Estimating Drivers of 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

April 19, 2016

Abstract

Cells are classified by based on their tissue of origin and their particular role within that tissue. Specific cellular states are associated with patterns in gene expression. These states are plastic, changing during development, or in the transition from health to disease. One relatively simple extension of this framework is to recognize that we can classify different cell-types by their active gene regulatory networks and that, consequently, transitions between cellular states can be modeled by changes in these underlying regulatory networks. Here we describe **MONSTER**, **MO**deling **N**etwork **S**tate **T**ransitions from **E**xpression and **R**egulatory data, a regression-based method for inferring transcription factor drivers of cell state conditions at the gene regulatory network level. As a demonstration, we apply MONSTER to four different studies of chronic obstructive pulmonary disease to identify transcription factors that alter the network structure as cell states changes toward the disease-state. Our results demonstrate the ability to find strong signals that persists across studies and tissues of the same disease and which are not detectable using conventional analysis methods based on differential expression.

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 that which 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.

Before modeling cell state transitions, the initial and final cell states must first be modelled. One might imagine that 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, in turn, characterizes 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 n target genes, is as a binary $n \times m$ "adjacency matrix," in which a 1 represents an active interaction between a transcription factor and a potential target, and a 0 represents the lack of a regulatory interaction.

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

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

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

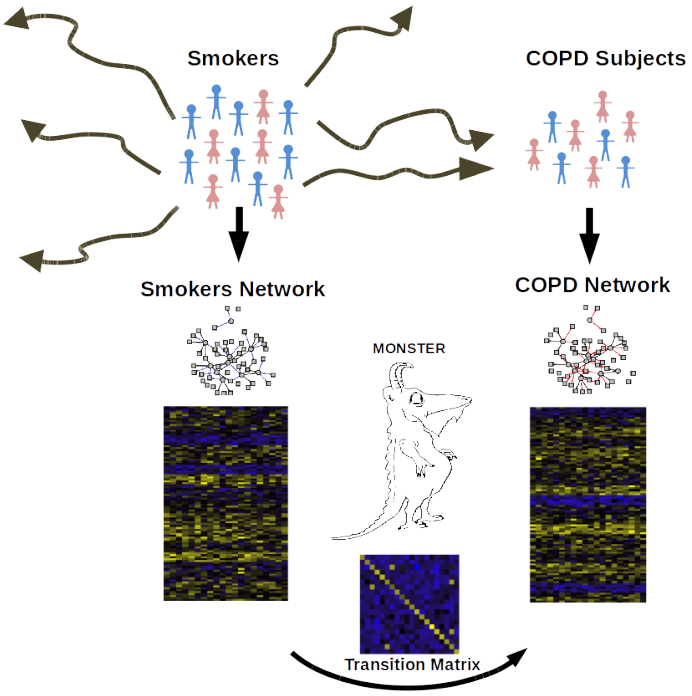

Figure 1: **Overview of the MONSTER approach, as applied to the transition between smokers and those suffering from chronic obstructive pulmonary disease (COPD).** MONSTER's approach seeks to find the $TF \times TF$ transition matrix that best characterizes the state change in network structure between an initial and final biological conditions. Subjects are first divided into two groups based on phenotype, COPD patients and non-COPD smokers, with network inference performed separately on each. A transition matrix is then computed which best characterizes the conversion from the consensus Smokers Network to the COPD Network.

While this framework, as depicted in , is intuitive, it is a bit simplistic in the sense that we have cast the initial and final states as discrete. However, the model can be generalized by recognizing that any phenotype consists of a collection of individuals, all of whom have slightly different manifestations of the state, and all of whom therefore have slightly different active gene regulatory networks. 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 as estimation of its presence in the population.

