Estimating Drivers of Cell State Transitions using Gene Regulatory Network Models

Daniel Schlauch[1,2], Kimberly Glass[2,3], Craig P. Hersh[2], Edwin K. Silverman[2,4], and 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; [4]Pulmonary and Critical Care Division, Brigham and Women's Hospital and Harvard Medical School, Boston, USA

This manuscript was compiled on September 12, 2016

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/dschlauch/MONSTER.

Gene Regulatory Network Inference | Chronic Obstructive Pulmonary Disease | Genomics

Cell state phenotypic transitions, such as those that occur during development, or as healthy tissue transforms into a disease phenotype, are fundamental processes that operate within biological systems. Understanding what drives these transitions, and modeling the processes, is one of the great open challenges in modern biology. One way to conceptualize the state transition problem is to imagine that each phenotype has its own characteristic gene regulatory network, and that there are a set of processes that are either activated or inactivated to transform the network in the initial state into one that characterizes the final state. Identifying those changes could, in principle, help us to understand not only the processes that drive the state change, but also how one might intervene to either promote or inhibit such a transition.

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

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

$$\mathbf{B} = \mathbf{AT} \quad [1]$$

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

While this framework, as depicted in 1, 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 or samples, 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

Significance Statement

Understanding how transcription factors - proteins which bind to DNA - regulate gene expression pattern is a goal of gene regulatory network inference. There are many existing methods for inferring these types of networks, but biological networks are extraordinarily complex and all methods inevitably leave a great deal of uncertainty in the individual edges they predict. This makes it challenging to make sense of these graphs, particularly when the goal is to compare two inferred networks in hopes of identifying key mechanistic differences. The method we present here, MONSTER, is a novel approach towards defining the changes in networks via key transcription factors which drive the change in a consistent manner. We apply our method to a set of four independent studies of Chronic Obstructive Pulmonary Disease and demonstrate strong agreement in implicated transcription factor drivers of disease.

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.

[2]To whom correspondence should be addressed. E-mail: author.twoemail.com

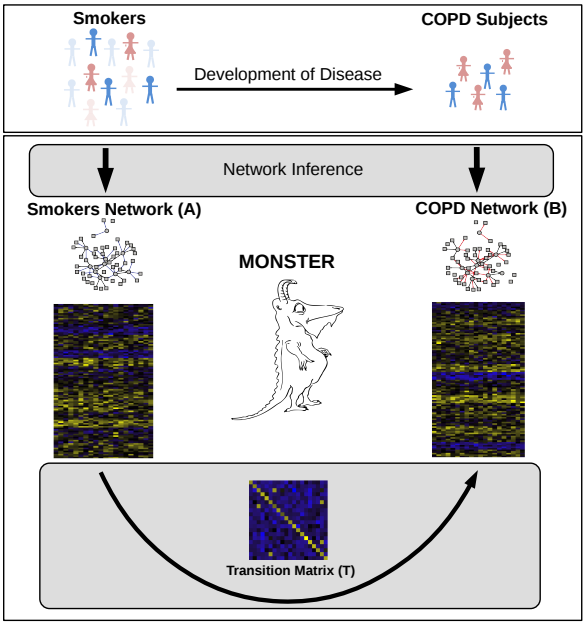

Fig. 1. Overview of the MONSTER approach, as applied to the transition between smokers and those suffering from chronic obstructive pulmonary disease (COPD). MONSTER's approach seeks to find the $TF \times TF$ transition matrix that best characterizes the state change in network structure between the 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.

with edges that are either "on" or "off," a phenotype should be represented by a network in which each edge has a weight that represents an estimation of its presence across the population. In other words, the initial and final state adjacency matrices are not comprised of 1's and 0's, but of continuous variables that estimate population-level regulatory network edge-weights. Consequently, the problem of calculating the transition matrix is generalized to solving $\mathbf{B} = \mathbf{AT} + \mathbf{E}$, where $\mathbf{E}$ is an $p \times m$ error matrix. In this expanded framework, modeling the cell state transition remains equivalent to estimating the appropriate transition matrix $\mathbf{T}$, and then identifying state transition drivers based on features of that matrix.

