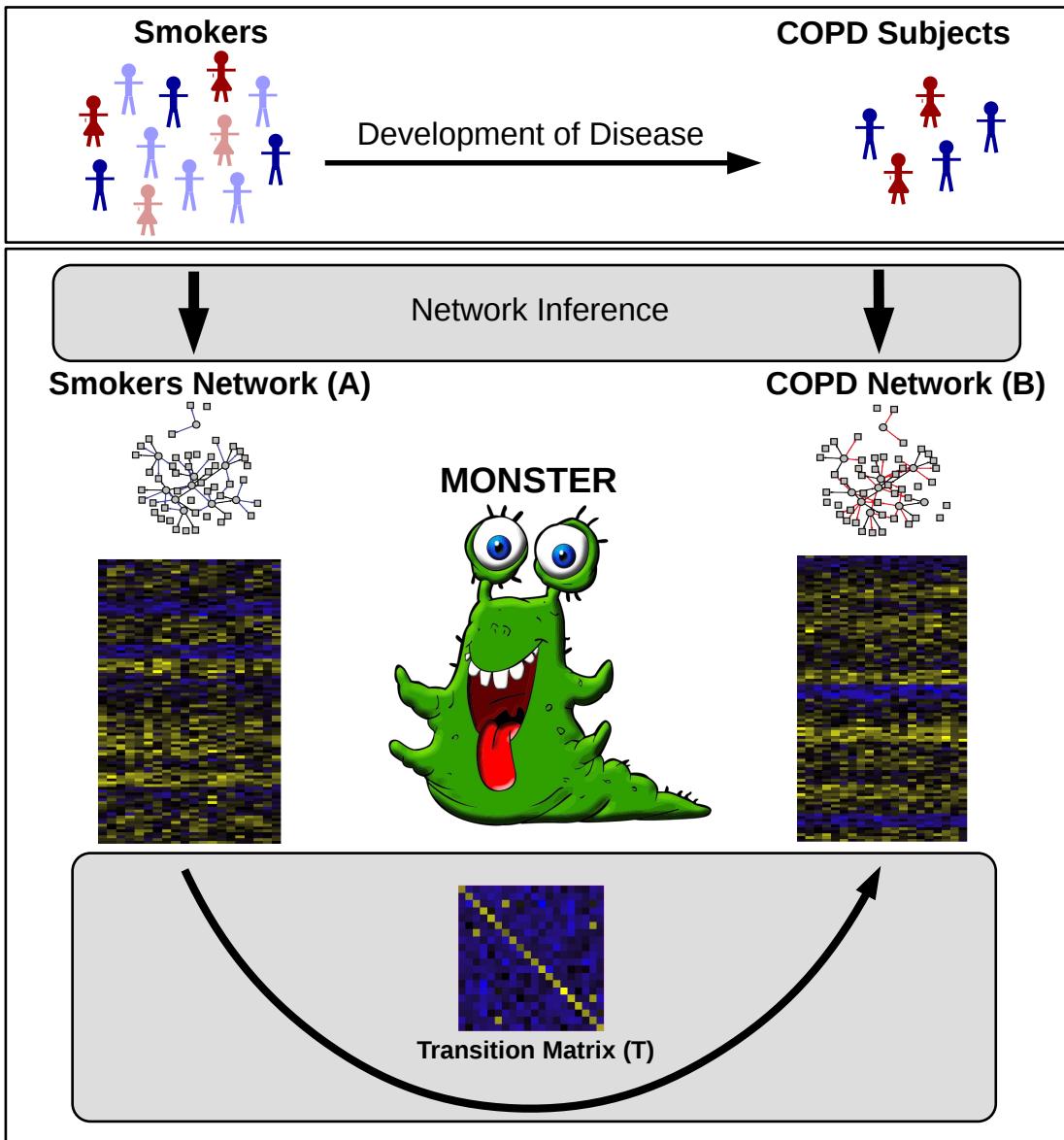


Figure 1.1: Overview of the MONSTER approach, as applied to the transition between smokers and those suffering from chronic obstructive pulmonary disease (COPD). MONSTER's approach seeks to find the $TF \times TF$ transition matrix that best characterizes the state change in network structure between 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.

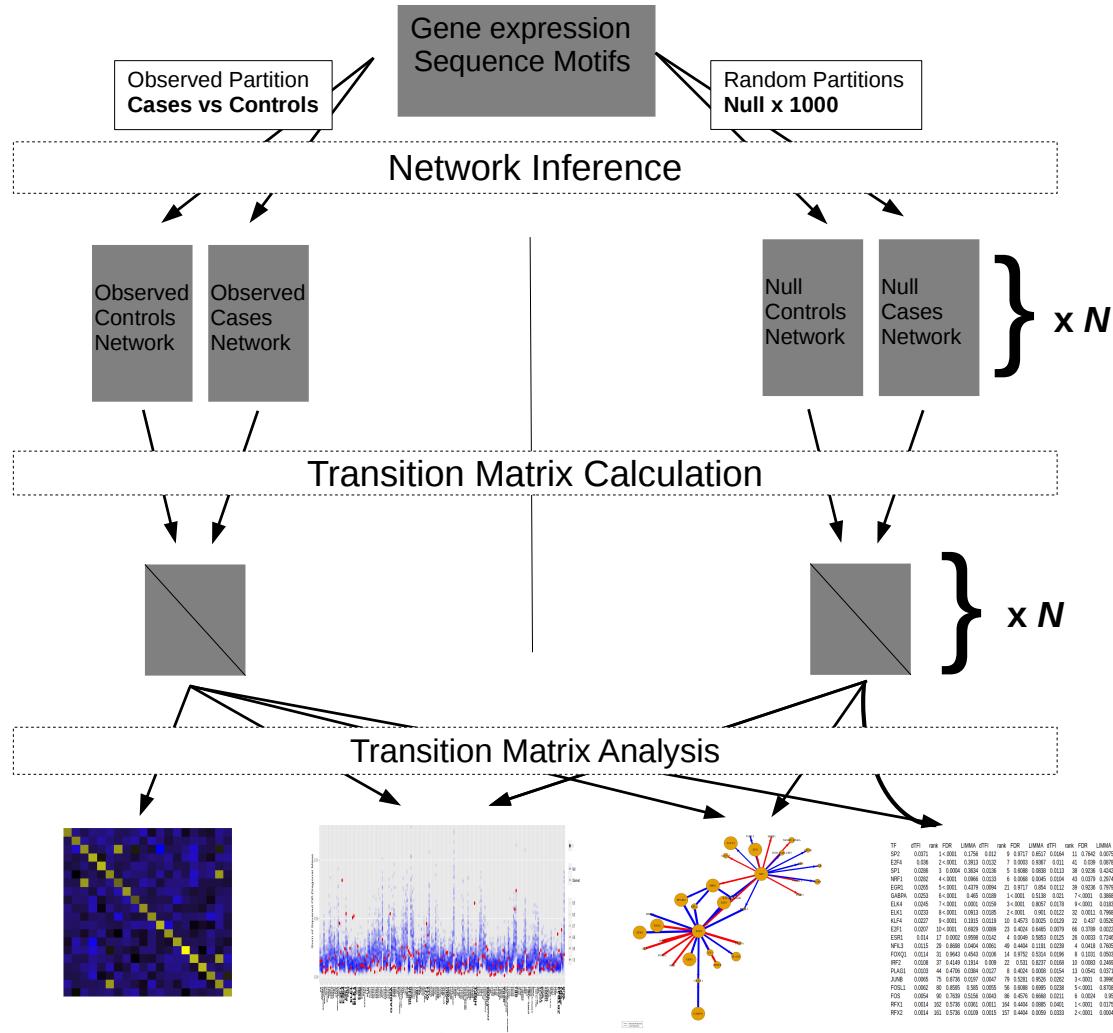


Figure 1.2: Overview of MONSTER analysis workflow. (1) Network inference is computed separately to subsets of the gene expression data including the case group, the control group and N permutations of the case and control labels. (2) The transition matrix is estimated between the cases and controls and 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.

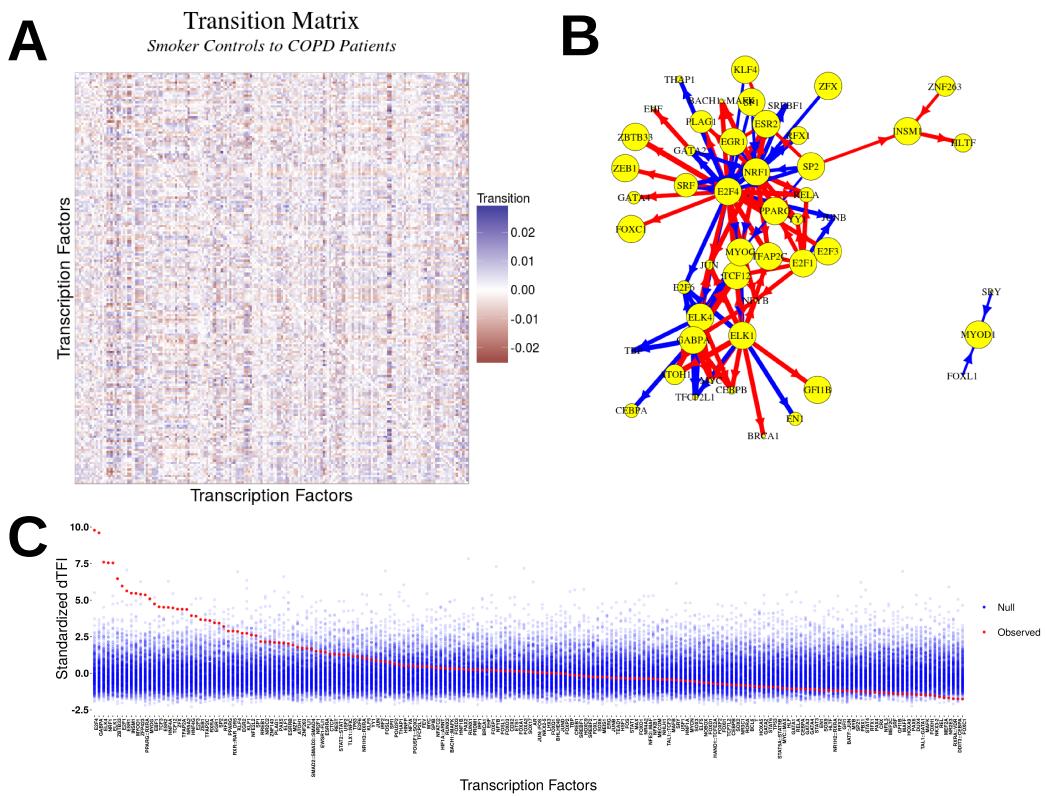


Figure 1.3: MONSTER analysis results in the ECLIPSE study. A Heatmap depicting the transition matrix calculated for smoker controls “transitioning” to COPD by applying MONSTER to ECLIPSE gene expression data. For the purposes of visualization, the magnitude of the diagonal is set to zero. B A network 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.