In the practice, what this means is that the initial and final state adjacency matrices are not comprised of 1's and 0's, but of continuous variables that truly estimate the phenotype-specific population estimates of the corresponding regulatory network edge weights. And consequently, the problem of estimating the transition matrix is generalized to solving $\mathbf{B} = \mathbf{AT} + \mathbf{E}$, where $\mathbf{E}$ is an $n \times m$ 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 by finding those transcription factors with the greatest "off diagonal mass" in $\mathbf{T}$. Large values in in the matrix that are found off of the diagonal represent a transition in the targeting behavior of the transcription factors linked to the row and column of those entries. We can estimate the total targeting variation by calculating the proportion of the total sum of squares of a column which is off of the diagonal.

MONSTER: Inferring state-specific gene regulatory networks, modeling the state transition matrix, and computing the transcription factor involvement

Before estimating the transition matrix, we must first estimate a gene regulatory starting point for each state. While there have been many methods developed to estimate such networks [REFS], we have found PANDA (*10*) to have features that are particularly amenable to interpretation in the framework of state transitions. PANDA begins by using genome-wide transcription factor binding data to postulate a starting network, and then uses a message-passing framework to use multiple data sources, including state-specific gene expression data, to iteratively update the starting network until converging on a final network model for each phenotype.

Although PANDA has been shown to provide highly informative gene regulatory network models, it can be computationally inefficient when applied to large data sets. For this reason we developed a classification-based network inference method that uses common motifs and coexpression patterns to estimate edges. We seek to generate a bipartite gene regulatory network connecting our set of transcription factors to our set of genes.

Inferring an expression-based gene regulatory network: This approach is motivated by the simple concept that genes which are affected by transcription factors will exhibit expression patterns that correlate with both the transcription factor and the other targets of that transcription factor. We begin with an initial transcription factor-target starting network derived from sets of known sequence binding motifs found in the vicinity of genes. Next, we calculate the direct evidence, defined as the squared partial correlation between each gene's expression and the transcription factor's gene expression, conditional on all other transcription factors.

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

Where g_i and g_j are the gene expression patterns across the N samples and $\mathbf{TF}$ represents the set of transcription factors for which we have gene expression data.

Simultaneously, for each phenotypic state, we fit a logistic regression model which predicts the probability of each gene being a suspected target of a transcription factor based the expression pattern across the N samples in each phenotypic class.

$$logit(M_i) = \beta_0 + \beta_1 g_{(1)} + \ldots + \beta_N g_{(N)}$$

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.

Combining these scores for the direct evidence and indirect evidence between each transcription factor-gene pairing yields estimated edgeweights for the gene regulatory network. (see supplementary materials and methods). The result of using this with gene expression data from two phenotypes is separate $m \times n$ gene regulatory adjacency

matrices for each phenotype, representing estimates of the targeting patterns of the m transcription factors onto the n genes.

This straightforward and computationally fast algorithm finds validated regulatory edges in In Silico, E. coli and Yeast (Saccharomyces cerevisiae) datasets (see supplementary materials and methods).

Computing drivers of state transition: Having gene regulatory network estimates for each of the starting phenotypes, we formulate the problem of estimating the transition matrix as a regression problem in which we solve for the mŒm matrix that best describes the transformation between phenotypes.

MONSTER formulates this problem in a regression framework whereby we predict the change in edgeweights for a transcription factor, i, in a network based on all of the the edgeweights in the baseline phenotype network.

$$E[b_i - a_i] = \tau_{i1}a_{1i} + \ldots + \tau_{im}a_{mi}$$

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

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

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

$$T = [\tau_1, \tau_2, \ldots, \tau_m]$$

The regression is performed m times corresponding to each of the known 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 supplementary materials and methods).

It is intuitive to see that this framework allows for the natural extension of constraints such as $L1$ and/or $L2$ regularization (see supplementary materials and methods). In this COPD analysis, we utilize the normal equations and do not impose a penalty on the regression coefficients.

In transitions between nearly identical states, we expect the transition matrix to approximate the identity matrix. As the initial and final states diverge, we would expect increasing differences in their gene regulatory networks and, consequently, we expect the transition matrix to increasingly diverge from the identity. 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 have very different targets between states and so are likely to be involved in the state transition process.. We define the "transcription factor involvement" (dTFI) to estimate the magnitude of the off-diagonal mass associated with each transcription factor, or,

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

where, 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 times across phenotypes (see supplementary materials and methods).