MONSTER: MOdeling Network State Transitions from Expression and Regulatory data

The MONSTER algorithm models the regulatory transition between two cellular states in three main 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, $\mathbf{T}$, we must first estimate a gene regulatory starting point for each state. While there have been many methods developed to infer such networks [1–7], we have found the bipartite framework used in PANDA [8] 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 network "prior", and then uses a message-passing approach to integrate multiple data sources, including state-specific gene coexpression data. Motivated by the approach used by PANDA, we developed a highly computationally-efficient, classification-based network inference method that uses common patterns between transcription factor targets and gene coexpression to estimate edges and to generate a bipartite gene regulatory network connecting transcription factors to their target genes.

This approach is motivated by the simple concept that genes that are affected by a common transcription factor will exhibit expression patterns that correlate. To begin, we 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 gene's expression, g_i, and the transcription factor's gene expression, g_j, conditional on all other transcription factors' gene expression, g_k :

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

where g_i and g_j are the gene expression patterns across the N samples and $\mathbf{S}$ represents the set of genes which are transcription factors.

We then use 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. In particular, we fit a logistic regression model which estimates the probability of each gene, indexed i, being a motif target of a transcription factor, indexed j, based on the expression pattern across the N samples in each phenotypic class:

$$logit(P\left[\mathbf{M}_{i,j} = 1\right]) = \beta_0 + \beta_1 g_i^{(1)} + \cdots + \beta_N g_i^{(N)}$$

$$\hat{e}_{i,j} = \frac{\hat{\beta}_0 + \hat{\beta}_1 g_i^{(1)} + \cdots + \hat{\beta}_N g_i^{(N)}}{1 + \hat{\beta}_0 + \hat{\beta}_1 g_i^{(1)} + \cdots + \hat{\beta}_N g_i^{(N)}}$$

where the response $\mathbf{M}$ is a binary $p \times m$ matrix indicating the presence of a sequence motif for the j^{th} transcription factor in the vicinity of each of the i^{th} gene. And where $g_{(k)}$ is a vector of length n specifying the gene expression for sample k over n genes. And where $g^{(q)}$ is a vector of length p specifying the gene expression for sample q over p genes. Thus, $\hat{e}_{i,j}$ represents our estimated indirect evidence. Combining the scores for the direct evidence, $\hat{d}_{i,j}$, and indirect evidence, $\hat{e}_{i,j}$, via weighted sum between each transcription factor-gene pair yields estimated edge-weights for the gene regulatory network (see Supplementary Materials and Methods).

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 straightforward and computationally fast algorithm finds validated regulatory edges in *E. coli* and Yeast (*Saccharomyces cerevisiae*) datasets (see Supplementary Materials and Methods).

Modelling 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). More specifically, MONSTER predicts the change in edge-weights for a transcription factor, indexed i, in a network

based on all of the edge-weights in the baseline phenotype network.

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

where b_i and a_i are column-vectors in $\mathbf{B}$ and $\mathbf{A}$ that describe the regulatory targeting of transcription factor i in the final and initial networks, respectively.

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

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

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

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

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

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

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 the initial and final states diverge, there would 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,

$$d\hat{TFI}_j = \frac{\sum_{i=1}^{m} I\,(i \neq j)\,\hat{\tau}_{i,j}^2}{\sum_{i=1}^{m} \hat{\tau}_{i,j}^2} \qquad [2]$$

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 $n = 1000$ 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 identify driving factors in disease, we applied the method to case-control gene expression datasets from four independent Chronic Obstructive Pulmonary Disease (COPD) cohorts: Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) [9][10] (2) , the COPDGene study [11][12] [13], Lung Genomics Research Consortium (LGRC) [14] and Lung Tissue from Channing Division of Network Medicine (LT-CDNM) [15]. 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 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 whether this is the case for our four COPD datasets using three commonly employed network inference methods - Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE)[16], Context Likelihood of Relatedness (CLR)[17], and Weighted Gene Correlation Network Analysis (WGCNA) [18] - 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 (Supplementary Figure S1A-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.

