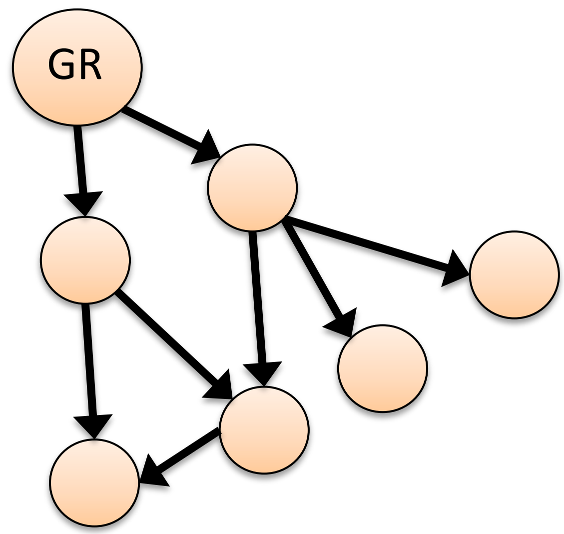
A Robust Causal Network Pipeline for Gene Expression Time Series



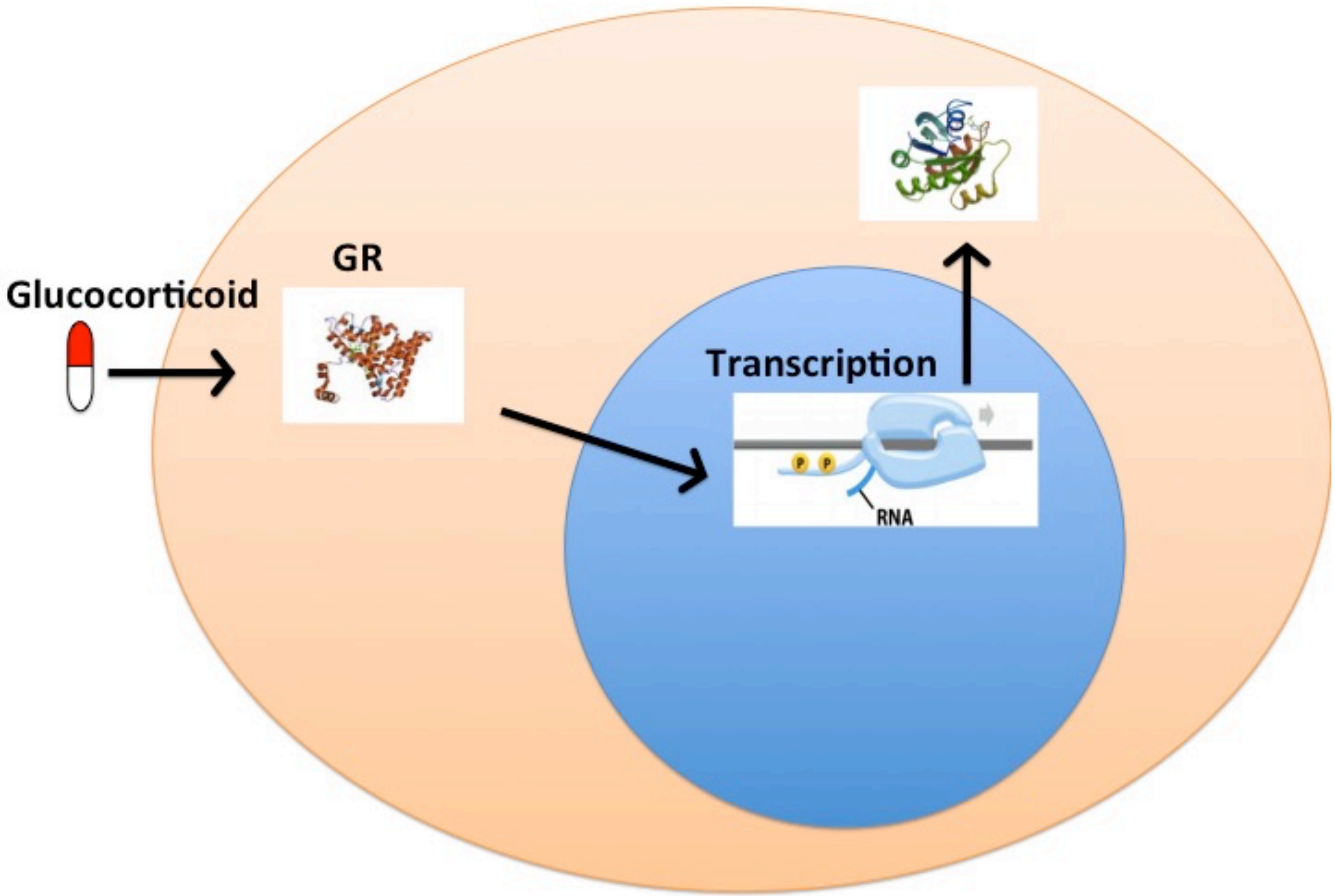
Jonathan Lu, Bianca Dumitrascu, Brian Jo, Ian McDowell
Prof. Barbara Engelhardt