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

As a demonstration of the power of MONSTER to identifying driving factors in disease, we to four independent case-control datasets for Chronic Obstructive Pulmonary Disease (COPD): Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) (*19*) (*20*) (2) , the COPDGene study (*2*) (*15*) (Supplemental Data), Lung Genomics Research Consortium (LGRC) (*1*)(Supplemental Data) and Lung Tissue Chronic Obstructive Pulmonary Disease (LTCOPD) [ltcopd] (Supplemental Data). Each study included gene expression data from

patients with COPD and a matched set of smoker controls. The tissue used in the ECLIPSE and COPDGene study was peripheral blood mononuclear cells (PBMCs), while lung tissue was sampled for LGRC and LTCOPD.

As a baseline comparison metric, we evaluated the efficacy of applying conventionally used network inference methods on these case-control studies. Commonly, networks are compared directly, with changes in the presence or weight of edges between key genes being of primary interest. It is therefore reasonable to assume that any reliable network results generated from a comparison of disease to controls will be reproducible in independent studies. We investigated this approach using three commonly used network inference methods - Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE) (*14*), Context Likelihood of Relatedness (CLR) (*8*), and Weighted Gene Correlation Network Analysis (WGCNA) (*21*) - computing the difference in edgeweights between cases and controls for each of the four studies. Interestingly, we found no meaningful correlation ($R^2 < .01$) of edgeweight difference across any of the studies regardless of network inference method or tissue type (supplemental Figure 1A-C). Edgeweight 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. However, the failure to replicate edge weight differences in independent datasets is further evidence that we need to rethink how we evaluate network state transitions. MONSTER provides a novel approach for making that comparison.

For each study, we applied MONSTER to identify the differential transcription factor involvement for each transcription factor and used permutation analysis to estimate their significance (Figure 2, supplemental figure 3). Out of 166 transcription factors used in this study, seven were among top 10 most differentially involved in both the ECLIPSE and COPDGene studies (Figure 3C). Furthermore, three of these seven transcription factors (GABPA, ELK4, ELK1) also appeared as significant in the LGRC results with FDR<.01 and each of the top five ECLIPSE results were among the top seven in the LTCOPD results. This agreement is quite striking considering that the there was almost no correlation in the edge weight differences across these same studies. There was significant correlation for each pairwise combination of studies ($p < 1e-15$).