The lack of replication in edge-weight differences between independent datasets representing similar study designs indicates that we need to rethink how we evaluate network state transitions. MONSTER provides a novel approach for making that comparison. Along these lines, for each study, we applied MONSTER to calculate the differential transcription factor involvement ($dTFI$, 2) for each transcription factor and used permutation analysis to estimate their significance (Figure 2, Supplementary Figure S3). 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 in common. Furthermore, three of these seven transcription factors (GABPA, ELK4, ELK1) also appeared as significant in the LGRC results with FDR<0.01 and each of the top five ECLIPSE results were among the top seven in the LT-CDNM results (Supplementary Table S2, Supplementary Figure S5). This agreement is quite striking considering that the there was almost no correlation in the edge-weight differences across these same studies.

Many of the top dTFI transcription factors, especially those identified by MONSTER in all studies, are highly relevant for COPD [Supplementary Table S2, Supplementary Figure S5]. For example, E2F4, is a transcriptional repressor important in airway development [19]. Recent work has pointed to the relevance of developmental pathways in COPD pathogenesis [20]. Additionally, we observed some of the highest effect sizes for SP1 and SP2 in the four studies. An additional member

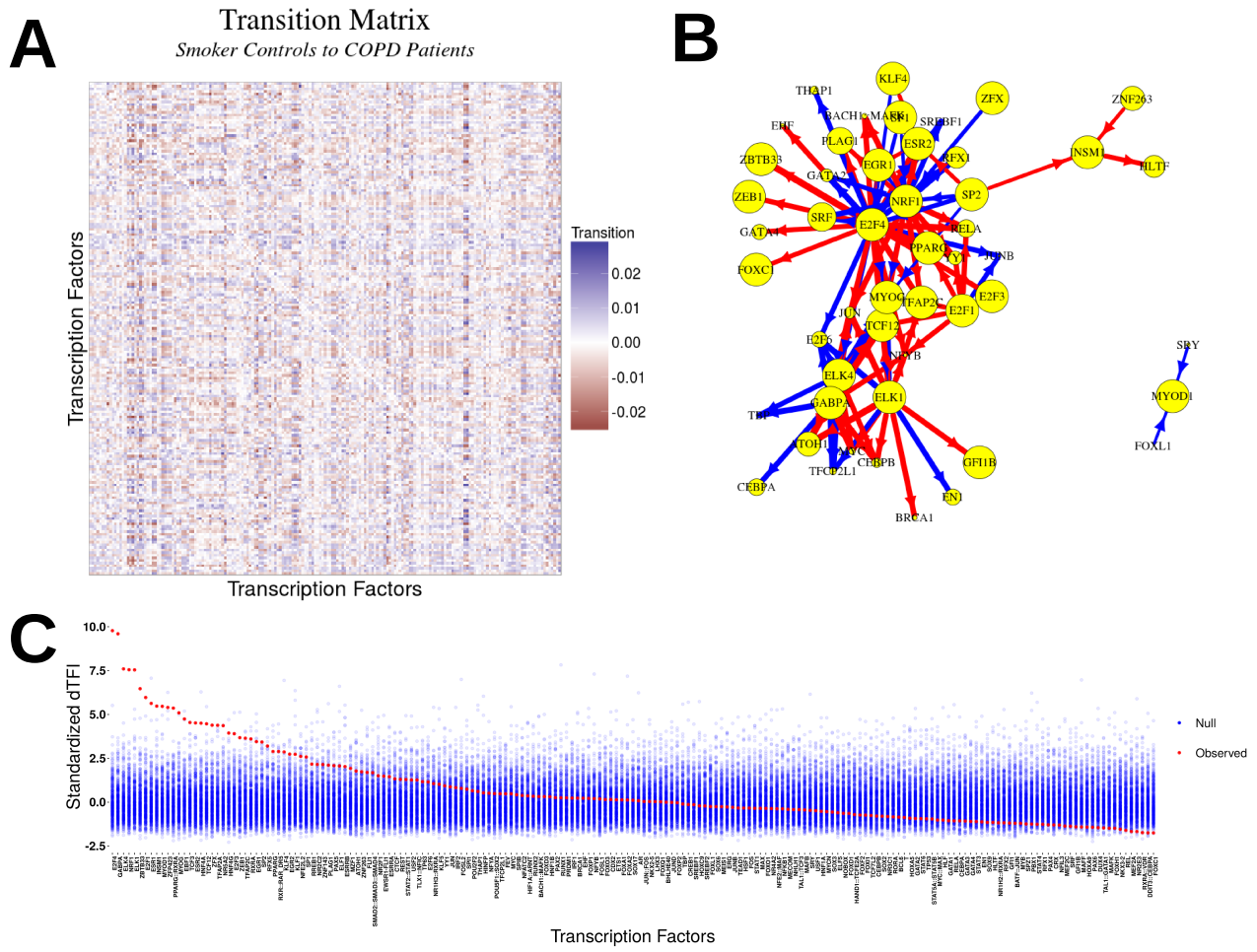

Fig. 2. MONSTER analysis results in the ECLIPSE study. A Heatmap depicting the transition matrix calculated from smoker controls to COPD cases by applying MONSTER to ECLIPSE gene expression data. For the purposes of visualization, the magnitude of the diagonal is set to zero. **B** A network visualization of the strongest 100 transitions identified based on the transition matrix shown 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. Edges are sized according to the magnitude of the transition and nodes (TFs) are sized by the dTFI for that 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 (red) and the background null distribution of dTFI values (blue) as estimated by 1000 random sample permutations of the data.

