Estimating Drivers of Cell State Transitions Using Gene Regulatory Network Models

Daniel Schlauch,[a,b] Kimberly Glass,[b,c] Craig P. Hersh,[b,c,d]
Edwin K. Silverman,[b,c,d] John Quackenbush,[a,b,c*]

[a]Department of Biostatistics and Computational Biology,
Dana-Farber Cancer Institute and Department of Biostatistics,
Harvard TH Chan School of Public Health, Boston, MA 02115,
[b]Channing Division of Network Medicine, Brigham and Women's
Hospital, Boston, MA 02115
[c]Department of Medicine, Harvard Medical School, Boston, MA 02115
[c]Pulmonary and Critical Care Division, Brigham and Women's
Hospital and Harvard Medical School, Boston, USA

*To whom correspondence should be addressed; E-mail: johnq@jimmy.harvard.edu

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 ob-

structive pulmonary disease to identify transcription factors that alter the network structure as the cell state progresses toward the disease-state. Our results demonstrate that MONSTER can find strong regulatory signals that persist across studies and tissues of the same disease and that are not detectable using conventional analysis methods based on differential expression. An R package implementing MONSTER is available at github.com/QuackenbushLab/MONSTER.

Introduction

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

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

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

$$\mathbf{B} = \mathbf{AT} \tag{1}$$

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

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

MONSTER: MOdeling Network State Transitions from Expression and Regulatory data

The MONSTER algorithm models the regulatory transition between two cellular states in three steps: (1) Inferring state-specific gene regulatory networks, (2) modeling the state transition matrix, and (3) computing the transcription factor involvement.

Inferring state-specific gene regulatory networks: Before estimating the transition matrix, **T**, we must first estimate a gene regulatory starting point for each state. While there have been many methods developed to infer such networks (*1*–*7*), we have found the bipartite framework used in PANDA (*8*) to have features that are particularly amenable to interpretation in the context of state transitions. PANDA begins by using genome-wide transcription factor binding data to postulate a network "prior", and then uses message-passing to integrate multiple data sources, including state-specific gene co-expression data.

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

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

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

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

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

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

where the response $\mathbf{M}$ is a binary $p \times m$ matrix indicating the presence of a sequence motif for the i^{th} transcription factor in the vicinity of each of the j^{th} gene. And where $g_j^{(k)}$ represents the gene expression measured for sample k at gene j. Thus, the fitted probability $\hat{\theta}_{i,j}$ represents our estimated indirect evidence. Combining the scores for the direct evidence, $\hat{d}_{i,j}$, and indirect evidence, $\hat{\theta}_{i,j}$, via weighted sum between each transcription factor-gene pair yields estimated edge-weights for the gene regulatory network (see Supporting Information).

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

Modeling the state transition matrix: Once we have gene regulatory network estimates for each phenotype, we can formulate the problem of estimating the transition matrix in a regression framework in which we solve for the $m \times m$ matrix that best describes the transformation between phenotypes (1). 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 Supporting Information).

This framework allows for the natural extension of constraints such as $L1$ and/or $L2$ regularization (see Supporting Information). 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 initial and final states diverge, there should be increasing differences in their corresponding gene regulatory networks and, consequently, the transition matrix will also increasingly diverge from the identity matrix. In this model, the transcription factors that most significantly alter their regulatory targets will have the greatest “off-diagonal mass” in the transition matrix, meaning that they will have very different targets between states and so are likely to be involved in the state transition process. We define the "differential transcription factor involvement" (dTFI) as the magnitude of the off-diagonal mass associated with each transcription factor, or,

