Improved Methods for Causal Inference and Experimental Prioritization in Gene Regulatory Networks

Jonathan Lu
2016-2017
Advised by Prof. Barbara Engelhardt

SUBMITTED TO PRINCETON UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

This Thesis represents my own work in accordance with University regulations.

Date of Submission: May 5th, 2017

Abstract

We address two important problems: causal inference of gene regulatory networks and experimental prioritization of genes for perturbation experiments. Successful solutions to both techniques can accelerate discovery of the pathways implicated in disease and drug response, and facilitate improved treatments to suppress adverse effects. Given the transcriptome-wide expression time series from a set of samples, the problem of causal inference in gene regulatory networks is to reconstruct a directed graph where nodes are protein-coding genes and edges denote a causal up or down regulation of expression. This requires solving several statistical problems, including high dimensionality, statistical significance, and validation. Prioritization of genes for followup experiments is also challenging without a system for hypothesizing the effect of perturbations on the global regulatory network. In this work, we develop a causal network inference pipeline (*VAR-GEN*) based on Vector Autoregression and apply it to transcriptional time series data in A549 cells exposed to glucocorticoids over a period of 12 hours. We validate the inferred causal network using genetic variants associated with pairs of connected genes. Finally, we develop a prioritization method, (*CCI*), based on Perturbation PageRank to rank genes by their causal influence in the global network context. We highlight key genes and relationships between genes which may play essential roles in the immune and metabolic effects of the glucocorticoid.

Contents

1. Acknowledgments

First and foremost, I wish to thank my primary mentors Bianca and Professor Engelhardt. Thank you for introducing me to the field of causal inference, for guiding me through the challenging statistics and implementation of this project, and being so kind, encouraging and patient past all mistakes and roadblocks on my part. I am so lucky to be able to work with you!

I would also like to thank Professor Troyanskaya for being willing to be the second reader for this thesis. Thank you for taking the time out of your busy schedule to be my reader!

I would like to thank my other mentors, Brian, Ian, and Professor Reddy. Brian, thank you for being so helpful and encouraging through every stage of this project. Ian and Professor Reddy, thank you for your guidance in helping us to understand the biological context and capture the fascinating biology behind the GR dataset.

I wish to thank the other members of the Engelhardt lab, including Derek, Greg D., Ari, Izzy, and Allison. Whether it was explaining a concept during our paper reviews, running jobs on the cluster, or helping me solve a tricky Linux problem, I am so grateful to you for always being there to help.

Finally, I would like to thank the data collection team at the Reddy Lab, including Alejandro, Linda, and Sarah, for gathering and preparing the fascinating GR dataset. I have learned so much from this project, and none of it would have been possible without the GR Data.

Author Contributions

Much of the material in this thesis is taken from a manuscript drafted with co-authors. For those sections of this thesis that I was not primarily responsible for implementing and writing, I have given appropriate credit to the authors at the beginning of that section.

Below, I include the information for the manuscript.

"Causal network inference from gene transcription response to glucocorticoids"

Bianca Dumitrascu,[1*] Jonathan Lu,[2*] Ian C. McDowell,[3] Brian Jo,[1] Alejandro Barrera,[4,5], Linda K. Hong,[4] Sarah M. Leichter,[4] Timothy E. Reddy,[3,†] Barbara E. Engelhardt[2,6†]

[1] Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

[2] Department of Computer Science, Princeton University, Princeton, NJ 08540, USA

[3] Department of Genome Sciences, Duke University, Durham, NC 27708, USA

[4] Center for Genomic and Computational Biology, Duke University, Durham, NC 27708, USA

[5] Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC 27710, USA

[6] Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08540, USA

* These authors contributed equally

,†To whom correspondence should be addressed; E-mail: bee@princeton.edu, tim.reddy@duke.edu.

TER and BEE designed and funded the study. LKH coordinated all genomic data production. LKH and SML collected RNA-seq data. JL, BD, BJ and BEE developed the methods and validation approaches. JL and BD applied these to data. ICM, TER, BJ, BD, JL, and AB analyzed the data. BJ performed validation in the GTEx data. BD, JL, and BEE drafted the manuscript, and all authors contributed to revision.

2. Introduction

Computational and statistical methods have become increasingly important to the analysis of biological datasets, in particular those derived from Next-Generation Sequencing (NGS). The rise of NGS now allows researchers to directly profile the genetics of living systems at high spatiotemporal resolution [1]. The data is rich enough so that we can gain insight into the causal mechanisms behind biological phenomena such as disease and drug response. I am specifically working to understand why glucocorticoids, which are drugs that control overactive immune reactions [2, 3, 4, 5], lead to metabolic side effects, such as diabetes and obesity [6, 7].

Though NGS data is rich, it also carries new challenges, such as high dimensionality with small sample size. Analyzing such data depends heavily on the development of effective statistical models and computational approaches. Furthermore, the insights from these methods should not be limited to data analysis, but also aid biologists in determining the optimal set of subsequent experiments to perform. **The goal of this project is to develop a framework for causal inference and experimental prioritization of gene regulatory networks, based on the GR dataset, the gene expression time series from glucocorticoid-stimulated cells.** Through this we aim to accelerate the process of scientific discovery, to enable development of treatments with the same beneficial effects on immunity as the glucocorticoid, but without the metabolic side effects such as diabetes.

In studying the glucocorticoid genetic response, the primary computational model we shall use is the Gene Regulatory Network (GRN). A GRN shows the directed network of causal relations where nodes are protein encoding genes and edges denote a causal up or down regulation of expression. The main problem is, given the transcriptome-wide time-series expression from a set of samples, to reconstruct the GRN. The problem of GRN inference has been extensively studied but faces major statistical challenges. Several effective models exist, including Vector Autoregression [8], Dynamic Bayesian Network [9], and Mutual Information [10]. In this work we shall focus on Vector Autoregression, due to its simplicity, speed, and empirical effectiveness in previous studies [8, 11] and our own simulation study. Though VAR is effective on small simulations, several problems must be addressed in order to apply it successfully to our high-dimensional glucocorticoid gene expression data. One must carefully design the methods for normalization, technical replicates, regularization, significance testing, false discovery control, and external validation. Though one or another of these problems have been addressed in previous work [12, 13, 14, 8], to our knowledge there has not been work that implements a comprehensive pipeline for all of these.

Our first contribution is to build a robust causal network pipeline, *VAR-GEN*, that addresses each step of the modelling process and rigorously validates on external data (Sections 5.1, 6.5.2). *VAR-GEN* performs high-dimensional Vector Autoregression with hyperparameter tuning. It addresses the problem of statistical significance by including multiple options for generating a permutation null and controlling the False Discovery Rate. We then validate our networks externally using gene association tests from the Genotype Tissue-Expression Consortium, demonstrating enrichment of dependent relations within our causal networks.

In addition, previous computational approaches over gene regulatory networks have focused on modelling and analysis of the underlying biology. However, biologists would also benefit greatly from experimental prioritization methods. In our particular project, we wish to use our inferred causal network to prioritize follow-up perturbation experiments with the ultimate goal of suppressing the metabolic response while maintaining the immune response. A naive approach is to simply list those genes with the most metabolically-related genes. However, this is insufficient in the case of a more complex network. An influential gene may regulate several genes that are not metabolically-related themselves, but each of which regulates metabolic genes. One must therefore consider the network context in measuring a gene's influence.

Our second contribution is thus to develop a contextual causal influence score (*CCI*) that considers network context to prioritize experiment interventions. (Section 5.2) This score is inspired by the PageRank algorithm [15], which provides a means to quantify the centrality of genes in the context of the global network topology. We leverage a variant known as Perturbation PageRank [15] to ranks network genes in terms of their strong causal influence on metabolic targets and weak causal influence on immune targets. We thus provide a method to identify candidate genes whose suppression will limit the adverse metabolic effects of glucocorticoids, while preserving the therapeutic immune effects. To our knowledge, such PageRank methods have not been previously applied to GRN networks.

Together, the causal pipeline *VAR-GEN* and the experimental prioritization method *CCI* form a coherent framework to analyze gene expression time series data and prioritize followup experiments. (Figure 1).

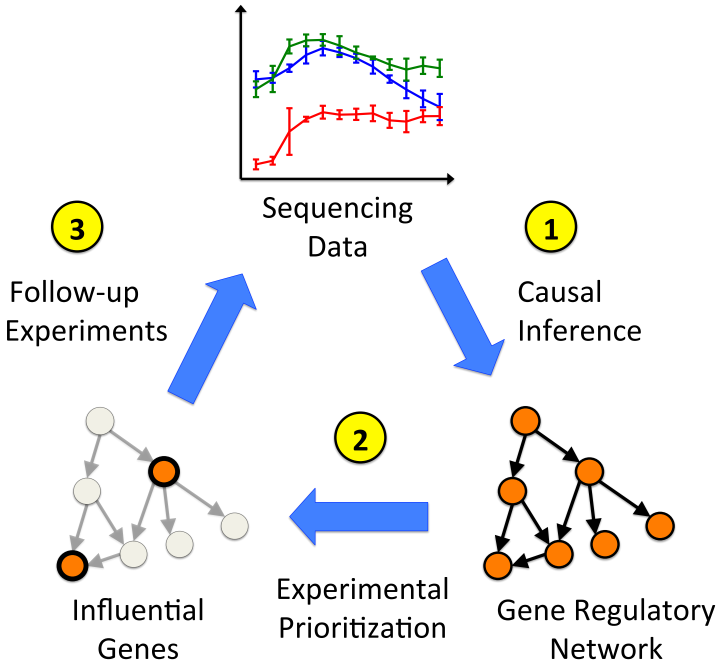

Figure 1: *Computational Framework for Gene Regulatory Network Analysis and Experimentation.* 1) A Gene Regulatory Network is generated from sequencing data using a causal inference model, *VAR-GEN*. 2) Genes are prioritized for experimental perturbation based on the desired biological response. In our case, the desired response is a limitation of metabolic effects while preserving immune effects. This uses the contextual causal influence score *CCI*. 3) Those genes' interactions are probed further in follow-up experiments. Data from those experiments can then be used to infer further more Gene Regulatory Networks.

First, in Section 3, we review related approaches to causal inference and experimental prioritization. In Section 4, we describe the GR gene expression dataset and our preprocessing procedures. In Section 5, we discuss the methods: the details of *VAR-GEN* and *CCI*, as well as the validation procedure and annotation methods. In Sections 6.1 - 6.4, we perform a preliminary analysis of the causal inference model, testing it on both a simulated gene dynamical system and on the real GR data, with multiple parameter settings, in order to choose appropriate settings for our main network. In Section 6.5, we perform the analysis of our main inferred GR network, with a focus on biologically relevant findings. We discuss the Validation (Subsection 6.5.2) and the Experimental Prioritization results (Subsection 6.5.3). We conclude in Section 7.

3. Problem Background and Related Work

3.1. Causal Inference Methods in Gene Regulatory Networks

A variety of approaches for causal inference in gene regulatory networks have been studied over the past decade. We review the approaches and discuss several comparison studies. A more detailed treatment can be found in [10, 16].

- **Vector Autoregression:** Model a gene's expression as a linear function of its and other genes' previous values
- **Differential Equations:** Like Vector Autoregression, except model the rate of change in a gene's expression instead of the absolute gene expression.
- **Dynamic Bayesian Networks:** Model the joint probability distribution of the gene expression values as the decomposition of conditional probability distributions.
- **Mutual Information:** Model the mutual information between a gene's expression values and other genes' previous values
- **Boolean Networks:** Model genes' binary values as a logical function of previous genes' binary values.
- **Non-parametric Dynamical Systems:** Model gene's expression as an unknown function of previous expression using a Gaussian Process prior

Vector Autoregressions (VAR) model the expression of a gene Y at time t as:

$$Y_t = \sum_{i=1}^{K} \alpha_i Y_{t-i} + \sum_{i=1}^{K} \beta_i X_{t-i} + \varepsilon_t$$

where $\varepsilon_t \overset{iid}{\sim} N(\mu, \sigma^2)$ and i denotes the lag of the causal effect, with K as the maximum lag. This is based on the principle of Granger Causality [17], in which $X \rightarrow Y$ if including information from X improves our prediction of Y. The causal edge $X \rightarrow Y$ is inferred when β_i is found to be significantly different from 0. We review previous VAR methods in Subsection 3.3, and compare with our method, *VAR-GEN*.

Differential Equations fit the change in the expression of a gene Y at time t [18, 19, 20].

$$\frac{dY_t}{dt} = \beta_0 + \sum_{i=1}^{n} \alpha_i Y_{t-i} + \sum_{i=1}^{n} \beta_i X_{t-i}$$

Since the derivative $\frac{dY_t}{dt}$ cannot be computed in real data with substantially spaced timepoints, it is often approximated by $\frac{Y_t - Y_{t-1}}{\Delta t}$. This results in the model:

$$\frac{Y_t - Y_{t-1}}{\Delta t} = \beta_0 + \sum_{i=1}^{n} \alpha_i Y_{t-i} + \sum_{i=1}^{n} \beta_i X_{t-i}$$

If we assume a fixed Δt and simply multiply it on both sides, then move the Y_{t-1} term to the right-hand side one can show that the differential equation model is actually equivalent to the Vector Autoregression model. The methods are closely related.

Dynamic Bayesian Networks (DBNs) model the global joint distribution as a breakdown of conditional probability distributions [21, 22, 9].

$$P(X_1^1, \ldots, X_1^n, X_2^1, \ldots, \ldots X_t^n) = P(X_1) \prod_{t=2}^{T} \prod_{i=1}^{n} P(X_t^i | pa(X_t^i))$$

Here, X_t^i is the expression of gene i at time t.

Two form are typically used. The first is the quantized DBN. Gene expression values are quantized into 3 values: $-1, 0, 1$ representing under-expression, baseline, and over-expression respectively. The conditional distribution is then discrete, constructed from the contingency table [9]. The second is the Gaussian DBN. Gene expression values are modelled as conditionally Gaussian given previous values [9]. In the Gaussian case, the model is quite similar to the Vector Autoregressive model, which also assumes conditional Gaussianity of effects Y_t given the causes X_{t-1}. An advantage of the

DBN is that they are commonly used to integrate external information, such as epigenetic data [23] or functional association networks [24]. One should note that because whole graphs are modelled jointly and the graph structure search space grows at least exponentially in the number of nodes, inference of DBNs is often highly computationally intensive [9].

DBNs have been used to model cell signaling networks from single-cell phosphoprotein data [25]. Several approaches have been developed to model gene regulatory networks from gene expression time series [22, 26, 21].

Mutual Information (MI) methods compute the mutual information between gene expression values. They can be extended to detect causal relations by taking the lagged values of the causal gene X [27].

$$I^k(X,Y) = -\sum_{t=k}^{T} P(X_{t-k},Y_t)\log\frac{P(X_{t-k},Y_t)}{P(X_{t-k})P(Y_t)}$$

Boolean Network (BN) methods model each gene as a boolean function of the values of previous genes. Here, each gene's expression has been discretized.

$$Y_t = f(X^1_{t-1},\ldots,X^n_{t-1}$$

The advantage of these methods is that they can be simple and provide an understanding of the network as a logical circuit. The main drawback is the discretization of gene expression, which may not be biologically realistic [16]. Furthermore, Boolean networks are challenging to compute, because of the large search space of network structure and logical functions [16].

Non-parametric Dynamical Systems (NDS) methods model each gene as a nonlinear function f of the values of previous genes, using a Gaussian process prior [28].

$$X(t+1) = f(X_1(t),\ldots,X_N(t))$$

where

$$p(X_t^i|X_{pa(x)}) \sim \mathcal{N}(X_t^i|\mu(X_{pa(x)}), K(X_{pa(x)}))$$

These provide an effective and flexible model for the gene dynamics [28]. Networks are then inferred by integrating over the distribution of parental sets, finding the model that maximizes the likelihood. The drawback is that Gaussian Process computation time grows exponentially in the number of possible causal genes, and requires constraints to be tractable [28].

3.2. Comparison of Causal Inference Methods

Several studies have evaluated the effectiveness of these causal inference methods in simulated and real gene expression data. We focus on those involving the Vector Autoregression. The main finding is that VAR performs effectively on data of similar time interval and high dimensionality to the GR data, which has 1-hour interval and thousands of genes (Section 4) [11, 8]. However, in other studies VAR was inferior to the Dynamic Bayesian Network (DBN) and Nonparametric Dynamical Systems (NDS) for small and shorter time series [21, 28].

Lopes [11] assess Vector Autoregressions (VAR), Dynamic Bayesian Networks (DBN), and Mutual Information methods (MI) on three microarray datasets: a 22-hour fly dataset with hour-long time intervals (primarily), a 5-hour E. Coli dataset with 10-50 minute time intervals, and a 2-hour Yeast dataset with 5 minute time intervals. The method accuracy was evaluated by comparison with known interactions in a database. The authors found that the lag-1 Vector Autoregressive models, "VAR(1) + lars" and "simone", performed the best on the Fly dataset with AUPRC of over 0.39. This suggests that VAR methods are effective for data of the hour-long time intervals, which is the case for our GR data.

Yao [8] also compare Vector Autoregressions (VAR), Dynamic Bayesian Networks (DBN) and Mutual Information methods (MI) on a simulated hierarchical gene network, where the number of genes, 1000, exceeded the number of timepoints, 20. They find that top two methods are their own developed prior-knowledge VAR and the lasso-penalized VAR. They also found that the DBN was unable to handle data of that scale, and that the MI methods did not perform as well as VAR. Yao's simulation work supports our use of VAR in the GR data: first, the simulated data is similar to our setting, with the same high-dimensionality and short time series, and second, it is computationally tractable.

Two studies [9, 28] found that VAR did not perform as well as other approaches. Zou specifically compare DBN and VAR, finding that DBN outperforms VAR on a short timeseries of 5 genes and has higher accuracy on a true clock network. Meanwhile, Penfold [28] compare Vector Autoregressions (VAR), Ordinary Differential Equations (ODE), Dynamic Bayesian Networks (DBN), and Nonparametric Dynamical Systems (NDS). On the simulated DREAM4 100-gene network with 21 timepoints, they found that NDS outperformed DBN, which outperformed VAR/ODE. For the in-vivo networks (a 5-gene yeast network and 7-gene clock network), Penfold again find that NDS outperformed DBN, which outperformed VAR/ODE; all performed better than random. These findings suggest that for small-sized networks, NDS or DBN are more effective than VAR. The main drawback is that they can be quite computationally challenging and require sophisticated implementation.

3.3. Comparison of Vector Autoregression Methods

In this section, we compare *VAR-GEN* to previous approaches that use vector autoregressions to infer causal networks from gene expression time series. We chose vector autoregression due to its simplicity, flexibility, interpretability, and proven efficacy on simulated [8] and real data [11]. Our primary finding is that *VAR-GEN* is the only one of the VAR methods to simultaneously consider high-dimensional gene expression time series, tune hyperparameters, incorporate technical replicates, use a statistical null, and control false discovries.

Mukhophadyay first applied VAR to the HeLa Cell Cycle gene expression data to find genetic modules and pathways [29]. Tam also applied VAR to the HeLa dataset. They show that it is important to consider as many predictor genes as possible during the model fit, in order to prevent confounding causal fit [30]. They perform a two-step fit: first pairwise to choose the predictors, then a full predictor fit. However, both Mukhophadyay and Tam rely on the F-test for assessing significance. This test is undefined for our cases in which the dimensionality (2768 genes) exceeds the sample size (44 samples, one for each timepoint-replicate relation).

To handle the dimensionality problem, Lozano [12] and Shojaie [13] introduce the lasso penalty to regularize the VAR fit. Both papers emphasize methods for adapting the lasso penalty so that it chooses the optimal lag. The lag is the total number of previous timepoints considered in performing the fit. For example, in a lag of 3, fitting X_t will use timepoints $t-1, t-2, t-3$. However, lasso alone may not perform favorably on our data. Our data is likely to contain many correlated predictors, as the gene temporal profiles may often exhibit similar trends, such as net increase followed by decrease. Because Lasso will select only one of the several correlated predictors [31], it may find an overly sparse network. Thus, we wished to test the ridge and elastic net penalties as well, which were not addressed. Furthermore, neither method mentioned the problem of false discovery control for the network.