of the Sp transcription factor family, Sp3, has been shown to regulate HHIP, a known COPD susceptibility gene o[21]. Both SP1 and SP2 have been found to form complexes with the E2F family [22, 23] and may play a key role in the alteration of E2F4 targeting behavior. Furthermore, E2F4 has been found to form a complex with EGR-1 (a highly significant transcription factor in ECLIPSE and LT-CDNM) in response smoke exposure, which may lead to autophagy, apoptosis and subsequently to development of emphysema [24].

Additionally, research has identified mitochondrial mechanisms associated with COPD progression [25]. It is therefore noteworthy that the two most highly significant transcription factors based on dTFI in the ECLIPSE study were NRF1 and GABPA (FDR<.001). These TFs had highly significant differential TF involvement (FDR<0.1) in all four studies. NRF1 regulates the expression of nuclear encoded mitochondrial proteins [26]. GABPA, also known as humanx nuclear respiratory factor-2 subunit alpha, may have a similar role in nuclear control of mitochondrial gene expression]. Furthermore, GABPA interacts with SP1 [27] 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 of note. Gene regulatory networks derived from gene expression data are notoriously difficult to replicate across studies [28] and it is of great interest that we have identified transcription factors whose regulatory mechanisms are suspected to play a role in COPD across multiple studies, including those assaying different tissues.

There are many aspects that are specific to each of the four studies we used in our analysis, including microarray platform, study demographics, location, time and tissue. MONSTER largely identified similar sets of transcription factors when defining the transition between cases and controls based on COPD diagnosis vs. smoking non-COPD patients. However, some transcription factors had different levels of $dTFI$ in the different studies. For example, in the LGRC dataset, we discovered a highly significant ($FDR < .0001$) differential targeting pattern involving the transcription factors RFX1 and RFX2 (Supplementary Table S2). However, these same TFs were not identified as potential drivers of the Smoker Control to COPD transition in either the ECLIPSE or COPDGene study - likely due the differences in tissue type. Transcription factors in the RFX family are known to regulate ciliogenesis [29]. This process is critical for clearing mucous from the airways in healthy lung tissue, but when disrupted can lead to infection and chronic obstruction [30–32].

