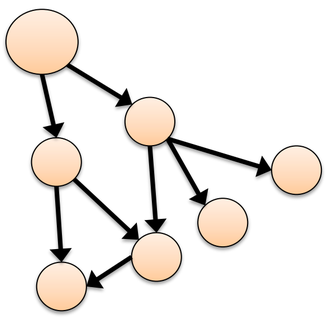
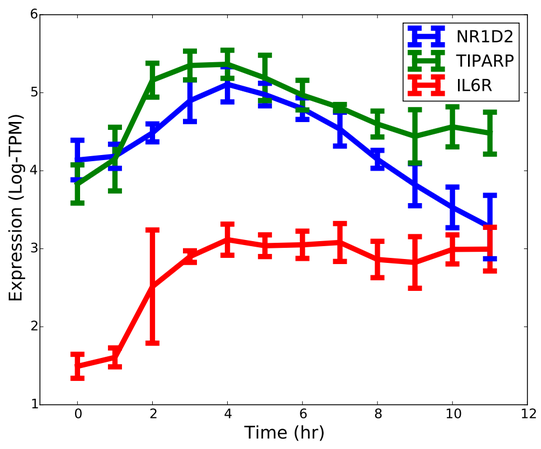


Improved Inference of Causal Networks from Gene Expression Time Series

Jonathan Lu
Advisors: Prof. Barbara Engelhardt, Bianca Dumitrascu and Brian Jo

Background

- Glucocorticoids (GCs) are immunosuppressant drugs that can lead to metabolic disorders such as diabetes and obesity
- How can we disentangle the desired immune effects from the adverse metabolic effects?
- GC binds to the glucocorticoid receptor (GR), which initiates transcription of a variety of genes, leading to a cascade of effects
- Can we infer the triggered causal network to pinpoint the immune and metabolic response?
- Use Gene expression time series from GC-stimulated lung cells
- 2768 differentially expressed genes, 12 timepoints, 4 replicates



Methodological challenges:

- <u>High dimensionality:</u> samples << predictors
- <u>Statistical significance:</u> F-test undefined[5]
- <u>Validation:</u> Ensure biological relevance

Goal

Build **<u>comprehensive pipeline for causal network inference</u>** that handles high dimensionality, maintains statistical significance, and validates on external biological data.

Prior Approaches

Previous methods either do 1) low-dimensional fit or 2) high-dimensional fit missing an effective statistical null or biological validation

	Mukhopadhyay 2007[1], Tam 2012[2], …	Lozano 2009[3], Shojaie 2010[4], Yao 2015[5], …	Our Work
High-dimensional causal fit	✗	✓	✓
Statistical significance	✓	~	✓
External validation	✗	~	✓

Vector Autoregression

Based on Granger Causality[6] principle:
X → Y if including past values of X helps to predict Y

- Fast, effective, flexible lags

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

$$H_0 : \beta_i = 0 \text{ for all } i$$

High-Dimensional Fit

Fit all causes simultaneously and regularize

$$Y_t = \sum_{i=1}^{k} \alpha_i Y_{t-i} + \sum_{g \in G} \sum_{i=1}^{k} \beta_i^g X_{t-i}^g + \varepsilon_t$$

