Supplementary Materials for
Estimating Drivers Cell State Transitions using Gene Regulatory Network Models

Daniel Schlauch[1] and Kimberly Glass[2,3] John Quackenbush[1,3]

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

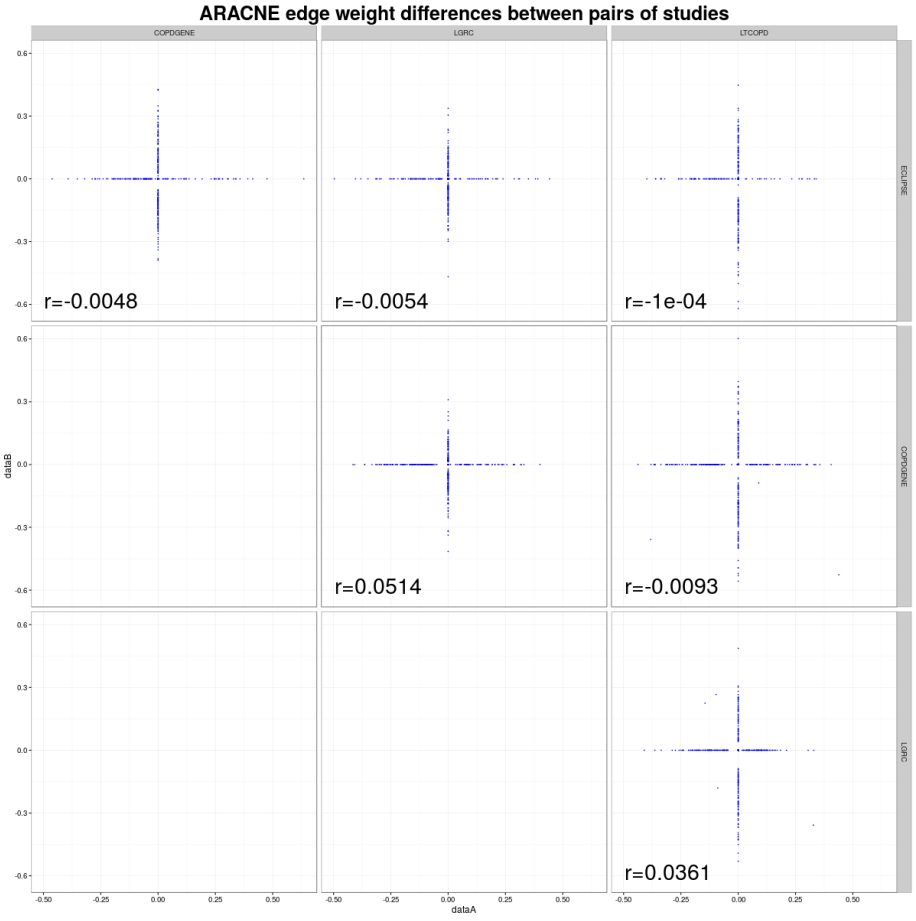
This PDF file includes:

- Materials and Methods
- Supplementary Text
- Figs. S1 to Sx
- Tables S1 to Sx

Materials and Methods

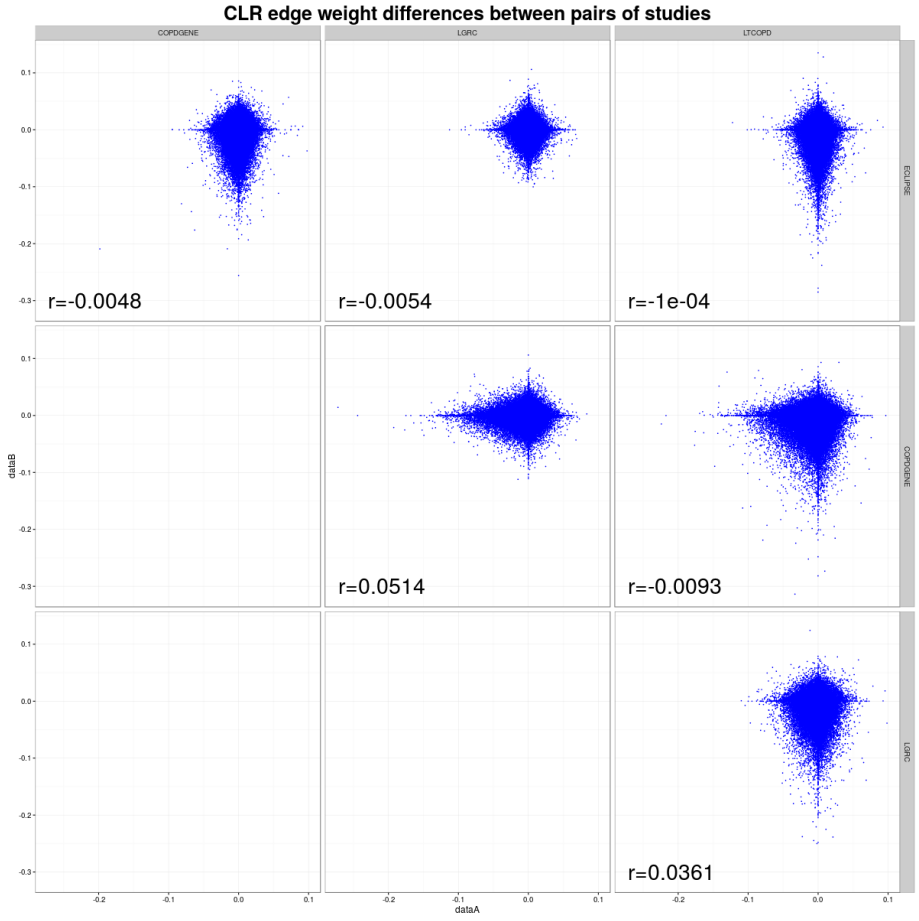
Common network inference methods do not produce reproducible results in COPD

