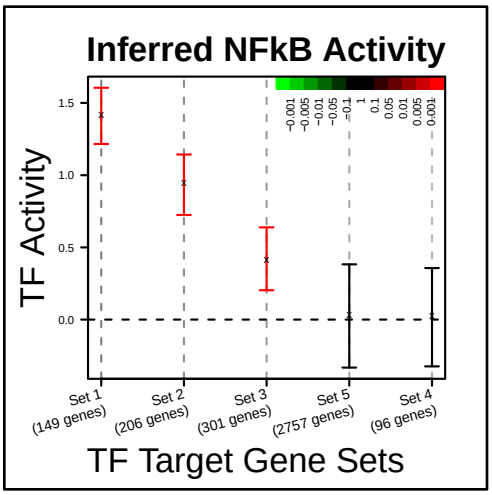
Introduction Transcriptional regulation is one of the crucial ways in which our cells control gene expression. Transcription factor (TF) activity itself is difficult to measure experimentally in high-throughput; however, many insights can be gained from applying statistical methods to infer activity from the expression of TF target genes. This indirect quantification of TF activity has been made possible by gene expression microarrays, which simultaneously profile thousands of genes. There has been extensive research on how to test for differential expression of *a priori* defined gene sets such as TF target genes.[1] One method recently developed in the Kleinstein lab, Quantitative Set Analysis of Gene Expression (QuSAGE), is unique in that it produces a probability density function for each set by convolution of the expression profiles of individual genes.[2] Still, the efficacy of all these methods relies on the chosen TF target gene set. The adjacent figure shows how the NFkB activity inferred after stimulation with TNF (an inducer of NFkB) is highly dependent on the choice of gene set. Thus, there is a need to improve methods for generating and refining such sets.



There are numerous ways to generate putative TF target gene sets both computationally and experimentally. Computational predictions provide candidate targets by scanning for a specific binding motif in promoter sequences genome-wide. However, this method is known to generate many false-positives. Protein-DNA binding experiments (e.g. ChIP-Seq, ChIP-ChIP) provide experimental evidence for TF-DNA interactions. However, there are a large number of binding interactions observed, many not in the promoter of known genes, and these interactions may be specific to the cell line used. In addition, the accuracy of both of these prediction methods suffers because the occurrence of a TF binding site or the actual binding of a TF to a gene promoter does not necessarily imply transcriptional regulation. Networks from pathway databases (e.g. KEGG) provide some additional information about which genes interact at a transcriptional level. However, the number of interactions in pathways is limited. Currently each TF is associated with a single target gene set. This is problematic because, in reality, TF target genes depend on the cellular and environmental context of the cell. To infer TF activity more accurately, candidate target gene sets from many sources can be refined to include only the genes under a TF's control in the specific context being studied. *I will develop a method for generating context-specific transcription factor target gene sets* (Aim 1), and *apply these gene sets to infer transcription factor activity during infection and vaccination responses* (Aim 2).

Aim 1: ***Develop a method for generating context-specific TF target gene sets.***
I will begin with a large set of proposed candidate TF target genes and then to utilize co-expression patterns from gene expression data to select candidate genes having a similar gene

expression pattern. One limitation of current approaches is that each gene is either a candidate of a TF, or it is not. I plan to integrate multiple information sources including computationally predicted binding sites from motif scanning algorithms, protein-DNA binding data, and pathway information to compute a prior probability for each gene being a candidate target of each TF. This prior probability represents the strength of the evidence for a certain gene being a target of a certain TF. Because the relative importance of each of the data sources is not known, I will estimate them as parameters in a Bayesian network. I hypothesize that this extra information will improve correct identification of TF targets. The method will allow for overlap of genes between target sets but I will explore whether this is necessary, since dependence between target sets is often problematic for quantification of TF activity. The proposed method will build on the framework proposed by Fertig et al.[3] Some TFs are transcriptionally regulated themselves, allowing for estimation of activity directly from gene expression measurements. I will evaluate my method by comparing the activities inferred from the proposed and published methods for these transcriptionally regulated TFs.

Aim 2: ***Apply these gene sets to infer transcription factor activity during infection and vaccination responses.***

One natural application of these TF target gene sets is to infer TF activity. Thus, I plan to apply the developed method to specific time-series gene expression data sets of influenza infection and vaccination responses. I have access to these data through the NIAID funded Program for Research on Immune Modeling and Experimentation (PRiME) and the Human Immunology Project Consortium (HIPC). In the case of the influenza infection data, the different contexts correspond to four different strains of *in vitro* influenza infection; while in the case of vaccination response data, the contexts correspond to vaccine responders or non-responders. Generating context-specific gene sets will allow us to answer two fundamental questions: Are there changes in which genes are regulated by certain TFs across contexts? And how do the activities of TFs differ between contexts? I will answer the first question by applying differential network analysis, a method to identify how the regulatory network is rewired in different contexts. To answer the second question, I will infer activity of each gene set using QuSAGE to find quantitative differences in TF activity between contexts.

Significance and Broader Impacts TFs are key regulators in development and disease. The ability to better characterize TF activity thus has implications for understanding disease states and the mechanisms underlying development. The proposed integrative approach to generate context-specific TF target gene sets will improve understanding of transcriptional regulation and allow for a more accurate inference of TF activity.

1. Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z. & DeLisi, C. Gene set enrichment analysis: performance evaluation and usage guidelines. Brief. Bioinform. 13, 281–91 (2012).
2. Yaari, G., Bolen, C. R., Thakar, J. & Kleinstein, S. H. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. Nucleic Acids Res. 41, e170 (2013).
3. Fertig, E. J., Favorov, A. V & Ochs, M. F. Identifying context-specific transcription factor targets from prior knowledge and gene expression data. IEEE Trans. Nanobioscience 12, 142–9 (2013).