Opgen-Rhein introduce a James-Stein shrinkage estimator for the regularized VAR coefficients [32]. Their work further accounts for hyperparameter tuning, significance and false discovery control. We plan to implement their method as part of a follow-up study and compare the results of their method with our own formulation of hyperparameter tuning, significance, and false discovery control. Furthermore, we incorporate technical replicates unlike those authors.

The most recent and related work to our method is from Yao, called CGC2SPR. They implements the ridge, lasso, and elastic net penalties to handle the high dimensionality [8]. The method offers two main contributions: a Bayesian approach to the ridge penalty which uses external transcription factor binding information, and a local null method to assess significance. However, the method's statistical null, which is based on uniform random vectors instead of permuted values of the data, has a very different distribution from the original expression, and so may result in excessively liberal null rejections.

3.4. Experimental Prioritization Methods

A number of approaches tackle the latter problem of ranking specific genes in a directed network for future experimental interventions, including Bayesian [33, 34, 35, 36] and classical techniques [37, 38]. These experimental design methods allow the estimation of directed acyclic graphs (DAGs) using interventions, but do not include time series observations. Moreover, the acyclic restriction of DAGs is a limitation in modelling signaling pathways [39]. While an acyclic graph is interpretable, relaxing the acyclic requirement enables our approach to encode realistic but complex biological phenomena such as regulatory feedback loops [40, 41, 42].

4. GR gene expression data.

The glucocorticoid receptor (GR) regulates the transcription of a variety genes controlling the metabolism and immune response [3]. It is activated via binding to glucocorticoids; the bound complex then enters the nucleus and activates or represses the transcription of a variety of genes, both on its own and as bound to other proteins [2, 3, 7, 4, 6, 5]. The GR dataset seeks the comprehensive characterization of the genomic response to glucocorticoids through the measurement of changes in chromatin accessibility, epigenetic state, transcription factor binding, chromatin looping, and gene expression at time points across 12 hours of glucocorticoid treatment [5]. We extracted the temporal profiles of the genes from the GR expression data set across the 12 time points: $\{0, 0.5, 1-8, 10, 12\}$ hrs from the initial treatment.

We select the temporal profiles of those genes whose average expression across time were higher than 2 TPM and that passed the edgeR [43] criteria for differential expression. To measure average expression, we first averaged the gene's expression value per timepoint, and then took the average of those timepoint averages. For differential expression, we used the same method as in [5]: for each timepoint, we tested each gene's expression against its basal expression at an FDR threshold of 0.05, such that the resulting selected genes had expression different from the basal expression for at least one time point. These steps lead to a processed data set of 2767 differentially expressed genes. Finally, we added *NR3C1*, which encodes the GR transcription factor, even though it was not found to be differentially expressed at the FDR threshold of 0.05. In the end, we had a set of 2768 genes, which included 226 transcription factors. The resulting temporal profiles were further log transformed (base 2) and corrected for surrogate variables using SVAseq[44].

There were 4 replicates of the GR gene expression dataset across time. We split the data by replicates. All replicates besides replicate 1 had a measurement for each timepoint. Replicate 1 was missing timepoints 5 and 6 hrs, so we performed a linear imputation for these values in the log-transformed, surrogate corrected space.

We considered two normalization schemes for the temporal profiles: zero-mean unstandardized and zero-mean unit-variance. Zero-mean unstandardized centered each gene temporal profile to have zero-mean across time. Zero-mean unit variance centered each gene temporal profile to have zero-mean, and then standardized it to have unit variance. By gene temporal profile, we mean the gene's expression values across time for a single replicate.

5. Methods

5.1. Causal Inference Framework: *VAR-GEN*

VAR-GEN is a vector-autoregressive approach to causal inference from gene expression time series data. It is based on the principle of Granger Causality [17], in which a gene $\vec{g}$ is said to be causal for another gene g if using information from gene $\vec{g}$ significantly improves our ability to predict gene g.

VAR-GEN handles the high dimensionality of gene expression time series data via regularization, for example with the elastic net penalty. Hyperparameters are chosen via cross-validation. Significant causal edges are inferred based off of a permutation null, while controlling for false discoveries. We now discuss the details of this method.

5.1.1. Vector Autoregressive Model

We used a vector autoregressive model (VAR) with lag $L \in \{1,2\}$ to fit temporal gene expression profiles across multiple replicates of G genes over $t \in \{1,2,\dots T\}$ time points. Let $g \in \{1,2,\dots G\}$ index the set of genes and let $\mathbf{g}-$ represent the set of all genes excluding gene g, namely $\{1,2,\dots g-$

$1, g+1, \dots G\}$. Let $X_t^g = \{X_{t,1}^g, X_{t,2}^g, X_{t,3}^g, X_{t,4}^g\}^T$ be the 4×1 vector of gene expression levels of gene g across $R = 4$ replicates at time t. We modeled each gene g as

$$X_t^g = \sum_{l=1}^{L} m_l^g X_{t-l}^g + \sum_{l=1}^{L} \sum_{g' \in \mathbf{g}-} \beta_l^{g',g} X_{t-l}^{g'} + \varepsilon_t \tag{1}$$

where $\varepsilon_t \sim \mathcal{N}(0,1)$. In other words, the expression of each gene g is modelled as a linear function of its and other genes' L previous expression values, under independent Gaussian noise. In Equation 1, m_l^g represents the (scalar) effect size of gene g's l-th previous value, X_{t-l}^g, on its current value, X_t^g. $\beta_l^{g',g}$ represents the (scalar) effect size of the l-th previous value of gene $g' \neq g$, $X_{t-l}^{g'}$ on gene g's current value, X_t^g. μ is the the intercept term. One should note Equation 1 requires that $t > l$ for the l-th previous value, X_{t-l}^g, to exist.

One should note that the VAR assumes equally spaced timepoints. The time interval for the GR data ranges from 0.5 up to 2, and is therefore technically in violation of this requirement. A counterpoint to this is that the short intervals are concentrated at the beginning (hours $0, 0.5, 1$) where there is more likely to be activity, and the long intervals are at the end (hours $8, 10, 12$), where there is likely decreased activity. Thus despite the theoretical violation, treating the timepoints as equally spaced may not be entirely problematic [11].

To demonstrate how our model is fit in practice, we reformulate Equation 1 using matrix notation. Here, each row represents one timepoint per replicate. There are $T - L$ timepoints with $t > L$ and R replicates, so there are $R(T - L)$ samples, or rows, in total. Let $N = R(T - L)$.

Let

$$\mathbf{X}_t^g = \begin{bmatrix} X_{L,1}^g \\ \vdots \\ X_{L,R}^g \\ X_{L+1,1}^g \\ \vdots \\ X_{L+1,R}^g \\ \vdots \\ \vdots \\ X_{T,R}^g \end{bmatrix} \tag{2}$$

$\mathbf{X}_t^g$ is a $N \times 1$ vector. We can similarly write $\mathbf{X}_{t-l}^g$ which is the same vector, but which replacing each entry with its l-th previous value. Now combine the L lagged vectors of gene g, $[\mathbf{X}_{t-1}^g, \dots, \mathbf{X}_{t-L}^g]$ into $\mathbb{X}_{t-l}^g$, a $N \times L$ matrix of the L lagged values of gene g. Finally, let $\mathbf{m}_l^g$ be a $L \times 1$ vector of the L lagged coefficients.

$$\begin{aligned} \mathbb{X}_{t-l}^g &= [\mathbf{X}_{t-1}^g \dots \mathbf{X}_{t-L}^g] \\ \mathbf{m}_l &= \begin{bmatrix} m_1^g \\ \vdots \\ m_L^g \end{bmatrix} \end{aligned} \tag{3}$$

Next, let us formulate the component of Equation 1 involving the other genes g' in matrix notation. Let $\mathbb{X}_{t-l}^{\mathbf{g}-}$ be a $N \times L(G-1)$ predictor matrix of the genes $g' \neq g$. Each column is of form $\mathbf{X}_{t-l}^{g'}$. Note the number of columns is $L(G-1)$, because there are $G-1$ genes g' and for each gene g', there are L lagged values: $X_{t-1}^{g'}, \dots X_{t-L}^{g'}$.

$$\mathbb{X}_{t-l}^{\mathbf{g}-} = \begin{bmatrix} \mathbf{X}_{t-1}^1 & \dots & \mathbf{X}_{t-L}^1 & \mathbf{X}_{t-1}^2 & \dots & \dots & \mathbf{X}_{t-L}^G \end{bmatrix} \tag{4}$$

Let β_l be a $L(G-1)\times 1$ vector of the causal coefficients $\beta_l^{g',g}$ where $g' \neq g$:

$$\beta_l = \begin{bmatrix} \beta_1^{1,g} \\ \vdots \\ \beta_L^{1,g} \\ \beta_1^{2,g} \\ \vdots \\ \vdots \\ \beta_L^{G,g} \end{bmatrix} \tag{5}$$

We then seek to fit the model:

$$\mathbf{X}_t^g = \mathbb{X}_{t-l}^g \mathbf{m}_l + \mathbb{X}_{t-l}^{\mathbf{g}-} \beta_l + \varepsilon_t \tag{6}$$

where ε_t is a $N \times 1$ vector with each element $\varepsilon_{t,r} \sim N(0,1)$

To write in the most compact form, we can write

$$\mathbb{X}_{t-l}^{\mathbf{g}} = [\mathbb{X}_{t-l}^g \mathbb{X}_{t-l}^{g-}], \qquad \bar{\beta} = \begin{bmatrix} \mathbf{m}_l \\ \beta_l \end{bmatrix}$$

. Note that $\mathbb{X}_{t-l}^{\mathbf{g}}$ is a $N \times LG$ matrix and $\bar{\beta}$ is a $LG \times 1$ vector.

Thus in final form we would fit:

$$\mathbf{X}_t^g = \mathbb{X}_{t-l}^{\mathbf{g}} \bar{\beta} + \varepsilon_t \tag{7}$$

With these equations prepared, we are ready to describe the penalized fitting procedure.

5.1.2. Penalized Regression

The ordinary least squares estimator fits the causal coefficients as:

$$\hat{\beta} = \underset{\bar{\beta} \in \mathbb{R}^{LG}}{\arg\min} \|\mathbf{X}_t^g - \mathbb{X}_{t-l}^{\mathbf{g}} \bar{\beta}\|_2^2 \tag{8}$$

Here $\|\cdot\|_2$ represents the l_2-norm of a vector, i.e. the square root of the sum of the vector's squared coordinates.

However, we are in the high-dimensional setting: the dimension, LG exceeds the sample size $N = R(T-L)$. For example, if $L = 1$, the dimension $LG = 2768$ whereas our sample size $N = R(T-L) = 44$. As a result, the ordinary least squares estimator is undefined. We must instead resort to the use of penalized approaches such as LASSO (Least Absolute Shrinkage and Selection Operator) [45], elastic net [31], and ridge regression [46]. These are designed for $\hat{\beta}$ to be sparse (only a few nonzero coefficients) and shrunk (reduced in magnitude).

We discuss the elastic net penalty, which is a more general case of the ridge and lasso penalties. The elastic net fits the following objective:

$$\hat{\beta} = \underset{\bar{\beta} \in \mathbb{R}^{LG}}{\arg\min} \|\mathbf{X}_t^g - \mathbb{X}_{t-l}^{\mathbf{g}} \bar{\beta}\|_2^2 + \lambda(\alpha\|\bar{\beta}\|_1 + (1-\alpha)\|\bar{\beta}\|_2^2) \tag{9}$$

Here $\|\cdot\|_1$ represents the l_1-norm and $\|\cdot\|_2$ represents the l_2-norm.

By setting $\alpha = 1$ in the above equation 9, we obtain the Lasso objective function. By setting $\alpha = 0$ in the above, we obtain the Ridge objective function.

For the Elastic Net, we used the following ranges of hyperparameter values:

$\lambda \in \{10^{-4}, 10^{-4}, \ldots, 1\}$, $\alpha \in \{0.1, 0.3, \ldots, 0.9\}$. For Lasso, we used $\lambda \in \{10^{-5}, \ldots, 1\}$. For Ridge, when we used $\{10^{-5}, \ldots, 1\}$, we found that the the optimal value selected in some cases was the max 1. We thus expanded the range to $\{10^{-5}, \ldots, 10^{6}\}$ to ensure that we were not missing more optimal hyperparameters at larger values. At this point, the optimal λ was found to be 100 (Table 4).

Our choice of the hyperparameters λ and α greatly affects the amount of penalization we apply to our coefficients, and thus the quality of our fit. In the next section, we discuss our method of hyperparameter tuning, via cross-validation.

5.1.3. Hyperparameter Tuning

Hyperparameters were selected using leave-one-out cross-validation (LOOCV). The hyperparameter (or pair of hyperparameters, for elastic net) that minimizes the mean-squared error on the held-out datapoints is selected.

More specifically, we first fix a hyperparameter (λ, α). Then, for a given gene g and row index i, extract the i-the row of $\mathbf{X}_t^g$ and $\mathbb{X}_{t-l}^{\mathbf{g}}$. Refer to this extracted validation set as $\left(\mathbf{X}_t^g\right)_i$ and $(\mathbb{X}_{t-l}^{\mathbf{g}})_i$. The remaining data is the training set, $\left(\mathbf{X}_t^g\right)_{-i}$, $(\mathbb{X}_{t-l}^{\mathbf{g}})_{-i}$.

First we fit our coefficient $\hat{\beta}_{\lambda,\alpha}^{g,i}$ over the training set.

$$\hat{\beta}_{\lambda,\alpha}^{g,i} = \underset{\bar{\beta}\in\mathbb{R}^{LG}}{\arg\min} \|(\mathbf{X}_t^g)_{-i} - (\mathbb{X}_{t-l}^{\mathbf{g}})_{-i}\bar{\beta}\|_2^2 + \lambda(\alpha\|\bar{\beta}\|_1 + (1-\alpha)\|\bar{\beta}\|_2^2) \tag{10}$$

We then compute the fit's prediction error on the validation set, $\| \left(\mathbf{X}_t^g\right)_i - (\mathbb{X}_{t-l}^{\mathbf{g}})_i \hat{\beta}_{\lambda,\alpha}^{g,i}\|_2^2$.

We repeat the fit $\hat{\beta}_{\lambda,\alpha}^{g,i}$ and error for every row index i of $\mathbf{X}_t^g$ and for every gene g.

The mean held-out cross-validation error for (λ, α) is:

$$MSE(\lambda,\alpha) = \sum_{g=1}^{G} \sum_{i=1}^{R(T-L)} \frac{1}{GR(T-L)} \| \left(\mathbf{X}_t^g\right)_i - (\mathbb{X}_{t-l}^{\mathbf{g}})_i \hat{\beta}_{\lambda,\alpha}^{g,i}\|_2^2 \tag{11}$$

The (λ, α) which minimizes the error in Equation 11 is selected.

5.1.4. Statistical Null

Borrowing from the language of econometrics, a gene g is *Granger-caused* by a gene $g' \in \mathbf{g}-$ if using the past values of g' can improve our prediction of gene g, given the information from all remaining genes. In the language of vector autoregression, this means that for at least one lag l, $\beta_l^{g',g}$ is significantly different from 0 [17]. The null hypothesis, where the $\beta_l^{g',g}$ is equal to 0 is evaluated using a permutation test.

We explored two possible choices of nulls: the "global" and "local" null. The global null permutes every possible causal gene, i.e. all of $\mathbf{g}-$, while the local null only permutes the particular causal gene g'. In both cases, the model is fit again over the permuted dataset to generate a null distribution of coefficients, under the case where the causal time structure ought to be removed.

In particular, we first generated a single permuted dataset $\widetilde{X_t^{\mathbf{G}}}$. For each gene, we independently shuffled the expression values of each gene $g \in \{1,\ldots,G\}$ across time. This is done separately for distinct replicates.

For the global null, we wish to model the hypothesis of no causal relations, from any gene $g' \in \mathbf{g}-$, upon a given effect gene g. Thus, we uses the unpermuted values of the effect gene X_t^g and the permuted values of all other causal genes $g' \in \mathbf{g}-$, as $\widetilde{\mathbf{X}_t^{\mathbf{g}-}}$. Permuting the effect gene X_t^g would allow us to test the significance of the gene's self-interaction, but we are only interested in testing significance of the causal relations of other genes on the given gene. Thus we do not permute the effect gene g.

Null causal coefficients $\tilde{\beta}^{\mathbf{g}-}$ are then fit as

$$X_t^g \sim \mathscr{N}\left(\sum_{l=1}^{L} m_l^g X_{t-l}^g + \sum_{l=1}^{L} \sum_{g' \in \mathbf{g}-} \beta_l^{g',g} \widetilde{X_{t-l}^{g'}}, 1\right) \tag{12}$$

For the local null, we wish to model the case of no causal relation from gene $\vec{g}$ upon gene g. Thus, we only use the permuted values of the causal gene $\vec{g}$, $\widetilde{X_t}^{\vec{g}}$, and use the unpermuted values of the effect gene g (X_t^g) and of all remaining genes $\mathbf{X}_t^{\mathbf{g}}$.

The null causal coefficient $\tilde{\beta}^{\vec{g}}$ is then taken from its fit in:

$$X_t^g \sim \mathscr{N}\left(\sum_{l=1}^{L} m_l^g X_{t-l}^g + \sum_{l=1}^{L} \sum_{g' \neq g, \vec{g}} \beta_l^{g',g} X_{t-l}^{g'} + \sum_{l=1}^{L} \beta_l^{\vec{g},g} \widetilde{X_{t-l}^{\vec{g}}}, 1\right) \tag{13}$$

In our final analyses, we chose to use the global null. We compare the empirical performance of the global and local null (6.4.1). We find that the local null is difficult to reject because the permuted values often result in higher causal coefficients (Figure 3). For further discussion see 6.4.1.

5.1.5. False Discovery Rate Control

Similar to the null model, there we consider two alternatives for controlling the False Discovery Rate: a global approach and a local one. For a fixed lag, the global FDR controls the rate of insignificant causal relations across the whole inferred causal network, while the local FDR controls for the causal relations conditioning on the specific effect gene. The local FDR may be more appropriate when there is a stringent threshold for one effect gene; i.e. the null coefficients for the effect gene. Under the global FDR, this would lead to a stringent threshold for *all* effect genes, while in the local FDR, it would only lead to a stringent threshold for that specific coefficient.

Let $\beta_l^{\cdot,g}$ refer to the set of all lag-l causal coefficients for the effect gene g. Let $\beta_l^{\cdot,\cdot}$ refer to the set of all lag-l causal coefficients. Define $\tilde{\beta}_l^{\cdot,g}$ and $\tilde{\beta}_l^{\cdot,\cdot}$ analogously for the null coefficients.

We control the global FDR by fixing a lag $l \in \{1,\ldots,L\}$ and finding the threshold T_l such that

$$\frac{|\{|\widetilde{\beta}_l^{\cdot,\cdot}\|>T_l\}|}{|\{|\widetilde{\beta}_l^{\cdot,\cdot}|>T_l\}|+|\{|\beta_l^{\cdot,\cdot}|>T_l\}|}<0.05 \tag{14}$$

For each gene pair $(\vec{g},g)$, $\vec{g} \in \mathbf{g}-$, a causal link $\vec{g} \rightarrow g$ exists if for at least one of the lags $l \in \{1,\ldots,L\}$, $|\beta_l^{\vec{g},g}| > T_l$.

We control the local FDR by fixing a lag $l \in \{1,\ldots,L\}$ and an effect gene g and finding the threshold T_l^g such that

$$\frac{|\{|\widetilde{\beta}_l^{\cdot,g}>T_l^g\}|}{|\{|\widetilde{\beta}_l^{\cdot,g}|>T_l^g\}|+|\{|\beta_l^{\cdot,g}|>T_l^g\}|}<0.05 \tag{15}$$

For each gene pair $(\vec{g},g)$, $\vec{g} \in \mathbf{g}-$, a causal link $\vec{g} \rightarrow g$ exists if for at least one of the lags $l \in \{1,\ldots,L\}$, $|\beta_l^{\vec{g},g}| > T_l^g$. The only difference from the global FDR is that there is a threshold T_l^g specific to the effect gene g.