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

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

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

As a demonstration of the power of MONSTER to identify driving factors in disease, we applied the method to case-control gene expression data sets from four independent Chronic Obstructive Pulmonary Disease (COPD) cohorts: Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) (*9*) (*10*) (2), COPDGene (*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 commonly used network inference methods on these case-control studies. In analyzing phenotypic changes, networks are generally compared directly, with changes in the presence or weight of edges between key genes being of primary interest. It is therefore reasonable to assume that any reliable network results generated from a comparison of disease to controls will be reproducible in independent studies. We investigated whether this is the case for our four COPD data sets using three widely used network inference methods - Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE) (*16*), Context Likelihood of Relatedness (CLR) (*17*), and Weighted Gene Correlation Network Analysis (WGCNA) (*18*) - computing the difference in edge weights between cases and controls for each of the four studies. We found no meaningful correlation ($R^2 < .01$) of edge weight difference across any of the studies regardless of network inference method or tissue type (Supporting Figure 3). Edge weight differences, even when very large in one study, did not reproduce in other studies. This suggests that a simple direct comparison of edges between inferred networks is insufficient for extracting reproducible drivers of network state transitions. This finding may be unsurprising given the difficulty in inferring individual edges in the presence of heterogeneous phenotypic states, technical and biological noise with a limited number of samples.

The lack of replication in edge-weight differences between independent data sets representing similar study designs indicates that we need to rethink how we evaluate network state transitions. MONSTER provides a unique approach for making that comparison. In each of the four COPD data sets, we used MONSTER to calculate the differential transcription factor involvement ($dTFI$, Equation 2) for each transcription factor and used permutation analysis to estimate their significance (Figure 2, Additional Figures 1-3). We observed strongly significant ($p < 1e - 15$) correlation in dTFI values for each pairwise combination of studies. In addition, out of the top 10 most differentially involved transcription factors in the ECLIPSE and COPDGene studies, we found 7 to be in common. Furthermore, three of these seven transcription factors (GABPA, ELK4, ELK1) also appeared as significant in the LGRC results with FDR¡0.01 and each of the top five ECLIPSE results were among the top seven in the LT-CDNM results (Additional Table 1, Additional Figure 3). This agreement is quite striking considering that the there was almost no correlation in the edge-weight differences across these same studies when we tested the other methods. But it is exactly what we should expect—that the same method applied to independent studies of the same phenotypes should produce largely consistent results.

Many of the top dTFI transcription factors, especially those identified by MONSTER across all four studies, are biologically plausible candidates to be involved in the etiology of COPD (Additional Table 1, Additional Figures 1-3). For example, E2F4 is a transcriptional repressor important in airway development (*19*) and studies have begun to demonstrate the relevance of developmental pathways in COPD pathogenesis (*20*).

Some of the greatest effect sizes across all four studies were found for SP1 and SP2. An additional member of the SP transcription factor family, SP3, has been shown to regulate HHIP, a known COPD susceptibility gene (*21*). Both SP1 and SP2 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*).

Mitochondrial mechanisms have also been associated with COPD progression (*25*). Two of most highly significant transcription factors based on dTFI in ECLIPSE were NRF1 and GABPA (FDR¡.001). Indeed, these TFs had highly significant dTFI (FDR¡0.1) in all four studies. NRF1 regulates the expression of nuclear encoded mitochondrial proteins (*26*). GABPA, also known as human nuclear respiratory factor-2 subunit alpha, may have a similar role in nuclear control of mitochondrial gene expression. Furthermore, GABPA interacts with SP1 (*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 notable as it suggests that the transition from smoker control to disease phenotype affects multiple tissues and supports the growing evidence for a role in immune response in COPD pathogenesis.

Gene regulatory networks, and results derived from their comparison, are notoriously difficult to replicate across studies (*28*). The four studies we used each has unique aspects, including the choice of microarray platform, study demographics, location, time, and tissue. Nevertheless, MONSTER identified similar sets of transcription factors associated with the transition between cases and controls. This consistency in biologically-relevant transcription factors, associated with the transition from the control phenotype to disease, in four independent studies suggests that MONSTER can provide not only robust network models, but also can identify reliable differences between networks.

Despite the overall consistency, some transcription factors had variable $dTFI$ across studies. For example, using the LGRC dataset, we discovered a highly significant ($FDR < .0001$) differential targeting pattern involving the transcription factors RFX1 and RFX2 (Additional Table 1). However, these same TFs were not identified as potential drivers of the control to COPD transition in either the ECLIPSE or COPDGene study. This difference is likely due the differences in tissue type as the RFX family transcription factors are known to regulate ciliogenesis (*29*). Cilia are critical for clearing mucous from the airways of healthy individuals, but disruption can lead to infection and potentially to chronic airflow obstruction (*30–32*).