4/26/17

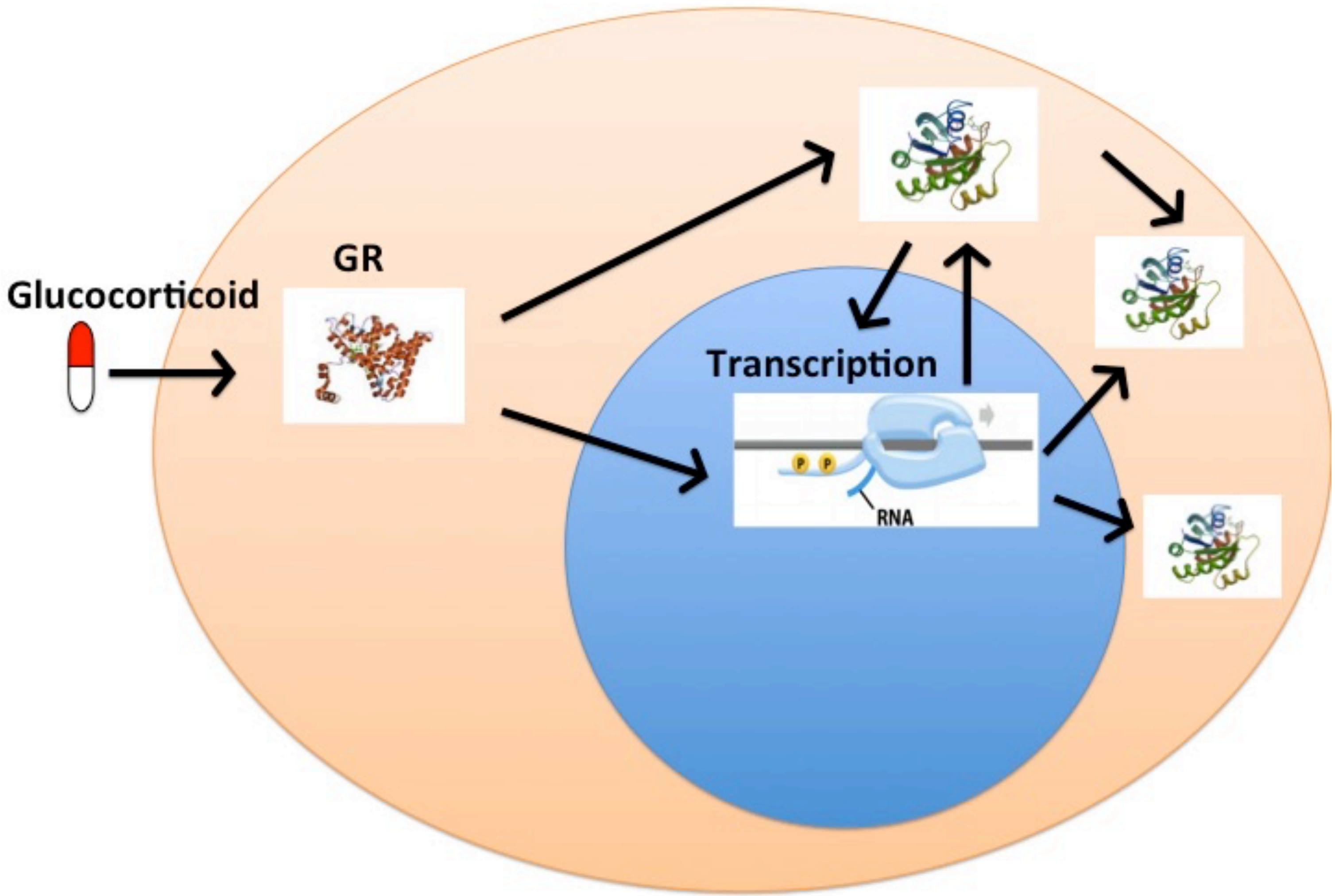
Goal: Understand Glucocorticoid Response