Our hypothesis is that transcription factors that alter their targets (and therefore have high $dTFI$ scores) are drivers of changes in phenotypic state. It would be reasonable to suspect that these transcription factors would differ across cases and controls at the transcriptional level. Therefore, we compared $dTFI$ to differential expression in ECLIPSE (Figure 4, other studies in Supplementary Figure S5). However, many of the transcription factors with high dTFI were not differentially expressed in these studies. This suggests that there may be other mechanisms, including epigenetics, phosphorylation and

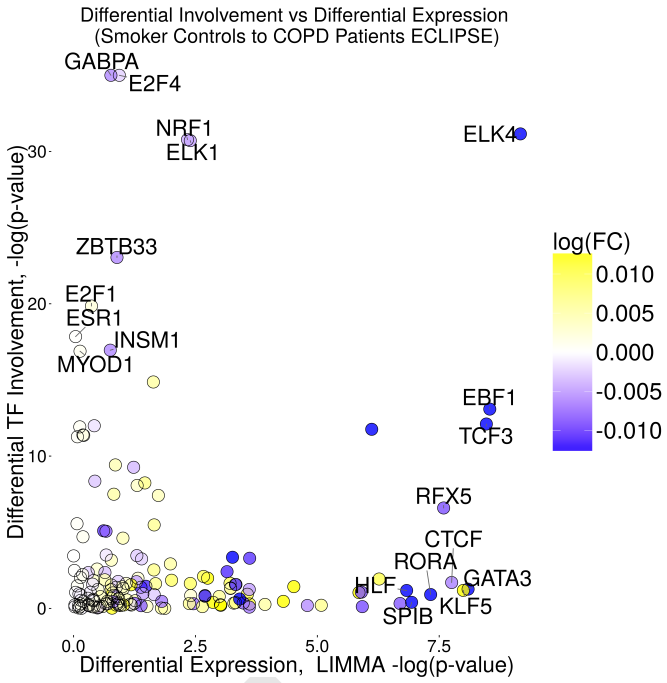

Fig. 4. Differentially involved transcription factors are not necessarily differentially expressed. MONSTER commonly finds transcription factors which are differentially involved but are expressed at similar levels across cases and controls. This change in involvement suggests that network rewiring occurs at a post-transcriptional stage. Importantly, these transcription factors would not have been identified using conventional differential expression methods.

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.

Discussion

One of the fundamental problems is biology is modeling the transition between biological states such as that which occurs during development or as a healthy tissue transforms into a disease state. As our ability to generate large-scale, integrative multi-omic datasets has grown, there has been an increased interest in using those data to infer gene regulatory networks to model fundamental biological processes. While there have been many network inference (NI) methods published, each of which uses a different approach to estimating the "strength" of interactions between genes (or between transcription factors and their targets), they 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.

With MONSTER we acknowledge uncertainty in individual regulatory edgeweights and instead focus on transcription factor drivers. The effects of these drivers are more readily apparent when taken over the entirety of the gene regulatory network. This effectively shines a spotlight on a reduced search space and thus contributes to the strength of the findings and reproducibility across studies.

The application of MONSTER to the four studies of COPD demonstrate the scientific value of the method. Numerous transcription factors that have been biologically implicated in COPD were found in agreement across the independent