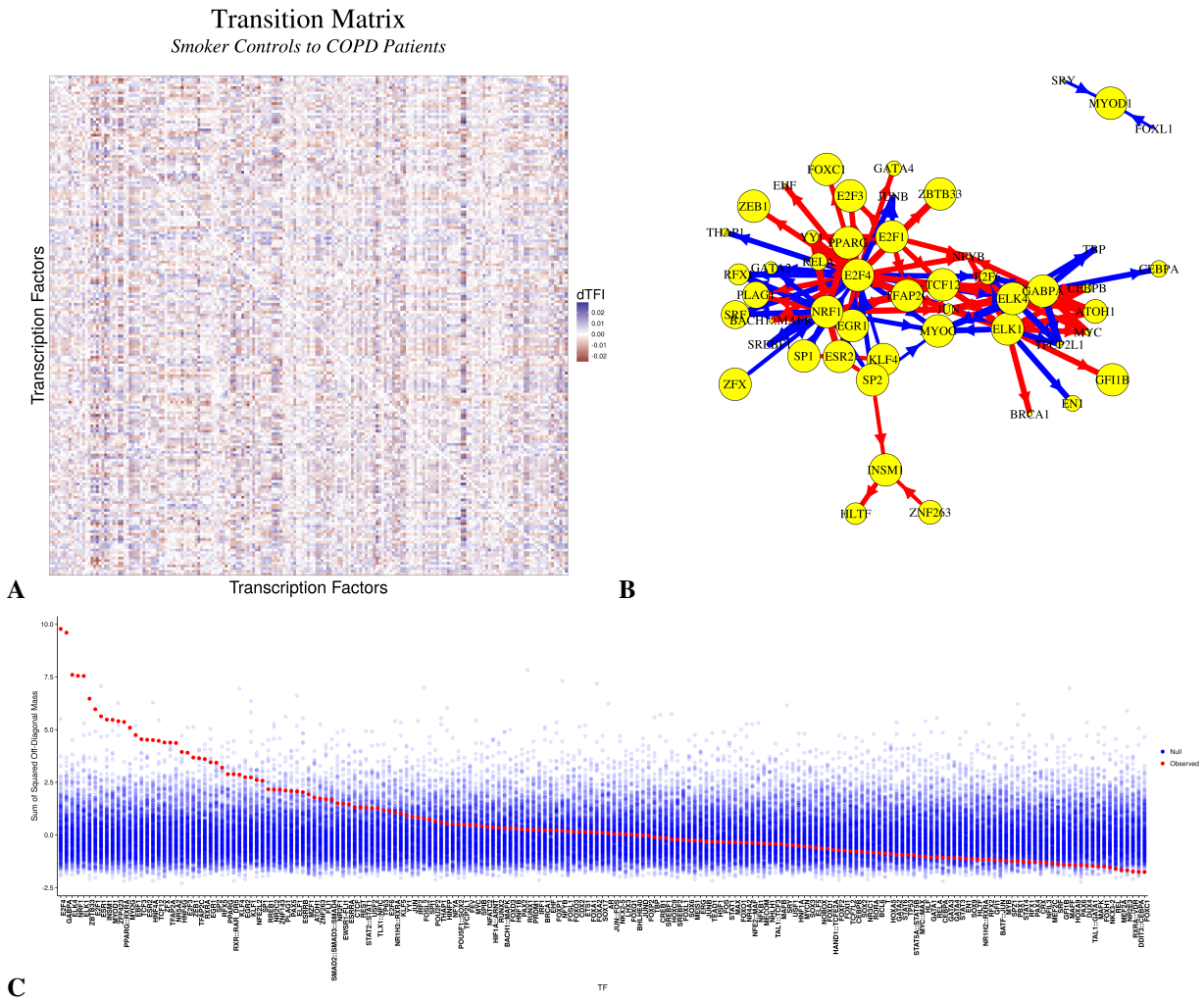

Figure 2: **Differential transcription factor involvement (dTFI) calculated in the ECLIPSE study. A** Heatmap depicting the transition matrix from smoker controls to COPD in ECLIPSE. For the purposes of visualization, the magnitude of the diagonal is not displayed. **B** Network transitions are depicted here with arrows indicating the flow of targeting patterns from one transcription factor to another. Edges are sized according to the magnitude of the transition and nodes (TFs) are sized by the overall dTFI for each TF. 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 is shown plotted in red against a background null distribution estimated by 1000 random sample permutations of the data shown in blue; significant dTFI scores are those rising above the null background and represent transcription factors that change targeting patterns between states.