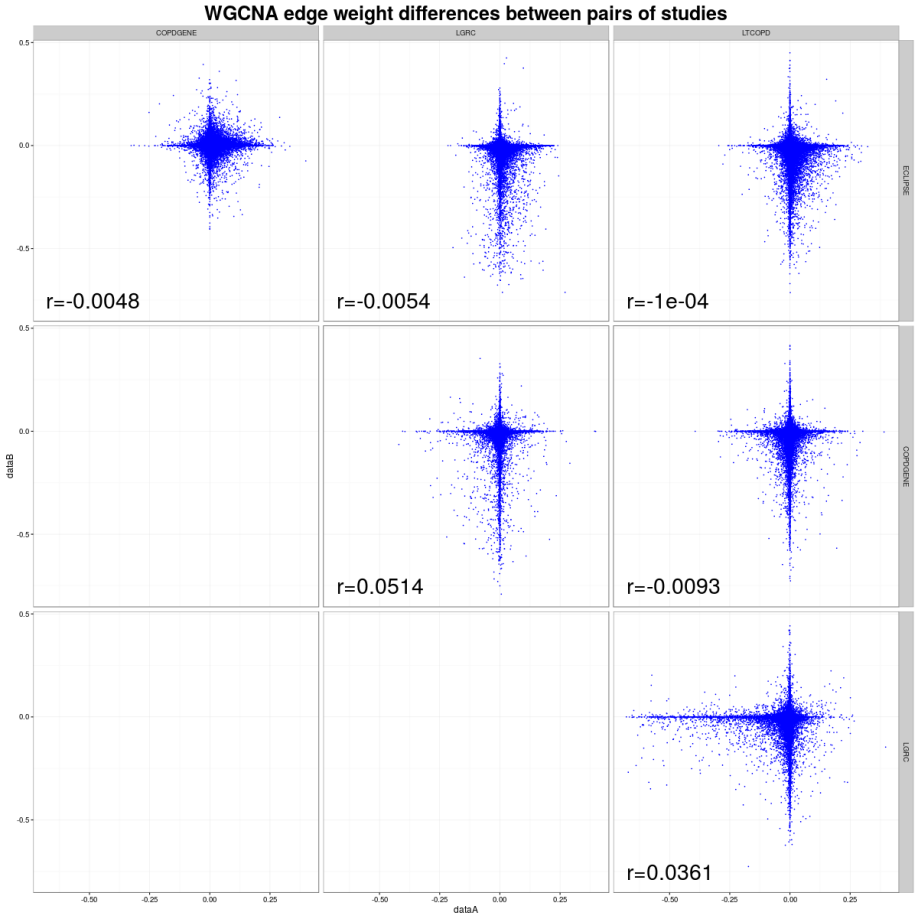
We evaluated the standard tools for COPD case-control studies at the network level by observing the agreement in differential network inference (NI) across studies. We utilizing four studies - Chronic Obstructive Pulmonary Disease (COPD)- Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE)[20], the COPDGene study [15], Lung Genomics Research Consortium (LGRC) [1] and Lung Tissue Chronic Obstructive Pulmonary Disease (LTCOPD) [?]. Our goal was to determine the ability of commonly used NI methods to find reproducible results across studies. We tested the methods Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE), Context Likelihood of Relatedness (CLR), and Weighted Gene Correlation Network Analysis (WGCNA). We computed differential edgeweights across cases and controls using each of the NI methods. Next, we plotted the differential edgeweights for each pairwise combination of studies [1]. We found no meaningful correlation for edgeweight differences between *any* two studies using *any* of the three methods.


ARACNE edge weight differences between pairs of studies
COPDGENE
LGRC
LTCOPD
ECLIPSE
COPDGENE
LGRC
r=-0.0048
r=-0.0054
r=-1e-04
r=0.0514
r=-0.0093
r=0.0361
dataB
dataA
0.6
0.3
0.0
-0.3
-0.6
-0.50
-0.25
0.00
0.25
0.50

A

B


CLR edge weight differences between pairs of studies
COPDGENE
LGRC
LTCOPD
ECLIPSE
COPDGENE
LGRC
dataB
dataA
r=-0.0048
r=-0.0054
r=-1e-04
r=0.0514
r=-0.0093
r=0.0361



C

Figure 1: Edgeweight differences between cases and controls do not correlate across studies. Using three commonly used network inference methods from gene expression data, network inference was performed separately on cases and controls. Here, the case-control difference is compared across studies of COPD. The methods tested were Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE) (**A**), Context Likelihood of Relatedness (CLR) (**B**), and Weighted Gene Correlation Network Analysis (WGCNA) (**C**). No detectable agreement between studies exist were found, regardless of tissue type.