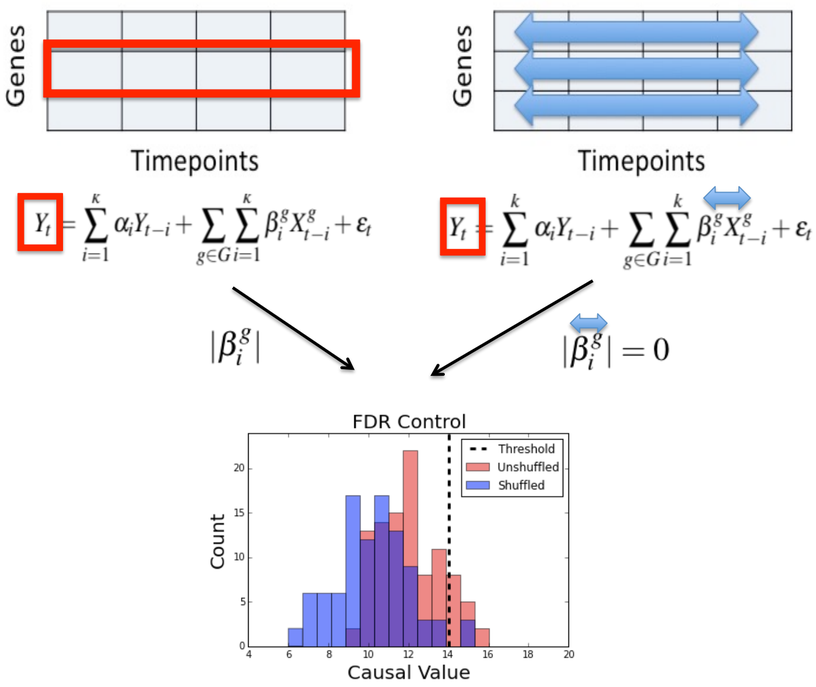
$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda f(\beta)$$

$$f_{\text{LASSO}}(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_1$$
$$f_{\text{RIDGE}}(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_2^2$$
$$f_{\text{ELASTIC}}(\boldsymbol{\beta}) = \alpha|\boldsymbol{\beta}|_1 + (1-\alpha)|\boldsymbol{\beta}|_2^2$$

$$H_0 : \beta_i^g = 0 \text{ for given } g \in G.$$

Statistical Significance

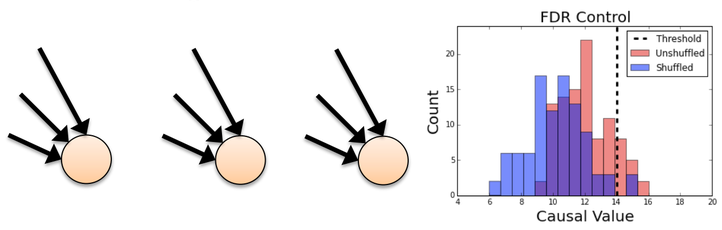
Use coefficients fit over permuted data for null



False Discovery Control

2 types of permutation FDR.

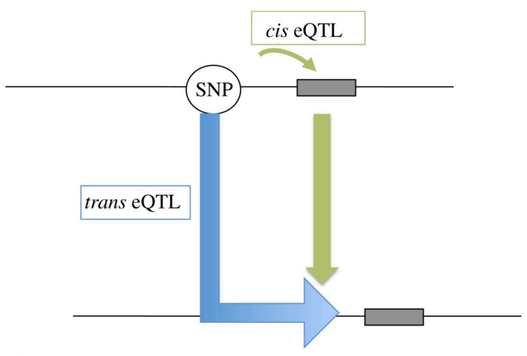
- Global FDR: global threshold over all coefficients



- Local FDR: threshold specific to effect gene
  - Finds networks with more consistent sizes over different settings (penalties, lag)

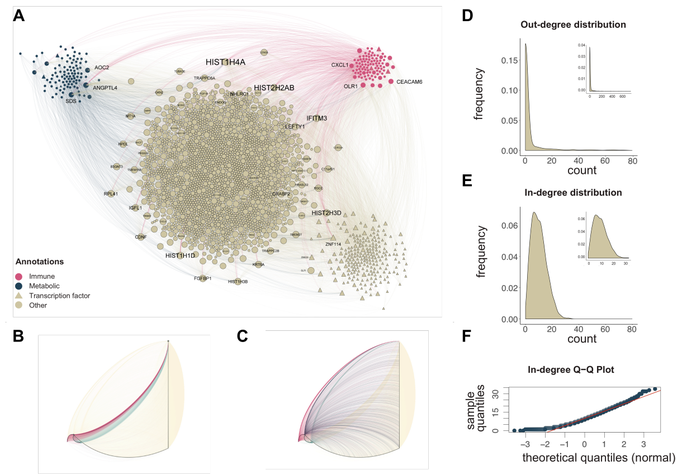
Validation

- X → Y means X, Y not independent
- Use Association test between Y's expression and SNP affecting X's expression in lung Genotype Tissue Expression Data