- Immunosuppressant drugs
  - Asthma, Eczema
  - Anti-inflammatory
  - Metabolic side effects
- <u>Complex genetic response</u>

Glucocorticoid Transcriptional Response is Complex

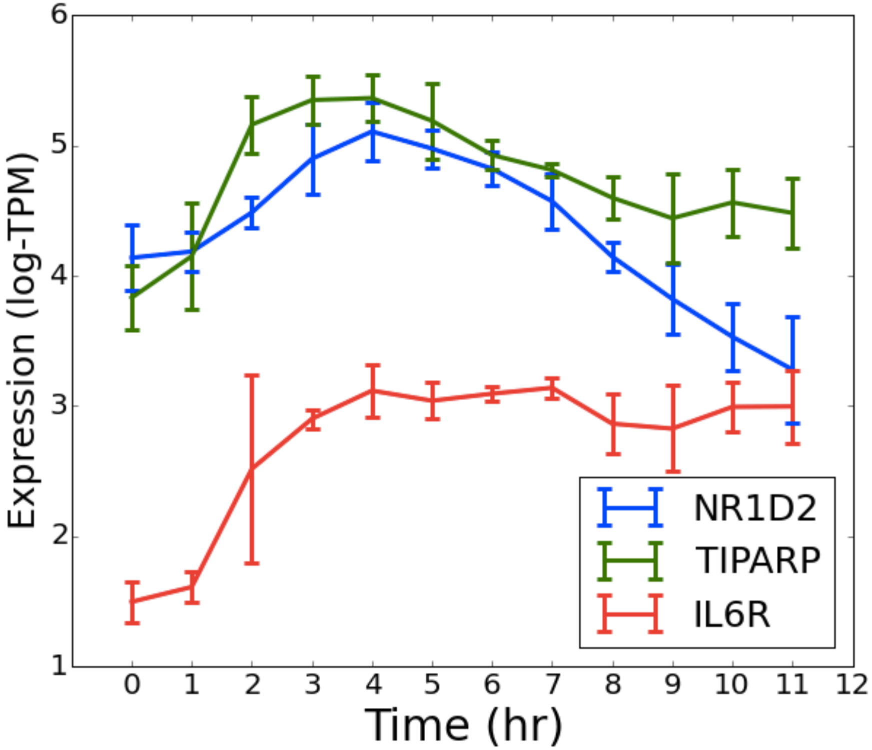


Glucocorticoid Transcriptional Response is Complex



Data

- Stimulated lung cell lines
- ~3-4 replicates/timepoint
- ~3k differentially expressed genes (~18k total)

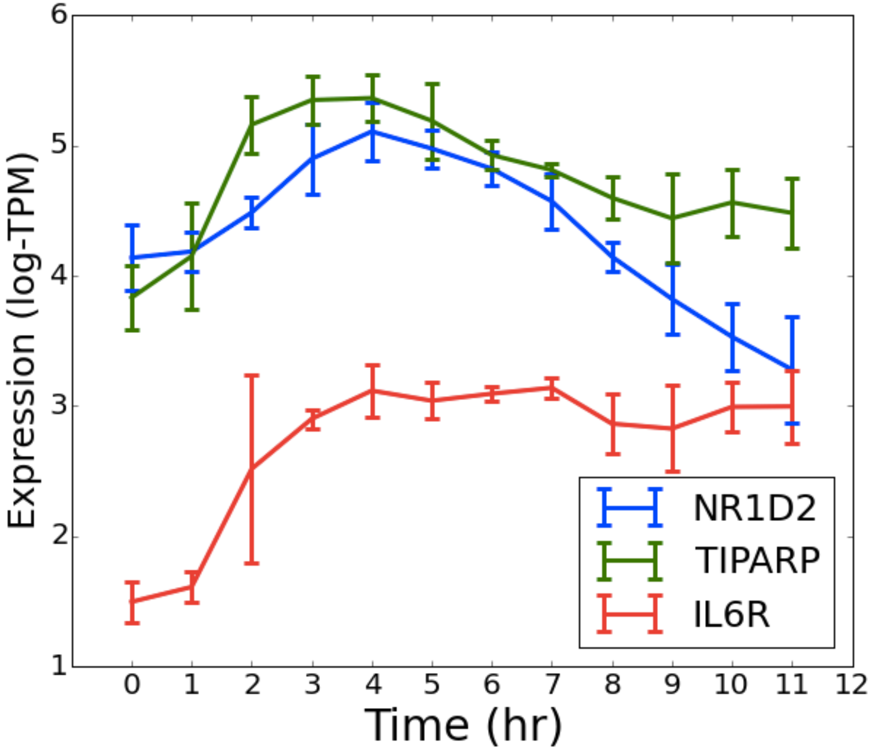


Data

- Stimulated lung cell lines
- ~3-4 replicates/timepoint
- ~3k differentially expressed genes (~18k total)