We conclude that identifying differences in COPD case-control studies using existing methods for NI do not generate reproducible results.

Inferring Gene Regulatory Networks

In 2013, we described PANDA [4], a method [14] for estimating gene regulatory networks that uses "message passing"ă [3] to integrate multiple types of genomic

data. PANDA begins with a prior regulatory network based on mapping transcription factor motifs to a reference genome and integrates other sources of data, such as protein-protein interaction and gene expression profiles, to estimate individual sample networks. While PANDA has proven to be very useful in a number of applications [10, 6, 5], its iterative approach to edge-weight optimization limits its utility in situations requiring a large number of network bootstrap estimations. To address this limitation, we developed an approach which considers the available evidence of an edge for each possible TF-gene pair. This evidence can be divided into two components, referred to here as direct and indirect. Consider the edge between a TF and a gene, referred to here as TF_i and g_j, respectively. The direct evidence, $d_{i,j}$, consists of the squared conditional correlation of the g_i and g_j given all other regulators of g_i. Where g_i is the gene which encodes TF_i

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

Naturally, the use of direct evidence inadequately captures regulatory relationships due to the impacts of technical noise and numerous biological external factors such as stable or transient protein-protein interactions, post-translational modifications, etc. which may confound or modify a regulatory effect. These sources of confounding and variability in the expression pattern of a gene coding a TF may obscure the effects it has on all of its target genes. Therefore it is of value if we can complement our estimate of the likelihood of a regulatory mechanism by aggregating the information from the gene expression patterns of all suspected targets of transcription factors. PANDA achieves its superior performance in part by convergence towards "agreement", whereby large collections of gene expression patterns must agree with the proposed regulatory structure in order to claim an interaction. Similarly, we look for agreement between the gene expression patterns of large sets of co-targeted genes. We refer to this feature as indirect evidence and can achieve this by again utilizing our set of regulatory priors. In this portion of the analysis we suspend the recognition of a TF as a member of the gene list and instead consider each of the m TFs to be binary classifications across the entire gene list. Class labels are determined by the presence or absence of a sequence binding motif for that TF in the vicinity of the gene.

The indirect evidence between the two nodes, $\theta_{i,j}$, represents the fitted probability that g_i belongs to the class of genes targeted by TF_j. g_i is considered to be a new observation placed into the n-dimensional space separated by transcription factor targets and non-targets. To divide up the space, we use a logistic regression on the gene expression data with outcome taken to be the existence or non-existence of a known sequence motif for TF_j upstream of g_i.

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

where the response M_j is a binary vector of length n indicating the of the presence of a sequence motif for transcription factor j in the vicinity of each of the n genes. And where $g_{(k)}$ is a vector of length n specifying the gene expression for sample k over n genes.

For some TF-gene pair, the fitted values for each $TF_j - g_i$ pair define the "indirect" evidence $\theta_{j,i}$.

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

By scoring each gene according to the strength of indirect evidence for a regulatory response to each of the TFs, we can combine this with the direct evidence of regulation (squared conditional correlation of expression for g_i and TF_j). The appropriate manner in which to combine direct and indirect evidence remains an open question. Though both measures are bounded by [0,1] their interpretation is quite different. The direct evidence can be considered in terms of it's conditional gene expression R^2 between nodes, while the indirect evidence is interpreted as a probability. We use a non-parametric approach to combine evidence. The targets of each TF are then ranked and combined as a weighted sum, $w_i = (1-\alpha)\left[rank\left(d_i\right)\right] + \alpha\left[rank\left(e_i\right)\right], i \in \{1, \ldots, n\}$. Our choice of the weight, α, here is based on empirical evaluation, and perhaps not surprisingly, is loosely correlated with organism complexity. In validation sets from Yeast, the optimal alpha was observed near $\alpha = .9$ while simpler E. coli datasets saw an optimal value of $\alpha = .6$ and an in silico dataset, optimal α was achieved at $\alpha < .5$ This naturally reflects the fact that the increased complexity of the network necessitates the use of larger scale agreement between genes, rather than a reliance on pairwise correlations between potentially noisier and more complex expression patterns.

Network edges are recovered in In Silico, E. coli and Yeast (*Saccharomyces cerevisiae*)

A common challenge in gene regulatory network inference is the difficulty of validating identified networks against a known, reliable gold standard. A number of validation methods have been used for in vivo samples, including ChIP-seq, ChIP-chip. Knockdowns have also been used in cell lines and have been shown to be effective at predicting in vivo responses [14]. Additionally, in silico methods have been used which simulate gene expression datasets based on a predefined set of regulatory mechanisms.