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

Discussion

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

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

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

Third, the change in the gene regulatory network structure between phenotypes will be driven by changes in the connectivity of the regulators—the transcription factors that alter when, how, and how strongly genes are expressed. A natural hypothesis in this model is that the transition between phenotype is likely associated with the transcription factors that experience the greatest change in their regulatory patterns between states, and that the activation or inactivation of their target genes, and the functions carried out by those genes, likely reflect the phenotypic differences between states.

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

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

Acknowledgements

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

Author Contributions

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

References and Notes

1. S. M. Hill, *et al.*, *Bioinformatics* **28**, 2804 (2012).

2. K. Glass, *et al.*, *BMC systems biology* **8**, 118 (2014).

3. K. Glass, J. Quackenbush, D. Spentzos, B. Haibe-Kains, G.-C. Yuan, *BMC bioinformatics* **16**, 115 (2015).

4. F. Eduati, J. De Las Rivas, B. Di Camillo, G. Toffolo, J. Saez-Rodriguez, *Bioinformatics* **28**, 2311 (2012).

5. W. W. Chen, *et al.*, *Molecular systems biology* **5** (2009).

6. E. J. Molinelli, *et al.*, *PLoS Comput Biol* **9**, e1003290 (2013).

7. J. Saez-Rodriguez, *et al.*, *Cancer research* **71**, 5400 (2011).

8. K. Glass, C. Huttenhower, J. Quackenbush, G.-C. Yuan, *PloS one* **8**, e64832 (2013).

9. D. Singh, *et al.*, *PloS one* **9**, e107381 (2014).

10. J. Vestbo, *et al.*, *European Respiratory Journal* **31**, 869 (2008).

11. E. A. Regan, *et al.*, *COPD: Journal of Chronic Obstructive Pulmonary Disease* **7**, 32 (2011).

12. T. M. Bahr, *et al.*, *American journal of respiratory cell and molecular biology* **49**, 316 (2013).

13. S. G. Pillai, *et al.*, *PLoS Genet* **5**, e1000421 (2009).

14. (2015). Accessed: 2016-02-02.

15. W. Qiu, *et al.*, *A30. BIG DATA: HARVESTING FRUITS FROM COPD AND LUNG CANCER* (Am Thoracic Soc, 2015), pp. A1253–A1253.

16. A. A. Margolin, *et al.*, *BMC bioinformatics* **7**, S7 (2006).

17. J. J. Faith, *et al.*, *PLoS Biol* **5**, e8 (2007).

18. B. Zhang, S. Horvath, *Statistical applications in genetics and molecular biology* **4** (2005).

19. P. S. Danielian, *et al.*, *Developmental biology* **305**, 564 (2007).

20. O. Boucherat, M. Morissette, S. Provencher, S. Bonnet, M. F, *American Journal of Respiratory and Critical Care Medicine* **193**, 362 (2016).