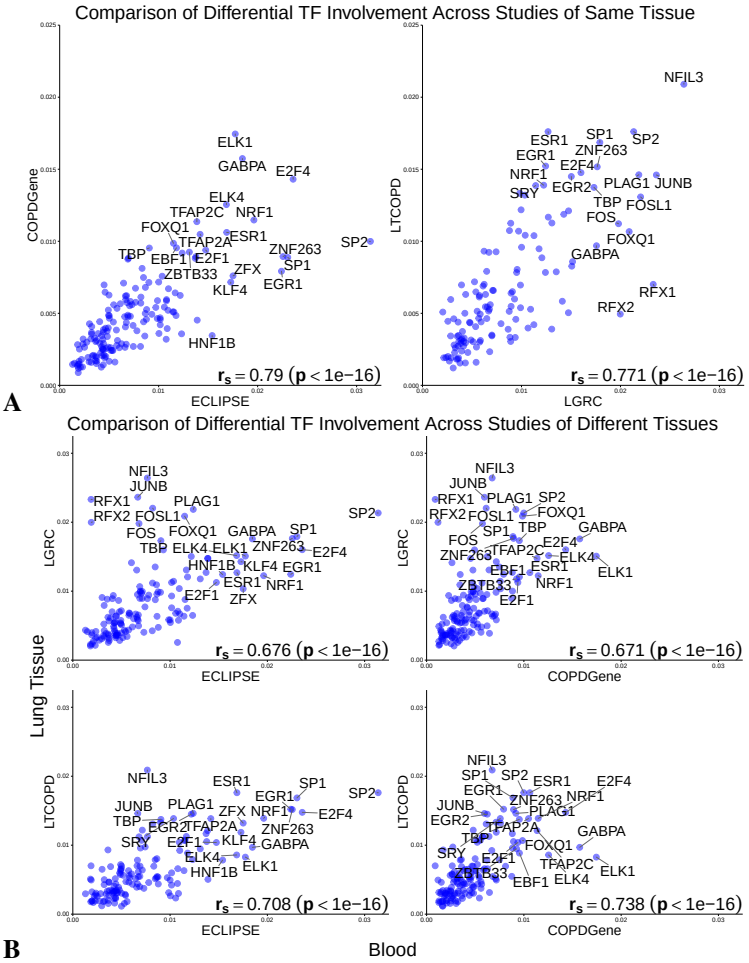

Figure 3: **Strong reproducibility in top differential transcription factor involvement found in case-control COPD studies**. ECLIPSE and COPDGene data were obtained via PBMC and LGRC and LTCOPD were obtained via lung tissue. Results for studies with gene expression data obtained from the same-tissue (**A**), PBMC (left) and lung tissue (right) each demonstrate very high spearman correlation of differential involvement. Correlations for the across-tissue study comparisons (**B**) demonstrated a weaker, but still meaningful agreement. Each of the four studies was most consistent with the study of the same tissue type.

Overall, we found a strong correlation in transcription factors identified as significantly differentially involved across case/control (Figure 3A-3B). It is reassuring that agreement is more strongly achieved between studies of the same tissue origin than across tissues. Each of the four studies was most closely correlated with studies of the same tissue. However it is quite notable that the we do see much of the same dTFI signal across studies involving different tissue types. Gene regulatory networks derived from gene expression data are notoriously difficult to replicate across studies and it is of great interest that we have identified suspected mechanisms which are correlated not only across studies but across tissues as well.

If we focus specifically on the transcription factors found by all studies [supplemental table 2, supplemental figure 5], we find interesting things. E2F4, found to be significant in all studies, is a transcriptional repressor important in airway development (*5*) and tumor suppression. Increasing evidence has emerged linking the pathogenesis of COPD and lung cancer (*12*) (*7*) including a substantially increased incidence of cancer in those with COPD (*6*).

Differential involvement of a tumor suppressor transcription factor is consistent with this research. Additionally, SP1 and SP2 were among the highest effect sizes observed in the four studies (although significance was reduced due to greater dTFI variance for those transcription factors) [supplemental table 2]. Both of these proteins have been found to form complexes with the E2F family (*13, 17*) and may play a key role in the alteration of E2F4 targeting behavior that we are observing in these studies. An additional member of the Sp transcription factor family, Sp3, has been shown to regulate HHIP, a known COPD susceptibility gene (*22*). Furthermore, E2F4 has been found to form a complex with EGR-1 (a top 5 hit in ECLIPSE and LTCOPD) in response smoke exposure, which may lead to autophagy and apoptosis, which may lead to development of emphysema (*3*).

Additionally, new research has identified mitochondrial mechanisms associated with COPD progression (*4,16*). It is therefore noteworthy that the most highly significant transcription factors in the ECLIPSE study were found to be NRF1 and GABPA (FDR<.001). Both of these transcription factors were also found to be significant at at least FDR<0.1 in each of the four studies and FDR<0.001 in the LGRC study. NRF1 regulates the expression of nuclear encoded mitochondrial proteins. (*11*). GABPA has similarly been linked to nuclear control of mitochondrial function and shares a subunit with nuclear respiratory factor 2 gene (NRF1), possibly indicating a role in cytochrome oxidase expression, a process linked to COPD progression (*18*). Furthermore, it is established that GABPA interacts with SP1 (*9*) providing additional evidence that a new mechanism is at play here among these common players.