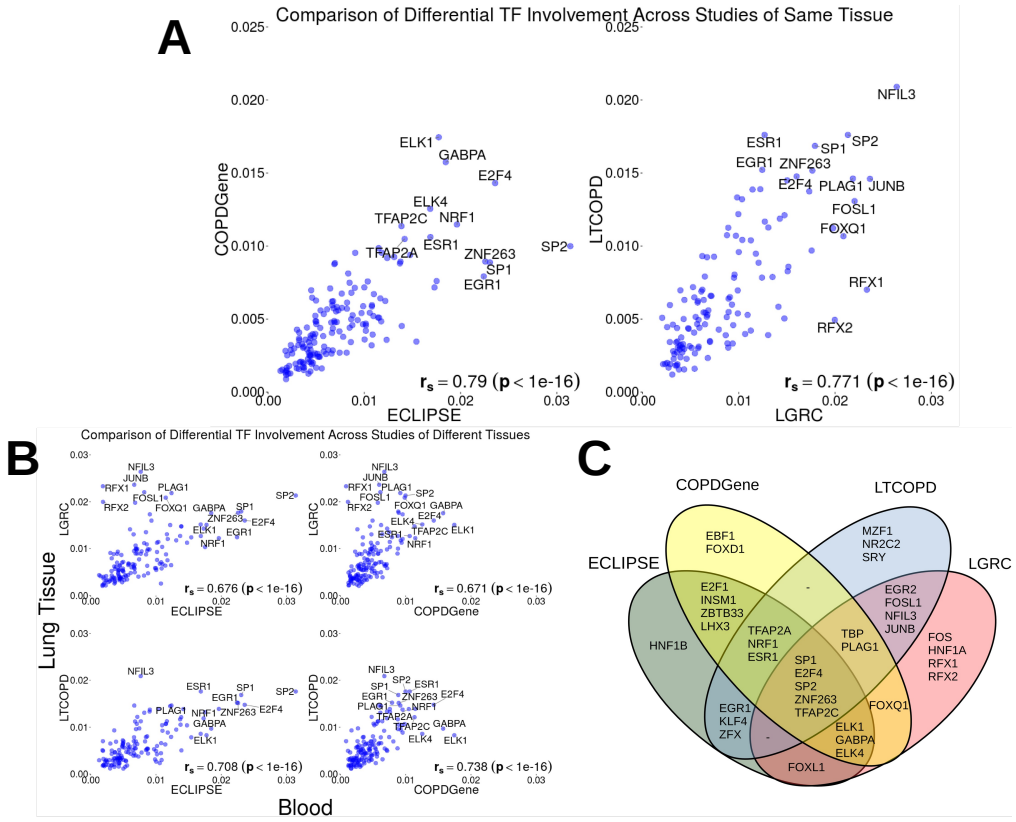

Fig. 3. Strong reproducibility in top differential transcription factor involvement found in case-control COPD studies. ECLIPSE and COPDGene data were obtained from whole-blood and PBMC while the gene expression data in LGRC and LT-CDNM were assayed in lung tissue. **A** Results for studies with gene 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 still found agreement between studies of different tissues. **C** Venn diagram depicting the top 20 transcription factors found in each study. Out of 166 original transcription factors the union of all four lists of top-20 hits yielded 36 total transcription factors.

studies. We also demonstrate the utility of MONSTER by showing that these transcription factors would not have been detected via differential gene expression analysis or conventional comparative network inference methods.

Although we apply MONSTER to a set of case-control studies, the method is not limited to such applications. Our method is suitable in identifying drivers of state change in any context involving a comparison of gene expression assays with regulatory data. Given the general nature of the method, MONSTER has the ability to shed light on possible biological mechanisms for cell state change that might otherwise have been undetected in the gene expression data across a wide range of applications.

ACKNOWLEDGMENTS. The project described was supported by Award Number R01 HL089897 and Award Number R01 HL089856 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.