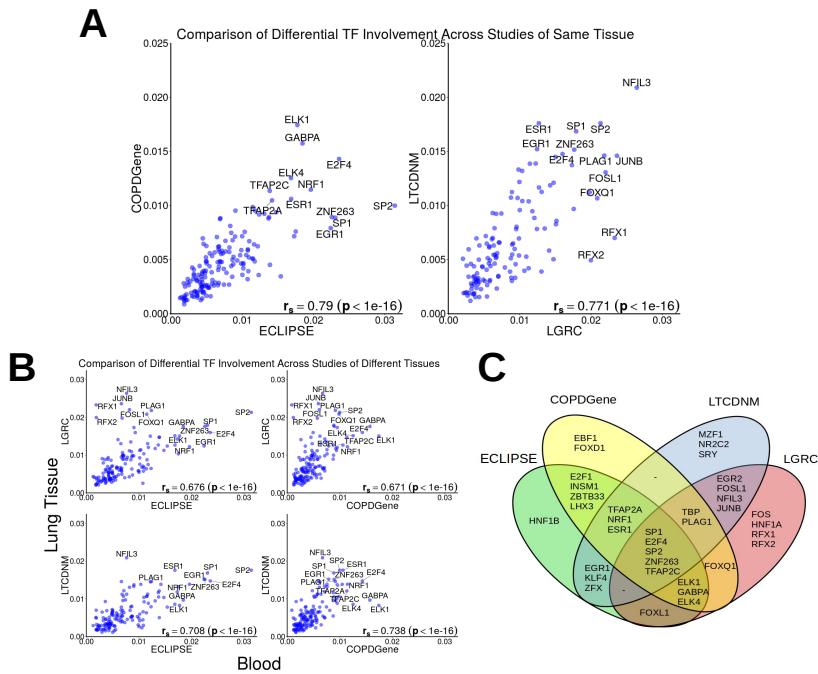


Figure 1.4: Strong reproducibility in top differential transcription factor involvement found in case-control COPD studies. ECLIPSE and COPDGene profiled gene expression in whole-blood and PBMC while the gene expression data in LGRC and LT-CDNM were assayed in lung tissue. **A** Results for studies with 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.

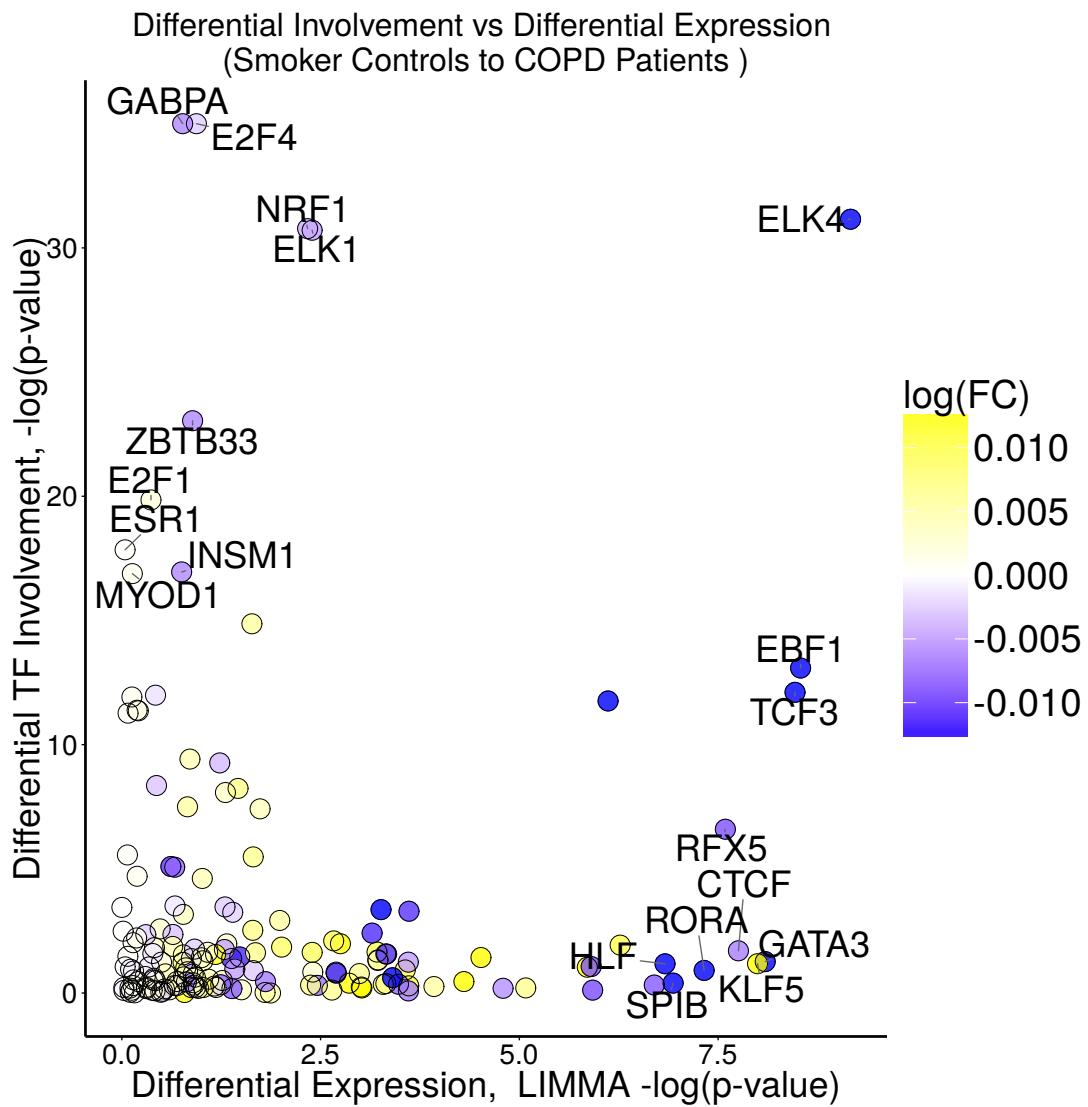


Figure 1.5: Differentially involved transcription factors are not necessarily differentially expressed. A plot of the differential expression versus the differential involvement for transcription factors based on our analysis of the ECLIPSE data. MONSTER commonly finds transcription factors which are 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.

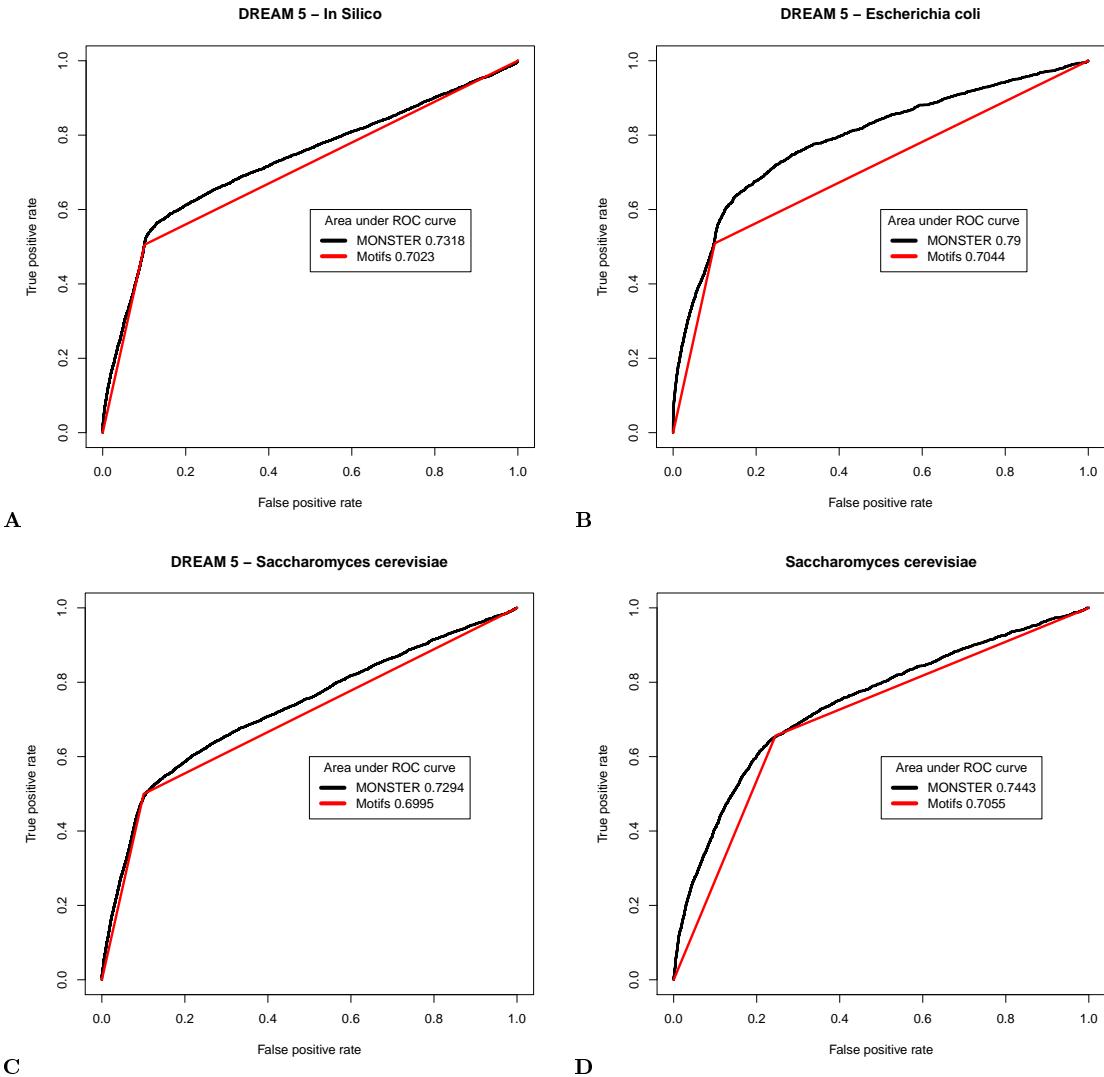


Figure 1.6: Receiver-Operator Characteristic curves for three DREAM 5 data sets (A) *in silico*, (B) *Escherichia coli*, (C) *Saccharomyces cerevisiae*, and an (D) additional *Saccharomyces cerevisiae* data set as described in Glass et. al.²⁸. The prior network for each of the DREAM5 data set analyses 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.

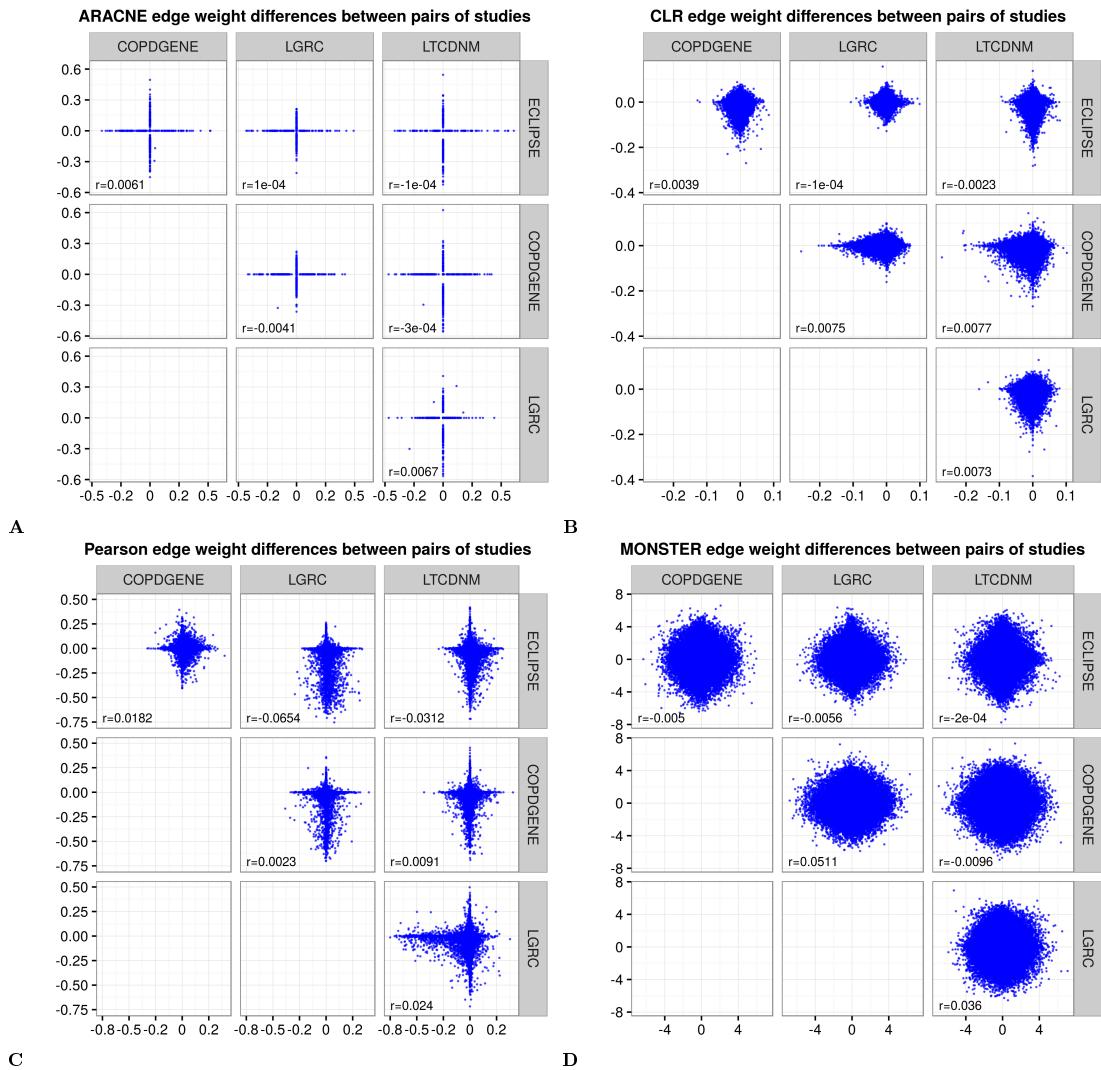


Figure 1.7: Edge weight differences between cases and controls do not correlate across studies. Using MONSTER and three other commonly used methods, we performed network inference separately on cases and controls in four COPD data sets. Here, the case-control difference is compared for each method in 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.