VAR-GEN is based on a local FDR calibration. We compared the empirical performance of the global and local FDR calibration (6.4.1). The local FDR calibration resulted in inferred networks with sizes of similar orders of magnitude across different settings. Meanwhile, the global FDR resulted in several overly sparse networks. As before, this was likely due to the following phenomenon: a high null coefficient for one effect gene g would result in a stringent threshold across all effect genes, not merely g.

5.2. Contextual Causal Influence Score

Though the writing in this section is my own, CCI was primarily conceived and implemented by Bianca Dumitrascu. I discuss it here for completeness. - Jonathan Lu

A primary goal of the GR study is to enable development of treatments with the same beneficial effects on immunity as the glucocorticoid, but without the metabolic side effects leading to diabetes or obesity [5, 2, 6, 4]. Disentangling these effects requires experimental perturbations upon the original GR network, for example suppression of a gene's expression via knockout. However, it is crucial to prioritize which of the thousands of genes is to be perturbed, given the limited experimental budget. One attempt is to simply select those genes whose immediate neighbors are highly metabolic. However, this is insufficient as an influential gene may regulate several genes that are not themselves metabolic, but each of which regulates metabolic genes. It is therefore necessary to consider the network context in measuring a gene's influence. Toward this end, we introduce a contextual causal influence score based off of PageRank [15]. The PageRank score quantifies node centrality (connectedness) in the context of overall network topology. They are most commonly known through their use in the Google search engines for ranking webpages. To our knowledge, such methods have not previously been used for GRNs.

We define a gene to have a high contextual causal influence score if its removal from the network minimizes the metabolic genes' network centrality, while also minimizing the change in the immune genes' network centrality. These intermediaries can be interpreted as belonging to a "cut set" that separates immune genes from metabolic genes, and are interesting candidates for experimental followup.

In particular, we considered the network $\mathscr{G}$ obtained through the *VAR-GEN* approach, and the set $\mathscr{M}$ of nodes corresponding to genes annotated as metabolic, and $\mathscr{I}$ the set corresponding to genes annotated as immune. Absolute values of the network edges' coefficients are normalized to create a Markov adjacency matrix of the network $\mathscr{G}$ (for each node, the weights of its out-going edges sum to 1). We then compute the Pagerank $pg(X)$, which is X's PageRank in network $\mathscr{G}$. Next, we perform the "perturbation": for any gene A, we considered the graph obtained by removing all the edges containing the node corresponding to A and re-normalizing the remaining weights appropriately ($\mathscr{G}^{A-}$). We then re-compute the PageRank $pg(X^{A-})$ over the perturbed network. A damping factor of 0.85 is used to compute the PageRank scores [15]. Finally, we define the contextual causal influence score of gene A as

$$CCI(A) = -\sum_{X \in \mathscr{I}} \|pg(X) - pg(X^{-A})\| - \sum_{X \in \mathscr{M}} \|pg(X^{-A})\|. \tag{16}$$

The term $\sum_{X \in \mathscr{I}} \|pg(X) - pg(X^{-A})\|$ represents the total absolute change in the PageRank of immune genes from perturbing A. The term $\sum_{X \in \mathscr{M}} \|pg(X^{-A})\|$ represents the PageRank of the metabolic genes after perturbing A. Thus, a gene A with a high CCI would, when perturbed, minimize both the change in the immune genes' network centrality, and the metabolic genes' absolute network centrality. This seeks to capture the biological goal of minimizing the change in immune effect, while also minimizing the metabolic genes.

5.3. Gene Annotations

In the Results Section 6, we perform a variety of analyses based on gene annotations. This section describes those annotations.

We used four main classifications in our analysis of genes: Immune, Metabolic, Transcription Factor, and Direct Target of Glucocorticoid Receptor (GR-Direct). We now describe the classification method.

Immune genes were called using two primary sources. The first is the Gene Ontology annotation "Immune" (*GO:0002376*) [47]. To emphasize experimentally verified annotations, we only used the evidence codes EXP, IDA, IGI, IMP, IPI, IC, TAS. The second is the Gene Ontology Consortium's curated, ranked list of immune-related genes based off of multiple databases and experimental evidence [48]. For the GO annotation, We selected all those genes with score greater than or equal to 7. This resulted in 616 immune genes overall, and 109 immune genes in our list of 2768 genes (differentially expressed + GR).

Metabolic genes were called using two primary sources. The first is the Gene Ontology annotation "carbohydrate metabolic process" *GO:0005975* [47]. We only used the evidence codes EXP, IDA, IGI, IMP, IPI, IC, TAS. The second is the Gene Set Enrichment Analysis' curated list of metabolic-related genes [49]. We searched only among those with experimental evidence: the Canonical, KEGG, BIOCARTA, and Reactome pathways. We used the following 4 search queries: "gluconeogenesis OR (glucose AND metabolism) OR glycolysis", "lipid AND metabolism", "Diabetes", "Obesity". We chose these queries to ensure we covered genes implicated in both metabolic processes and disorders, which may be affected by *GR*. Combining these, we found 544 metabolic genes overall and 120 in our list of 2768 genes. Finally, 65 genes were both immune and metabolic overall, and 12 were both immune and metabolic in our geneset.

Transcription Factors were called using the Bioguo database of Human Transcription Factors [50]. There were 1463 factors overall and 226 present in our list of 2768 genes.

GR direct targets were called based on the binding data of GR [5]. These genes were found to be up-regulated at timepoints 0.5, 1, or 2 hours after initial treatment with dexamethasone, and had GR binding within 10 kb of the transcription start site. Up-regulation was called based on a differential expression test between a gene's expression at one timepoint with the basal timepoint at FDR 0.01. This calling method included several positive controls such as *DUSP1* [51] and *PER1* [52]. The method resulted in 111 genes. All 111 of these genes were included in our set of differentially expressed genes by definition.

6. Results

6.1. Preliminary Analysis: Simulations

The goal of this section was to test the Vector Autoregression on a simplified gene dynamical system to ensure it could recover the true underlying relations. We found that lags of 2 and greater were able to do so. This informed our decision to use the lag 2 network in the analysis of the GR data (Section 6.5).

6.2. Simulated Dynamical System

We applied the Vector Autoregression method to 2 simulated gene dynamical systems as a first test of whether it could recover true causal relations. We based our system off the system used by Mugler [53], which they successfully use to study the competence circuit in Bacillus subtilis. Their system models the copy number of each gene via differential equations that represent transcription and degradation. Activation/repression dynamics are further modelled using Hill kinetic equations.

The general form of these equations is in Equation 17. In this example, A is activated by B, and both compete to be degraded by an enzyme. As the authors in [53] show, these equations can produced stimulation and oscillation circuit dynamics; similar dynamics are evident in our GR data. An explanation of the parameters is in Table 1.

$$\frac{dA}{dt} = (\alpha_A + \beta_A \frac{1}{1+(K_A/[B])^{n_A}} - \lambda_A - \frac{\delta_A}{1+[A]/\gamma_A+[B]/\gamma_B})[A] \tag{17}$$

We tested Vector Autoregression on two dynamical systems: $Z \to X \to Y$ and $Z \to X, Z \to Y$ (Figure 2 A-B), where $\to$ represents an activation relationship. Our motivation was to ensure that VAR could detect the basic presence and absence of regulatory relationships. The complete equations for these are listed in Tables 2 and 3. As the authors do in [53] did, we included a gene A involved in a feedback loop with Z, to introduce oscillation dynamics to the circuit.

The simulation data was gathered using StochPy [54]. After simulating the system for 20000 seconds, we sampled the data by averaging all the expression values over each 30-minute interval, thus producing a 111 timepoint dataset (Figure 2 C-D).

Parameter	Description
α_A	Basal expression rate for A
β_A	Maximum activated expression rate of A
K_A	Concentration of A producing half the maximum expression rate.
n_A	Hill Coefficient describing the binding cooperativity.
λ_A	Basal degradation rate constant for A
δ_A	Maximal degradation rate of A due to enzyme
γ_A	Michaelis-Menten constant of A's binding with enzyme
γ_B	Michaelis-Menten constant of B's binding with enzyme

Table 1: ***Parameters in Simulated Gene Dynamical System.***

Reaction	Rate Constant	Additional Parameters	Notes
$Pol_Z \rightarrow Z$	$\alpha_Z + \beta_Z \frac{[Z]^{n_Z}}{K_Z^{n_Z} + [Z]^{n_Z}}$	$\alpha_Z = 0.9 \text{ hr}^{-1}$, $\beta_Z = 0.03, n_Z = 2, K_Z = 20$	Z self-activates
$Z \rightarrow pool$	$\frac{\delta_Z}{1+[Z]/\gamma_Z+[A]/\gamma_A} + \lambda_Z$	$\delta_Z = 0.001, \lambda_Z = 0.0001, \gamma_Z = 100, \gamma_A = 1$	Z & A competitively degraded
$Pol_X \rightarrow X$	$\alpha_X + \beta_X \frac{[Z]^{n_X}}{K_X^{n_X} + [Z]^{n_X}}$	$\alpha_X = 0, \beta_X = 0.003, n_X = 5, K_X = 20$	Z activates X
$X \rightarrow pool$	λ_X	$\lambda_X = 0.0001$	X linearly degraded
$Pol_Y \rightarrow Y$	$\alpha_Y + \beta_Y \frac{[X]^{n_Y}}{K_Y^{n_Y} + [X]^{n_Y}}$	$\alpha_Y = 0, \beta_Y = 0.003, n_Y = 5, K_Y = 10$	X activates Y
$Y \rightarrow pool$	λ_Y	$\lambda_Y = 0.0001$	Y linearly degraded
$Pol_A \rightarrow A$	$\alpha_A + \beta_A \frac{1}{1+([Z]/K_A)^{n_A}}$	$\alpha_A = 0, \beta_A = 0.003, n_A = 5, K_A = 3.3$	Z represses A.
$A \rightarrow pool$	$\frac{\delta_A}{1+[Z]/\gamma_Z+[A]/\gamma_A} + \lambda_A$	$\delta_A = 0.001, \lambda_A = 0.0001, \gamma_Z = 100, \gamma_A = 1$	Z & A competitively degraded

Table 2: ***Complete reaction dynamics of*** $Z \rightarrow X \rightarrow Y$ ***circuit.*** **Here, gene A adds oscillation to the circuit by being repressed by Z and competitively degraded with Z. The same model as in [53] was used.**

Reaction	Rate Constant	Additional Parameters	Notes
$Pol_Z \rightarrow Z$	$\alpha_Z + \beta_Z \frac{[Z]^{n_Z}}{K_Z^{n_Z} + [Z]^{n_Z}}$	$\alpha_Z = 0.9 \text{ hr}^{-1}$, $\beta_Z = 0.03, n_Z = 2, K_Z = 20$	Z activates its own expression
$Z \rightarrow pool$	$\frac{\delta_Z}{1+[Z]/\gamma_Z+[A]/\gamma_A} + \lambda_Z$	$\delta_Z = 0.001, \lambda_Z = 0.0001, \gamma_Z = 100, \gamma_A = 1$	Degradation of Z
$Pol_X \rightarrow X$	$\alpha_X + \beta_X \frac{[Z]^{n_X}}{K_X^{n_X} + [Z]^{n_X}}$	$\alpha_X = 0, \beta_X = 0.003, n_X = 5, K_X = 20$	Z activates X
$X \rightarrow pool$	λ_X	$\lambda_X = 0.0001$	Linear degradation of X
$Pol_Y \rightarrow Y$	$\alpha_Y + \beta_Y \frac{[Z]^{n_Y}}{K_Y^{n_Y} + [Z]^{n_Y}}$	$\alpha_Y = 0, \beta_Y = 0.003, n_Y = 5, K_Y = 10$	Z activates Y
$Y \rightarrow pool$ $Pol_A \rightarrow A$	λ_Y $\alpha_A + \beta_A \frac{1}{1+([Z]/K_A)^{n_A}}$	$\lambda_Y = 0.0001$ $\alpha_A = 0, \beta_A = 0.003, n_A = 5, K_A = 3.3$	Linear degradation of Y Z represses A.
$A \rightarrow pool$	$\frac{\delta_A}{1+[Z]/\gamma_Z+[A]/\gamma_A} + \lambda_A$	$\delta_A = 0.001, \lambda_A = 0.0001, \gamma_Z = 100, \gamma_A = 1$	Degradation of A

Table 3: ***Complete reaction dynamics of*** $Z \rightarrow X, Z \rightarrow Y$ ***circuit.*** **The only difference from $Z \rightarrow X \rightarrow Y$ in Table 2 is that Z directly activates Y instead of via X. Here, gene A adds oscillation to the circuit by being repressed by Z and competitively degraded with Z. The same model as in [53] was used.**

6.3. Performance of Causal Tests on Simulated System

We performed pairwise VAR tests for all possible cause-effect pairs of the set $\{Z, X, Y, A\}$. Significance of the Causal relation was evaluated by performing an F-test, assessing if the model of the

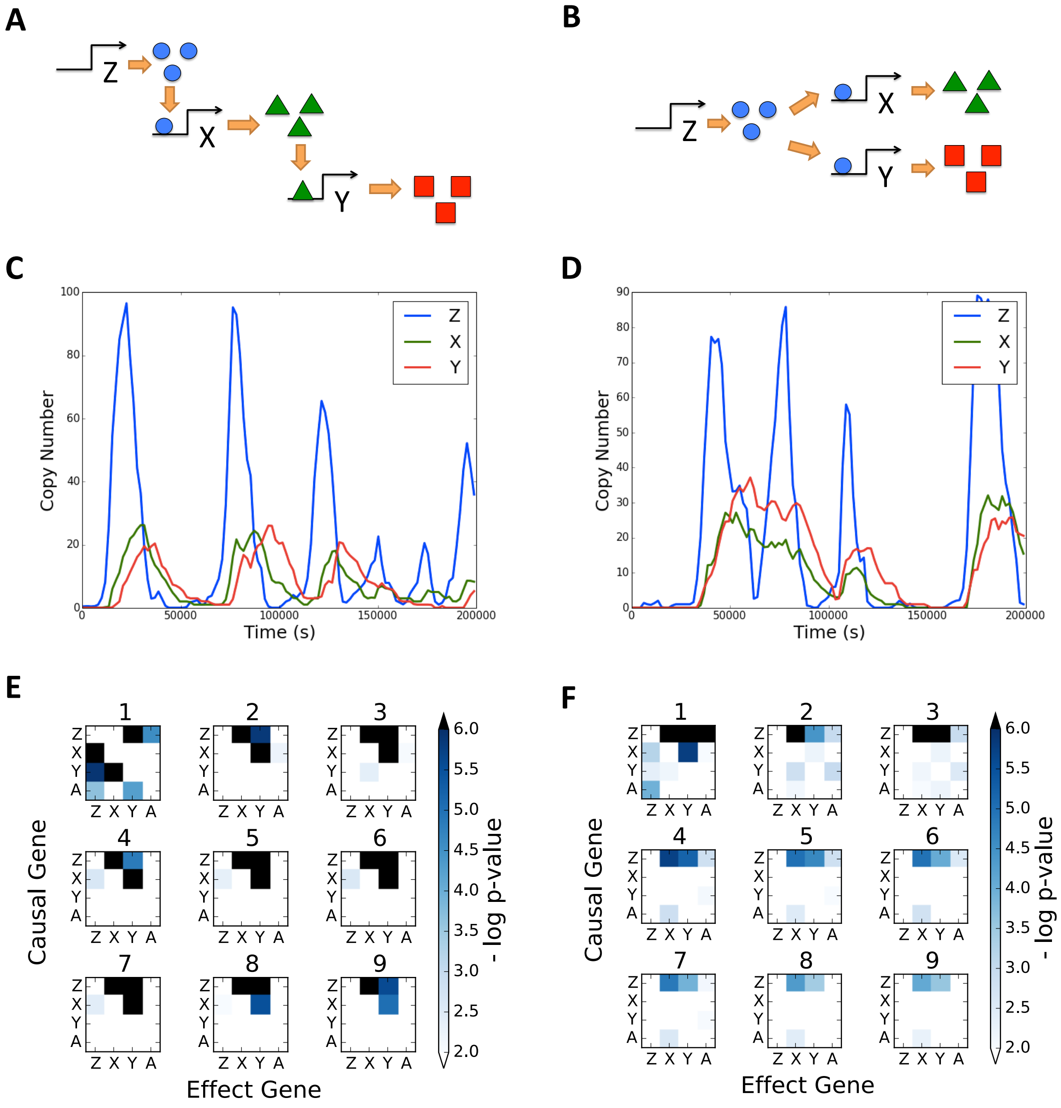

Figure 2: ***Causal Test Performance on Simulated Gene Dynamical system.*** **(A-B) Schematic of regulatory relationships for $Z \rightarrow X \rightarrow Y$ and $Z \rightarrow X, Z \rightarrow Y$, respectively. Transcribed proteins (colored shapes) are shown as promotors for subsequent genes. (C-D) Simulated time series for $Z \rightarrow X \rightarrow Y$ and $Z \rightarrow X, Z \rightarrow Y$, respectively. 111 time-points were taken at intervals of 1800 seconds. (E-F) P-value matrices of causal VAR fits between genes in the $Z \rightarrow X \rightarrow Y$ and $Z \rightarrow X, Z \rightarrow Y$ systems, respectively. Each matrix corresponds with the fit from a chosen lag K in $\{1, \ldots, 9\}$. The darkness of the box is the significance of the causal relation, $\log_{10}$ p-value of the causal fit from performing an F-test.**

effect gene with the causal gene as a predictor had a significantly better fit than the model of the effect gene without the causal gene as a predictor.

We found that the test was able to accurately recover the underlying dynamics for lags greater than 1 (Figure 2 E-F) under an FDR of 0.05. In $Z \rightarrow X \rightarrow Y$, Z is detected as causal for X and Y, and X is detected as causal for Y (Figure 2 E), In $Z \rightarrow X, Z \rightarrow Y$, Z is indeed detected as causal for X and for Y, and there is correctly no causal relation between X and Y (Figure 2 F).

6.4. Preliminary Analysis: GR Networks

6.4.1. Comparison Run

We quantified the effects of different parameter settings from our pipeline on the resulting directed networks. We ran the data through 8 different parameter settings, representing changes to normalization, null, and FDR thresholds (Table 4). For each setting, we used three types of regularization—lasso, elastic net, and ridge—and lags 1 and 2. We compared the network results to determine the appropriate parameter settings. Our main finding was that use of the global null and local FDR setting provided the greatest power and stability of the edge numbers.

Our first finding was that the local null was very difficult to reject, resulting in networks of size 0 (Table 4 E-H). This was because the local null coefficients had very large magnitude (Figure 3, top)) compared with the true coefficients. This may be partially explained by the fact that the many of 2768 gene temporal profiles are highly correlated. Removal of one of the predictor genes would not result in a drastic loss of predictive power because the gene temporal profiles were so similar. On the other hand, the randomized set of gene expression values could produce a unique gene temporal profile which could fit some aspect of the target gene better than the original values. This could explain the larger null coefficients. In contrast, under the global null, the randomized distribution had a narrower spread than the true coefficient distribution (Figure 3, bottom). The decreased spread allows normal coefficients to be declared significant.