Consistency across in dTFI results was not universal for all transcription factors. While each of the four groups separated cases and controls based on COPD diagnosis vs. smoking non-COPD patients, many differences exist for each of the four studies including microarray platform, study demographics, location, time and tissue. For example, in the LGRC dataset, we discovered a differential targeting pattern involving the transcription factors transcription factors RFX1 and RFX2 [supplemental table 2]. Both of these transcription factors transcription factors were highly statistically significant (FDR<.0001) and ranked as the top two results in the LGRC study. However, their signal was muted in the ECLIPSE and COPDGene studies, neither of which identified these transcription factors transcription factors as drivers of the Smoker Control to COPD transition. We emphasize that significant transcription factors are attributed to differences in the case/control differences, subjecting the results to possible confounding. This highlights the importance of utilizing case control designs which are properly matched along known covariates.

Although our hypotheses is that transcription factors that alter their targets (and therefore have high dTFI scores) are drivers of changes in phenotypic state, many of the transcription factors that we identified are not themselves differentially expressed in comparing control and COPD populations (Figure 4). This suggests that there may be other mechanisms, including epigenetic and protein interaction factors, affecting the structure of gene regulatory networks and that the master regulators of phenotypic state change may have differentiated targeting behavior in patients in the COPD group compared to the control group.

These results demonstrate the scientific value of MONSTER. Numerous biologically sensible results were found across several independent studies. The reproducibility of the top transcription factor hits is compelling. We demonstrate that these findings would not have been possible via differential gene expression analysis or conventional comparative network inference methods.

References

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

2. Timothy M Bahr, Grant J Hughes, Michael Armstrong, Rick Reisdorph, Christopher D Coldren, Michael G Edwards, Christina Schnell, Ross Kedl, Daniel J LaFlamme, Nichole Reisdorph, et al. Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease. *American journal of respiratory cell and molecular biology*, 49(2):316–323, 2013.

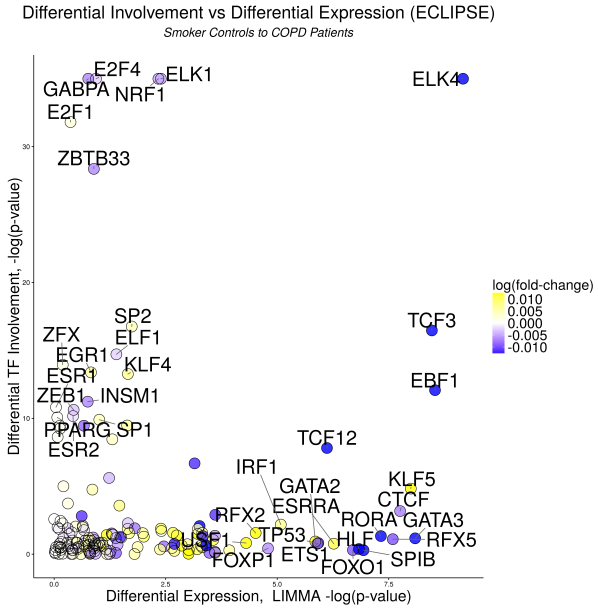

Figure 4: **Comparison of significant results for TF differential involvement vs differential expression in ECLIPSE.** TFs which are differentially involved are not necessarily differentially expressed, and vice versa. Many TFs can be observed which have significantly different targeting patterns, but which are not statistically significantly differentially expressed. This suggests that our method finds transcription factors which are differentially affected at a post-transcriptional stage.

3. Zhi-Hua Chen, Hong Pyo Kim, Frank C Sciurba, Seon-Jin Lee, Carol Feghali-Bostwick, Donna B Stolz, Rajiv Dhir, Rodney J Landreneau, Mathew J Schuchert, Samuel A Yousem, et al. Egr-1 regulates autophagy in cigarette smoke-induced chronic obstructive pulmonary disease. *PloS one*, 3(10):e3316, 2008.