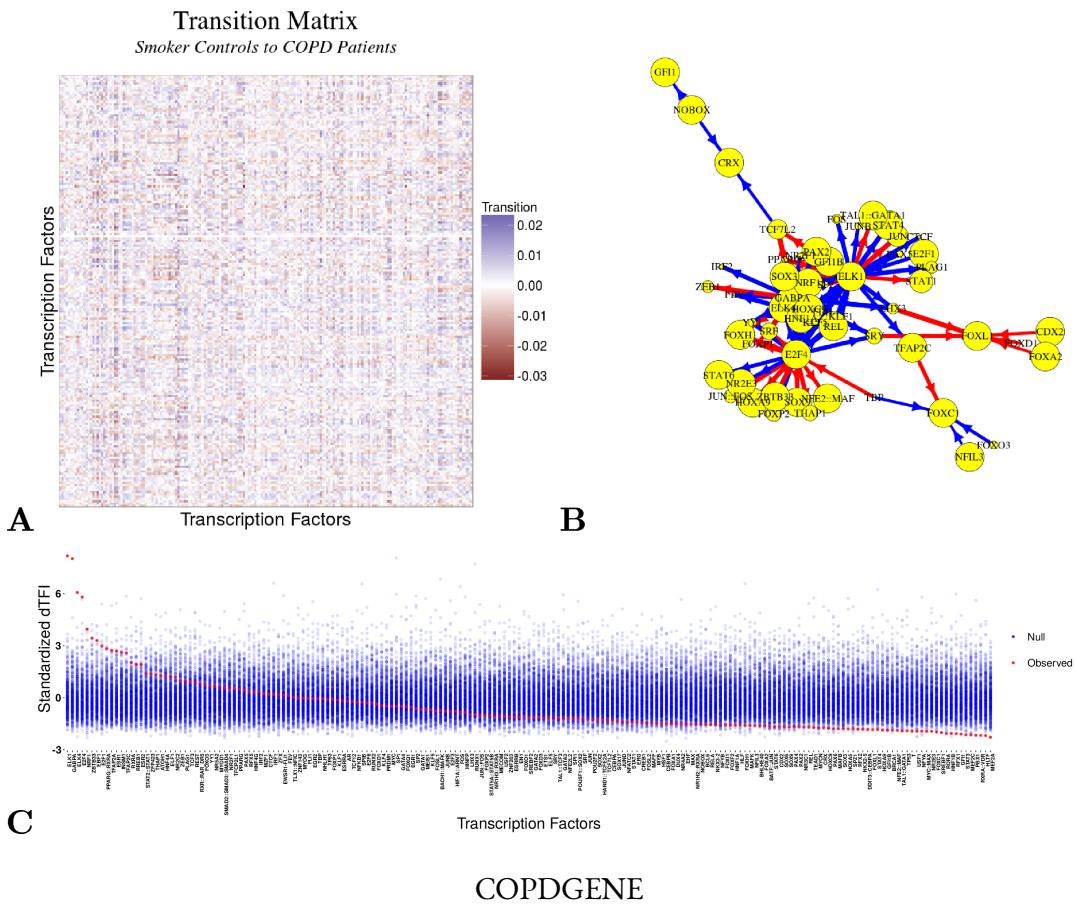


Figure 1.8: MONSTER analysis results for COPDGENE study. **A** Heatmap depicting the transition matrix calculated from smoker controls to COPD cases by applying MONSTER to the COPDGene study. For the purposes of visualization, the magnitude of the diagonal is set to zero. **B** A network visualization 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.

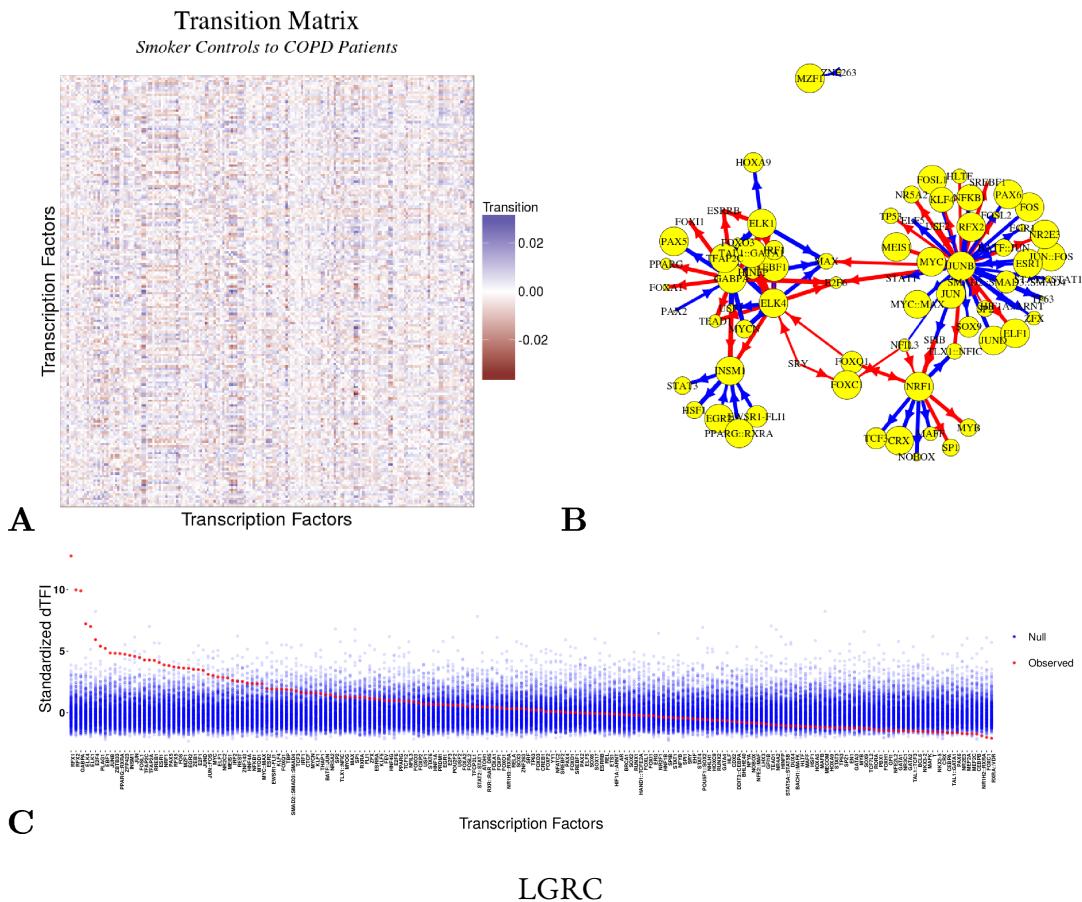


Figure 1.9: MONSTER analysis results for LGRC study. **A** Heatmap depicting the transition matrix calculated from smoker controls to COPD cases by applying MONSTER to the LGRC study. For the purposes of visualization, the magnitude of the diagonal is set to zero. **B** A network visualization of the 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.

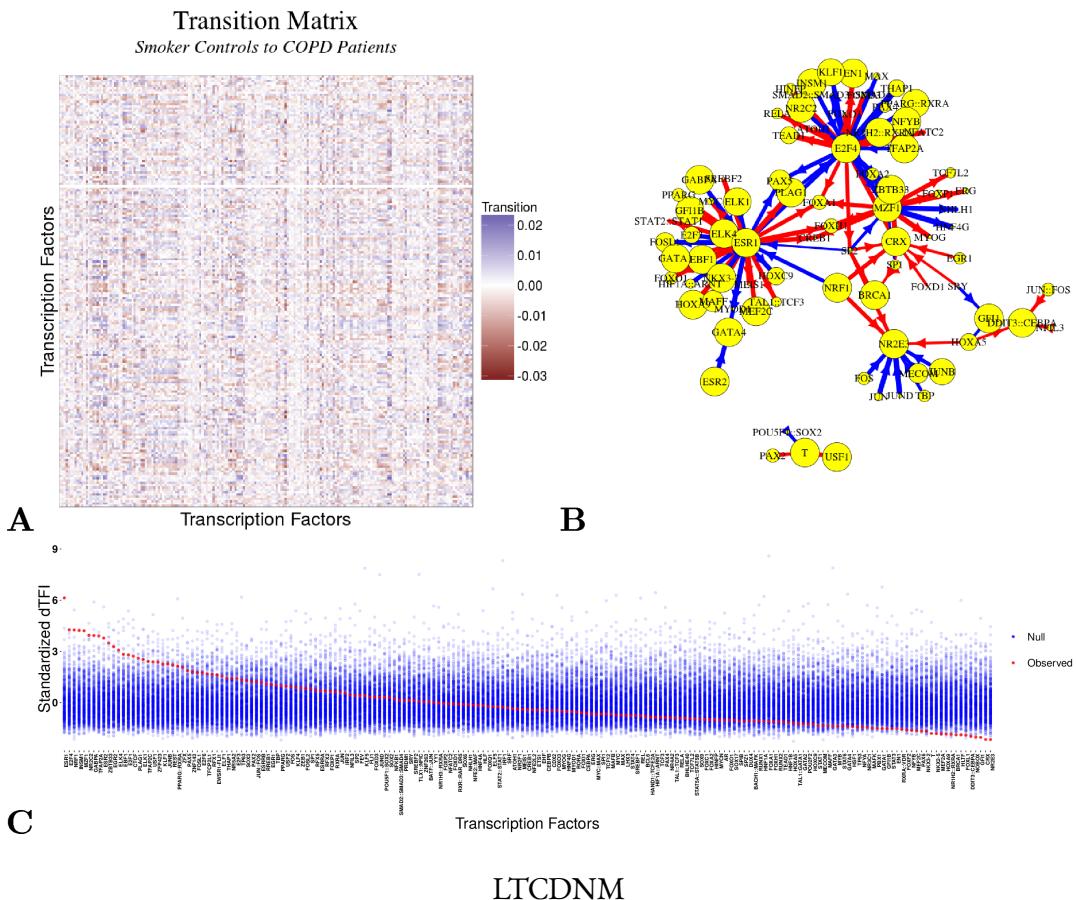


Figure 1.10: MONSTER analysis results for LTCDNM study. **A** Heatmap depicting the transition matrix calculated from smoker controls to COPD cases by applying MONSTER to the LTCDNM study. For the purposes of visualization, the magnitude of the diagonal is set to zero. **B** A network visualization of 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.