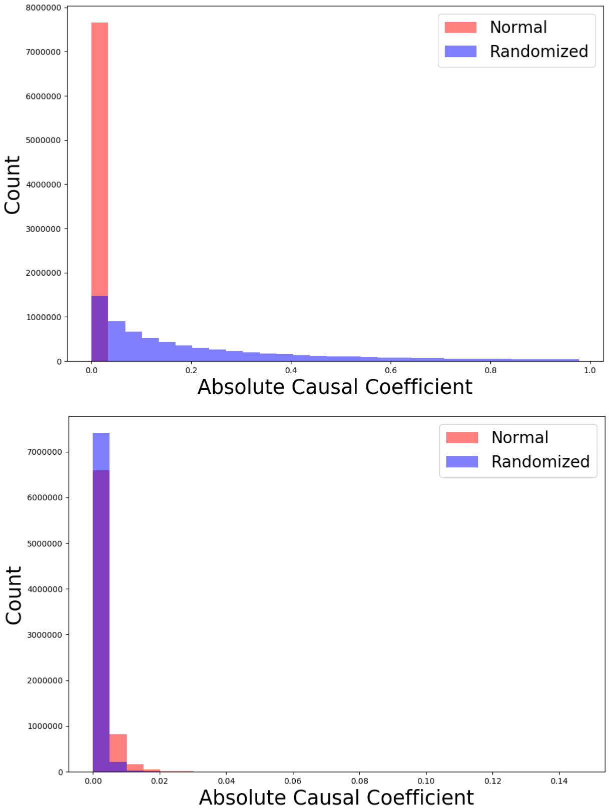

Figure 3: Causal coefficient distribution under local and global nulls. Top: under the local null, the randomized coefficients have much larger spread than normal coefficients, preventing any normal coefficients from being declared significant. Bottom: under the global null, the randomized coefficients have much narrower spread, allowing normal coefficients to be declared significant.

To select an FDR type, we found that using a global FDR was not robust, as a single permuted coefficient with high value could prevent a majority of true coefficients from being declared significant. These size issues are evident in the global null, global FDR network for zero-mean unit-variance normalization (Table 4 A). Here, the elastic net lag-1 network is found to be sparser than the lasso lag-1 network (170 and 841 edges, respectively), which runs contrary to existing theory that the lasso should be the sparsest network [31]. A similar problem was in the global null, global FDR network for zero-mean unstandardized normalization (Table 4 C) in which the ridge-1 network has only 82 edges, two orders of magnitude below the elastic net, despite not encouraging explicit sparsity in the coefficients [31].

The local FDR was less vulnerable to significance testing effects. The networks had on the order of $10^3 - 10^5$ edges, which is proportional to the number of genes 2,768, as compared with the 10^2 edges found by the global FDR approach (Table 4 B, D). Therefore, we chose the parameter setting of global null, local FDR for analysis of the GR networks.

6.4.2. Selecting Causal Networks for Downstream Analysis

After the comparison run, we chose four final networks to analyze in more detail. These networks shared the global null and local FDR based on our findings from the previous section. The network each used the elastic net penalty out of the lasso, elastic net, and ridge penalties. We chose the elastic net penalty over lasso because it allowed the selection of correlated predictors [31] whereas Lasso will only select one. As many genes are highly correlated over the short time series and the goal is to uncover possible interactions, we chose to use the elastic net penalty over lasso. Finally, we chose the elastic net over ridge because of the elastic net penalty fitted sparser, parsimonious models. The elastic net networks found on the order of $10,000$ edges, compared with the ridge on the order of $100,000$ edges.

Each network used one of the 4 unique combinations of two run options: normalization and lag number. The two options for normalization include: 1) gene temporal profiles were standardized to zero-mean unit-variance, or 2) gene temporal profiles were centered at zero-mean (zero-mean unstandardized). The lag number is the number of previous time points included to infer causal effects in the model; it was set to either 1 or 2 time points because of the sparse sampling of these observations across time.

We shall refer to the networks using the following abbreviations:

- Zero-mean Standardized, elastic net lag 1 (ZS-1)
- Zero-mean Standardized, elastic net lag 2 (ZS-2)
- Zero-mean Unstandardized, elastic net lag 1 (ZU-1)
- Zero-mean Unstandardized, elastic net lag 2 (ZU-2)

6.4.3. Individual Network Analysis

We ran *VAR-GEN* on the networks and computed several basic statistics of our networks (Table 5, the frequency of various edge types (Table 6), and the odds ratios of those edge types (Table 7,

Batch	Network	Test	Lag	Normalization	Null	FDR	Hyper	Sig. Edges	Sig. Edges (%)	Causal Genes	Effect Genes
A	lasso-1	lasso	1	zero-mean unit-variance	global	global	1.00E-04	841	0.01	311	705
	enet-1	enet	1	zero-mean unit-variance	global	global	(1E-03, 1E-01)	170	0	78	157
	ridge-1	ridge	1	zero-mean unit-variance	global	global	1.00E-04	502440	6.56	2358	2768
	lasso-2	lasso	2	zero-mean unit-variance	global	global	1.00E-02	9	0	9	9
	enet-2	enet	2	zero-mean unit-variance	global	global	(1E-01, 1E-01)	1494	0.02	302	962
	ridge-2	ridge	2	zero-mean unit-variance	global	global	1.00E+02	465810	6.08	2258	2749
B	lasso-1	lasso	1	zero-mean unit-variance	global	local	1E-04	7882	0.1	828	2444
	enet-1	**enet**	**1**	**zero-mean unit-variance**	**global**	**local**	**(1E-03, 1E-01)**	**10546**	**0.14**	**758**	**2195**
	ridge-1	ridge	1	zero-mean unit-variance	global	local	1E-04	549627	7.18	2463	2744
	lasso-2	lasso	2	zero-mean unit-variance	global	local	1E-02	8976	0.12	1063	2577
	enet-2	**enet**	**2**	**zero-mean unit-variance**	**global**	**local**	**(1E-01, 1E-01)**	**13879**	**0.18**	**860**	**2525**
	ridge-2	ridge	2	zero-mean unit-variance	global	local	1E+02	638575	8.34	2680	2678
C	lasso-1	lasso	1	zero-mean unstandardized	global	global	1E-04	12171	0.16	711	2254
	enet-1	enet	1	zero-mean unstandardized	global	global	(1E-03, 1E-01)	1479	0.02	149	539
	ridge-1	ridge	1	zero-mean unstandardized	global	global	1E-06	82	0.0	21	45
	lasso-2	lasso	2	zero-mean unstandardized	global	global	1E-02	295	0.0	66	124
	enet-2	enet	2	zero-mean unstandardized	global	global	(1E-03, 1E-01)	41	0.0	14	26
	ridge-2	ridge	2	zero-mean unstandardized	global	global	1E+00	43	0.0	7	27
D	lasso-1	lasso	1	zero-mean unstandardized	global	local	1E-04	27330	0.36	867	2759
	enet-1	**enet**	**1**	**zero-mean unstandardized**	**global**	**local**	**(1E-03, 1E-01)**	**23025**	**0.3**	**708**	**2721**
	ridge-1	ridge	1	zero-mean unstandardized	global	local	1E-06	233931	3.05	2653	2643
	lasso-2	lasso	2	zero-mean unstandardized	global	local	1E-02	4587	0.06	184	2443
	enet-2	**enet**	**2**	**zero-mean unstandardized**	**global**	**local**	**(1E-03, 1E-01)**	**27781**	**0.36**	**617**	**2744**
	ridge-2	ridge	2	zero-mean unstandardized	global	local	1E+00	158340	2.07	2502	2627
E	lasso-1	lasso	1	zero-mean unit-variance	local	global	1E-04	0	0.0	0	0
	ridge-1	ridge	1	zero-mean unit-variance	local	global	1E-04	0	0.0	0	0
	lasso-2	lasso	2	zero-mean unit-variance	local	global	1E-02	0	0.0	0	0
	enet-2	enet	2	zero-mean unit-variance	local	global	(1E-01, 1E-01)	0	0.0	0	0
	ridge-2	ridge	2	zero-mean unit-variance	local	global	1E+02	0	0.0	0	0
F	lasso-1	lasso	1	zero-mean unit-variance	local	local	1E-04	1220	0.02	418	323
	ridge-1	ridge	1	zero-mean unit-variance	local	local	1E-04	6457	0.08	2724	4
	lasso-2	lasso	2	zero-mean unit-variance	local	local	1E-02	2639	0.03	663	1184
	enet-2	enet	2	zero-mean unit-variance	local	local	(1E-01, 1E-01)	180	0.0	90	136
	ridge-2	ridge	2	zero-mean unit-variance	local	local	1E+02	4	0.0	4	4
G	lasso-1	lasso	1	zero-mean unstandardized	local	global	1E-04	0	0.0	0	0
	enet-1	enet	1	zero-mean unstandardized	local	global	(1E-03, 1E-01)	0	0.0	0	0
	ridge-1	ridge	1	zero-mean unstandardized	local	global	1E-06	0	0.0	0	0
	lasso-2	lasso	2	zero-mean unstandardized	local	global	1E-02	0	0.0	0	0
	ridge-2	ridge	2	zero-mean unstandardized	local	global	1E+00	0	0.0	0	0
H	lasso-1	lasso	1	zero-mean unstandardized	local	local	1E-04	1046	0.01	327	561
	enet-1	enet	1	zero-mean unstandardized	local	local	(1E-03, 1E-01)	169	0.0	71	130
	ridge-1	ridge	1	zero-mean unstandardized	local	local	1E-06	2300	0.03	2254	3
	lasso-2	lasso	2	zero-mean unstandardized	local	local	1E-02	2859	0.04	101	1680
	ridge-2	ridge	2	zero-mean unstandardized	local	local	1E+00	0	0.0	0	0

Table 4: *Network run results under various parameter settings.* Each of the 8 batches uses a unique combination of normalization (zero-mean unit-variance or zero-mean unstandardized), null (global or local), and FDR (global or local). Each batches has 6 runs, with the lasso, elastic net, or ridge penalties, and lag 1 or 2.
The batch of local null networks are very spare (E-H). Note that the networks that use global FDR (A, C) sometimes have the elastic net and ridge networks more sparse than the lasso network of corresponding setting; this is inconsistent with the trend that the lasso should be the sparsest penalty. The bolded rows were those analyzed in-depth in Section 6.4.2.

Figure 4) to detect enrichment of edge types. We now performing the following analysis of each individual network:

1. the presence and enrichment of Transcription Factor-Causal Edges (edges where the causal gene is a Transcription Factor),Transcription Factor-Effect Edges, Immune-Causal Edges, Immune-Effect Edges, Metabolic-Causal Edges, and Metabolic-Effect Edges,
2. transcription factors with the highest out-degree,
3. immune and metabolic genes with the highest number of TF Causes and Effects.

The motivation for analysis 1 is to get a sense of the overall biological function of network edges. The motivation for analysis 2 is to find transcription factors, which have causal roles in regulation, which may play broad roles in the network. The motivation for analysis 3 is to find immune and metabolic gnes that may be highly affected in the networks as dwonstream targets.

Network	Test	Lag	Normalization	Null	FDR	Hyper	Sig. Edges	Sig. Edges (%)	Causal Genes	Effect Genes
ZS-1	enet	1	zero-mean unit-variance	global	local	(1E-03, 1E-01)	10546	0.14	758	2195
ZS-2	enet	2	zero-mean unit-variance	global	local	(1E-01, 1E-01)	13879	0.18	860	2525
ZU-1	enet	1	zero-mean unstandardized	global	local	(1E-03, 1E-01)	23025	0.3	708	2721
ZU-2	enet	2	zero-mean unstandardized	global	local	(1E-03, 1E-01)	27781	0.36	617	2744

Table 5: *Run Summary for GR Causal Networks.* Parameter settings and basic network statistics for the network runs are listed here. All networks use an FDR threshold of 0.05. "Hyper" refers to the hyperparameter used in the penalty. For Elastic net, hyper is of form (λ, α) where λ controls regularization and α is the l1-ratio.

Network	Edges	TF-Causal Edges	TF-Effect Edges	Immune-Causal Edges	Immune-Effect Edges	Metabolic-Causal Edges	Metabolic-Effect Edges
ZS-1	10546	828 (7.9%)	898 (8.5%)	422 (4%)	415 (3.9%)	372 (3.5%)	435 (4.1%)
ZS-2	13879	1136 (8.2%)	1096 (7.9%)	367 (2.6%)	545 (3.9%)	636 (4.6%)	629 (4.5%)
ZS-1	23025	1677 (7.3%)	1900 (8.3%)	2410 (10.5%)	834 (3.6%)	1309 (5.7%)	1001 (4.3%)
ZS-2	27781	1931 (7%)	2393 (8.6%)	2119 (7.6%)	1047 (3.8%)	2271 (8.2%)	1211 (4.4%)

Table 6: *Frequency of Edge Types in Causal Networks.* Percentages are calculated relative to the individual network's total edges.

Network	Edges	TF-Causal Edges	TF-Effect Edges	Immune-Causal Edges	Immune-Effect Edges	Metabolic-Causal Edges	Metabolic-Effect Edges
ZS-1	10546	0.96	1.05	1.02	1	0.81	0.95
ZS-2	13879	1	0.96	0.66	1	1.06	1.05
ZU-1	23025	0.88	1.01	**2.87**	0.92	1.33	1
ZU-2	27781	0.84	1.06	**2.02**	0.95	**1.97**	1.01

Table 7: *Odds Ratios of Edge Type in Causal Networks.* Odds ratios significantly larger than 1 **indicate enrichment for that edge type within the network. Note enrichment of immune-causal edges in ZU-1 and ZU-2, and of metabolic-causal edges in ZU-2. This corresponds with Figure 4.**

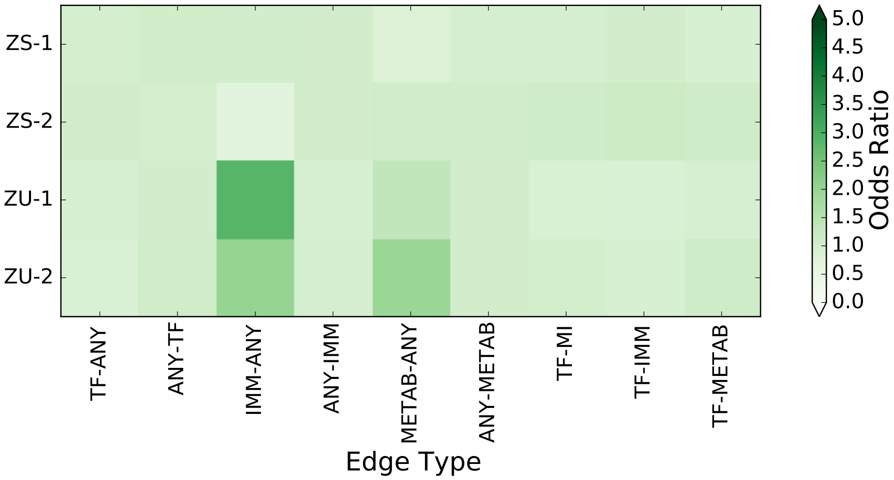

Figure 4: Odds Ratio enrichment for Edge types across the 4 **GR causal networks. "TF", "IMM', "METAB", "ANY" refer to transcription factor, immune, metabolic, and any gene, respectively. For example, "TF-ANY" refers to edges with transcription factor as cause. A baseline odds ratio is** 1.0**; larger (darker) indicates enrichment. Note enrichment of immune-causal edges ("IMM-ANY") in ZU-1 and ZU-2, and of metabolic-causal edges ("METAB-ANY") in ZU-2.**

We find that the zero-mean standardized lag 1 (ZS-1) causal network contains 10546 edges, among which 828(7.9%) have a transcription factor as cause and 898(8.5%) have a transcription factor as effect. We found that the causal transcription factor with the greatest out-degree of 139 was *RXRB*, which is highly related to metabolism. *RXRB* encodes a member of a family of the retinoid X receptor nuclear receptors, several of which interact with the *PPAR*α protein to trigger transcription of genes involved in gluconeogenesis and lipid metabolism [55]. Two other high out-degree TFs were *PRDM1* and *POU5F1*, which had out-degrees of 65 and 28, respectively. Both are *GR* direct targets (Section 5.3, and *PRDM1* is involved in inducing the maturation of B cells [56]. The network had a modest number of edges that involved a metabolic/immune gene as cause or effect, ranging from 372(3.5%) to 435(4.1%). The network had no enrichment of any edge type, with odds ratios very close to 1. Among immune genes, *BIRC3* had the highest number of Transcription Factor causes, 3. *BIRC3* encodes *cIAP2*, an inhibitor of apoptosis found to be under-expressed in T cells from multiple sclrosis patients [57, 58]. It is also a direct target of *GR*. Among metabolic genes, *RXRA* was one of the genes with the highest number of Transcription Factor causes 2. It is itself a transcription factor whose interaction with *PPARα* regulates expression of genes involved in lipid metabolism [55]

We find that the zero-mean unit-variance lag 2 (ZS-2) causal network contains 13879 edges, among which 1136(8.2%) have a transcription factor as cause and 1096(7.9%) have a transcription factor as effect. *ETS1* had the highest out-degree among transcription factors, 86. *ETS1* is found to be highly immune-related . It serves as an important transcription factor controlling the expression of cytokine and chemokine genes, regulating the differentiation of a variety of leukocytes, and is involved in the B-Cell Receptor Signaling Pathway [59, 55]. There were modest number of edges involving a metabolic/immune gene as cause or effect, ranging from 367(2.6%) to 636(4.6%). Among immune genes, *CXCL2* had the highest number of Transcription Factor causes, 4. *CXCL2* helps to encode chemokines, a family of proteins that are involved in inflammatory and immune response [60, 55]. Among metabolic genes, *ALDH7A1* had the highest number of Transcription Factor causes, 3. This gene is involved in glycolysis/gluconeogenesis via its role in metabolizing toxic aldehydes generated by oxidative processes such as alcohol metabolism [61, 62]. The network had no enrichment for any edge type, with odds ratios very close to 1.

We find that the zero-mean unstandardized lag 1 (ZU-1) causal network contains 23025 edges, among which 1677(7.3%) have a transcription factor as cause and 1900(8.3%) have a transcription factor as effect. The TF with highest out-degree was *POU5F1*, with out-degree 263. Recall *POU5F1* is likely a direct target of GR [5]; thus, its causal relations may pay a key role in the overall glucocorticoid response. Interestingly, in the network we found enrichment of edges with an immune gene as cause: 2410(11%), with odds-ratio 2.9. The immune gene with highest out-degree was *CEACAM6* at 577. 25 of *CEACAM6*'s effects are Immune-related. *CEACAM6* encodes a type of carcinoembryonic antigen. Such proteins are involved in cell adhesion, and are widely used as tumor markers in serum immunoassay determinations of carcinoma [63, 64]. There were a modest number of edges with immune effects or metabolic causes/effects 834(3.6%) to 1309(5.7%). *RXRA* was among the metabolic genes with highest number of Transcription Factor Causes, 4. Again, *RXRA* is itself a Transcription Factor, whose interaction with *PPAR*α regulates expression of genes involved in lipid metabolism [55]. Among immune genes, *TCIRG1* had the highest number of Transcription Factor causes, 4. *TCIRG1* encodes *TIRC7*, which has a negative inhibitory role on immune and inflammatory response [65, 66].

Finally, we find that the zero-mean unstandardized lag 2 (ZU-2) causal network contains 27781 edges, among which 1931(7%) have a transcription factor as cause and 2393(8.6%) have a transcription factor as effect. The TF with highest out-degree was *ZNF114*, with out-degree 462. We were unable to find annotations of *ZNF114* related to our immune/metabolic effects of interest. Interestingly, there are enrichment of edges with immune causes: 2119(7.6%), with odds ratio 2.0. There was also enrichment of edges with metabolic causes: 2271(8.2%), with odds ratio 1.97. The immune gene with highest out-degree was *OLR1*, with out-degree 398. *OLR1* is a downstream target of the *PPAR* signaling pathway which affect lipid metabolism [55]. *OLR1* is specifically involved in Fatty Acid Transport as a receptor of oxidized low-density lipoprotein; moreover it also plays a role in inflammation [67, 68, 55]. The metabolic gene with highest out-degree was *ANGPTL4* with out-degree 615. *ANGPTL4* is also a direct target of GR. *ANGPTL4* is a target of the PPAR receptors and regulates glucose and lipid metabolism [69, 70, 55]. There were a modest number of edges with immune effects or metabolic effects: 1047(3.8%) to 1211(4.4%). Among immune genes, *FOS* had the highest number of Transcription Factor Causes, 5. *FOS* is a member of the glucocorticoid pathway and encodes a component of the transcription factor complex $AP-1$, which regulates cell differentiation and proliferation in response to a variety of stimuli such as growth factors and cytokines [71]. *ANGPTL4* was one of the metabolic genes with the most Transcription

Factor Causes, 4.