To test our methods, we used four test datasets of increasing biological complexity- (1) in silico, (2) E. coli, and (3) Yeast with simulated motif priors and (4) Yeast with biological motif priors. Data from (4) was collected from *Saccharomyces cerevisiae* with TF knock-out and stress conditions [7]. Data from the first three sources was obtained from the publicly available DREAM5 challenge[12]. This challenge asked contestants to infer gene networks from expression data alone, using a gold standard for evalution. Instead, we started with the gold standard and swapped a number of edges to create the type I and type II error rates consistent with Yeast motif priors and evaluated the performance of MONSTER to refine its predictions of the true gold standard. The estimated edges in MONSTER utilizing the gene expression data were demonstrated to be superior to those of the edge prior alone.

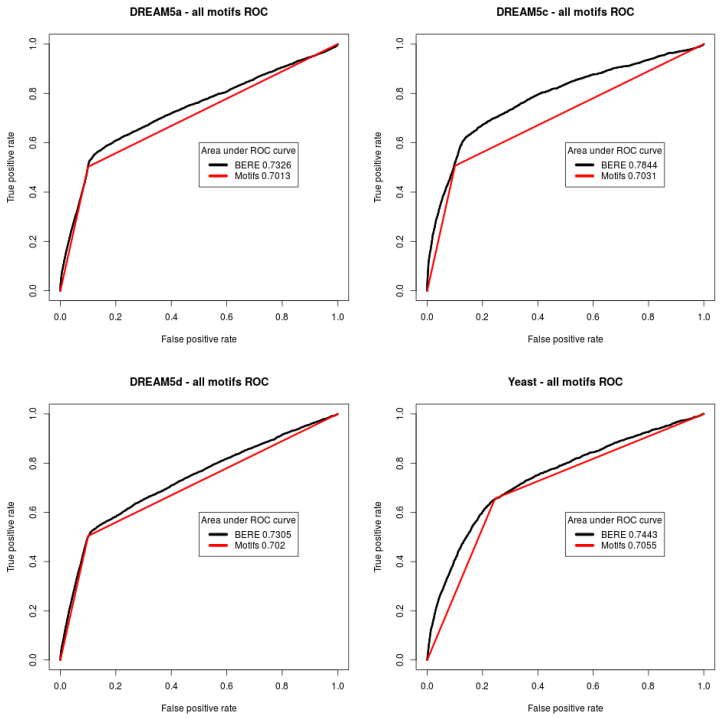


Figure 2: ROC curves for (1) in silico, (2) E. coli, (3) Yeast, demonstrating at least comparable prediction performance for both MONSTER and PANDA in refining edges beyond what is obtained using sequence motif priors.

Computation of MONSTER's transition matrix

A primary purpose of reconstructing GRNs is the understanding of the biological mechanisms which characterize disease. Identifying differential TF targeting may suggest a therapeutic target, uncover a disease substructure or identify biomarkers for early detection. However, significant limitations exist with respect to generating reliable context-specific GRNs. Substantial advances have been made in this area with algorithms like SEREND and PANDA, but these methods rely more heavily on the static sequence motif data rather than gene expression data collected across groups, such as in a case-control study. Both methods, along with MONSTER's network inference approach, use gene expression data as a refinement of the more accurate motif priors. It is therefore a side effect that any GRN inference method that relies on motifs directly for edgeweight calculation will be susceptible to having the bulk of its predictive power stripped when making comparisons between networks. Consequently, it is of critical importance to choose a network inference method that uses the context-specific data to most accurately refine above what information is gleaned from our static data.

The task of identifying meaningful network transitions then becomes an evaluation of the relative refinement of edgeweights. Since the majority of the predictive power for each edge is contained in the motif contribution, we are left with relative edgeweight refinement that have a low signal, high noise. In other words, we have a very large number of individually unreliable edgeweights. In effort to extract the maximum effect, we seek to combine the information contained in each edge via a novel dimension reduction approach.

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

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

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

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

or alternatively expressed

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

where $E\left[e_{ij}\right] = 0$

This can be solved with normal equations,

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

Which produces the least squares estimate. I.e. loss function $L(\mathbf{T}) = \sum_{gene=1}^{N} ||\mathbf{B}_{gene} - \mathbf{A}_{gene}\mathbf{T}||^2$ is minimized. We further extend this method to include a penalty term[19]. An L_1 regularization is used by creating 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 priority for the k^{th} regression coefficient in the k^{th} regression model.

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

This solution is obtained using the "penalized" library in R as the minimization of the penalized squared loss function

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

This example illustrates a key feature of this method. Specifically, that the transition matrix 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 easily interpreted. We can think of a column, τ_i, on the matrix $\mathbf{T}$ as containing the linear combination of regulatory targets of TF_i in $\mathbf{A}$ that best approximates the regulatory targets of TF_i in $\mathbf{B}$. As one would expect, a large proportion of the matrix "mass" would be on the diagonal for those TFs which do not change regulatory behavior between case and control. It is therefore of interest to evaluate values off of the diagonal as indications of a network transition.