21. X. Zhou, *et al.*, *Human molecular genetics* **21**, 1325 (2012).

22. H. Rotheneder, S. Geymayer, E. Haidweger, *Journal of molecular biology* **293**, 1005 (1999).

23. J. Karlseder, H. Rotheneder, E. Wintersberger, *Molecular and cellular biology* **16**, 1659 (1996).

24. Z.-H. Chen, *et al.*, *PloS one* **3**, e3316 (2008).

25. S. M. Cloonan, *et al.*, *Nature medicine* (2016).

26. L. Gopalakrishnan, R. C. Scarpulla, *Journal of Biological Chemistry* **270**, 18019 (1995).

27. F. Galvagni, S. Capo, S. Oliviero, *Journal of molecular biology* **306**, 985 (2001).

28. A. Sîrbu, H. J. Ruskin, M. Crane, *BMC bioinformatics* **11**, 1 (2010).

29. S. P. Choksi, G. Lauter, P. Swoboda, S. Roy, *Development* **141**, 1427 (2014).

30. J. Hessel, *et al.*, *PloS one* **9**, e85453 (2014).

31. J. C. Hogg, *The Lancet* **364**, 709 (2004).

32. J. V. Fahy, B. F. Dickey, *New England Journal of Medicine* **363**, 2233 (2010).

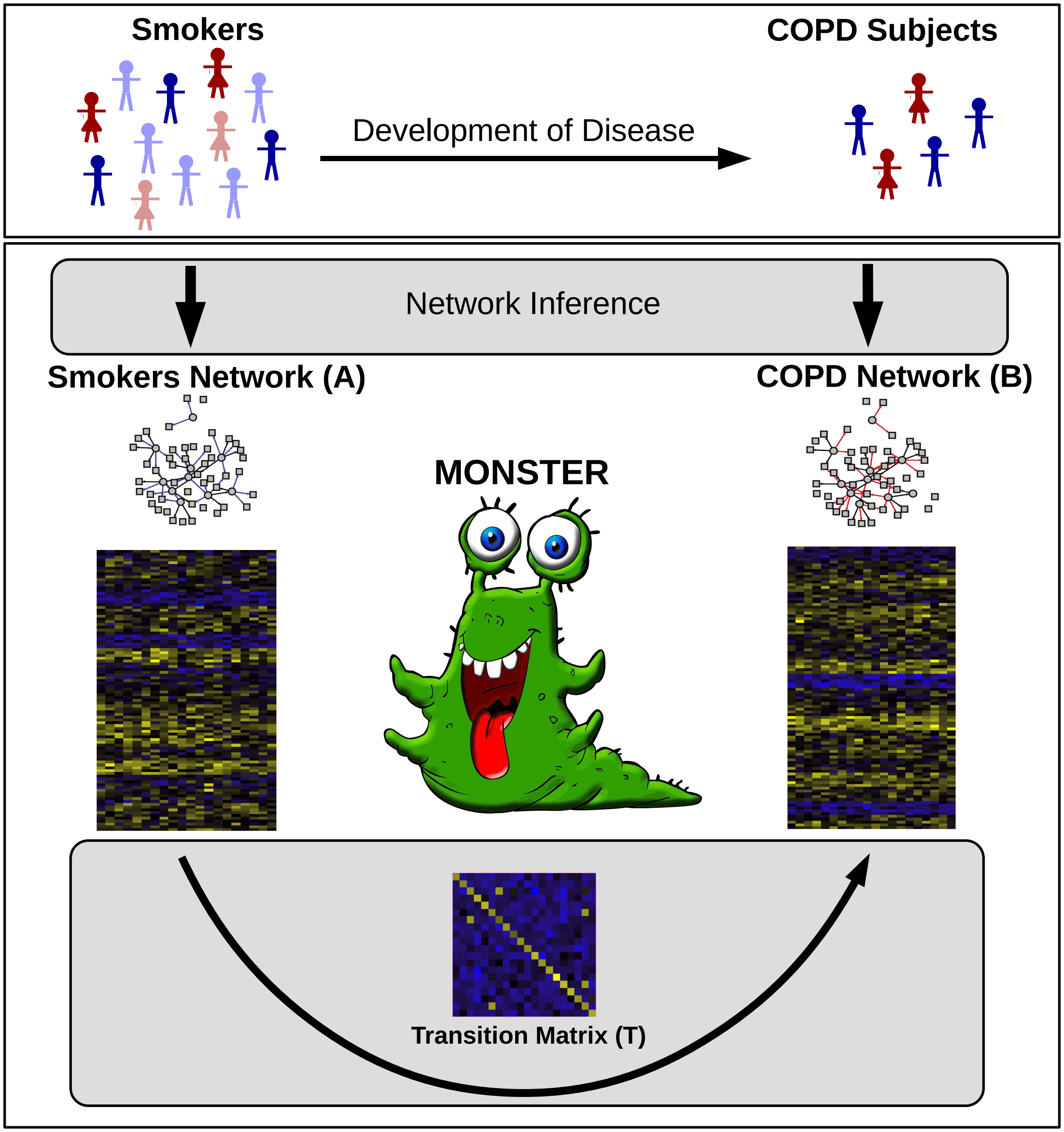

Figure 1: **Overview of the MONSTER approach, as applied to the transition between smokers and those suffering from chronic obstructive pulmonary disease (COPD).** MONSTER's approach seeks to find the $TF \times TF$ transition matrix that best characterizes the state change in network structure between the initial and final biological conditions. Subjects are first divided into two groups based on whether they have COPD or are smokers that have not yet developed clinical COPD. Network inference is then performed separately on each group, yielding a bipartite adjacency matrix connecting transcription factors to genes. Finally, a transition matrix is computed which characterizes the conversion from the consensus Smokers Network to the COPD Network.