6.4.4. Cross-Network Analysis

We then performed analysis across the 4 networks. We evaluated the intersections and odds ratios between networks. We then discuss edges of biological interest that were preserved across the networks.

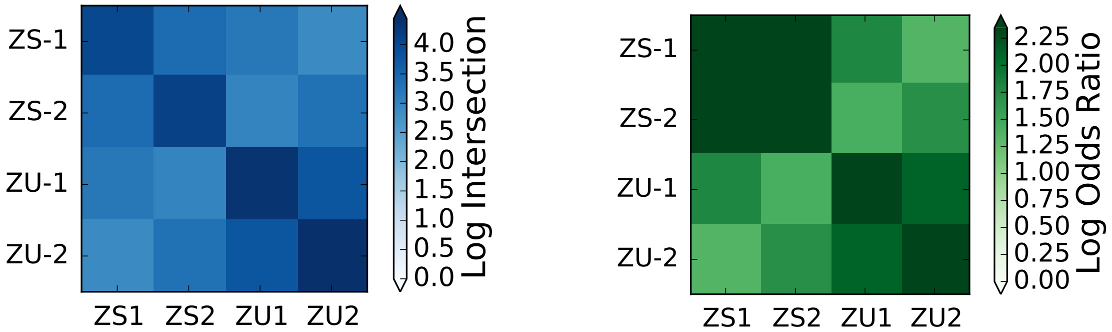

Figure 5: Overlap between the 4 GR causal networks. Left: $\log_{10}$ values of the number of intersecting edges among the networks. Right: $\log_{10}$ values of the odds ratio of two networks' intersection. There is substantial overlap between every pair of networks.

There are high intersection and odds ratios among the different parameter settings of the network runs (Figure 5).

The highest intersections were between the networks with the same normalization. In particular, the zero-mean Standardized Lag 1 (ZS-1) and 2 (ZS-2) networks share 2598 edges with an odds ratio of 221. The zero-mean Unstandardized Lag 1 (ZU-1) and 2 (ZU-2) networks share 6203 edges with an odds ratio of 130. Across the normalizations, the number of shared edges ranged from 784 to 6203, with odds ratios from 22.7 to 61.8.

We next investigated several of the top causal interactions preserved across these 4 networks (Table 8).

We found the transcription factor *POU5F1* to be strongly causal to the cytokine *CXCL1*; *CXCL1*'s expression is dramatically reduced following increase of the *POU5F1* expression level (not shown). This interaction suggests that *POU5F1* may play a key causal role in inflammation reduction through the transcriptional repression of *CXCL1*. Several consistent interactions between transcription factors and metabolism-related genes were also observed. In particular, the TF *TEF* is inferred as causal to *PLD1*, a phospholipase which regulates cytosolic lipid droplet formation [72], an interaction likely related to GC response-induced lipolysis or fat breakdown. An intriguing finding is the interaction between the TF *PRDM1* and the circadian clock related gene *CRY2*. However, further research is required to investigate the mechanism through which the beta-interferon repressor *PRDM1* interacts with the transcriptional repressor *CRY2*, a main circadian clock regulator which plays a key role in glucose and lipid metabolism modulation [73, 74].

Similarly, several causal relations between immune and metabolic genes illustrate the substantial overlap between metabolic and immune pathways following the GC response. Among the top such interactions we found the genes *IFITM2*, which mediates innate immune response to a variety of viruses, and *ALDH7A1*, a gene needed to metabolize toxic aldehydes and involved in stress response [75]. Moreover, *IFITM3*, an interferon-induced membrane protein, was found to cause the *PRKAB2* protein-coding gene, a regulatory subunit of the AMP-activated protein kinase, which is heavily involved in the regulation of intracellular and whole-body energy homeostasis [76, 77]. *OLR1*, an endothelial receptor for oxidized low-density lipoprotein, was found to be causal to the change of expression of the gene *IL1R1*, a key receptor involved in innate immune response and inflammation [78], suggesting a possible feedback mechanism. *CFD*, which encodes the important adipokine adipsin needed to maintain the function of pancreatic beta cells which control normal insulin storage and secretion [79, 80], is causal for Major Histocompatibility Complex 1 ($HLA-C$), which presents antigens to cytotoxic T cells [81].

Causal Relation	**Cause Annotation**	**Target Annotation**	**Coefficient (average)**	**Frequency (across networks)**	**Networks**
POU5F1 → *CXCL1*	Encodes component of Oct4, a Transcription Factor complex regulating Embryonic Stem Cell pluripotency pathways [82]	Chemokine that controls the early stage of neutrophile recruitment during tissue inflammation [60]	-0.21	2	ZS-1, ZU-1
IFITM3 → *PRKAB2*	An interferon-inducible trans-membrane protein, conferring resistance to influenza A H1N1, West Nile, and dengue viruses [83]	Encodes β subunit 2 of AMPK, which regulates intracellular and whole-body energy homeostasis [76, 77]	-0.09	4	ZS-1, ZS-2, ZU-1, ZU-2
CXCL2 → *ARX*	Chemokine that controls the early stage of neutrophile recruitment during tissue inflammation [60]	Homeobox Transcription Factor that regulates migration of interneurons in the brain; Mutations lead to a variety of X-linked intellectual disorders [84, 85]	-0.08	4	ZS-1, ZS-2, ZU-1, ZU-2
OLR1 → *IL1R1*	An endothelial receptor for oxidized low-density lipoprotein, it is involved with inflammation, antigen cross-presentation, and atherosclerosis [67, 68]	Interleukin-1 receptor type 1, involved in the innate immune response and inflammation [78]	-0.08	3	ZS-2, ZU-1, ZU-2
PRDM1 → *CRY2*	Encodes Blimp1, Transcription Factor that induces maturation of B cells into Ab-secreting plasma cells through regulation of multiple tunable pathways [56]	A key transcriptional regulator of the mammalian circadian clock [86, 87]	0.08	2	ZS-1, ZU-1
IFITM2 → *ALDH7A1*	An interferon-inducible trans-membrane protein, which mediates innate immune response to influenza A H1N1 virus, West Nile virus, and dengue virus [83]	Metabolizes toxic aldehydes generated by oxidative processes such as alcohol metabolism; defends against hyperosmotic stress [62, 61]	-0.07	4	ZS-1, ZS-2, ZU-1, ZU-2
TEF → *PLD1*	Transcription Factor that promotes cell survival and inhibits cell growth by downregulating the expression of the βc chain [88]	Phospoholipase that regulates cytosolic lipid droplet formation. [72]	0.07	3	ZS-1, ZS-2, ZU-1
CFD → *HLA-C*	Encodes adipsin, a serine protease and adipokine that maintains function of pancreatic beta cells responsible for insulin storage and secretion [79, 80]	A Major Histocompatability Complex 1 molecule that presents antigens to cytotoxic T cells and is required for regulation of natural killer cell function [81]	0.06	3	ZS-1, ZU-1, ZU-2
AOC2 → *FANCD2*	Encodes an amine oxidase that is upregulated during adipocyte maturation and may be involved in regulation of growth, differentiation, and apoptosis [89, 90]	The protein is involved in DNA damage repair through interaction with BRCA1 and may be needed to prevent chromosome instability [91, 92]	-0.06	4	ZS-1, ZS-2, ZU-1, ZU-2
FGG → *C1RL*	Forms the gamma chain of fibrinogen, a crucial protein for blood clot formation [93]	Serine protease that mediates cleavage of a proform of haptoglobin, which regulates T-cell mediated immune responses [94, 95]	0.06	3	ZS-1, ZS-2, ZU-1

Table 8: Top causal relations preserved across the 4 GR Causal Networks.

6.5. Analysis: GR Causal Network.

6.5.1. Biological Analysis

The work and writing in this section was a collaborative effort between me, Bianca Dumitrascu, and Prof. Barbara Engelhardt. My primarily role was to gather statistics and annotations, and create the tables. I wrote initial paragraphs on the annotations, though these were substantially revised by Bianca and Prof. Engelhardt. Figures 6 - 7 were created by Bianca.

Of the 4 possible networks that we analyzed at first (Section 6.4.2-6.4.4, we chose the zero-mean unstandardized lag 2 (ZU-2) network for more in-depth analysis. We chose lag 2 since it can capture both lag 1 and lag 2 relationships; furthermore, in previous simulations (not shown), lag 2 successfully recovered the a higher proportion of true relationships. We chose unstandardized because the variance of gene temporal profiles had a wide range, from almost constant profiles to drastic increases and decreases. Normalizing these profiles over-represents weak causal effects of the genes with higher variance.

We applied *VAR-GEN* to the GC-mediated expression responses to infer a causal network (Figure 6A). The network contains 27,781 directed edges and 2,747 nodes representing distinct genes. Transcription factors (TFs) represent 8.2% of nodes; TFs are causal in 1,931 pairwise interactions (7%) and effects in 2,393 interactions (8.6%). The inferred network follows a power law-like distribution on edge degree that has been observed in similar biological networks [96, 97], with the in-degree distribution appearing approximately Gaussian (Figure 6D,E,F).

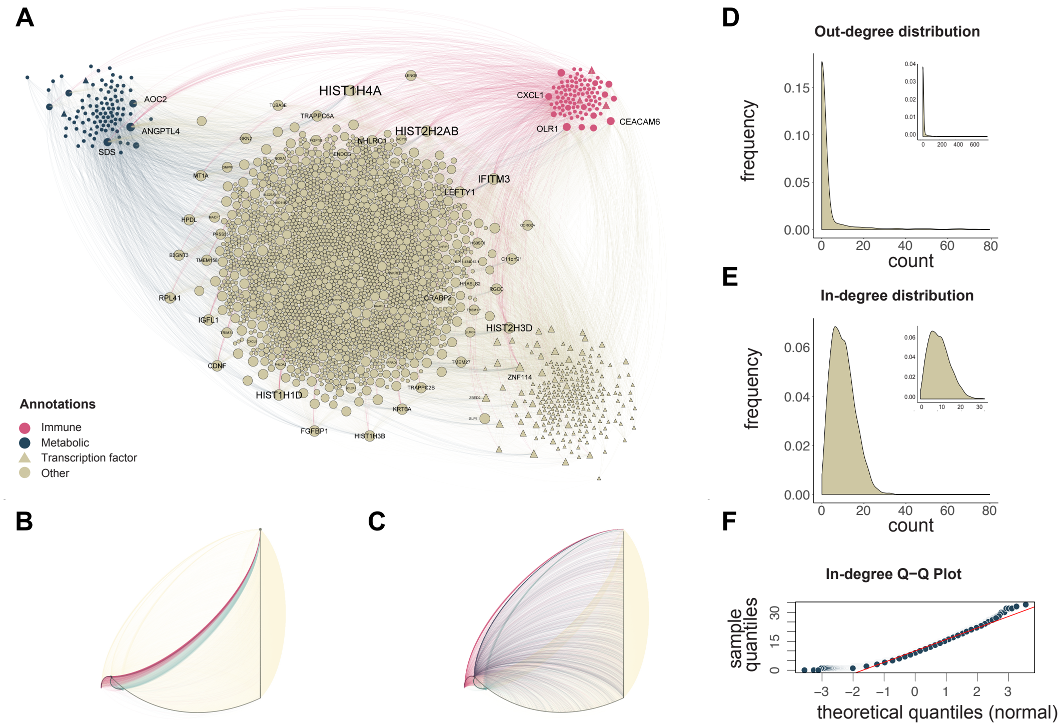

Figure 6: Network summary of the regulatory dynamics of the GC response A: Directed gene network illustrating the interplay between immune-related genes (red) and metabolic-related genes (blue); B-C: Radial plot views of the *VAR-GEN* output with edge color reflecting the annotation of the effect node (B) and causal node (C); D: histogram illustrating the distribution of in-degree across nodes in the network; E: histogram illustrating the distribution of out-degrees across the inferred network; F: a quantile-quantile plot showing the quantiles of the empirical distribution of out-degree and quantiles of a normal distribution.

Using Gene Set Enrichment Analysis (GSEA) and Gene Ontology annotations, we labeled genes in the network with immune or metabolic function [49, 47]. We found 97 and 108 genes with immune and metabolic functions, respectively, and 12 genes associated with both immune and metabolic functions (Figure 6B,C). The annotated immune and metabolic genes, despite making up less than 10% of nodes, were twice as likely to be causes rather than effects in the network (Table 7, Figure 4), suggesting that the diverse transcriptional effects of GC exposure have yet to be characterized.

Among the most well connected genes in our network are *OLR1* and *ANGPTL4*. (Figure 7C,D). *OLR1* is an immune-related gene involved in fatty acid transport [55] with 398 predicted effect genes. Although annotated as an immune-related gene, *OLR1* (Figure 7C) is also involved in metabolic processes as a downstream target of the *PPAR* signaling pathway, which affects lipid metabolism [55]. Similarly, *ANGPTL4* is a metabolism-related gene with 615 predicted effect genes. *ANGPTL4* was previously found to be a direct target of the GC receptor [70]. Two genes with predicted effects from *ANGPTL4*, *ACSL1* and *SLC2A14*, are related to fatty-acid metabolism [98] and glucose transport [99], respectively. Together with *OLR1*, *ANGPTL4* is a target of the *PPAR* receptors [100, 101].

Next, we investigated the network genes with metabolic effects. We found that *SOCS3*, *CXCL1*, and *CCL2* were among the network genes with the largest number of effects that were metabolism-related. *SOCS3* (Figure 7A) is a protein-coding gene involved in the inhibition of cytokine signal transduction that has been associated with the GC response pathways [102]. Among its targets, the multifunctional cytokines *IL11* and *EDN1* have both been functionally associated with glycemic control and signaling pathways related to diabetes [103]. Moreover, *CXCL1* is a chemokine with a role in controlling the early stage of neutrophil recruitment during tissue inflammation [104, 105]. This gene has been linked to a variety of autoimmune diseases, including Type 1 Diabetes Mellitus (T1DM) [106, 105], and Type 2 Diabetes Mellitus (T2DM) [107]. Similarly, *CCL2* is a chemokine identified as a probable mediator between immune function and metabolic dysfunctions [108, 109]. The most well-connected transcription factor in the network, *ZNF114*, was predicted to have 25 metabolic effects, including *SLC7A11*, a gene involved in cysteine-glutamate transport [110]; its overexpression is associated with altered cellular metabolism in glioma cells [111]. Together, these results suggest a complex interplay between immune response pathways and metabolic pathways.

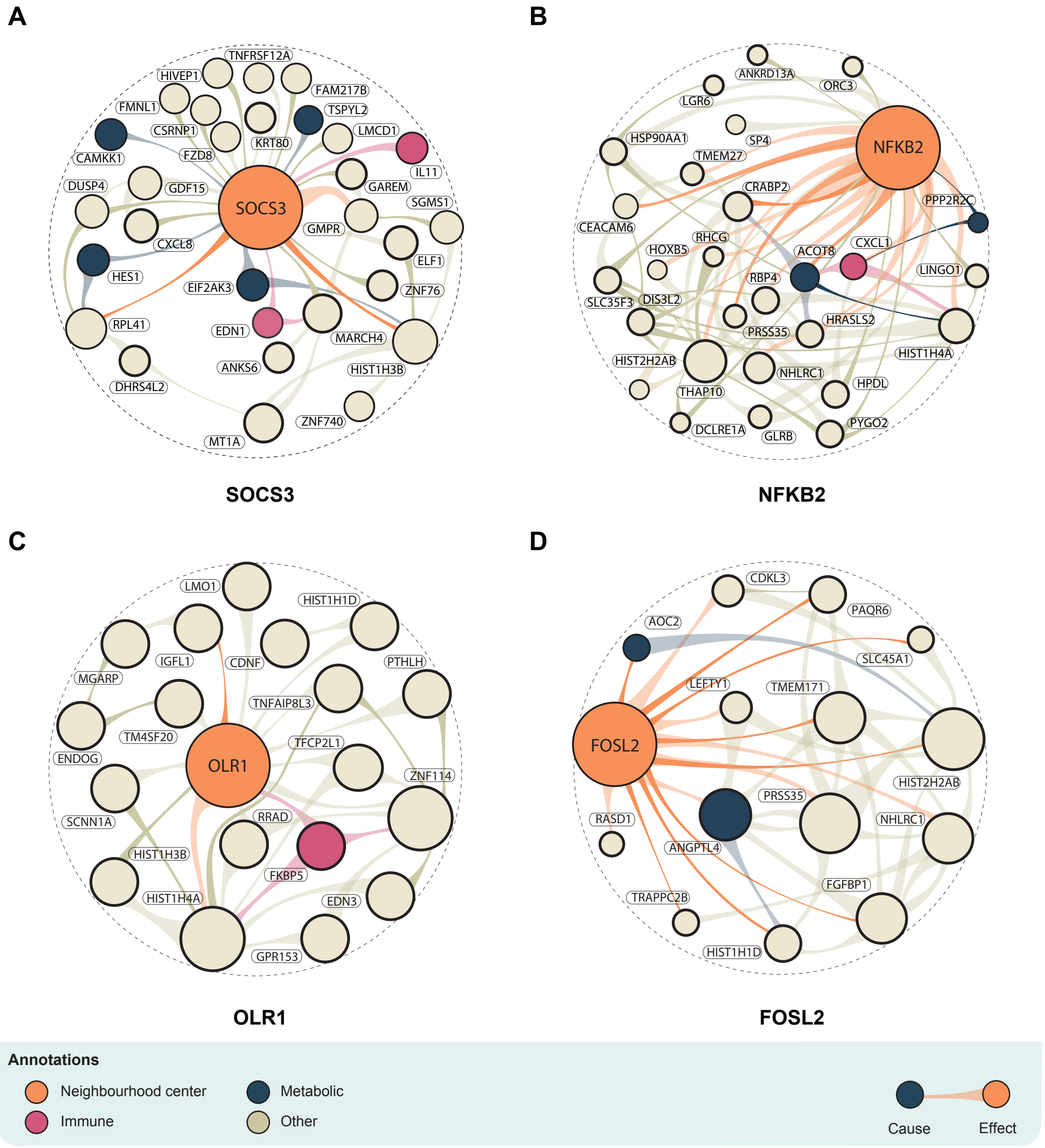

Figure 7: Gene-specific regulatory subnetworks. Top causes and effects (causal coefficient > 0.05) of well-connected genes in the network. A: *SOCS3*; B: predicted direct GR target gene *NFKB2*; C: *OLR1*; and D: *FOSL2*.

Causal relation	Cause annotation	Effect annotation
NHLRC1 → *CALB2*	Codes for the E3 protein-ubiquitin ligase malin, which downregulates glycogen synthesis, and is regulated by *AMPK* [112]	A calcium-binding protein primarily expressed in neurons [113]
ANGPTL4 → *SDS*	Regulates lipid and glucose metabolism, and is a direct target of the glucocorticoid receptor [69, 70]	Involved in metabolism of amino acids such as serine and glycine [114]
CRABP2 → *HIST1H1D*	Involved in the metabolism and transport of retinoic acid [115]	Codes for a member of the histone H1 family [116]
ANGPTL4 → *CILP*	Regulates lipid and glucose metabolism, and is a direct target of the glucocorticoid receptor [69, 70]	An essential cartilage matrix protein that is associated with diseases such as lumbar disc disease [117]
ANGPTL4 → *SLC2A14*	Regulates lipid and glucose metabolism, and is a direct target of the glucocorticoid receptor [69, 70]	Encodes the *GLUT14*, a major glucose transporter [99]
ANGPTL4 → *ACSL1*	Regulates lipid and glucose metabolism, and is a direct target of the glucocorticoid receptor [69, 70]	Encodes an acyl-CoA synthetase that plays a key role in lipid biosynthesis and fatty acid degradation [98]
ZNF114 → *SLC7A11*	A zinc finger transcription factor [50]	A cysteine-glutamate antiporter; its dysfunction leads to a variety of central nervous system disorders [110]

Table 9: Top causal relationships predicted by *VAR-GEN* in the GC response data. Relationships with causal coefficients ranking in the top 1% were selected. The second and third column describe the functional annotations of the cause gene and effect gene, respectively.