Evaluating the Transition Matrix

Many mechanisms which may be differentially present, such as RNA degredation, post-translational modification, protein-level interactions and epigenetic alterations have the ability to impact downstream targeting without impacting the expression level of the TF itself. It may be of particular scientific or therapeutic interest to identify those TFs which have undergone significant overall changes in behavior between controls and cases. With that objective in mind, we express the statistic- differential Transcription Factor Involvement (DTFI), as a measure for quantifying this property.

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

DTFI can be loosely interpreted as the proportion of TF targeting patterns which is explained by the targeting patterns of other available TFs. This measure, a statistic on the interval $[0, 1]$ seeks to elucidate transitions which are systematic, informative, and non-arbitrary in nature by capturing only the edgeweight signal for which there is an attributable regulatory pattern. The distribution of this statistic under the null has a mean and standard deviation which depend on the motif structure. In particular, both mean and standard deviation are increased for TFs which have fewer prior regulatory targets. From a statistical perspective, TFs with relatively more targets are able to generate more stable targeted expression patterns, which leads to more consistent estimates in "agreement" algorithms such as PANDA. From a biological perspective, increased motif presence may indicate that the TFs are more likely to be ubiquitous housekeeping proteins that do not meaningfully alter their involvement between cases and controls. The dependence of the null distribution on the motif structure is addressed via the following resampling procedure.

1. Gene expression samples are randomly assigned to case and control forming the null-case and null-control with group sizes preserved.
2. GRNs are reconstructed for the null-case and null-control with the same prior regulatory structure.
3. The transition matrix algorithm is applied for the two null networks.
4. The differential TFI is calculated for each TF.
5. Repeat 1-4 1000 times.

MONSTER significantly improves TF-TF edge estimation from *in silico* gene expression data

To evaluate the ability of our method to recover edges between transcription factors, we generated simulated gene expression data. We began by generating a true controls adjacency matrix, $M_{0(p\times q)}$, describing the weighted edges between q transcription factors and p genes. A state transition was generated by sampling 100 TF-TF pairs and adjusting the edgeweight at the corresponding point on the true cases adjacency matrix, $M_{1(p\times q)}$. These TF-TF pairs ultimately represent the edges that we seek to recover and the size of the adjustments are the parameters of interest. We sampled from a multivariate Gaussian distribution with the off-diagonal of the variance-covariance matrix, Σ, defined as the $M_0 M_0'$. Furthermore, we scaled the magnitude of the diagonal of Σ to achieve the desired proportion of noise. We aimed for an area under the curve of the receiver-operator characteristic of approximately 0.70 as this has reasonably been achieved in existing biological studies [4]. We note that the two simulated regulatory priors have AUC-ROC of .570 and .547, which feeds the MONSTER and PANDA algorithms with priors which are substantially less predictive than sequence motif priors commonly used for network inference methods.

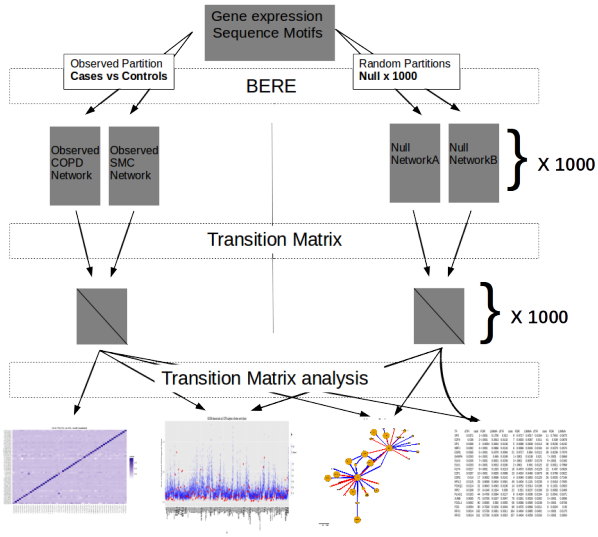


Figure 3: **Overview of Transition Matrix analysis workflow.** (1) Network inference is computed separately to subsets of the gene expression data including the cases group, the controls group and 1000 permutations of the cases and controls labels. (2) Transition matrix is estimated between the cases and controls and each of the pair or permuted "cases" and "controls".

This sampling represented our simulated control samples. The adjusted adjacency matrix, M_1, was similarly used to generate simulated expression data for the cases group. Next, we reconstructed the networks from our expression data using a set of commonly used network inference methods - Weighted Gene Correlation Network Analysis (WGCNA) [8] [9], Topological Overlap Measure (TOM) [17], Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE) [13], Context Likelihood of Relatedness (CLR) [2], Passing Attributes between Networks for Data Assimilation (PANDA) [4] and simple Pearson correlation (PC).