4. Suzanne M Cloonan, Kimberly Glass, Maria E Laucho-Contreras, Abhiram R Bhashyam, Morgan Cervo, Maria A Pabón, Csaba Konrad, Francesca Polverino, Ilias I Siempos, Elizabeth Perez, et al. Mitochondrial iron chelation ameliorates cigarette smoke-induced bronchitis and emphysema in mice. *Nature medicine*, 2016.

5. Paul S Danielian, Carla F Bender Kim, Alicia M Caron, Eliza Vasile, Roderick T Bronson, and Jacqueline A Lees. E2f4 is required for normal development of the airway epithelium. *Developmental biology*, 305(2):564–576, 2007.

6. Juan P de Torres, David O Wilson, Pablo Sanchez-Salcedo, Joel L Weissfeld, Juan Berto, Arantzazu Campo, Ana B Alcaide, Marta García-Granero, Bartolome R Celli, and Javier J Zulueta. Lung cancer in patients with chronic obstructive pulmonary disease. development and validation of the copd lung cancer screening score. *American journal of respiratory and critical care medicine*, 191(3):285–291, 2015.

7. AL Durham and IM Adcock. The relationship between copd and lung cancer. *Lung Cancer*, 90(2):121–127, 2015.

8. 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.

9. Federico Galvagni, Sabrina Capo, and Salvatore Oliviero. Sp1 and sp3 physically interact and co-operate with gabp for the activation of the utrophin promoter. *Journal of molecular biology*, 306(5):985–996, 2001.

10. 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.

11. Lekha Gopalakrishnan and Richard C Scarpulla. Structure, expression, and chromosomal assignment of the human gene encoding nuclear respiratory factor 1. *Journal of Biological Chemistry*, 270(30):18019–18025, 1995.

12. Leda Guzmán, María Soledad Depix, Ana María Salinas, Rosa Roldán, Francisco Aguayo, Alejandra Silva, and Raul Vinet. Analysis of aberrant methylation on promoter sequences of tumor suppressor genes and total dna in sputum samples: a promising tool for early detection of copd and lung cancer in smokers. *Diagn Pathol*, 7(87):1596–7, 2012.

13. Jan Karlseder, Hans Rotheneder, and Erhard Wintersberger. Interaction of sp1 with the growth-and cell cycle-regulated transcription factor e2f. *Molecular and cellular biology*, 16(4):1659–1667, 1996.

14. 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.

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. Luis Puente-Maestu, José Pérez-Parra, Raul Godoy, Nicolás Moreno, Alberto Tejedor, Federico González-Aragoneses, José-Luis Bravo, F Villar Álvarez, Sonia Camaño, and Alvar Agustí. Abnormal mitochondrial function in locomotor and respiratory muscles of copd patients. *European Respiratory Journal*, 33(5):1045–1052, 2009.

17. Hans Rotheneder, Sibylle Geymayer, and Eva Haidweger. 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, 1999.

18. Jaume Sauleda, Francisco Garcia-Palmer, Rudolf J Wiesner, Salvador Tarraga, Inga Harting, Purificación Tomás, Cristina Gomez, Carles Saus, Andreu Palou, and Alvar GN Agusti. Cytochrome oxidase activity and mitochondrial gene expression in skeletal muscle of patients with chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 157(5):1413–1417, 1998.

19. Dave Singh, Steven M Fox, Ruth Tal-Singer, Stewart Bates, John H Riley, and Bartolome Celli. Altered gene expression in blood and sputum in copd frequent exacerbators in the eclipse cohort. *PloS one*, 9(9):e107381, 2014.

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.

21. Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

22. Xiaobo Zhou, Rebecca M Baron, Megan Hardin, Michael H Cho, Jan Zielinski, Iwona Hawrylkiewicz, Pawel Sliwinski, Craig P Hersh, John D Mancini, Ke Lu, et al. Identification of a chronic obstructive pulmonary disease genetic determinant that regulates hhip. *Human molecular genetics*, 21(6):1325–1335, 2012.