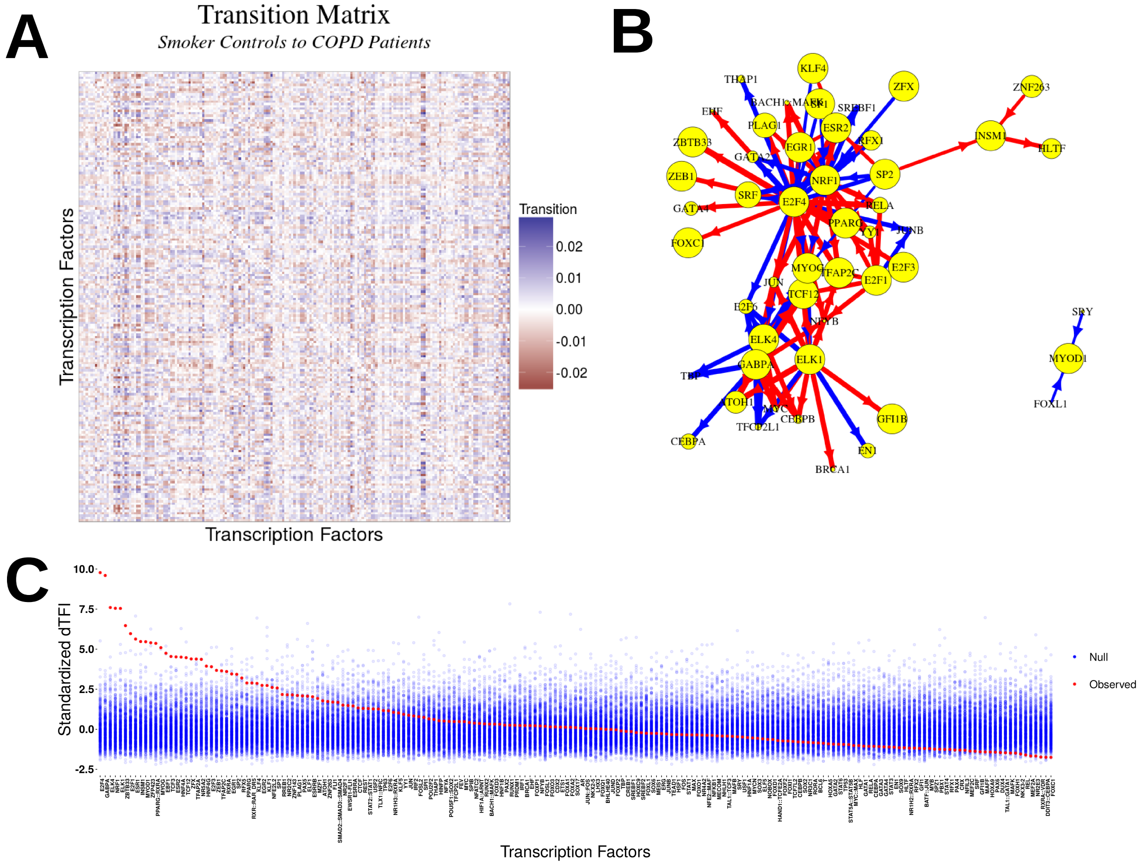

Figure 2: **MONSTER analysis results in the ECLIPSE study.** **A** Heatmap depicting the transition matrix calculated for smoker controls "transitioning" to COPD by applying MONSTER to ECLIPSE gene expression data. For the purposes of visualization, the magnitude of the diagonal is set to zero. **B** A network visualization of the 100 largest transitions identified based on the transition matrix in (A). Arrows indicate a change in edges from a transcription factor in the Smoker-Control network to resemble those of a transcription factor in the COPD network. Edge thickness represents the magnitude of the transition and node (TFs) sizes represent the dTFI for that TF. Blue edges represent a gain of targeting features and red represents the loss. **C** The dTFI score from MONSTER (red) and the background null distribution of dTFI values (blue) as estimated by 400 random sample permutations of the data.