We applied the transition matrix with default parameters on each case-control pair of networks. For comparison, we estimated the difference from case to control in edgeweights derived from the direct edge prediction using each network inference method. The predictions for the TM approach and the direct approach were evaluated by the area-under-the-curve of the receiver-operator-characteristic (AUCROC) with the true transition adjustments taken as the gold standard. For each of the network inference methods tested, we found substantial improvement in the predicted transitions over the direct network inference method. In many cases, the edgeweight difference (column 2) was not statistically significant for predicting transitions, but when the TM was applied (column 3) a strong predictive signal appeared. In other cases, an existing signal was observed using the direct approach, but was dramatically improved with the application of the TM.

The intuition behind the improvement is simple. While the estimation of a TF-TF edge is typically evaluated via some pairwise gene expression pattern which may be rife with technical and biological noise, the TM approach borrows information from all downstream targets in estimating the relative change in relationship between the TFs.

Transition matrix finds significant protein-protein interaction

As noted above, there are numerous biological regulatory mechanisms which may yield detectable transitions. Of particular interest are those which are less readily detectable via conventional methods, such as differential gene expression analysis. One mechanism studied here involves one TF binding to another TF to promote, suppress or alter one or both of their regulatory patterns. These multi-protein interactions create combinatorial complexity that can explain much of the variation in organism complexity which is unexplained by number of genes alone [11].

To test the ability to detect protein level interactions, we compared our estimated transitions to a set of known protein-protein interactions [16]. This set contained 223 pairwise interactions between a total of 189 transcription factors. Of interest was the effectiveness of identifying these interactions via the transition from one phenotypic state to another. This is a challenging task for several reasons, (1) protein-protein interaction is merely one of a myraid of detectable transition mechanisms, (2) it is reasonable to assume that only a small subset of the known PPI are actually differentially present between case and control and

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

NI Method	Network AUC	Edgeweight differences	Transition Matrix
Pearson	.704	.510 (p=.72)	.802 (p<.0001)
WGCNA(6)	.704	.512 (p=.61)	.688 (p<.0001)
WGCNA(12)	.704	.52 (p=.10)	.589 (p=.02)
ARACNE	.515	.523 (p=.58)	.566 (p=.09)
CLR	.694	.57 (p=.19)	.814 (p<.0001)
TOM	.703	.51 (p=.62)	.689 (p<.0001)
PANDA*	.747	.520 (p=.13)	.793 (p<.0001)
PANDA**	.652	.509 (p=.43)	.66 (p<.0001)

Table 1: **Comparison of edgeweight 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. For PANDA* and PANDA**, which additionally utilizes motif prior information, motif priors with AUC-ROC of .570 and .547 were used. The naive TF-TF transitions were calculated as the difference in TF-TF edgeweight between cases and controls. The transition matrix TF-TF transitions used the absolute transition matrix values.

(3) technological limitations in the active field of proteomics cannot be expected to identify all interactions with a reasonable degree of certainty.

For the transition from Smoker control to COPD, we compared for the transition estimates to those generated by the transition from randomly resampling the phenotypic labels. The AUCROC for the prediction of the 223 known protein-protein interactions was $0.5695\,(p = 0.018)$, suggesting that our approach is successful at detecting this highly obscured signal.

Impact of homogeneity of cases and controls on statistical inference

Many network inference methods use a measurement of pairwise co-expression as a sufficient statistic for building gene networks based on gene expression data. The ability to detect coexpression in a sample is a function of both the level of extraneous noise and the biological variability within the sample. In our analysis, we compared the transition from the observed controls to observed cases and compared it to the null distribution estimated from a randomly sampled partition of mixed phenotypes. It is therefore of interest to consider the impact of generating gene regulatory networks from samples with heterogeneous versus homogeneous phenotypes. We consider whether transitions between *any* two homogeneous networks yields increased variance of test statistics and inflation of type I error when using heterogeneous populations for estimating the null distribution of these test statistics.

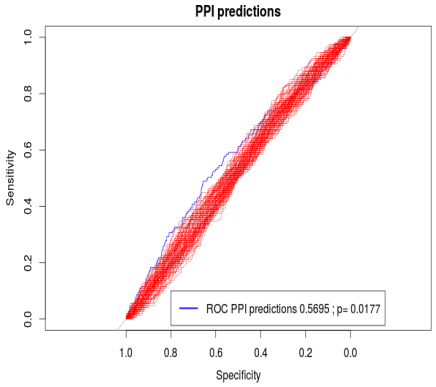


Figure 4: ROC curve for prediction of PPI based on a transition from Smoker control to COPD (blue) compared to a random case-control partition of the ECLIPSE data.

To explore, we sampled 100 cases and 100 controls. Within both cases and controls, we split each into two groups, denoted Cases_A, Cases_B, Controls_A, Controls_B, each of size 50. We performed network inference on each group and ran the transition matrix for each pairwise network transition. The null networks were generated via a heterogeneous sampling without replacement of 50 samples for the null cases and null controls.

Of interest is the relative distribution of test statistics in transitions across-phenotypes compared to within-phenotype, relative to the null (heterogeneous) transitions.

We find that the distribution of dTFI for homogeneous within-phenotype transitions (Cases_A $\rightarrow$Cases_B, Controls_A $\rightarrow$ Controls_B) closely matched the distribution of dTFI under the heterogeneous null. Comparatively, across-phenotype transitions showed strongly significant results compared to the heterogeneous null. We view these results as strong evidence that our findings in this manuscript are not inflated due to the comparison of homogeneous observed networks relative to heterogeneous null networks.