Differential Involvement vs Differential Expression (Smoker Controls to COPD Patients)

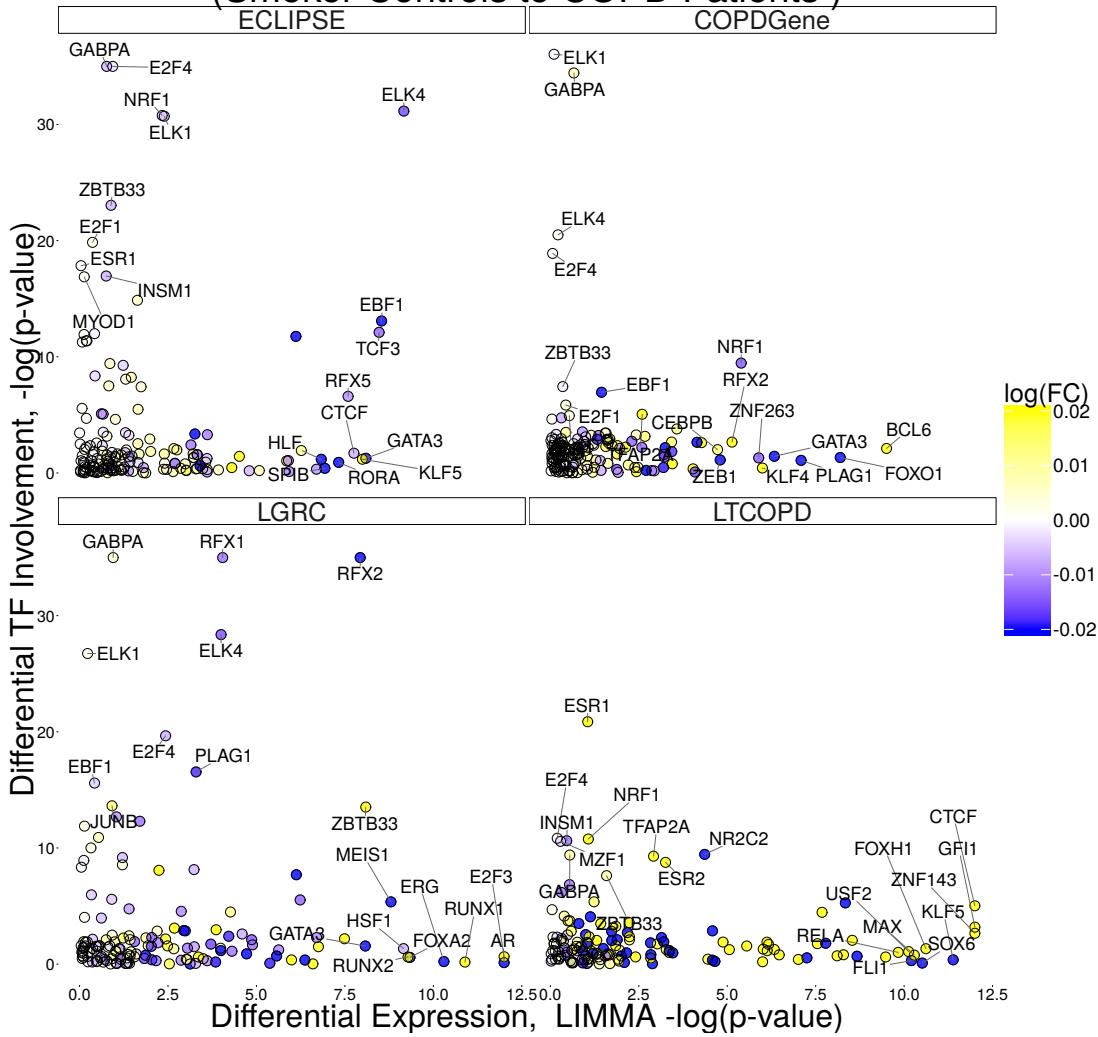


Figure 1.11: Differentially transcription factor involvement vs differential gene expression in four studies of COPD.
 Plots of the differential expression of transcription factors based on LIMMA, and their different involvement (dTF1) based on MONSTER. We observe much higher consistency between the 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.

This is some random quote to start off the chapter.

Firstname lastname

2

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 associa-

tion 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>.

2.1 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 (^{10,60,75}). 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 (^{55,38,59,73}). 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) (^{17,3}), 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 (^{55,57}). 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 (^{44,43,77}). 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 (⁴⁹).

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 eigenvalue 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 (⁵²), 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 (^{60,75,38,10}). These methods have been shown to be biased in the context of population heterogeneity (⁴⁵). More recently, methods have been developed which attempt to estimate relatedness with population structure (^{70,45}). 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 EIGENSTRAT (⁵⁸). 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 em-

phasis of our approach on small allele frequencies.

2.2 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 (⁴⁹), 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.

2.2.1 SIMILARITY MEASURE AMONG HAPLOID GENOMES

Consider a matrix of n individuals ($2n$ haploid genomes), with N independent variants described by the genotype matrix $G_{2n \times N}$. 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,

$$s_{i,j}$$

$$s_{i,j} = \frac{\sum_{k=1}^N w_k G_{i,k} G_{j,k}}{\sum_{k=1}^N I \left[\sum_{l=1}^{2n} G_{l,k} > 1 \right]} \quad (2.1)$$

where

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

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 $\sum_{l=1}^{2n} G_{l,k} > 1$, w_k is monotonically decreasing as the minor allele count increases (See Supplemental Materials).

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

$$E(s_{i,j}) = 1 \quad (2.2)$$

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.

$$s_{i,j} \sim N(1, \sigma_{i,j}^2)$$

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

$$\hat{\sigma}_{i,j}^2 = \hat{Var}(s_{i,j}) = \frac{\sum_{k=1}^N (w_k - 1)}{\left(\sum_{k=1}^N I\left[\sum_{l=1}^{2n} G_{l,k} > 1\right]\right)^2} \quad (2.3)$$

The similarity index $s_{i,j}$ provides an easily interpreted statistical test for evaluating possible relat-

edness between individuals in a purportedly homogeneous dataset of unrelated individuals. Note that this formulation (2.3) is independent of the samples i, j and depends only on the allele counts for each variant across the study group. See Supplemental Methods 1.2-1.3.

2.2.2 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

$$s_{i,j}^{(diploid)} = \frac{\sum_{k=1}^N \sum_{l=1}^2 \sum_{m=1}^2 [w_k G_{i_l,k} G_{j_m,k}] / 4}{\sum_{k=1}^N I[(\sum_{l=1}^n [G_{i_l,k} + G_{j_l,k}]) > 1]}$$

where $G_{i_2,k}$ refers to the 2nd 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 $H_{n \times N}$, where H contains the number of minor alleles, $\{0, 1, 2\}$, for a subject at a particular variant.

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

This formulation will have the same mean

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

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

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

Which yields the asymptotic result

$$s_{i,j} \sim N(1, \hat{\sigma}_{i,j}^2) \quad (2.4)$$

2.2.3 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 (2.4). Under the null, K follows the Kolmogorov distribution (74), so we define K as the value such that $P(K > K^* | H_0) = \alpha$, and reject homogeneity in the data if $K > K^*$.

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 (67) or the approximately equivalent Bonferroni procedure. We reject the null at the α level when we obtain similarity scores in the rejection region

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

2.2.4 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 $\varphi + (1 - \varphi) \times p$, 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 ($\varphi > 0$) compared to unrelated individuals ($\varphi = 0$). 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 , $\varphi_{i,j} > 0$ 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, $s_{i,j,k}$ is

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

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

$$\frac{\sum_{k=1}^N I\left[\sum_{l=1}^{2n} G_{l,k} > 1\right] \left[1 + \sum_{i,j} p_k \frac{\binom{2n(2n-1)}{(p_k(2n-2)+2)(p_k(2n-2)+1)}}{\binom{p_k(2n-2)+2}{2}} - 1 \right]}{\sum_{k=1}^N I\left[\sum_{l=1}^{2n} G_{l,k} > 1\right]} \quad (2.5)$$

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

$$E(s_{i,j} | \varphi_{i,j} > 0) > 1$$

With $\sum_{i=1}^{2n} G_{i,k}$ as the maximum likelihood estimator for $p_k n$, by the invariance principle, 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

$$\hat{\varphi}_{i,j} = \frac{s_{i,j} - 1}{\left[\frac{\sum_{k=1}^N \hat{p}_k w_k}{\sum_{k=1}^N I[\sum_{l=1}^{2n} G_{l,k} > 1]} - 1 \right]} \quad (2.6)$$

with

$$\hat{Var}(\hat{\varphi}_{i,j}) = \frac{\hat{\sigma}_{ij}^2}{\left[\frac{\sum_{k=1}^N \hat{p}_k w_k}{\sum_{k=1}^N I[\sum_{l=1}^{2n} G_{l,k} > 1]} - 1 \right]^2}$$

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

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

2.2.5 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 $\varphi \geq \gamma$ allows for the calculation of the probability of obtaining a pair of samples inside the

rejection region given two unacceptably closely related individuals.

$$P\left(\text{Reject } H_0 | \varphi_{i,j} = \gamma\right) = \alpha + (1 - \alpha) \left(1 - \Phi\left(\frac{\mu_{i,j} - 1}{\sqrt{\hat{\sigma}_{i,j}^2}}\right)\right) \quad (2.7)$$

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 .

2.2.6 EXAMPLE OF $\hat{s}_{i,j}$ COMPUTATION

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

The matrix 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 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	o	o	o	o	o	o	i	o	o	o	o	o	o	i	o	o	o	o	o	
variant 2	i	o	i	o	i	o	i	o	i	o	i	o	i	o	i	o	i	o	i	
variant 3	i	i	i	i	i	i	i	i	i	i	o	o	o	o	o	o	o	o	o	
variant 4	o	o	o	o	o	o	o	o	o	o	i	i	i	i	i	i	i	i	i	

For the purposes of this explanatory example, note that columns g and n share the low frequency

allele (i), 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 , $k \in \{1, 2, 3, 4\}$

$$w_k = \frac{\binom{20}{2}}{\left(\sum_{l=1}^{20} G_{l,k}\right)}$$

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

2.2.7 EXPECTATION OF $s_{i,j}$

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

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

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 $\sum_{l=1}^{2n} G_{l,k} > 1$, and let k^* index those variants. We condition on the minor allele count for each variant, $a_{k^*} = \sum_{l=1}^{2n} G_{l,k^*}$, where $n \geq a > 1$,

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

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.

2.2.8 VARIANCE OF $s_{i,j}$

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

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

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

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

2.2.9 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 Supplemental Materials 1.6). 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.

2.2.10 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 computa-

tion 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 (Supplemental Figure 2.II). 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 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 $\mathcal{O}(pN^2)^2$, and therefore each of the three methods compared have the same asymptotic computational complexity.

2.2.II 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.

2.2.12 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 PCA⁵⁵. 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 (Supplemental Figure 2.12). 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 population 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.

2.2.13 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 $1 - 4\varphi$, 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

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

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, $H_0 : \varphi = 0$, 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. Supplemental Figure 8 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 Supplemental Materials 1.7).

2.3 RESULTS

2.3.1 IDENTIFICATION OF RELATEDNESS AND STRUCTURE IN 1000GP DATA

We applied our method to data from the 1000 Genomes Project (TGP) (^{13,14}), 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 (¹¹). However, subsequent analyses have discovered numerous inferred relationships closer than first cousins (^{27,2,23}).

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 (⁵⁶) and increased the independence of consecutive measurements (See Supplemental Methods). 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% (Supplemental Methods 1.4, Supplemental Figure 5). STEGO was then run on each of these populations separately to test for heterogeneity and relatedness within population groups (Figure 2.1, 2.3A).

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 individuals, 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 (2.3), 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, $(.2 < \hat{\phi}_{i,j} < .3)$, a set of pairings which corresponds identically with the conclusions of Gazal et al(²⁷).