We further investigated the causal neighbourhoods of immune- and metabolism-related genes (Table 10). We were particularly interested in *NFKB2* (Figure 7B) and *FOSL2* (Figure 7D), which are direct targets to the glucocorticoid receptor, GR [5] . The gene *ACOT8* is predicted to effect *NFKB2* and metabolic genes *ENO3* and *PRKAB2*. *ACOT8*, *ENO3*, and *PRKAB2* play critical roles in the regulation of energy metabolism and homeostasis. *ACOT8* catalyzes the hydrolysis of acyl-CoAs into fatty acids and the coenzyme A, an important step for lipid metabolism. Meanwhile, *ENO3* catalyzes the interconversion of 2-phosphoglycerate and phospholenolpyruvate, a critical function for in both glycolysis and glucoeneogenesis. *PRKAB2* encodes a component of the *AMPK* protein kinase, which plays a broad role in regulating both intracellular and intercellular metabolism. For example, phosphorylation of *AMPK* triggers production of energy-generation pathways such as glycolysis and fatty acid oxidation [76]. *PRKAB2* is an effect of *IFITM3*, which plays a role in interferon (IFN) signaling; this interaction suggests another specific mechanism by which GC exposure triggers a response to inflammation and a downstream metabolic response. *ANGPTL4*, *AOC2*, and the glucose transporter *SLC45A1* all are causal to transcription factor *FOSL2*. No effects of *FOSL2* were predicted by the network. This may be due to the mechanisms through which *FOSL2* regulates transcription: as a leucine zipper protein, it is likely to dimerize with proteins from other transcription factor families, leading to nonlinear dependencies that are not be captured by *VAR-GEN*.

The network predicts several relationships involving genes that are involved in cytokine signaling. *SOCS3* (Figure 7A), a cytokine-signalling suppressor [118], is predicted to be causal to *IL11*, an interleukin related to clinical conditions including T1DM and T2DM [119]. *CD14* is predicted to affect another cytokine, *SOCS1*. *CD14* is a surface co-receptor for lipopolysaccharides, which enables immune response in various organisms [120].

Causal Relation	**Relation Type**	**Cause Annotation**	**Target Annotation**
ACOT8 → *ENO3*	*M* → *M*	An acyl-CoA thioesterase involved in fatty acid oxidation [121]	An enolase isoenzyme, a key catalyst for the glycolytic pathway in muscle [122]
ACOT8 → *PRKAB2*	*M* → *M*	An acyl-CoA thioesterase involved in fatty acid oxidation [121]	Encodes β subunit 2 of AMPK, which regulates intracellular and whole-body energy homeostasis [76, 77]
SDS → *CUL7*	*M* → *M*	Involved in metabolism of amino acids such as serine and glycine [114]	A ligase that targets Insulin receptor substrate 1 for degradation [123]
CD14 → *SOCS1*	*I* → *I*	A co-receptor for detection of bacterial lipopolysaccharide, [124]	A suppressor of cytokine signaling as well as proliferative signaling; its expression is induced by inflammation [125, 126]
SOCS3 → *IL11*	*I* → *I*	A suppressor of cytokine signaling; its expression is induced by inflammation [118, 126]	An important interleukin that also functions as an adipogenesis inhibitory factor [127]
OLR1 → *IL1R1*	*I* → *I*	An endothelial receptor for oxidized low-density lipoprotein, it is involved with inflammation, antigen cross-presentation, and atherosclerosis [67, 68]	Interleukin-1 receptor type 1, involved in the innate immune response and inflammation [78]
CFD → *HLA-C*	*I* → *I*	Encodes adipsin, a serine protease and adipokine that maintains function of pancreatic beta cells responsible for insulin storage and secretion [79, 80]	A Major Histocompatability Complex 1 molecule required for regulation of natural killer cell function [81]
IFITM2 → *ALDH7A1*	*I* → *M*	An interferon-inducible trans-membrane protein, which mediates innate immune response to influenza A H1N1 virus, West Nile virus, and dengue virus [83]	Metabolizes toxic aldehydes generated by oxidative processes such as alcohol metabolism; defends against hyperosmotic stress [62, 61]
ANGPTL4 → *IRAK2*	*M* → *I*	Regulates lipid and glucose metabolism, and is a direct target of the glucocorticoid receptor [69, 70]	A kinase for the interleukin-1-receptor, which mediates the widespread inflammatory effects of IL-1 signaling [128]
OLR1 → *TRAF2*	*I/M* → *I/M*	An endothelial receptor for oxidized low-density lipoprotein, it is involved with inflammation, antigen cross-presentation, and atherosclerosis [67, 68]	Mediates TNF-directed activation of NF-kappa B and immune response; enhances glucagon signaling to promote gluconeogenesis [129, 130]

Table 10: Immune and metabolic causal interactions predicted by *VAR-GEN* in the GC response data. We highlight metabolic (*M*) and immune-related genes (*I*).

6.5.2. Validation

The work behind this section was primarily a collaborative effort between Brian Jo, Bianca

Dumitrascu, and Prof. Barbara Engelhardt. I summarize the work below. The complete section, written by those authors, can be found in the Appendix, Section A.2.

We validated our network using association tests on external gene expression data. Briefly, if there exists a true cause-effect relationship $X \rightarrow Y$, the expression of X and Y should be dependent. We should be able to detect this relationship by performing an association test. Broadly, a Single Nucleotide Polymorphism (SNP) is identified within 20 kb of the the causal gene X. This SNP is tested for association with the effect gene Y's expression values; a significantly associated SNP is referred to as an expression Quantitative Trait Locus (eQTL). If for $X \rightarrow Y$, we find a significant association between a cis-SNP of X and the expression of Y, we consider the edge validated. A full description can be found in the Appendix, Section A.1.

We performed this test for our network edges in expression data of a variety of tissues, including lung; this was from the Genotype Tissue Expression (GTEx) Consortium [131]. We found that an enrichment of low p-values of our edges compared to permuted values, which suggests that the network edges have a higher likelihood of finding real siginal (Figure 8 A). We also found that 280 edges validated in the held-out tissues at an FDR threshold of 0.2, and in particular 81 edges were validated from the lung tissues, . Considering a contingency table where edges validated in lung are one group and edges validated in all other tissues are another group, a Fisher's exact test showed enrichment of network edges validated in lung samples. This enrichment reflects the match of the A549 cells used in the GR data with the GTEx lung samples.

6.5.3. Experimental Prioritization

The work in this section was a collaborative effort between myself, Bianca Dumitrascu and Barbara Engelhardt. My main contribution was in compiling the annotation information in Table 11, which laid the factual basis for the paragraph. Bianca ran the CCI score on the network.

When we apply the *CCI* score to the inferred network, we find that the top ranked genes are enriched for genes with DNA metabolic processes, DNA replication, and DNA conformation change processes (GSEA analysis, Bonferroni-corrected p-value $p \leq 0.02$). The top ranking gene, the insulin receptor *IRS2*, is a cytoplasmic signaling molecule known to mediate the effect of insulin on downstream targets, which, when absent, has been linked to diabetic conditions in mice [132, 133]. Several top ranked genes, including E3 ubiquitin-protein ligase *CBLB*, the krüppel-like factor *KLF6*, the connective tissue growth factor *CTGF*, the heme oxygenase *HMOX1*, and the glucose transporter *SLC2A1*, have been routinely associated with metabolic disorders such as wound healing [134, 135] and obesity [136]. Moreover, the top ranked genes are enriched for those involved in carbohydrate metabolic processes, including the glucosidase *ATHL1*, the galactosyltransferase *B4GALT4*, the cholesterol transcriptional repressor *KLB*, lactate dehydrogenase A *LDHA*, and two glycoproteins *MUC1* and *TM4SF4*. Additional research is required to further characterize and evaluate the mechanisms through which these top genes modulate the metabolic response to glucocorticoids while circumventing control of the immune response.

Gene	**Gene Annotation**
IRS2	An essential signaling intermediate from insulin receptors to metabolic and mitogenic pathways; its dysregulation in beta cells results in obesity and diabetes [137]
TAF5L	Encodes a component of the PCAF histone acetylase complex, which regulates transcription, cell cycle progression, and differentiation [138]
PHKA1	A phosphorylase kinase which is a key regulatory enzyme of glycogen metabolism [139]
MUC1	A mucin that lines epithetlial cells and protects the body from pathogen infection [140]
TM4SF4	A transmembrane glycoprotein that mediates signal transduction and can regulate cell density-dependent proliferation [141]
ATF5	A member of the CREB family of transcription factors, involved in intracellular signal transduction [142]
SLC2A1	Encodes the GLUT1, a major glucose transporter [143]
CTGF	A connective-tissue growth factor involved in type 1 diabetes nephropathy; its inactivation leads to defects in pancreatic beta cell proliferation [144]
COL4A3	Encodes a component of type IV collagen, which forms a key part of basement membranes [145]
ACSL	Encodes an acyl-CoA synthetase that plays a key role in lipid biosynthesis and fatty acid degradation [98]

Table 11: Several genes with *CCI* score ranking in the top 1%. A high *CCI* rank signifies that suppression of the gene will likely limit broad adverse metabolic effects of glucocorticoids while preserving the desired immune effects.

7. Conclusion

Summary

In this thesis, we have described a coherent framework for causal inference and experimental prioritization in gene regulatory networks. First, we have developed *VAR-GEN*, a robust vector-autoregression method that infers causal networks from high-dimensional gene expression time-series while still maintaining statistical significance. Second, we have developed *CCI*, a contextual causal influence score that prioritizes genes for experimental interventions. Though the score is tailored to our specific biological context, to preserve immune response while minimizing the metabolic response, the Perturbation PageRank method on which *CCI* is based is general enough to apply to a variety of biological and other network settings.

Before our final run over our biological data, we performed a set of preliminary analyses to justify our choice of parameter settings. The success of higher order lags from our simulated system analysis informed our decision to choose lag 2 for our final analysis. The comparison of different run settings allowed us to select the global null, local FDR setting for the pipeline.

Finally, we demonstrate the validity and utility of the causal network by performing a variety of biologically relevant analyses. We validate the network on the external gTeX data and find enrichment of dependent relationships within our network. This strongly suggests that the method does not simply find noise. We compute a variety of network-wide statistics while also isolating biologically relevant genes and edges for discussion with our experimental collaborators.

We should acknowledge the limitations of both our work in particular and gene expression time series analysis as a whole. Our method makes the strong assumption that the causal relationship is linear. This may prevent us from capturing more complex relationships. Furthermore, we do not address the possibility of a nonstationary (i.e. only temporarily) relationship. We also have not compared our method to methods that are not Vector Autoregressions, such as Dynamic Bayesian Networks.

Gene expression timeseries analysis itself has important limitations toward understand cellular biology. It approximates protein levels with transcript levels, but there may be significant relations that do not involve any transcripts. Furthermore, key aspects of cell control systems such as transcription-factor binding, chromatin state and epigenetic state are not considered. Above all, it is difficult to learn causal relations with high accuracy when the data length (around 44) is much less than the number of possible edges (several million).

There are many ways we can improve and extend the work presented here. On the VAR model, we currently assume independent and identically distributed noise at each timepoint, but this is likely false, especially at timepoints of high activity. We ought to allow heteroscedastic noise instead. On the network side, we ought to develop a method that can integrate information from followup experiments to refine our understanding of the network. We could, for example, introduce a Bayesian belief distribution over network edges and parameters, which could incorporate the new data using the Bayesian formalism.

Finally, we plan to compare *VAR-GEN* to non-VAR techniques, such as the Dynamic Bayesian Network, in their accuracy at reconstructing a simulated network from data similar to the GR data.

Appendices

A. Validation of Network Edges

A.1. Methods

The validation work in this section was primarily conceived of, written, and implemented by Brian Jo. I include it here for completeness. - Jonathan Lu

The out-of-sample validation study was performed with autosomal GTEx RNA-seq expression data version v6p with the following five tissues: lung ($n = 278$), subcutaneous adipose ($n = 297$), transformed fibroblasts ($n = 272$), tibial artery ($n = 285$), and thyroid ($n = 278$). We also generated a set of permuted expression values to generate the null statistics. The set of SNPs tested were selected from both genotyped and imputed SNPs from the same dataset, and lying in the vicinity of the causal genes ($<$ 20 KB). In order to reduce the rate of false positive discovery, SNPs that have MAF < 0.05 and SNPs that are annotated as belonging to repeat regions by RepeatMasker were excluded from analysis. The gene pairs tested were restricted to lie on different chromosomes. Association testing was performed using Matrix-eQTL, and the model included top three genotype principal components (calculated from genotyped variants using EIGENSTRAT), genotyping platform, sex, and PEER factors estimated from expression data. [131] The correlation between variant and gene expression levels was evaluated using the estimated t-statistic from this model, and corresponding FDR was estimated using the R q-value package.

A.2. Results

The work from this section was primarily conceived, implemented, and written by Brian Jo, Bianca Dumitrascu, and Prof. Barbara Engelhardt. I include it here for completeness. - Jonathan Lu

In order to validate our directed network edges, we tested the set of single nucleotide polymorphisms (SNPs) in proximity ($<$ 20 kb) to the causal gene for association with effect genes for every directed edge in the network [146]. If the cause-effect gene pairs in the network represent a directed, dosage-dependent relationship, an enrichment of association p-values between the SNPs and the genes representing the effects of the cis-eQTL targets in the directed network should be observed To do this, we used gene expression levels quantified in lung and four other tissues from the Genotype Tissue Expression (GTEx) project version v6p [131]. These SNPs ought to be enriched for associations with the local gene as expression quantitative trait loci (eQTLs). . However, it is worth noting that the GTEx samples are entirely distinct from the cell line used for the GC study, and the expression data is not in the form of a time series. Although much of the cis-eQTL signal tends to be shared across tissues [131], we also compared the degree of enrichment of the distal associations in matched and unmatched out-of-sample eQTLs with similar sample sizes using this approach.

The five tissues (Figure 8B) that we considered from the GTEx data were lung ($n = 278$), subcutaneous adipose ($n = 297$), transformed fibroblasts ($n = 272$), tibial artery ($n = 285$), and thyroid ($n = 278$). The distal-eQTL mapping was performed by taking the set of SNPs within 20 kb of a causal gene and computing the p-value for linear association of each SNP with the corresponding effect gene using MatrixEQTL [147] (Supplementary Materials). The matched null distribution was generated by permuting the effect gene expression values, across all cause-effect gene pairs. FDR over test statistics was calculated using q-value [148]. While this approach will validate specific edges that have low p-values, lack of a low p-value does not imply that the edge is false: if the causal cis-gene does not have an eQTL within 20 kb, for example, or the causal relationship between a pair of genes is induced by GC exposure, then we will not see a low p-value for that edge.

With this design, we found a strong enrichment of association p-values for the distal associations in our network, globally validating our approach(Figure 8A). When we compared tissue-specific p-value enrichments across the five tested tissues (Figure 8B), we found substantial enrichment across other tissues as well, likely because many of these gene pair interactions are shared across tissues. At an FDR threshold of 0.2, we were able to validate 280 network edges across the five tissues with this approach. Of these 280 validated edges, 81 edges were validated with eQTLs from lung tissue samples. With the exception of transformed fibroblasts, the number of edges with significant eQTL associations in other tissues was substantially lower than for lung samples. Across all five tissues, there were 81, 13, 130, 32, 25 edges validated in samples from lung, subcutaneous adipose, transformed fibroblasts, tibial artery, and thyroid tissues, respectively. Considering a contingency table where edges validated in lung are one group and edges validated in all other tissues are another group, where the total number of edges tested is 23,317, a Fisher's exact test shows an enrichment of edges validated in lung samples ($p \leq 4.23 \times 10^{-4}$; Supplementary Materials). This enrichment reflects the match of GTEx lung samples and A549 cells, which are a model of human lung tissue.

The top two distal associations were found among the gene pairs (*FAM19A5*, *IGFBP4*), and (*NR4A1*, *IFITM3*), both of which were specific to lung. Another top TF-specific interactions that replicated in the GTEx study among was found among the genes transcription factor *SNAI2*, and the immune gene *VEGFA* (Figure 8B, $p \leq 6.7 \times 10^{-6}$, q-value FDR ≤ 0.16). A steroid-thyroid hormone with antagonistic interaction with GC targets, *NR4A1* plays an important role in the transforming growth factor-beta (*TFG-beta1*) pathway [149]. Its causal gene, *IFITM3*, is an active component in the interferon-gamma signaling (*IFN-alpha*). As previously noted, *NR4A1* is likely a down-stream effect of *IFN* signaling [150, 151], however, further investigation is required to validate a possible direct interaction between the two genes. Finally, the most significant hit, the interaction between the genes *FAM19A5* and *IGFBP4* is not well characterized, but commends attention due to the pharmacological importance of the insulin-like growth factor-binding protein *IGFBP4*[152, 153].

This validation approach also allows the discovery of distal-eQTLs. Using these associations, we found 203 distal-eQTL pairs in lung, and 756 over all five tissues (q-value FDR ≤ 0.2). The number of identified distal-eQTLs using these directed networks improve on the number identified in the GTEx distal-eQTL study [154], which found 2 distal-eQTLs in lung at FDR ≤ 0.2. This suggests that these directed networks capture meaningful regulatory pathways even across study participants and unmatched tissue types.

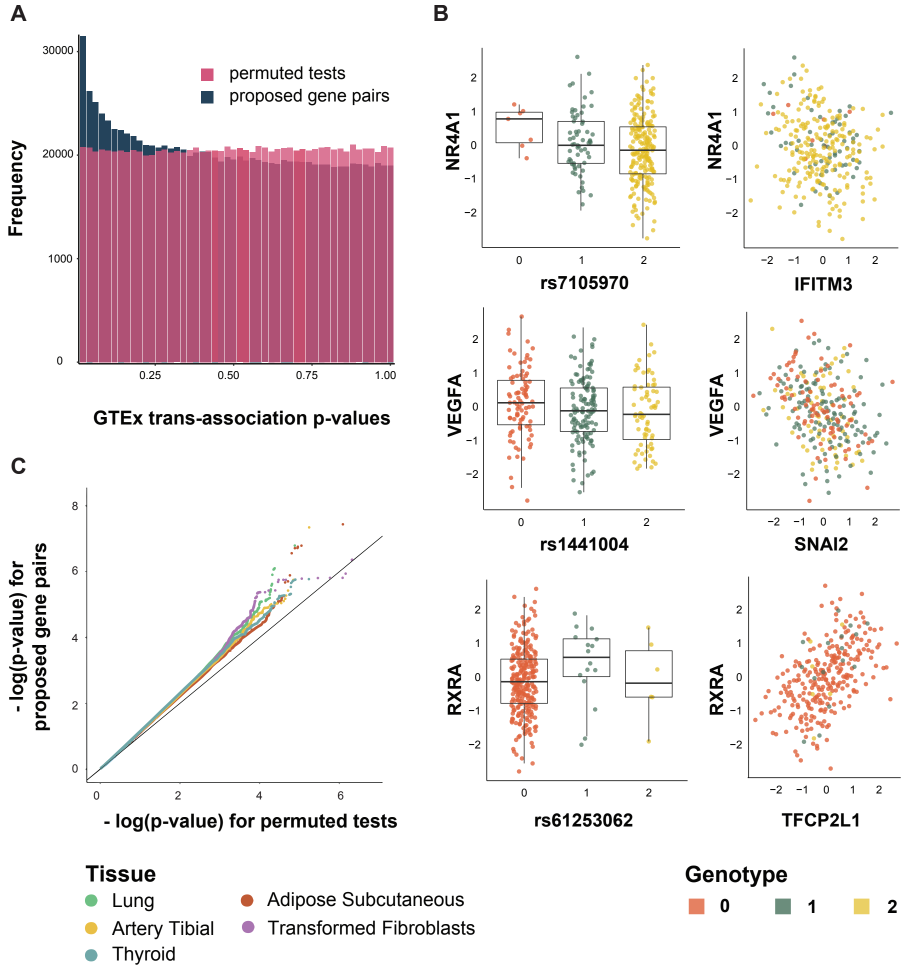

Figure 8: Network edge validation using known cis-regulatory elements from GGR and lung cis-eQTLs from GTEx. **A: The set of p-values for distal associations corresponding to the edges recovered by *VAR-GEN* shows enrichment when compared to p-values quantified from permutations; B: SNPs associated with inferred gene pairs: Genotype-phenotype plots corresponding to the cis-effect (left column), correlation in the GTEx data between cause (y-axis) and effect (x-axis) gene pairs (right column); C: quantile-quantile plot of validated edges shows signal enrichment in lung samples when compared to signals from four other tissues in the GTEx study**

References

[1] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, 01 2009.