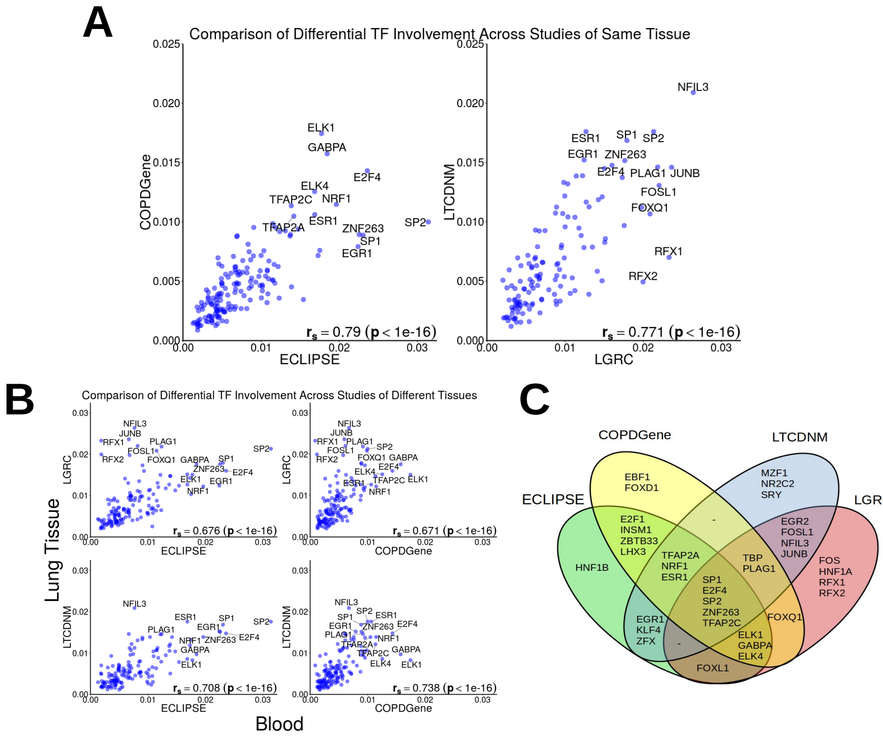

Figure 3: **Strong reproducibility in top differential transcription factor involvement found in case-control COPD studies**. ECLIPSE and COPDGene profiled gene expression in whole-blood and PBMC while the gene expression data in LGRC and LT-CDNM were assayed in lung tissue. **A** Results for studies with gene expression data obtained from the same-tissue. Both the blood based (left) and lung tissue studies (right) demonstrate very high Spearman correlation of differential involvement. **B** Despite using data from different sources we found agreement between studies of different tissues. **C** Venn diagram depicting the top 20 transcription factors found in each study. The union of all top 20 lists contains 36 transcription factors.