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 populations 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 ($p < 10^{-30}$), whereas no other pairing exceeded the family-wise cutoff of 1.3 (Figure 2.1). Using the formula above, the estimated coefficient of kin-

ship 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 2.2) 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 (^{5,20}) methods. This results in a reduced set of 2261 individuals which are assumed to be no more closely related than half first cousins ($\phi = .0312$)⁽²⁷⁾. We applied this filter and re-analyzed each of the 26 populations again to test for heterogeneity and cryptic relatedness.(Figure 2.3B, Supplemental Table 1, Supplemental Figure 1)

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 ad-mixed 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) ($p = 7 \times 10^{-8}$), PUR (Puerto Ricans from Puerto Rico) ($p = 3 \times 10^{-31}$), and PEL (Peruvians from Lima, Peru) ($p = 2 \times 10^{-27}$). It is therefore reassuring that these groups of individuals would exhibit the greatest amount of structure among the populations surveyed.

2.3.2 POPULATION DIFFERENTIATION IN 1000 GENOMES PROJECT

There are many methods for detecting population structure. Most commonly, Principal Components Analysis (^{55,57}) 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 estimate genetic similarity (⁵⁸). This approach provides a higher resolution picture of the genetic landscape by exploiting the co-

Distribution of similarity statistic within population subgroups from 1000 Genomes Project

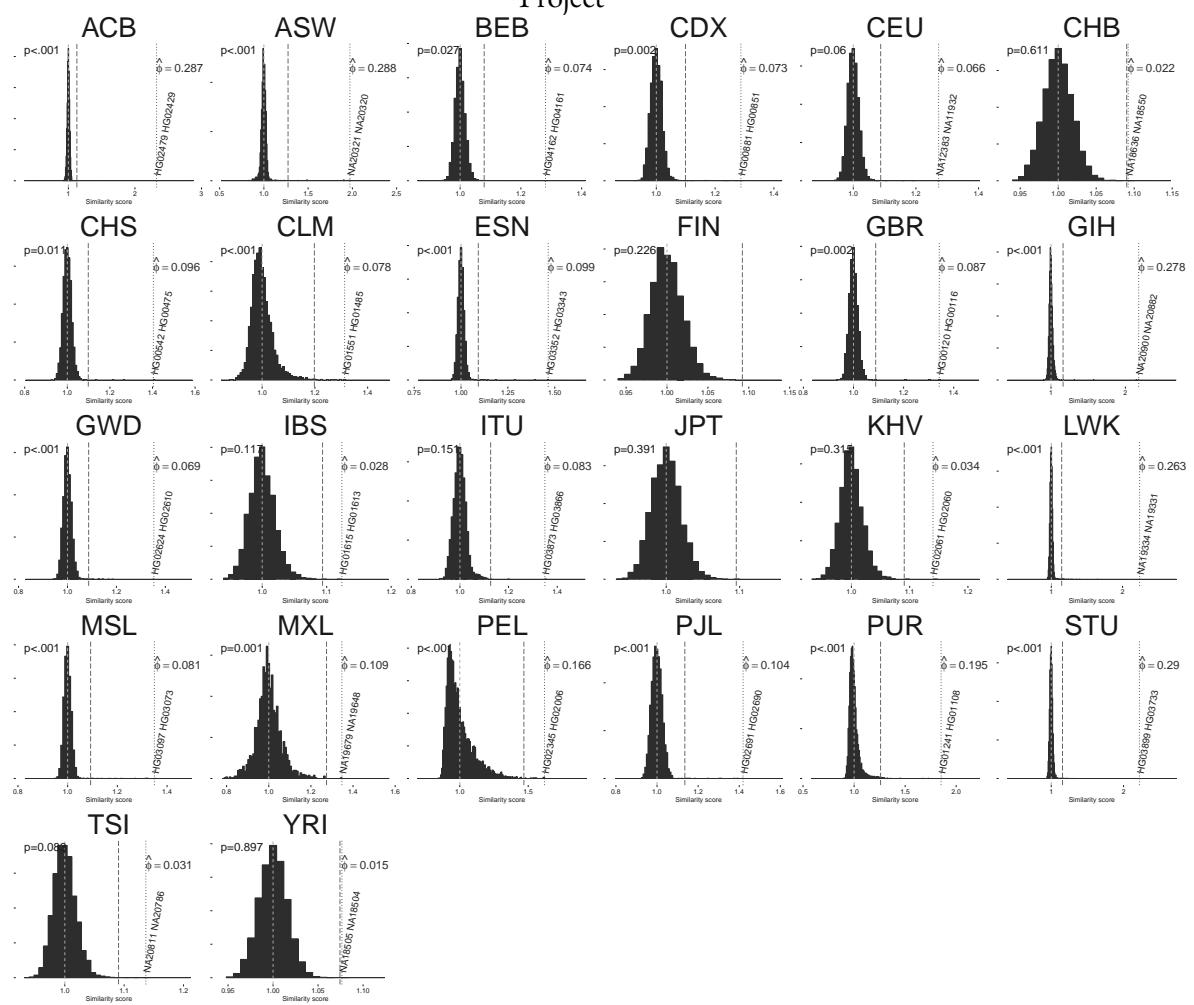


Figure 2.1: Distribution of similarity coefficients for each of the 26 populations in the 1000 Genomes Project. Homogeneous populations lacking cryptic relatedness should be expected to exhibit distributions centered around 1 with no outliers. The red dotted vertical line on each plot indicates the family-wise $\alpha = .01$ 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)

ITU with HG03998

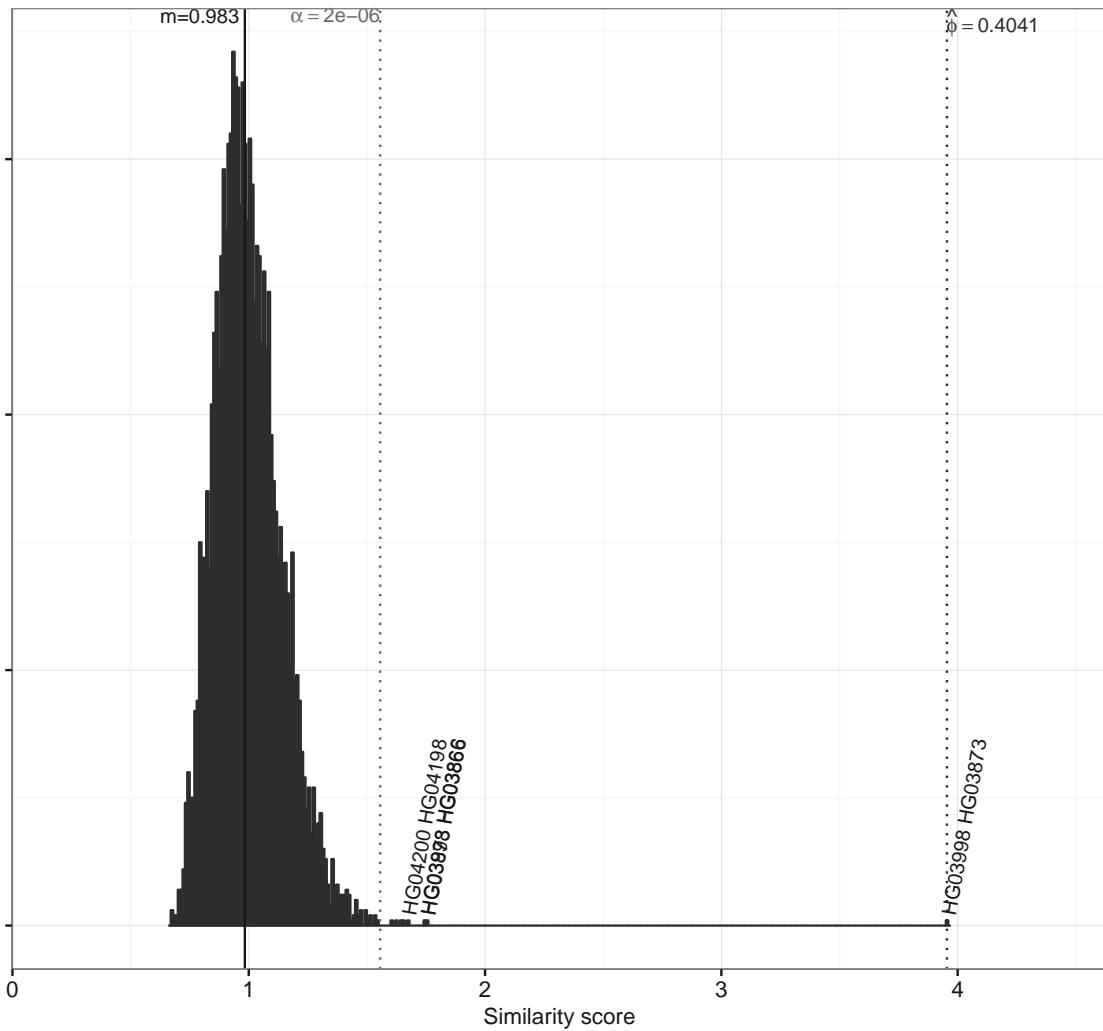


Figure 2.2: Distribution of all pairwise s statistics for population Indian Telugu from the UK (ITU) with individual HG03998 included. HG03998 is now believed to be related to HG03873, despite being labeled in the Sri Lankan Tamil from the UK (STU) population. The family-wise $\alpha = .01$ cutoff is indicated by the dotted red vertical line and the s statistic for HG03998 and HG03873 is seen as an extreme outlier at 3.97.

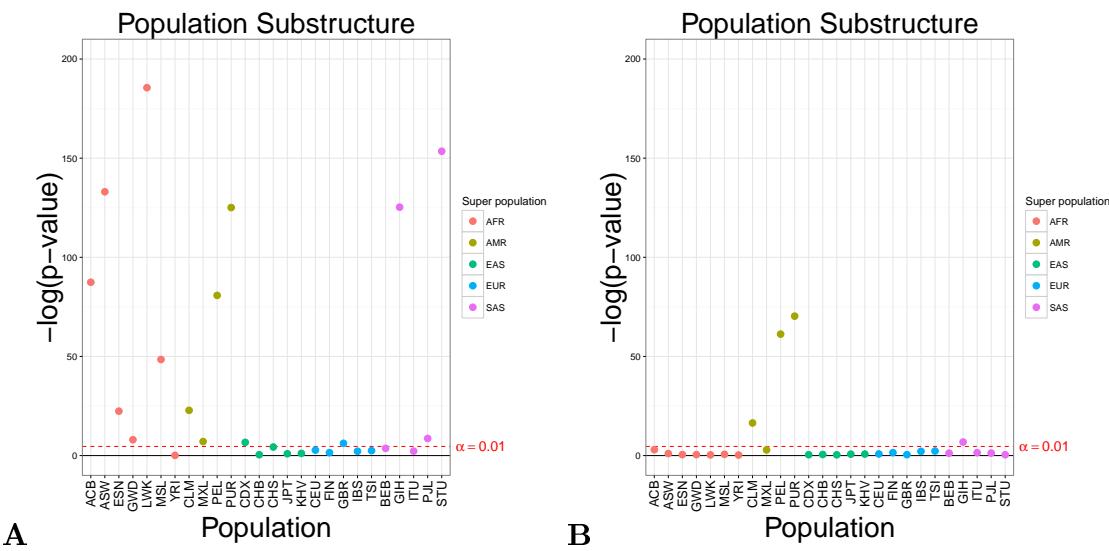
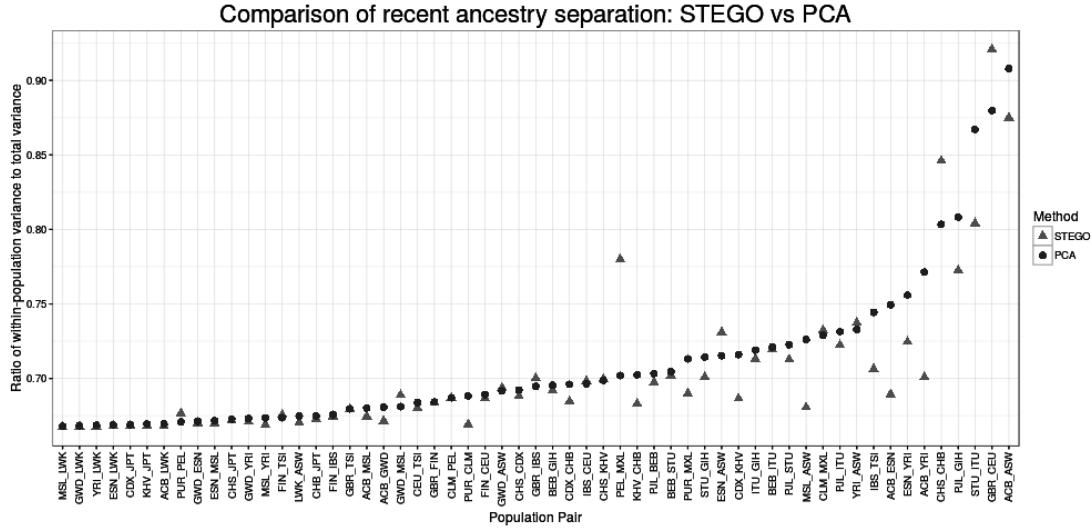


Figure 2.3: Significance of population heterogeneity in 26 populations of the TGP. Detection of population structure was found at $p < .01$ in 15 of the 26 populations using the full dataset (A). Upon removal of suspected related individuals, four populations (CLM, PEL, PUR and GIH) violated homogeneity in the relatedness-removed populations (B).