1. Hill SM et al. (2012) Bayesian inference of signaling network topology in a cancer cell line. *Bioinformatics* 28(21):2804–2810.
2. Glass K et al. (2014) Sexually-dimorphic targeting of functionally-related genes in copd. *BMC systems biology* 8(1):118.
3. Glass K, Quackenbush J, Spentzos D, Haibe-Kains B, Yuan GC (2015) A network model for angiogenesis in ovarian cancer. *BMC bioinformatics* 16(1):115.
4. 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.
5. Chen WW et al. (2009) Input–output behavior of erbb signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular systems biology* 5(1).
6. Molinelli EJ et al. (2013) Perturbation biology: inferring signaling networks in cellular systems. *PLoS Comput Biol* 9(12):e1003290.
7. Saez-Rodriguez J et al. (2011) Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer research* 71(16):5400–5411.
8. Glass K, Huttenhower C, Quackenbush J, Yuan GC (2013) Passing messages between biological networks to refine predicted interactions. *PloS one* 8(5):e64832.
9. Singh D et al. (2014) Altered gene expression in blood and sputum in copd frequent exacerbators in the eclipse cohort. *PloS one* 9(9):e107381.
10. Vestbo J et al. (2008) Evaluation of copd longitudinally to identify predictive surrogate endpoints (eclipse). *European Respiratory Journal* 31(4):869–873.
11. Regan EA et al. (2011) Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease* 7(1):32–43.
12. Bahr TM 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.
13. Pillai SG 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.
14. (year?) Lung genomics research consortium (lgrc). Accessed: 2016-02-02.
15. Qiu W 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*. (Am Thoracic Soc), pp. A1253–A1253.
16. Margolin AA et al. (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* 7(Suppl 1):S7.
17. Faith JJ et al. (2007) Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 5(1):e8.
18. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* 4(1).
19. Danielian PS et al. (2007) E2f4 is required for normal development of the airway epithelium. *Developmental biology* 305(2):564–576.
20. 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.
21. Zhou X et al. (2012) Identification of a chronic obstructive pulmonary disease genetic determinant that regulates hhip. *Human molecular genetics* 21(6):1325–1335.
22. 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.
23. 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.
24. Chen ZH et al. (2008) Egr-1 regulates autophagy in cigarette smoke-induced chronic obstructive pulmonary disease. *PloS one* 3(10):e3316.
25. Cloonan SM et al. (2016) Mitochondrial iron chelation ameliorates cigarette smoke-induced bronchitis and emphysema in mice. *Nature medicine*.
26. Gopalakrishnan L, Scarpulla RC (1995) Structure, expression, and chromosomal assignment of the human gene encoding nuclear respiratory factor 1. *Journal of Biological Chemistry* 270(30):18019–18025.
27. Galvagni F, Capo S, Oliviero S (2001) Sp1 and sp3 physically interact and co-operate with gabp for the activation of the utrophin promoter. *Journal of molecular biology* 306(5):985–996.
28. Sîrbu A, Ruskin HJ, Crane M (2010) Comparison of evolutionary algorithms in gene regulatory network model inference. *BMC bioinformatics* 11(1):1.
29. Choksi SP, Lauter G, Swoboda P, Roy S (2014) Switching on cilia: transcriptional networks regulating ciliogenesis. *Development* 141(7):1427–1441.
30. Hessel J et al. (2014) Intraflagellar transport gene expression associated with short cilia in smoking and copd. *PloS one* 9(1):e85453.
31. Hogg JC (2004) Pathophysiology of airflow limitation in chronic obstructive pulmonary disease. *The Lancet* 364(9435):709–721.
32. Fahy JV, Dickey BF (2010) Airway mucus function and dysfunction. *New England Journal of Medicine* 363(23):2233–2247.
33. Irizarry RA et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2):249–264.
34. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy-analysis of affymetrix genechip data at the probe level. *Bioinformatics* 20(3):307–315.
35. Dai M et al. (2005) Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic acids research* 33(20):e175–e175.
36. Mathelier A et al. (2013) Jaspar 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research* p. gkt997.
37. Pinello L, Xu J, Orkin SH, Yuan GC (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.
38. Heinz S et al. (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.
39. Olsen C et al. (2014) Inference and validation of predictive gene networks from biomedical literature and gene expression data. *Genomics* 103(5):329–336.
40. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. *science* 315(5814):972–976.
41. Lao T et al. (2015) Haploinsufficiency of hedgehog interacting protein causes increased emphysema induced by cigarette smoke through network rewiring. *Genome medicine* 7(1):1.
42. Harbison CT et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431(7004):99–104.
43. Marbach D et al. (2012) Wisdom of crowds for robust gene network inference. *Nature methods* 9(8):796–804.
44. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.
45. Langfelder P, Horvath S (2008) Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics* (1):559.
46. Langfelder P, Horvath S (2012) Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software* 46(11):1–17.
47. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. *science* 297(5586):1551–1555.
48. Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424(6945):147–151.
49. Ravasi T et al. (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140(5):744–752.
50. Ritchie ME et al. (2015) limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research* p. gkv007.