Efficiency of estimation

Let $\mathbf{x}_p$ be a Gaussian p-vector representing a sample of gene expression data containing q transcription factors and $p-q$ non-transcription factor genes.

$$\mathbf{x}_p \sim N(\mu, \Sigma)$$

where μ is the p-vector of mean gene expression values and Σ is the $p \times p$ variance-covariance matrix. In this scenario, Σ may be regarded as a combination of two independent variance-covariance sources- (1) biological signal, (2) biological noise and technical noise.

In investigating gene regulation, many network inference methods are constructed for the estimation of the $p \times q$ subset of Σ pertaining to the effect of the q TFs on the p genes. In identifying drivers of state transitions, we seek to focus on the $q \times q$ matrix of TF-TF effects. We show that our method vastly ourperforms commonly used network inference methods in estimating these specific effects.

Consider a state change between two experimental conditions, A and B, characterized by an alteration of size δ to the biological signal component of the TF-TF variance-covariance matrix at point $\Sigma_{i,j}$ where i and j are indices for two TFs in Σ.

Using a univariate coexpression calculation (the basis for Pearson networks and WGCNA estimates), the estimated variance of our estimate of δ can be calculated:

$$-\rho_A < \delta < \rho_A, \delta + \rho_A \leq 1$$

$$\begin{aligned} Var\left(\hat{\rho}_{i,j,A} - \hat{\rho}_{i,j,B}\right) &= Var\left(\hat{\delta}_{cor}\right) \\ & Var\left(\hat{\rho}_{i,j,A}\right) + Var\left(\hat{\rho}_{i,j,B}\right) \\ &= \frac{1-\rho_{i,j,A}^2}{n_A - 2} + \frac{1-\rho_{i,j,B}^2}{n_B - 2} \\ &= \frac{1}{n_A - 2} + \frac{1}{n_B - 2} - \frac{\rho_{i,j,A}^2}{n_A - 2} - \frac{\rho_{i,j,B}^2}{n_B - 2} \end{aligned}$$

Meanwhile, in condition B the new correlation of TF_i with some gene, $gene_k$ $k \in 1, 2 \ldots p$, denoted cor^*, becomes

$$cor^*\left(TF_i, gene_k\right) = cor\left(TF_i, gene_k\right) + \delta cor\left(TF_j, gene_k\right)$$

where the term $\delta cor\left(TF_j, gene_k\right)$ is the change due to the interaction of TF_i and TF_j.

The variance of our estimate using the transition matrix can be expressed

as follows:

$$\begin{aligned}
Var\left(TM_{i,j}\right) &= Var\left(\hat{\delta}_{TM}\right) \\
&= \frac{\left(\frac{1}{p}\right)\sum_{k=1}^{p} vVar\left(\hat{\rho}_{i,k,A}-\hat{\rho}_{i,k,B}\right)}{\sum_{k=1}^{p}\left(\rho_{j,k}-\bar{\rho}_{j}\right)^{2}} \\
&= \frac{\left(\frac{1}{p}\right)\sum_{k=1}^{p}\left[Var\left(\hat{\rho}_{i,k,A}\right)+Var\left(\hat{\rho}_{i,k,B}\right)\right]}{\sum_{k=1}^{p}\left(\rho_{j,k}-\bar{\rho}_{i}\right)^{2}} \\
&\leq \frac{\left(\frac{1}{p}\right)\sum_{k=1}^{p}\left[\frac{1}{n_A-2}+\frac{1}{n_B-2}\right]}{\sum_{k=1}^{p}\left(\rho_{j,k}-\bar{\rho}_{i}\right)^{2}} \\
&\leq \frac{\frac{1}{n_A-2}+\frac{1}{n_B-2}}{\sum_{k=1}^{p}\left(\rho_{j,k}-\bar{\rho}_{i}\right)^{2}} \\
&\leq \frac{Var\left(\hat{\delta}_{cor}\right)+\frac{\rho_{i,j,A}^{2}}{n_A-2}+\frac{\rho_{i,j,B}^{2}}{n_B-2}}{\sum_{k=1}^{p}\left(\rho_{j,k}-\bar{\rho}_{i}\right)^{2}} \\
&\leq Var\left(\hat{\delta}_{cor}\right)+\frac{Var\left(\hat{\delta}_{cor}\right)\left(1-\sum_{k=1}^{p}\left(\rho_{j,k}-\bar{\rho}_{i}\right)^{2}\right)+\frac{\rho_{i,j,A}^{2}}{n_A-2}+\frac{\rho_{i,j,B}^{2}}{n_B-2}}{\sum_{k=1}^{p}\left(\rho_{j,k}-\bar{\rho}_{i}\right)^{2}}
\end{aligned}$$