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 eigen-decomposition of each GSM was performed and each individual in the study was plotted against the top two eigenvectors for each method.



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 2.4). 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 2.5). 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 (Supplementary Figure 3).

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

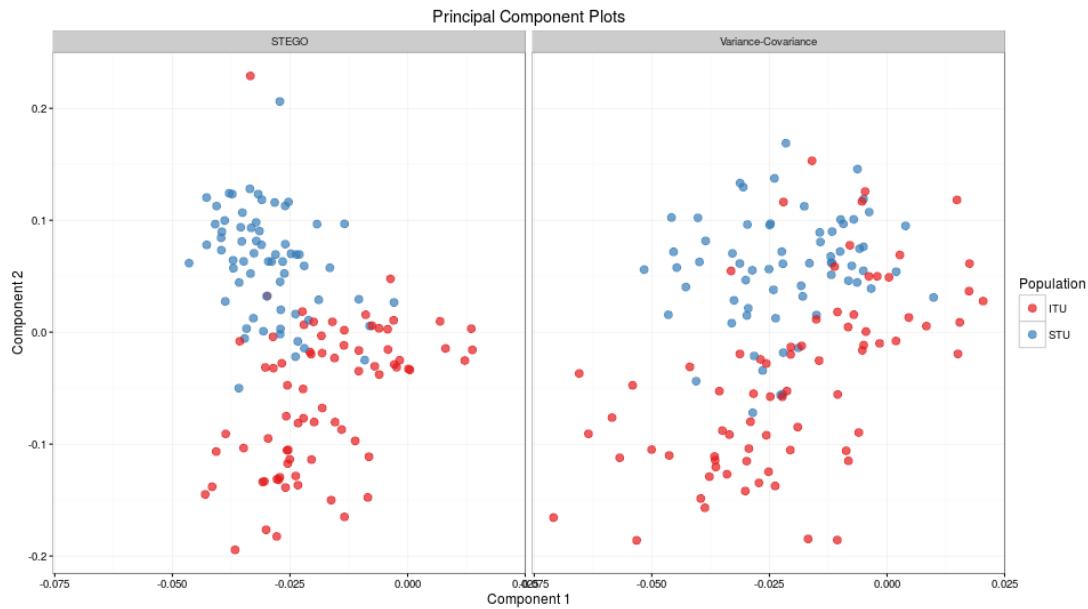


Figure 2.5: Example: ITU vs STU. Two populations of Southern Asian origin, Indian Telugu from the UK (ITU) and Sri Lankan Tamil from the UK (STU). A genetic similarity matrix was computed using STEGO and standard correlation on the same set of variants (See Supplemental Methods 1.4). An eigendecomposition 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.

(See Supplementary Materials 1.6). It stands to reason that these alleles would therefore be the most informative of recently related populations. By appropriately recognizing the increased information contained in the co-occurrences of less frequent alleles, we achieve superior separation of recently related populations.

2.4 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 assumption that genetic variants are independent or uncorrelated. However, this assumption is often violated in real-world datasets due to linkage disequilibrium (LD) between variants. LD occurs when variants are physically close on a chromosome and tend to be inherited together. This can lead to spurious associations if not properly accounted for.

tions that population homogeneity holds in the study. Checking for violations of these assumptions typically involves a qualitative assessment 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 (7) 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 1000, 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 quantification of the specific gains achieved in controlling type I error and power in the context of rare variant association studies. Higher resolu-

tion 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

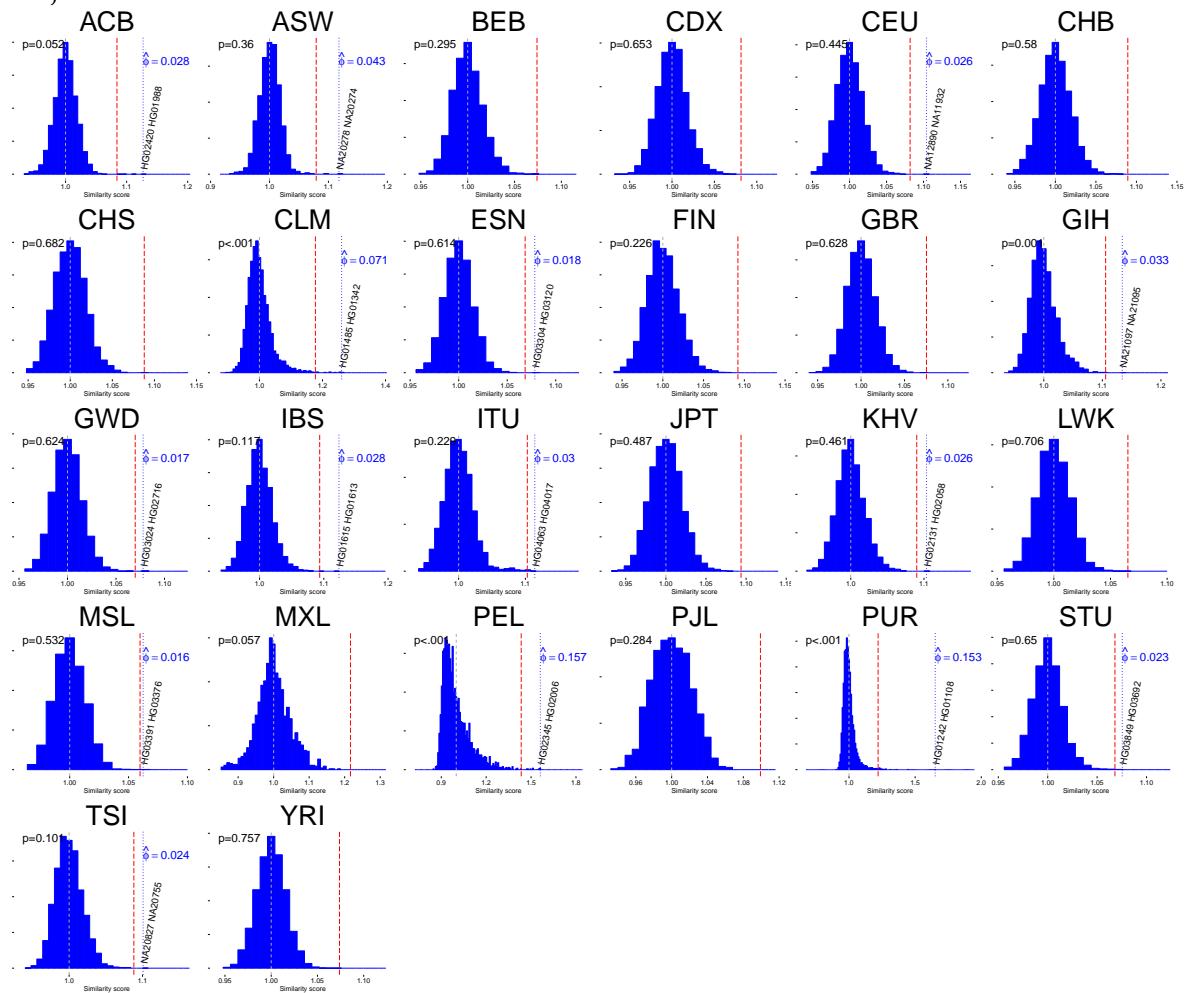


Figure 2.6: Distribution of similarity coefficients for each of the 26 populations in the 1000 Genomes Project after the removal of suspected related individuals. Homogeneous populations lacking cryptic relatedness should be expected to exhibit distributions centered around 1 with no outliers. A heterogeneous 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.

Genetic Similarity Matrix

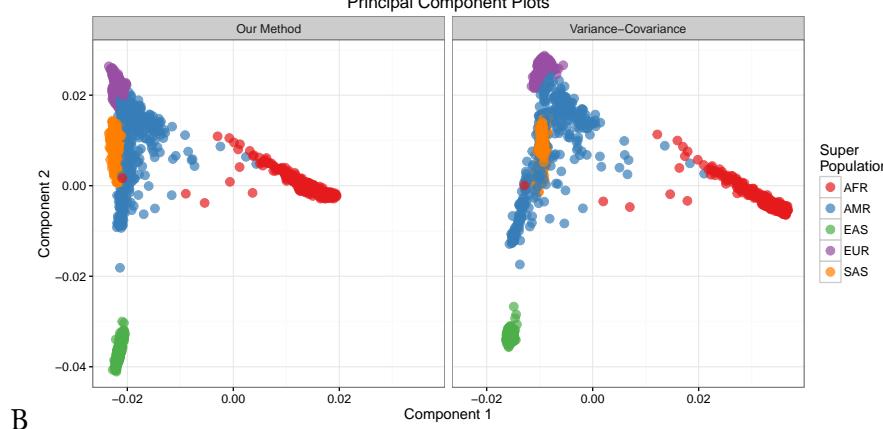
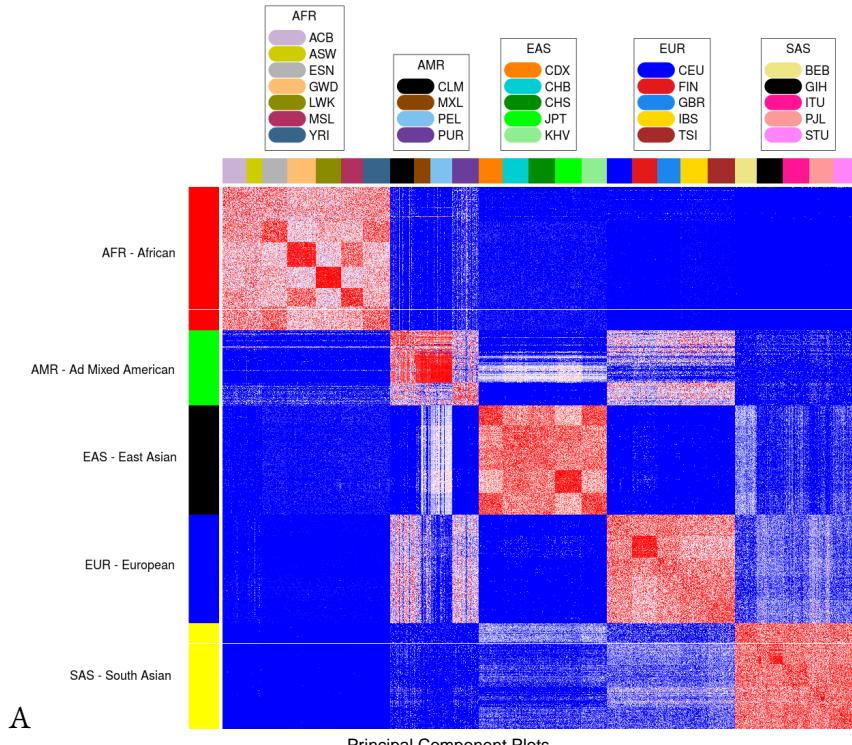


Figure 2.7: Population structure in 2504 samples from 1000 Genomes Project. (A) Heatmap of the GSM generated by STEGO using 80,000 LD-sampled variants. The vertical colorbar indicates membership in one of the five superpopulations, while the horizontal colorbar indicates membership in one of the 26 populations. (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.