[2] Timothy E Reddy, Florencia Pauli, Rebekka O Sprouse, Norma F Neff, Kimberly M Newberry, Michael J Garabedian, and Richard M Myers. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Research*, 19(12):2163–2171, 2009.

[3] Turk Rhen and John A. Cidlowski. Antiinflammatory action of glucocorticoids — new mechanisms for old drugs. *New England Journal of Medicine*, 353(16):1711–1723, 2005. PMID: 16236742.

[4] Derek W Cain and John A Cidlowski. Immune regulation by glucocorticoids. *Nature Reviews Immunology*, 2017.

[5] Ian C. McDowell, Alejandro Barrera, Linda K. Hong, Sarah M. Leichter, Bill. Majoros, Bianca Dumitrascu, Kaixuan Luo, Anthony M. D' Ippolito, Lingyun Song, Alexias Safi, Christopher M. Vockley, Jonathan Lu, Dewran D. Kocak, Luke C. Bartelt, Charles A. Gersbach, Alexander J. Hartemink, Barbara E. Engelhardt, Gregory E. Crawford, and Timothy E. Reddy. Highly coordinated dynamics of the genomic response to glucocorticoids. *in submission*, 2017.

[6] Eliza B Geer, Julie Islam, and Christoph Buettner. Mechanisms of glucocorticoid-induced insulin resistance: focus on adipose tissue function and lipid metabolism. *Endocrinology and Metabolism Clinics of North America*, 43(1):75–102, 2014.

[7] Sarah J. Spencer and Alan Tilbrook. The glucocorticoid contribution to obesity. *Stress*, 14(3):233–246, 2011.

[8] Shun Yao, Shinjae Yoo, and Dantong Yu. Prior knowledge driven granger causality analysis on gene regulatory network discovery. *BMC Bioinformatics*, 16:273, 2015.

[9] Cunlu Zou and Jianfeng Feng. Granger causality vs. dynamic bayesian network inference: a comparative study. *BMC Bioinformatics*, 10(1):122, 2009.

[10] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene van Someren, and Reinhard Guthke. Gene regulatory network inference: Data integration in dynamic models—a review. *Biosystems*, 96(1):86 – 103, 2009.

[11] Miguel Lopes and Gianluca Bontempi. Experimental assessment of static and dynamic algorithms for gene regulation inference from time series expression data. *Frontiers in Genetics*, 4:303, 2013.

[12] Aurélie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. Grouped graphical granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, 25(12):i110–i118, 2009.

[13] Ali Shojaie and George Michailidis. Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 09 2010.

[14] Rainer Opgen-Rhein and Korbinian Strimmer. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, 8(2):1, 2007.

[15] David F Gleich. Pagerank beyond the web. *SIAM Review*, 57(3):321–363, 2015.

[16] Zhi-Ping Liu. Reverse engineering of genome-wide gene regulatory networks from gene expression data. *Current Genomics*, 16(1):3–22, 02 2015.

[17] C.W.J. Granger. Testing for causality. *Journal of Economic Dynamics and Control*, 2:329 – 352, 1980.

[18] Aviv Madar, Alex Greenfield, Eric Vanden-Eijnden, and Richard Bonneau. Dream3: Network inference using dynamic context likelihood of relatedness and the inferelator. *PLoS ONE*, 5(3):e9803, 2010.

[19] Hulin Wu, Tao Lu, Hongqi Xue, and Hua Liang. Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109(506):700–716, 04 2014.

[20] Shizhe Chen, Ali Shojaie, and Daniela M. Witten. Network reconstruction from high dimensional ordinary differential equations. *Journal of the American Statistical Association*, 0(ja):0–0, 0.

[21] Min Zou and Suzanne D. Conzen. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71, 2005.

[22] Kevin Murphy, Saira Mian, et al. Modelling gene expression data using dynamic bayesian networks. Technical report, Technical report, Computer Science Division, University of California, Berkeley, CA, 1999.

[23] Haifen Chen, DAK Maduranga, Piyushkumar A Mundra, and Jie Zheng. Integrating epigenetic prior in dynamic bayesian network for gene regulatory network inference. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2013 IEEE Symposium on*, pages 76–82. IEEE, 2013.

[24] Matthew E Studham, Andreas Tjärnberg, Torbjörn EM Nordling, Sven Nelander, and Erik LL Sonnhammer. Functional association networks as priors for gene regulatory network inference. *Bioinformatics*, 30(12):i130–i138, 2014.

[25] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[26] Alexander J Hartemink, David K Gifford, Tommi S Jaakkola, Richard A Young, et al. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Pacific symposium on biocomputing*, volume 6, page 266, 2001.

[27] Pietro Zoppoli, Sandro Morganella, and Michele Ceccarelli. Timedelay-aracne: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinformatics*, 11(1):154, 2010.

[28] Christopher A. Penfold and David L. Wild. How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6):857–870, 2011.

[29] Nitai D. Mukhopadhyay and Snigdhansu Chatterjee. Causality and pathway search in microarray time series experiment. *Bioinformatics*, 23(4):442, 2007.

[30] Gary H.F. Tam, Chunqi Chang, and Yeung Sam Hung. Application of granger causality to gene regulatory network discovery. In *IEEE 6th International Conference on Systems Biology (ISB)*, ISB '12, 2012.

[31] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[32] Rainer Opgen-Rhein and Korbinian Strimmer. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[33] Hyunghoon Cho, Bonnie Berger, and Jian Peng. Reconstructing causal biological networks through active learning. *PloS One*, 11(3):e0150611, 2016.

[34] Kevin P. Murphy. Active learning of causal bayes net structure. Technical report, 2001.

[35] Marloes H Maathuis, Markus Kalisch, Peter Bühlmann, et al. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

[36] Andrea Rau, Florence Jaffrézic, and Grégory Nuel. Joint estimation of causal effects from observational and intervention gene expression data. *BMC Systems Biology*, 7(1):1, 2013.

[37] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014.

[38] Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov):2523–2547, 2008.

[39] Karen Sachs, David Gifford, Tommi Jaakkola, Peter Sorger, and Douglas A Lauffenburger. Bayesian network approach to cell signaling pathway modeling. *Sci STKE*, 148:e38, 2002.

[40] Alexander Y Mitrophanov and Eduardo A Groisman. Positive feedback in cellular control systems. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 30(6):542–555, 06 2008.

[41] Joseph R Pomerening. Positive feedback loops in cell cycle progression. *FEBS letters*, 583(21):3388–3396, 11 2009.

[42] Sandeep Krishna, Anna M C Andersson, Szabolcs Semsey, and Kim Sneppen. Structure and function of negative feedback loops at the interface of genetic and metabolic networks. *Nucleic Acids Research*, 34(8):2455–2462, 2006.

[43] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 01 2010.

[44] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genetics*, 3(9):1–12, 09 2007.

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

[46] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[47] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[48] Gene ontology consoritum's curated list of immune genes. `http://wiki.geneontology.org/index.php/Immunology`, 2014. Accessed: 2017-04-22.

[49] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[50] Hong-Mei Zhang, Hu Chen, Wei Liu, Hui Liu, Jing Gong, Huili Wang, and An-Yuan Guo. Animaltfdb: a comprehensive animal transcription factor database. *Nucleic Acids Research*, 40(Database issue):D144–D149, 01 2012.

[51] Lauren E Shipp, Joyce V Lee, Chi-Yi Yu, Miles Pufall, Pili Zhang, Donald K Scott, and Jen-Chywan Wang. Transcriptional regulation of human dual specificity protein phosphatase 1 (dusp1) gene by glucocorticoids. *PLoS ONE*, 5(10):e13754, 2010.

[52] Timothy E Reddy, Jason Gertz, Gregory E Crawford, Michael J Garabedian, and Richard M Myers. The hypersensitive glucocorticoid response specifically regulates period 1 and expression of circadian genes. *Molecular and Cellular Biology*, 32(18):3756–3767, 09 2012.

[53] Andrew Mugler, Mark Kittisopikul, Luke Hayden, Jintao Liu, Chris H Wiggins, Gurol M Suel, and Aleksandra M Walczak. Noise expands the response range of the bacillus subtilis competence circuit. *PLOS Comp. Bio.*, 2016.

[54] Bruggeman FJ Maarleveld TR, Olivier BG. Stochpy: A comprehensive, user-friendly tool for simulating stochastic biological processes. *PLoS ONE*, 8(11), 2013.

[55] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*, 45(D1):D353–D361, Jan 2017.

[56] Roger Sciammas and Mark M. Davis. Modular nature of blimp-1 in the regulation of gene expression during b cell maturation. *The Journal of Immunology*, 172(9):5427–5440, 2004.

[57] Han-kuei Huang, Claudio A. P. Joazeiro, Emanuela Bonfoco, Shinji Kamada, Joel D. Leverson, and Tony Hunter. The inhibitor of apoptosis, ciap1, functions as a ubiquitin-protein ligase and promotes in vitro ubiquitination of caspases-3 and -7. *Journal of Biological Chemistry*, 2000.

[58] M.K Sharief, M.A Noori, and Y Zoukos. Reduced expression of the inhibitor of apoptosis proteins in t cells from patients with multiple sclerosis following interferon-beta therapy. *Journal of Neuroimmunology*, 129(1–2):224 – 231, 2002.

[59] Lisa Russell and Lee Ann Garrett-Sinha. Transcription factor ets-1 in cytokine and chemokine gene regulation. *Cytokine*, 51(3):217 – 226, 2010.

[60] Katia De Filippo, Anne Dudeck, Mike Hasenberg, Emma Nye, Nico van Rooijen, Karin Hartmann, Matthias Gunzer, Axel Roers, and Nancy Hogg. Mast cell and macrophage chemokines cxcl1/cxcl2 control the early stage of neutrophil recruitment during tissue inflammation. *Blood*, 121(24):4930–4937, 2013.

[61] Chad Brocker, Natalie Lassen, Tia Estey, Aglaia Pappa, Miriam Cantore, Valeria V Orlova, Triantafyllos Chavakis, Kathryn L Kavanagh, Udo Oppermann, and Vasilis Vasiliou. Aldehyde dehydrogenase 7a1 (aldh7a1) is a novel enzyme involved in cellular defense against hyperosmotic stress. *The Journal of Biological Chemistry*, 285(24):18452–18463, 06 2010.

[62] Satori A Marchitti, Chad Brocker, Dimitrios Stagos, and Vasilis Vasiliou. Non-p450 aldehyde oxidizing enzymes: the aldehyde dehydrogenase superfamily. *Expert Opinion on Drug Metabolism & Toxicology*, 4(6):697–720, 06 2008.

[63] Yunqiang Zhang, Mingde Zang, Jianfang Li, Jun Ji, Jianian Zhang, Xiaolei Liu, Ying Qu, Liping Su, Chen Li, Yinyan Yu, Zhenggang Zhu, Bingya Liu, and Min Yan. Ceacam6 promotes tumor migration, invasion, and metastasis in gastric cancer. *Acta Biochimica et Biophysica Sinica*, 46(4):283, 2014.

[64] Joan S Lewis-Wambi, Heather E Cunliffe, Helen R Kim, Amanda L Willis, and V Craig Jordan. Overexpression of ceacam6 promotes migration and invasion of oestrogen deprived breast cancer cells. *European Journal of Cancer*, 44(12):1770–1779, 08 2008.

[65] N Utku, T Heinemann, and EL Milford. T-cell immune response cdna 7 in allograft rejection and inflammation. *Current Opinion In Investigational Drugs*, 8(5):401–410, 2007.

[66] Grit-Carsta Bulwin, Stephanie Wälter, Mirko Schlawinsky, Thomas Heinemann, Anke Schulze, Wolfgang Höhne, Gerd Krause, Wiltrud Kalka-Moll, Patricia Fraser, Hans-Dieter Volk, Jürgen Löhler, Edgar L Milford, and Nalân Utku. Hla-dr alpha 2 mediates negative signalling via binding to tirc7 leading to anti-inflammatory and apoptotic effects in lymphocytes in vitro and in vivo. *PLoS ONE*, 3(2):e1576, 2008.

[67] Tatsuya Sawamura, Noriaki Kume, Takuma Aoyama, Hideaki Moriwaki, Hajime Hoshikawa, Yuichi Aiba, Takeshi Tanaka, Soichi Miwa, Yoshimoto Katsura, Toru Kita, and Tomoh Masaki. An endothelial receptor for oxidized low-density lipoprotein. *Nature*, 386(6620):73–77, 03 1997.

[68] Sayoko Ogura, Akemi Kakino, Yuko Sato, Yoshiko Fujita, Shin Iwamoto, Kazunori Otsui, Ryo Yoshimoto, and Tatsuya Sawamura. Lox-1, the multifunctional receptor underlying cardiovascular dysfunction. *Circulation Journal*, 73(11):1993–1999, 2009.

[69] Pengcheng Zhu, Yan Yih Goh, Hwee Fang Alison Chin, Sander Kersten, and Nguan Soon Tan. Angiopoietin-like 4: a decade of research. *Bioscience Reports*, 32(3):211–219, 2012.

[70] Suneil K. Koliwad, Taiyi Kuo, Lauren E. Shipp, Nora E. Gray, Fredrik Backhed, Alex Yick-Lun So, Robert V. Farese, and Jen-Chywan Wang. Angiopoietin-like 4 (angptl4, fasting-induced adipose factor) is a direct glucocorticoid receptor target and participates in glucocorticoid-regulated triglyceride metabolism. *Journal of Biological Chemistry*, 284(38):25593–25601, 2009.

[71] M Karin, Liu Zg, and Zandi E. Ap-1 function and regulation. *Current Opinions in Cell Biology*, 9(2):240–246, 1997.

[72] Linda Andersson, Pontus Boström, Johanna Ericson, Mikael Rutberg, Björn Magnusson, Denis Marchesan, Michel Ruiz, Lennart Asp, Ping Huang, Michael A. Frohman, Jan Borén, and Sven-Olof Olofsson. Pld1 and erk2 regulate cytosolic lipid droplet formation. *Journal of Cell Science*, 119(11):2246–2257, 2006.

[73] Leena Kovanen, Kati Donner, Mari Kaunisto, and Timo Partonen. Cry1, cry2 and prkcdbp genetic variants in metabolic syndrome. *Hypertension Research*, 38(3):186–192, 2015.

[74] Fausto Machicao, Andreas Peter, Jürgen Machann, Ingmar Königsrainer, Anja Böhm, Stefan Zoltan Lutz, Martin Heni, Andreas Fritsche, Fritz Schick, Alfred Königsrainer, et al. Glucose-raising polymorphisms in the human clock gene cryptochrome 2 (cry2) affect hepatic lipid content. *PloS One*, 11(1):e0145563, 2016.

[75] Sarah Søs Poulsen, Nicklas R Jacobsen, Sarah Labib, Dongmei Wu, Mainul Husain, Andrew Williams, Jesper P Bøgelund, Ole Andersen, Carsten Købler, Kristian Mølhave, et al. Transcriptomic analysis reveals novel mechanistic insight into murine biological responses to multi-walled carbon nanotubes in lungs and cultured lung epithelial cells. *PloS One*, 8(11):e80452, 2013.

[76] David Carling. The amp-activated protein kinase cascade: a unifying system for energy control. *Trends in Biochemical Sciences*, 29(1):18 – 24, 2004.

[77] David Stapleton, Ken I. Mitchelhill, Guang Gao, Jane Widmer, Belinda J. Michell, Trazel Teh, Colin M. House, C. Shamala Fernandez, Timothy Cox, Lee A. Witters, and Bruce E. Kemp. Mammalian amp-activated protein kinase subfamily. *Journal of Biological Chemistry*, 271(2):611–614, 1996.

[78] Aisling Dunne and Luke A. J. O'Neill. The interleukin-1 receptor/toll-like receptor superfamily: Signal transduction during inflammation and host defense. *Science Signaling*, 2003(171):re3–re3, 2003.

[79] JamesÂ C. Lo, Sanda Ljubicic, Barbara Leibiger, Matthias Kern, IngoÂ B. Leibiger, Tilo Moede, MollyÂ E. Kelly, Diti ChatterjeeÂ Bhowmick, Incoronata Murano, Paul Cohen, AlexanderÂ S. Banks, MelinÂ J. Khandekar, Arne Dietrich, JeffreyÂ S. Flier, Saverio Cinti, Matthias BlÃŒher, NikaÂ N. Danial, Per-Olof Berggren, and BruceÂ M. Spiegelman. Adipsin is an adipokine that improves beta cell function in diabetes. *Cell*, 158(1):41 – 53, 2014.

[80] J E Volanakis and S V Narayana. Complement factor d, a novel serine protease. *Protein Science : A Publication of the Protein Society*, 5(4):553–564, 04 1996.

[81] Mar Valés-Gómez1998, Hugh T Reyburn, Michal Mandelboim, and Jack L Strominger. Kinetics of interaction of hla-c ligands with natural killer cell inhibitory receptors. *Immunity*, 9(3):337 – 344, 1998.

[82] Dana Zeineddine, Aya Abou Hammoud, Mohamad Mortada, and Hélène Boeuf. The oct4 protein: more than a magic stemness marker. *American Journal of Stem Cells*, 3(2):74–82, 2014.

[83] Abraham L Brass, I-Chueh Huang, Yair Benita, Sinu P John, Manoj N Krishnan, Eric M Feeley, Bethany Ryan, Jessica L Weyer, Louise van der Weyden, Erol Fikrig, David J Adams, Ramnik J Xavier, Michael Farzan, and Stephen J Elledge. Ifitm proteins mediate the innate immune response to influenza a h1n1 virus, west nile virus and dengue virus. *Cell*, 139(7):1243–1254, 12 2009.

[84] Cheryl Shoubridge, Tod Fullston, and Jozef Gécz. Arx spectrum disorders: making inroads into the molecular pathology. *Human Mutation*, 31(8):889–900, 2010.

[85] Gaia Colasante, Patrick Collombat, Valentina Raimondi, Dario Bonanomi, Carmelo Ferrai, Mario Maira, Kazuaki Yoshikawa, Ahmed Mansouri, Flavia Valtorta, John L. R. Rubenstein, and Vania Broccoli. Arx is a direct target of dlx2 and thereby contributes to the tangential migration of gabaergic interneurons. *Journal of Neuroscience*, 28(42):10674–10686, 2008.

[86] Gijsbertus T. J. van der Horst, Manja Muijtjens, Kumiko Kobayashi, Riya Takano, Shin-ichiro Kanno, Masashi Takao, Jan de Wit, Anton Verkerk, Andre P. M. Eker, Dik van Leenen, Ruud Buijs, Dirk Bootsma, Jan H. J. Hoeijmakers, and Akira Yasui. Mammalian cry1 and cry2 are essential for maintenance of circadian rhythms. *Nature*, 398(6728):627–630, 04 1999.

[87] Edmund A. Griffin, David Staknis, and Charles J. Weitz. Light-independent role of cry1 and cry2 in the mammalian circadian clock. *Science*, 286(5440):768–771, 1999.

[88] Takeshi Inukai, Toshiya Inaba, Jinjun Dang, Ryoko Kuribara, Keiya Ozawa, Atsushi Miyajima, Wenshu Wu, A. Thomas Look, Yojiro Arinobu, Hiromi Iwasaki, Koichi Akashi, Keiko Kagami, Kumiko Goi, Kanji Sugita, and Shinpei Nakazawa. Tef, an antiapoptotic bzip transcription factor related to the oncogenic e2a-hlf chimera, inhibits cell growth by down-regulating expression of the common Î chain of cytokine receptors. *Blood*, 105(11):4437–4444, 2005.