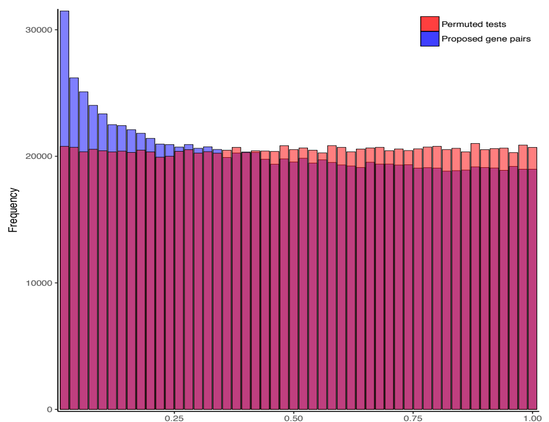


Results

- 27,781 edge network
- 617 causal genes, 2744 effect genes
- Power-law out-degree distribution
- Normal in-degree distribution



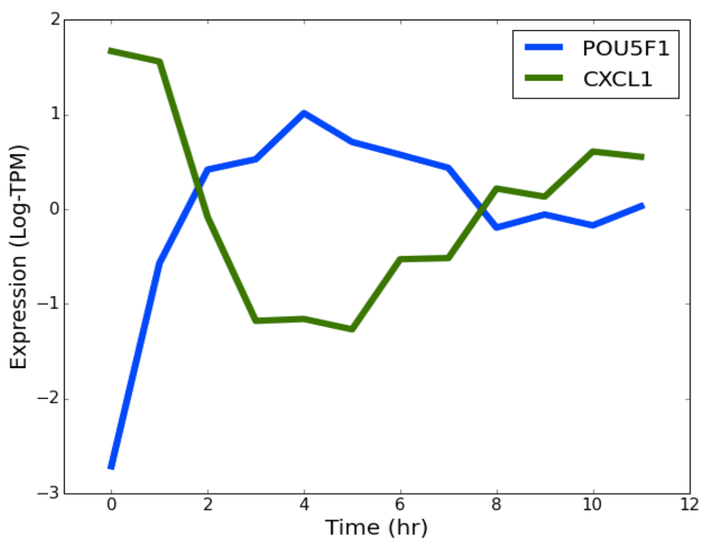
- 280 validated edges (FDR 0.2)
- 81 edges in lung



- Enrichment for Immune-Causal and Metabolic-Causal Edges

Network Edge Statistics			
Edge Type	**Total**	**%**	**Odds Ratio**
TF-Causal	1931	7	0.8
TF-Effect	2393	8.6	1.1
Immune-Causal	**2119**	**7.6**	**2**
Immune-Effect	1047	3.8	1
Metabolic-Causal	**2271**	**8.2**	**2**
Metabolic-Effect	1211	4.4	1

- Strong repressive relation between Transcription Factor POU5F1 and Immune-related gene CXCL2



Conclusion & Future Work

- We have developed an improved pipeline for causal network inference that validates on external data.
- <u>Extension 1:</u> Use the network to suggest genes for perturbation in follow-up experiments
- <u>Extension 2:</u> Extend model to incorporate causal relations learned from data under multiple conditions

References
1. Nitai D. Mukhopadhyay and Snigdhansu Chatterjee. Causality and pathway search in microarray time series experiment. Bioinformatics , 23(4):442, 2007.
2. 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.
3. 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
4. Ali Shojaie and George Michailidis. Discovering graphical granger causality using the truncating lasso penalty. Bioinformatics , 26(18):i517–i523, 09 2010.
5. Shun Yao, Shinjae Yoo, and Dantong Yu. Prior knowledge driven granger causality analysis on gene regulatory network discovery. BMC Bioinformatics , 16:273, 2015.
6 C.W.J. Granger. Testing for causality. Journal of Economic Dynamics and Control , 2:329 – 352, 1980.