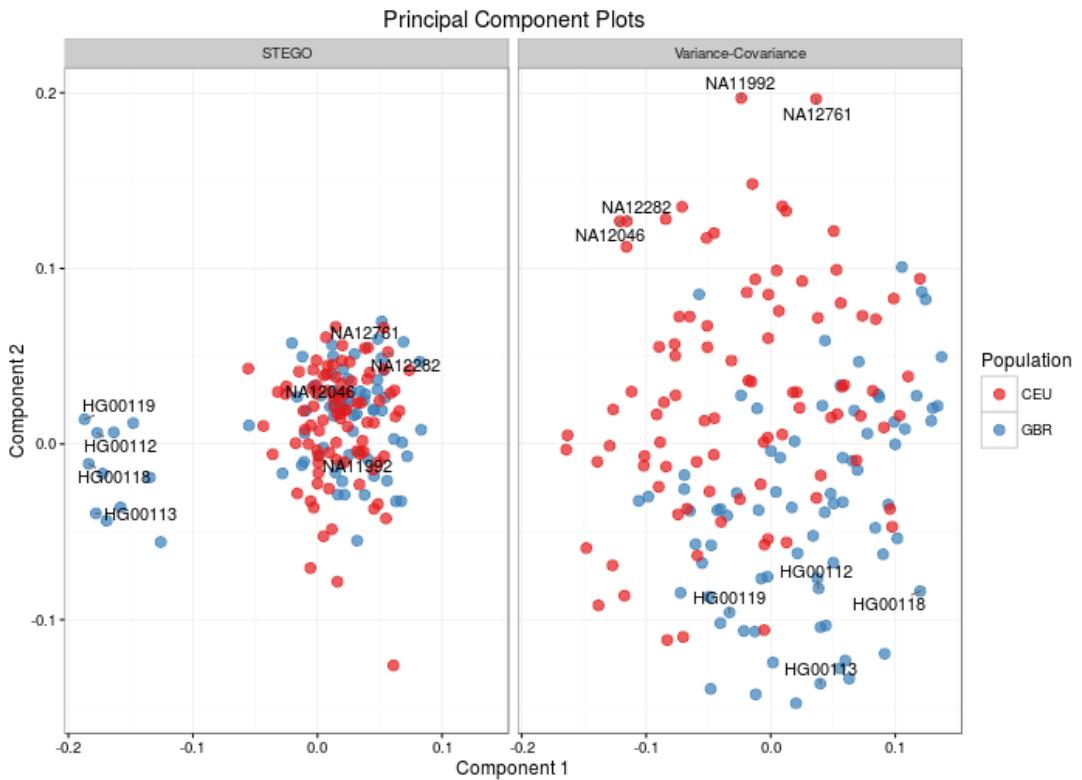


Figure 2.8: Despite a clear trend of superior performance with STEGO, notable exceptions occur. For example, by this measure, the populations GBR and CEU were more clearly divided by PCA (Right) than by STEGO (Left). Closer inspection revealed that the first eigenvector from STEGO isolates 11 samples 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-occurrences of rare variants, which were not observed separately by PCA.

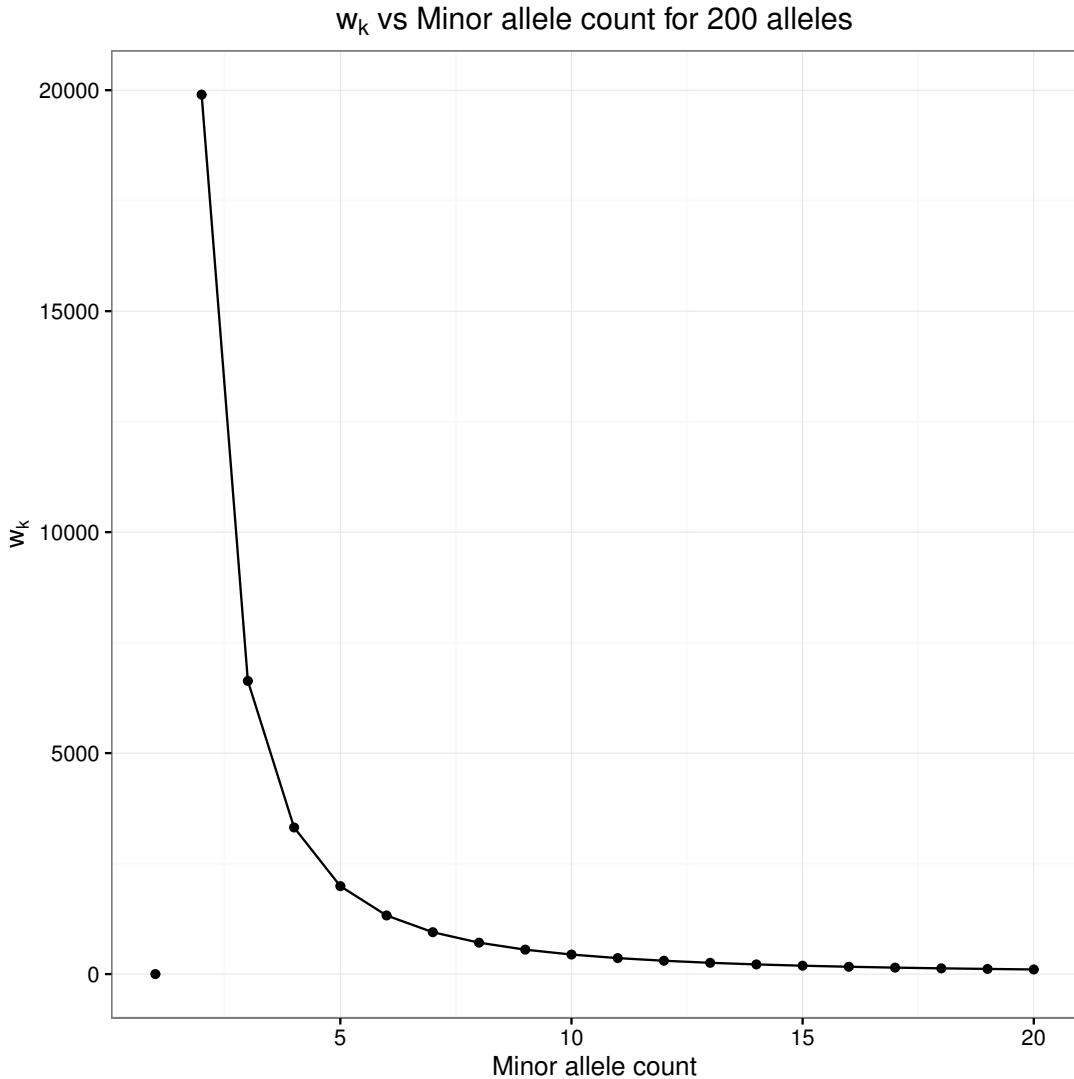


Figure 2.9: The weight factor, w_k , is shown here as a function of the minor allele count in an example of 100 individuals (200 alleles) for each locus, k . w_k is monotonically decreasing for minor allele counts greater than 1, lending greater power to lower frequency variants. Typically, there will be a minimum minor allele count such that the largest values for w_k are never obtained in practice

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 2.1: 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.

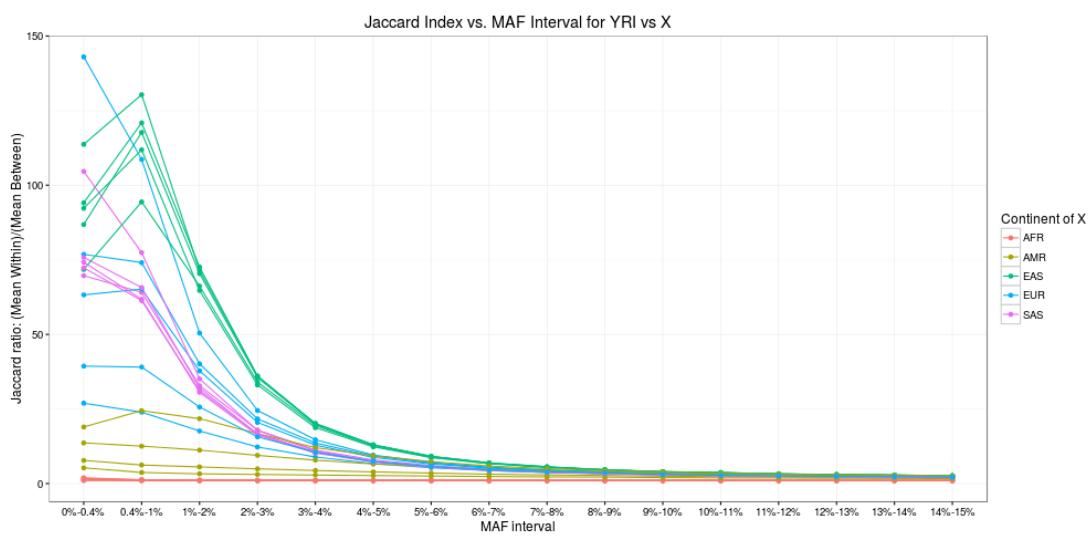


Figure 2.10: Lower frequency alleles are more informative of ancestry. For the Yoruban population (YRI), this plot compares the average unweighted Jaccard Index between individuals within group the to individuals in all other populations of the 1000 Genomes Project. When filtering by each minor allele 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.

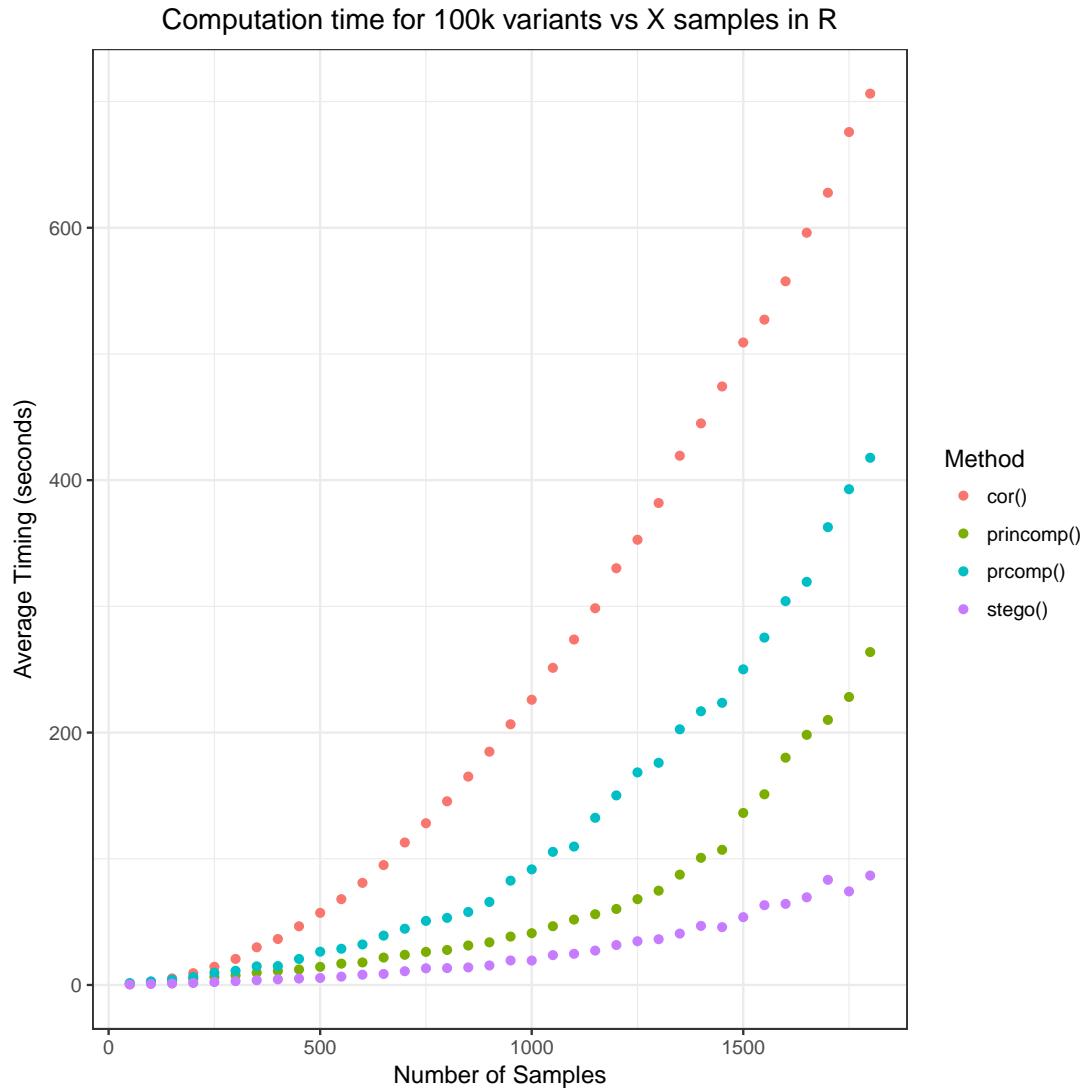


Figure 2.11: Average running time of STEGO is compared to the default implementations of `cor()` and `princomp()` in R. For each sample size on the x-axis, 100,000 variants were randomly generated across the samples. R functions for STEGO, correlation, and two implementations of PCA (`prcomp` and `princomp`) 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.