[89] Sam Kaitaniemi, Heli Elovaara, Kirsi Grön, Heidi Kidron, Janne Liukkonen, Tiina Salminen, Marko Salmi, Sirpa Jalkanen, and Kati Elima. The unique substrate specificity of human aoc2, a semicarbazide-sensitive amine oxidase. *Cellular and Molecular Life Sciences*, 66(16):2743, 2009.

[90] Melvin Anyasi Ambele, Carla Dessels, Chrisna Durandt, and Michael Sean Pepper. Genome-wide analysis of gene expression during adipogenesis in human adipose-derived stromal cells reveals novel patterns of gene expression during adipocyte differentiation. *Stem Cell Research*, 16(3):725 – 734, 2016.

[91] Irene Garcia-Higuera, Toshiyasu Taniguchi, Shridar Ganesan, M.Stephen Meyn, Cynthia Timmers, James Hejna, Markus Grompe, and Alan D D'Andrea. Interaction of the fanconi anemia proteins and {BRCA1} in a common pathway. *Molecular Cell*, 7(2):249 – 262, 2001.

[92] Alan D. D'Andrea and Markus Grompe. The fanconi anaemia/brca pathway. *Nat Rev Cancer*, 3(1):23–34, 01 2003.

[93] Birger Blomback, Birgit Hessel, Desmond Hogg, and Lisbeth Therkildsen. A two-step fibrinogen-fibrin transition in blood coagulation. *Nature*, 275(5680):501–505, 10 1978.

[94] Krzysztof B. Wicher and Erik Fries. Prohaptoglobin is proteolytically cleaved in the endoplasmic reticulum by the complement c1r-like protein. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14390–14395, 2004.

[95] M Arredouani, P Matthijs, E Van Hoeyveld, A Kasran, H Baumann, J L Ceuppens, and E Stevens. Haptoglobin directly affects t cells and suppresses t helper cell type 2 cytokine release. *Immunology*, 108(2):144–151, 02 2003.

[96] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[97] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[98] Hannah Schneider, Sarah Staudacher, Margarete Poppelreuther, Wolfgang Stremmel, Robert Ehehalt, and Joachim Füllekrug. Protein mediated fatty acid uptake: Synergy between cd36/fat-facilitated transport and acyl-coa synthetase-driven metabolism. *Archives of Biochemistry and Biophysics*, 546:8 – 18, 2014.

[99] Hans-Georg Joost, Graeme I. Bell, James D. Best, Morris J. Birnbaum, Maureen J. Charron, Y. T. Chen, Holger Doege, David E. James, Harvey F. Lodish, Kelle H. Moley, Jeffrey F. Moley, Mike Mueckler, Suzanne Rogers, Annette Schürmann, Susumu Seino, and Bernard Thorens. Nomenclature of the glut/slc2a family of sugar/polyol transport facilitators. *American Journal of Physiology - Endocrinology and Metabolism*, 282(4):E974–E976, 2002.

[100] Till Adhikary, Dominique T Brandt, Kerstin Kaddatz, Josefine Stockert, Simone Naruhn, Wolfgang Meissner, Florian Finkernagel, Julia Obert, Sonja Lieber, Maren Scharfe, et al. Inverse pparβ/δ agonists suppress oncogenic signaling to the angptl4 gene and inhibit cancer cell invasion. *Oncogene*, 32(44):5241–5252, 2013.

[101] Marius R Robciuc, Paulina Skrobuk, Andrey Anisimov, Vesa M Olkkonen, Kari Alitalo, Robert H Eckel, Heikki A Koistinen, Matti Jauhiainen, and Christian Ehnholm. Angiopoietin-like 4 mediates ppar delta effect on lipoprotein lipase-dependent fatty acid uptake but not on beta-oxidation in myotubes. *PloS One*, 7(10):e46212, 2012.

[102] Anna Dittrich, Christina Khouri, Sara Dutton Sackett, Christian Ehlting, Oliver Böhmer, Ute Albrecht, Johannes G Bode, Christian Trautwein, and Fred Schaper. Glucocorticoids increase interleukin-6–dependent gene induction by interfering with the expression of the suppressor of cytokine signaling 3 feedback inhibitor. *Hepatology*, 55(1):256–266, 2012.

[103] Haitao Li, Janice WC Louey, Kwong Wai Choy, David TL Liu, Wai Man Chan, Yiu Man Chan, Nicholas SK Fung, Bao Jian Fan, Larry Baum, Juliana CN Chan, et al. Edn1 lys198asn is associated with diabetic retinopathy in type 2 diabetes. *Molecular Vision*, 2008.

[104] Mark J Cowley, Anita Weinberg, Nathan W Zammit, Stacey N Walters, Wayne J Hawthorne, Thomas Loudovaris, Helen Thomas, Tom Kay, Jenny E Gunton, Stephen I Alexander, et al. Human islets express a marked proinflammatory molecular signature prior to transplantation. *Cell Transplantation*, 21(9):2063–2078, 2012.

[105] Suparna A Sarkar, Catherine E Lee, Francisco Victorino, Tom T Nguyen, Jay A Walters, Adam Burrack, Jens Eberlein, Steven K Hildemann, and Dirk Homann. Expression and regulation of chemokines in murine and human type 1 diabetes. *Diabetes*, 61(2):436–446, 2012.

[106] T Tuller, S Atar, E Ruppin, M Gurevich, and A Achiron. Common and specific signatures of gene expression and protein–protein interactions in autoimmune diseases. *Genes and Immunity*, 14(2):67–82, 2013.

[107] Susan J Burke, Danhong Lu, Tim E Sparer, Thomas Masi, Matthew R Goff, Michael D Karlstad, and J Jason Collier. Nf-κb and stat1 control cxcl1 and cxcl2 gene transcription. *American Journal of Physiology-Endocrinology and Metabolism*, 306(2):E131–E149, 2014.

[108] Anna Rull, Fernando Rodríguez, Gerard Aragonès, Judit Marsillach, Raúl Beltrán, Carlos Alonso-Villaverde, Jordi Camps, and Jorge Joven. Hepatic monocyte chemoattractant protein-1 is upregulated by dietary cholesterol and contributes to liver steatosis. *Cytokine*, 48(3):273–279, 2009.

[109] Andrew V Turnbull and Catherine L Rivier. Regulation of the hypothalamic-pituitary-adrenal axis by cytokines: actions and mechanisms of action. *Physiological Reviews*, 79(1):1–71, 1999.

[110] Hideyo Sato, Michiko Tamba, Tetsuro Ishii, and Shiro Bannai. Cloning and expression of a plasma membrane cystine/glutamate exchange transporter composed of two distinct proteins. *Journal of Biological Chemistry*, 274(17):11455–11458, 1999.

[111] Monika D. Polewski, Rosyli F. Reveron-Thornton, Gregory A. Cherryholmes, Georgi K. Marinov, Kaniel Cassady, and Karen S. Aboody. Increased expression of system xc- in glioblastoma confers an altered metabolism and chemoresistance. *Molecular Cancer Research*, 2016.

[112] Maria Carmen Solaz-Fuster, José Vicente Gimeno-Alcañiz, Susana Ros, Maria Elena Fernandez-Sanchez, Belen Garcia-Fojeda, Olga Criado Garcia, David Vilchez, Jorge Dominguez, Mar Garcia-Rocha, Maribel Sanchez-Piris, Carmen Aguado, Erwin Knecht, Jose Serratosa, Joan Josep Guinovart, Pascual Sanz, and Santiago Rodriguez de Córdoba. Regulation of glycogen synthesis by the laforin–malin complex is modulated by the amp-activated protein kinase pathway. *Human Molecular Genetics*, 17(5):667, 2008.

[113] Vincent Guinard-Samuel, Arnaud Bonnard, Pascal De Lagausie, Pascale Philippe-Chomette, Corine Alberti, Alaa El Ghoneimi, Michel Peuchmaur, and Dominique Berrebi-Binczak. Calretinin immunohistochemistry: a simple and efficient tool to diagnose hirschsprung disease. *Mod Pathol*, 22(10):1379–1384, 07 2009.

[114] Lei Sun, Mark Bartlam, Yiwei Liu, Hai Pang, and Zihe Rao. Crystal structure of the pyridoxal-5'-phosphate-dependent serine dehydratase from human liver. *Protein Science*, 14(3):791–798, 03 2005.

[115] A Aström, U Pettersson, and J J Voorhees. Structure of the human cellular retinoic acid-binding protein ii gene. early transcriptional regulation by retinoic acid. *Journal of Biological Chemistry*, 267(35):25251–25255, 1992.

[116] Werner Albig, Efterpi Kardalinou, Birgit Drabent, Andreas Zimmer, and Detlef Doenecke. Isolation and characterization of two human {H1} histone genes within clusters of core histone genes. *Genomics*, 10(4):940 – 948, 1991.

[117] Masaki Mori, Masahiro Nakajima, Yasuo Mikami, Shoji Seki, Masaharu Takigawa, Toshikazu Kubo, and Shiro Ikegawa. Transcriptional regulation of the cartilage intermediate layer protein (cilp) gene. *Biochemical and Biophysical Research Communications*, 341(1):121 – 127, 2006.

[118] Yoh-ichi Seki, Hiromasa Inoue, Naoko Nagata, Katsuhiko Hayashi, Satoru Fukuyama, Koichiro Matsumoto, Okiru Komine, Shinjiro Hamano, Kunisuke Himeno, Kyoko Inagaki-Ohara, Nicholas Cacalano, Anne O'Garra, Tadahilo Oshida, Hirohisa Saito, James A Johnston, Akihiko Yoshimura, and Masato Kubo. Socs-3 regulates onset and maintenance of th2-mediated allergic responses. *Nat Med*, 9(8):1047–1054, 08 2003.

[119] Mark A Febbraio. Role of interleukins in obesity: implications for metabolic disease. *Trends in Endocrinology & Metabolism*, 25(6):312–319, 2014.

[120] Jianjin Shi, Yue Zhao, Yupeng Wang, Wenqing Gao, Jingjin Ding, Peng Li, Liyan Hu, and Feng Shao. Inflammatory caspases are innate immune receptors for intracellular lps. *Nature*, 514(7521):187–192, 2014.

[121] Mary C. Hunt, Marina I. Siponen, and Stefan E.H. Alexson. The emerging role of acyl-coa thioesterases and acyltransferases in regulating peroxisomal lipid metabolism. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1822(9):1397 – 1410, 2012. Metabolic Functions and Biogenesis of Peroxisomes in Health and Disease.

[122] Lucchiari S Bordoni A Prelle A Jann S Keller A Ciscato P Galbiati S Chiveri L Torrente Y Scarlato G Bresolin N Comi GP, Fortunato F. Beta-enolase deficiency, a new metabolic myopathy of distal glycolysis. *Annals of Neurology*, 50(2):202–207, 08 2001.

[123] Xinsong Xu, Antonio Sarikas, Dora C Dias-Santagata, Georgia Dolios, Pascal J Lafontant, Shih-Chong Tsai, Wuqiang Zhu, Hidehiro Nakajima, Hisako O Nakajima, Loren J Field, Rong Wang, and Zhen-Qiang Pan. The cul7 e3 ubiquitin ligase targets insulin receptor substrate 1 for ubiquitin-dependent degradation. *Molecular Cell*, 30(4):403–414, 05 2008.

[124] RL Kitchens. Role of cd14 in cellular recognition of bacterial lipopolysaccharides. *Chemical Immunology*, 74:61–82, 12 2004.

[125] Roland P. Bourette, Paulo De Sepulveda, Sylvie Arnaud, Patrice Dubreuil, Robert Rottapel, and Guy Mouchiroud. Suppressor of cytokine signaling 1 interacts with the macrophage colony-stimulating factor receptor and negatively regulates its proliferation signal. *Journal of Biological Chemistry*, 276(25):22133–22139, 2001.

[126] Liangyou Rui, Minsheng Yuan, Daniel Frantz, Steven Shoelson, and Morris F. White. Socs-1 and socs-3 block insulin signaling by ubiquitin-mediated degradation of irs1 and irs2. *Journal of Biological Chemistry*, 277(44):42394–42398, 2002.

[127] Jun Ohsumi, Kenji Miyadai, Ichiro Kawashima, Hiroko Ishikawa-Ohsumi, Sachiko Sakakibara, Katsuko Mita-Honjo, and Yo Takiguchi. Adipogenesis inhibitory factor a novel inhibitory regulator of adipose conversion in bone marrow. *FEBS Letters*, 288(1-2):13–16, 1991.

[128] Philip E. Auron. The interleukin 1 receptor: Ligand interactions and signal transduction. *Cytokine & Growth Factor Reviews*, 9(3–4):221 – 237, 1998.

[129] Véronique Baud, Zheng-Gang Liu, Brydon Bennett, Nobutaka Suzuki, Ying Xia, and Michael Karin. Signaling by proinflammatory cytokines: oligomerization of traf2 and traf6 is sufficient for jnk and ikk activation and target gene induction via an amino-terminal effector domain. *Genes & Development*, 13(10):1297–1308, 05 1999.

[130] Zheng Chen, Liang Sheng, Hong Shen, Yujun Zhao, Shaomeng Wang, Robert Brink, and Liangyou Rui. Hepatic traf2 regulates glucose metabolism through enhancing glucagon responses. *Diabetes*, 61(3):566–573, 03 2012.

[131] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[132] Tetsuya Kubota, Naoto Kubota, and Takashi Kadowaki. Imbalanced insulin actions in obesity and type 2 diabetes: Key mouse models of insulin signaling pathway. *Cell Metabolism*, 25(4):797–810, 2017.

[133] Alan R Saltiel and C Ronald Kahn. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature*, 414(6865):799–806, 2001.

[134] Stephan Barrientos, Olivera Stojadinovic, Michael S Golinko, Harold Brem, and Marjana Tomic-Canic. Growth factors and cytokines in wound healing. *Wound Repair and Regeneration*, 16(5):585–601, 2008.

[135] Wen-jun Liu, Wei-qiang Gao, and Xiao-ni Kong. Polarization and functional plasticity of macrophages in regulating innate immune response. *Journal of Shanghai Jiaotong University (Science)*, 19(6):646–650, 2014.

[136] Alexander Jais, Maite Solas, Heiko Backes, Bhagirath Chaurasia, André Kleinridders, Sebastian Theurich, Jan Mauer, Sophie M Steculorum, Brigitte Hampel, Julia Goldau, et al. Myeloid-cell-derived vegf maintains brain glucose uptake and limits cognitive impairment in obesity. *Cell*, 165(4):882–895, 2016.

[137] Xueying Lin, Akiko Taguchi, Sunmin Park, Jake A Kushner, Fan Li, Yedan Li, and Morris F White. Dysregulation of insulin receptor substrate 2 in beta cells and brain causes obesity and diabetes. *Journal of Clinical Investigation*, 114(7):908–916, 10 2004.

[138] Vasily V Ogryzko, Tomohiro Kotani, Xiaolong Zhang, R.Louis Schiltz, Tazuko Howard, Xiang-Jiao Yang, Bruce H Howard, Jun Qin, and Yoshihiro Nakatani. Histone-like {TAFs} within the {PCAF} histone acetylase complex. *Cell*, 94(1):35 – 44, 1998.

[139] María M Adeva-Andany, Manuel González-Lucán, Cristóbal Donapetry-García, Carlos Fernández-Fernández, and Eva Ameneiros-Rodríguez. Glycogen metabolism in humans. *BBA Clinical*, 5:85–100, 06 2016.

[140] Darcy M. Moncada, Srinivas J. Kammanadiminti, and Kris Chadee. Mucin and toll-like receptors in host defense against intestinal parasites. *Trends in Parasitology*, 19(7):305 – 311, 2003.

[141] Burton M. Wice and Jeffrey I. Gordon. A tetraspan membrane glycoprotein produced in the human intestinal epithelium and liver that can regulate cell density-dependent proliferation. *Journal of Biological Chemistry*, 270(37):21907–21918, 1995.

[142] Tsonwin Hai and Matthew G Hartman. The molecular biology and nomenclature of the activating transcription factor/camp responsive element binding family of transcription factors: activating transcription factor proteins and homeostasis. *Gene*, 273(1):1 – 11, 2001.

[143] M Mueckler, C Caruso, SA Baldwin, M Panico, I Blench, HR Morris, WJ Allard, GE Lienhard, and HF Lodish. Sequence and structure of a human glucose transporter. *Science*, 229(4717):941–945, 1985.

[144] Laura A Crawford, Michelle A Guney, Young Ah Oh, R Andrea DeYoung, David M Valenzuela, Andrew J Murphy, George D Yancopoulos, Karen M Lyons, David R Brigstock, Aris Economides, and Maureen Gannon. Connective tissue growth factor (ctgf) inactivation leads to defects in islet cell lineage allocation and beta-cell proliferation during embryogenesis. *Molecular Endocrinology*, 23(3):324–336, 03 2009.

[145] N Turner, P J Mason, R Brown, M Fox, S Povey, A Rees, and C D Pusey. Molecular cloning of the human goodpasture antigen demonstrates it to be the alpha 3 chain of type iv collagen. *Journal of Clinical Investigation*, 89(2):592–601, 02 1992.

[146] Chuan Gao, Shiwen Zhao, Ian C McDowell, Christopher D Brown, and Barbara E Engelhardt. Context-specific and differential gene co-expression networks via Bayesian biclustering models. *PLOS Computational Biology*, 12:e1004791, 2016.

[147] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[148] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[149] Xing Guo and Xiao-Fan Wang. Signaling cross-talk between tgf-β/bmp and other pathways. *Cell Research*, 19(1):71–88, 2009.

[150] N Papac-Milicevic, JM Breuss, J Zaujec, L Ryban, T Plyushch, and GA Wagner. The interferon stimulated gene 12 inactivates vasculoprotective functions of 729 nr4a nuclear receptors. *Circulation Research*, 110:e50–e63, 2012.

[151] Mohamed A Tantawy, Bastian Hatesuer, Esther Wilk, Leonie Dengler, Nadine Kasnitz, Siegfried Weiß, and Klaus Schughart. The interferon-induced gene ifi27l2a is active in lung macrophages and lymphocytes after influenza a infection but deletion of ifi27l2a in mice does not increase susceptibility to infection. *PloS One*, 9(9):e106392, 2014.

[152] Tianfu Wu, Chun Xie, Jie Han, Yujin Ye, Sandeep Singh, Jinchun Zhou, Yajuan Li, Huihua Ding, Quan-zhen Li, Xin Zhou, et al. Insulin-like growth factor binding protein-4 as a marker of chronic lupus nephritis. *PloS One*, 11(3):e0151491, 2016.

[153] Weiwen Li, Debin Sun, Zhuqing Lv, Yueqiu Wei, Liyun Zheng, Tingting Zeng, and Jialu Zhao. Insulin-like growth factor binding protein-4 inhibits cell growth, migration and invasion, and downregulates cox-2 expression in a549 lung cancer cells. *Cell Biology International*, 41(4):384–391, 2017.

[154] Brian Jo, Yuan He, Benjamin J Strober, Princy Parsana, Francois Aguet, Andrew A Brown, Stephane E Castel, Eric R Gamazon, Ariel Gewirtz, Genna Gliner, Buhm Han, Amy Z He, Eun Yong Kang, Ian C McDowell, Xiao Li, Pejman Mohammadi, Christine B Peterson, Gerald Quon, Ashis Saha, Ayellet V Segre, Jae Hoon Sul, Timothy J Sullivan, Kristin G Ardlie, Christopher D Brown, Donald F Conrad, Nancy J Cox, Emmanouil T Dermitzakis, Eleazar Eskin, Manolis Kellis, Tuuli Lappalainen, Chiara Sabatti, , Barbara E Engelhardt, and Alexis Battle. Distant regulatory effects of genetic variation in multiple human tissues. *bioRxiv*, 2016.