So we have that $Var\left(TM_{i,j}\right) < Var\left(\hat{\delta}_{cor}\right)$ when

$$Var\left(\hat{\delta}_{cor}\right)\left(1-\sum_{k=1}^{p}\left(\rho_{j,k}-\bar{\rho}_{i}\right)^{2}\right) < \frac{\rho_{i,j,A}^{2}}{n_A-2}+\frac{\rho_{i,j,B}^{2}}{n_B-2}$$

Since each term except $\left(1-\sum_{k=1}^{p}\left(\rho_{j,k}-\bar{\rho}_{i}\right)^{2}\right)$ is strictly non-negative, we see that this inequality holds when

$$\sum_{k=1}^{p}\left(\rho_{j,k}-\bar{\rho}_{i}\right)^{2} < 1$$

Thus, we have a more efficient estimator of δ when

$$p > \frac{1}{Var\left(\rho_{j,k}\right)}$$

In practice, we typically have a large number of genes, p, so that our transition matrix estimator will be expected to be dramatically more efficient than the commonly used Pearson or WGCNA estimators.

References

[1] Lung genomics research consortium (lgrc). Accessed: 2016-02-02.

Top significantly differentially involved transcription factors

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

Differential gene expression for significantly involved transcription factors.

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

Table 2: **Top Transcription Factor Hits. A** Combined list of TFs which were among the top 10 hits (out of 166 available TFs) in any of the 4 studies, ordered by the dTFI in the ECLIPSE study. For each study, columns indicate the TF's (1) differential TF Involvement, (2) dTFI Rank within list of TFs, (3) and Significance of dTFI by false discovery rate. **B** The same list of top transcription factors evaluated for differential gene expression with p-value for differential gene expression analysis using LIMMA [18]. Notably, a substantial number of differentially involved transcription factors do not exhibit gene expression differentiation, highlighting the ability of MONSTER to identify key features distinguising phenotypes which are not detectable via gene expression analysis.

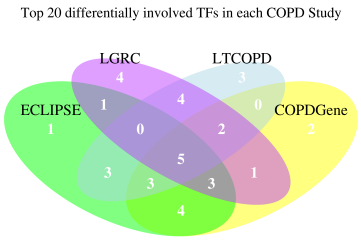


Figure 5: **Differentially involved transcription factors were found in agreement across datasets**. Five TFs were discovered in the top 20 in each of the four studies. An additional 21 TFs were found in the top 20 TFs of at least two studies.

[2] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8, 2007.

[3] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[4] Kimberly Glass, Curtis Huttenhower, John Quackenbush, and Guo-Cheng Yuan. Passing messages between biological networks to refine predicted interactions. *PloS one*, 8(5):e64832, 2013.

[5] Kimberly Glass, John Quackenbush, Edwin K Silverman, Bartolome Celli, Stephen I Rennard, Guo-Cheng Yuan, and Dawn L DeMeo. Sexually-dimorphic targeting of functionally-related genes in copd. *BMC systems biology*, 8(1):118, 2014.

[6] Kimberly Glass, John Quackenbush, Dimitrios Spentzos, Benjamin Haibe-Kains, and Guo-Cheng Yuan. A network model for angiogenesis in ovarian cancer. *BMC bioinformatics*, 16(1):115, 2015.

[7] Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

[8] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, (1):559, 2008.

[9] Peter Langfelder and Steve Horvath. Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software*, 46(11):1–17, 2012.

[10] Taotao Lao, Kimberly Glass, Weiliang Qiu, Francesca Polverino, Kushagra Gupta, Jarrett Morrow, John Dominic Mancini, Linh Vuong, Mark A Perrella, Craig P Hersh, et al. Genome medicine. 2015.

[11] Michael Levine and Robert Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–151, 2003.

[12] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, Gustavo Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

[13] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo D Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

[14] Catharina Olsen, Kathleen Fleming, Niall Prendergast, Renee Rubio, Frank Emmert-Streib, Gianluca Bontempi, Benjamin Haibe-Kains, and John Quackenbush. Inference and validation of predictive gene networks from biomedical literature and gene expression data. *Genomics*, 103(5):329–336, 2014.

[15] Sreekumar G Pillai, Dongliang Ge, Guohua Zhu, Xiangyang Kong, Kevin V Shianna, Anna C Need, Sheng Feng, Craig P Hersh, Per Bakke, Amund Gulsvik, et al. A genome-wide association study in chronic obstructive pulmonary disease (copd): identification of two major susceptibility loci. *PLoS Genet*, 5(3):e1000421, 2009.

[16] Timothy Ravasi, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B Bajic, Kai Tan, Altuna Akalin, Sebastian Schmeier, Mutsumi Kanamori-Katayama, Nicolas Bertin, et al. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, 2010.

[17] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002.

[18] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, page gkv007, 2015.

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

[20] Jørgen Vestbo, Wayne Anderson, Harvey O Coxson, Courtney Crim, Ffyona Dawber, Lisa Edwards, Gerry Hagan, Katharine Knobil, David A Lomas, William MacNee, et al. Evaluation of copd longitudinally to identify predictive surrogate end-points (eclipse). *European Respiratory Journal*, 31(4):869–873, 2008.