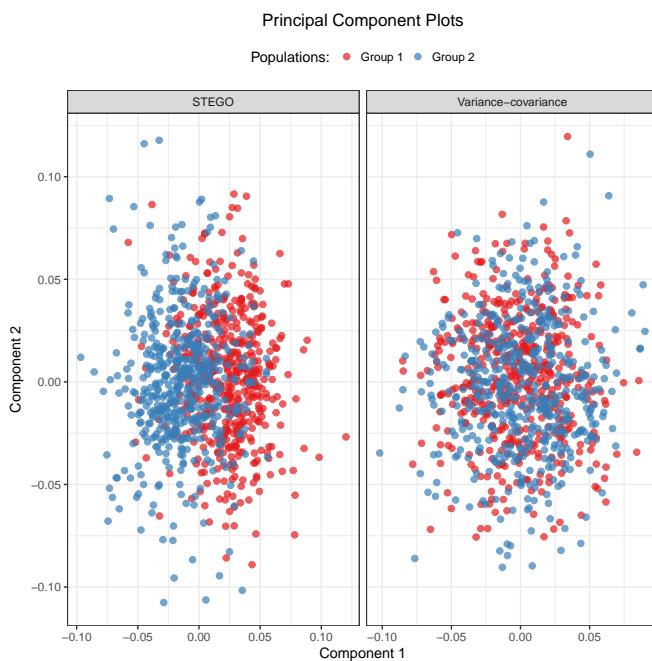


Figure 2.12: Principal Component plots for two methods for generating the genetic similarity matrix. On the left, the GSM is generated via STEGO and on the right the GSM uses the normalized covariance matrix. The STEGO method makes more efficient use of the more ancestry informative rare variants, providing 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.

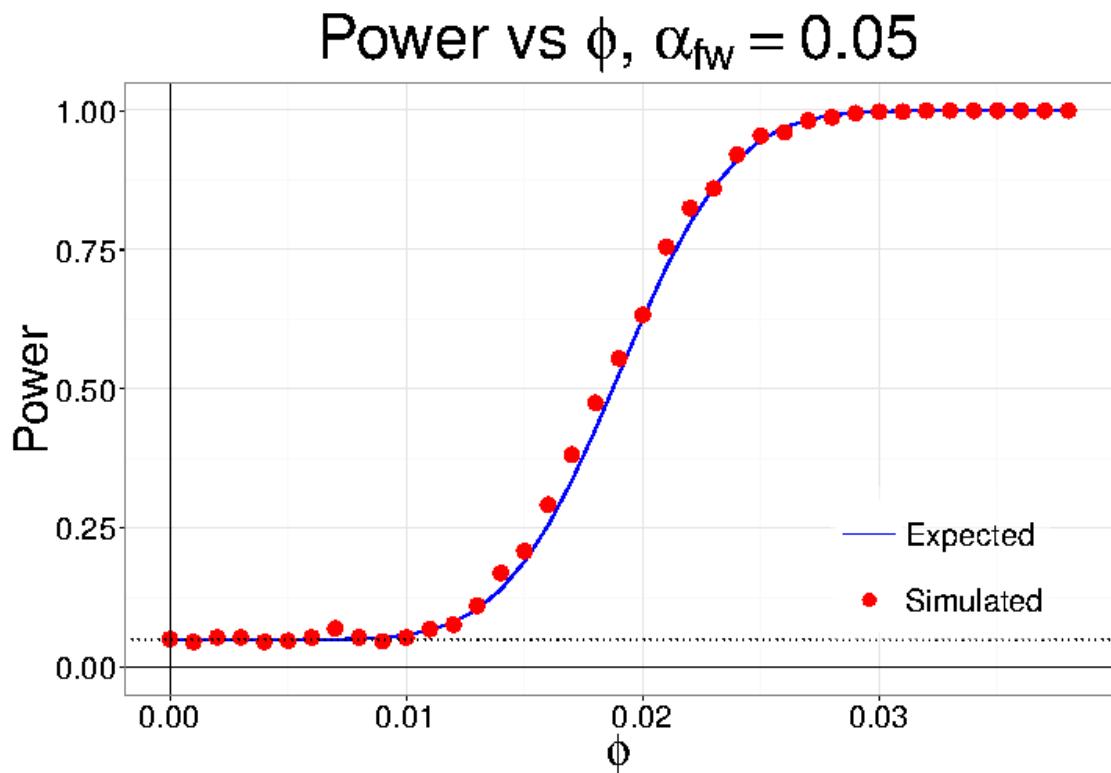


Figure 2.13: The probability of rejecting the null hypothesis given a simulated set of 301 homogeneous individuals containing a single related pair with coefficient of kinship, ϕ . The simulated power curve aligns with the analytically derived expectation demonstrating the clearly defined power of the method.

Nulla facilisi. In vel sem. Morbi id urna in diam dignissim feugiat. Proin molestie tortor eu velit. Aliquam erat volutpat. Nullam ultrices, diam tempus vulputate egestas, eros pede varius leo.

Quoteauthor Lastname

3

Consectetuer adipiscing elit

Some stuff about networks and batch effect... sdfsdg sdg ag afgfgfdgag

4

Conclusion

 Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra
 gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Sus-
 pendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla
 facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc
 eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ip-

sum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

A

Some extra stuff

 Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra
 gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Sus-
 pendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla
 facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc
 eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ip-

sum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

References

- [1] (2015). Lung genomics research consortium (lgrc). Accessed: 2016-02-02.
- [2] 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.
- [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] 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.
- [6] 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.
- [7] 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).
- [8] 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).
- [9] Chen, Z.-H., Kim, H. P., Sciurba, F. C., Lee, S.-J., Feghali-Bostwick, C., Stoltz, 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.

- [10] 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.
- [11] Choksi, S. P., Lauter, G., Swoboda, P., & Roy, S. (2014). Switching on cilia: transcriptional networks regulating ciliogenesis. *Development*, 141(7), 1427–1441.
- [12] Cloonan, S. M., Glass, K., Lauchó-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*.
- [13] Consortium, . G. P. et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), 56–65.
- [14] Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74.
- [15] 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.
- [16] 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.
- [17] Devlin, B., Roeder, K., & Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theoretical population biology*, 60(3), 155–166.
- [18] Du P, Kibbe WA, L. S. (2007). *Using lumi, a package processing Illumina Microarray*. Bioconductor R package.
- [19] 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.
- [20] 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.
- [21] Fahy, J. V. & Dickey, B. F. (2010). Airway mucus function and dysfunction. *New England Journal of Medicine*, 363(23), 2233–2247.

- [22] 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.
- [23] 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.
- [24] Filosi, M., Visintainer, R., & Riccadonna, S. (2014). *nettools: A Network Comparison Framework*. R package version 1.0.1.
- [25] Galvagni, F., Capo, S., & Oliviero, S. (2001). Spi and sp3 physically interact and co-operate with gabp for the activation of the utrophin promoter. *Journal of molecular biology*, 306(5), 985–996.
- [26] 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.
- [27] 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.
- [28] Glass, K., Huttenhower, C., Quackenbush, J., & Yuan, G.-C. (2013). Passing messages between biological networks to refine predicted interactions. *PloS one*, 8(5), e64832.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] 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.

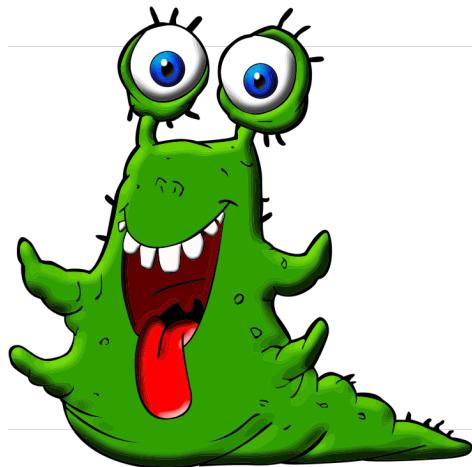
- [33] Hessel, J., Heldrich, J., Fuller, J., Staudt, M. R., Radisch, S., Hollmann, C., Harvey, B.-G., Kaner, R. J., Salit, J., Yee-Levin, J., et al. (2014). Intraflagellar transport gene expression associated with short cilia in smoking and copd. *PLoS one*, 9(1), e85453.
- [34] Hill, S. M., Lu, Y., 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.
- [35] Hogg, J. C. (2004). Pathophysiology of airflow limitation in chronic obstructive pulmonary disease. *The Lancet*, 364(9435), 709–721.
- [36] 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.
- [37] Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1), 118–127.
- [38] 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.
- [39] 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.
- [40] Langfelder, P. & Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1), 1.
- [41] Langfelder, P. & Horvath, S. (2012). Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software*, 46(11), 1–17.
- [42] Levine, M. & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945), 147–151.
- [43] 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.

- [44] 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.
- [45] 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.
- [46] 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.
- [47] 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.
- [48] 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).
- [49] Mathieson, I. & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, 44(3), 243–246.
- [50] 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.
- [51] Nemesh, J. & McCarroll, S. (2012). Addressing cryptic relatedness in candidate samples for ikkg. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phaseI/analysis_results/supporting/cryptic_relation_analysis. Accessed: 2016-06-06.
- [52] Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS genet*, 2(12), e190.
- [53] 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.

- [54] 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.
- [55] 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.
- [56] 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.
- [57] 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.
- [58] 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), i366–i372.
- [59] 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.
- [60] 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.
- [61] 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.
- [62] 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.

- [63] 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.
- [64] 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.
- [65] 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.
- [66] 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.
- [67] Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318), 626–633.
- [68] 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.
- [69] Sîrbu, A., Ruskin, H. J., & Crane, M. (2010). Comparison of evolutionary algorithms in gene regulatory network model inference. *BMC bioinformatics*, 11(1), 1.
- [70] 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.
- [71] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 267–288).
- [72] 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.
- [73] Voight, B. F. & Pritchard, J. K. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genet*, 1(3), e32.

- [74] Wang, J., Tsang, W. W., & Marsaglia, G. (2003). Evaluating kolmogorov's distribution. *Journal of Statistical Software*, 8(18).
- [75] 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.
- [76] Zhang, B. & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1).
- [77] 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.
- [78] 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.



THIS THESIS WAS TYPESET using L^AT_EX,
originally developed by Leslie Lamport
and based on Donald Knuth's T_EX.

The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under [CC BY-NC-ND 3.0](#). A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.