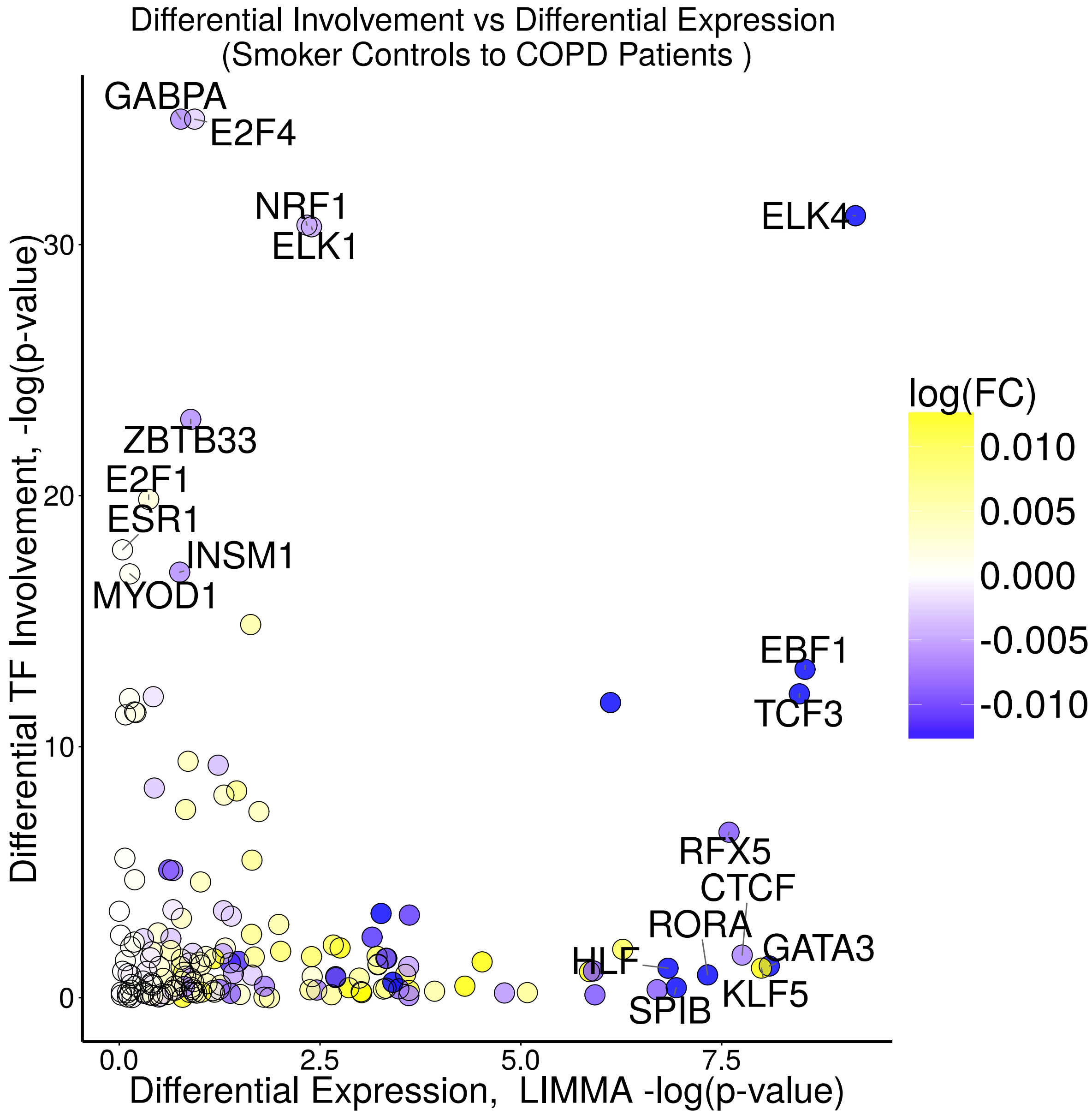

Figure 4: **Differentially involved transcription factors are not necessarily differentially expressed.** A plot of the differential expression versus the differential involvement for transcription factors based on our analysis of the ECLIPSE data. MONSTER commonly finds transcription factors which are differentially involved but are expressed at similar levels across cases and controls. Importantly, these transcription factors would not have been identified using conventional differential expression methods. This demonstrates the unique potential MONSTER has for discovery beyond standard gene expression analysis.