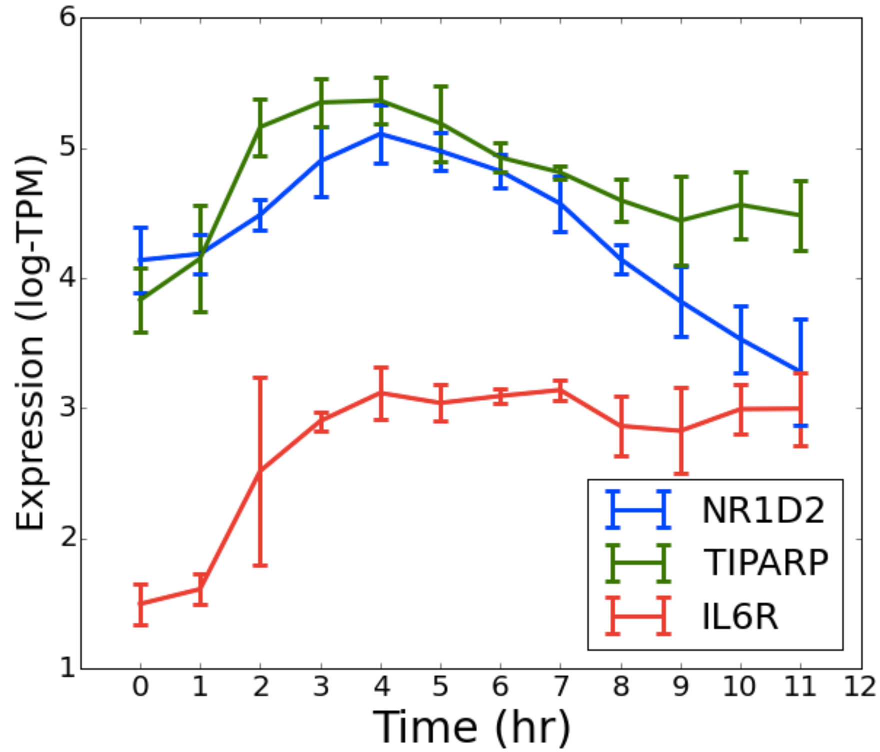
Challenges

- Causal Inference
- High Dimensionality
- Statistical Significance



Data

- Stimulated lung cell lines
- ~3-4 replicates/timepoint
- ~3k differentially expressed genes (~18k total)



Challenges

- Causal Inference
- High Dimensionality
- Statistical Significance

Goal

1. Build robust causal pipeline that overcomes challenges
2. Validate causal networks using external data

Previous Work

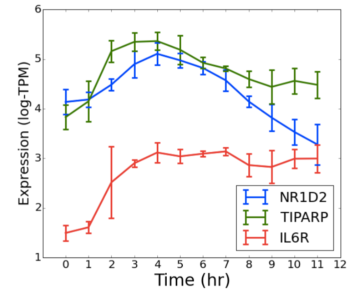
	Mukhophadyay 2007, Tam 2012, …	Lozano 2009, Shojaie 2010, Yao 2015,…	Our Work
High-Dimensional Causal Fit	✗	✓	✓
Statistical Significance	✓	~	✓
External Validation	✗	~	✓

Approach

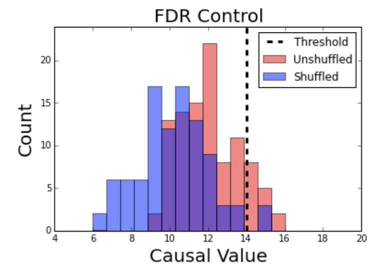
	Our Work
High-Dimensional Causal Fit	Regularized Vector Autoregression
Statistical Significance	Statistical Null and False Discovery Control from Permuted Data
External Validation	Association Test in Lung Gene Expression Data

Pipeline Workflow

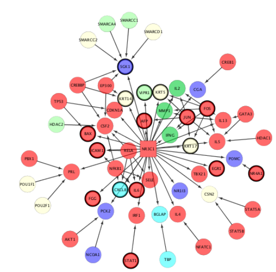
Preprocess Data



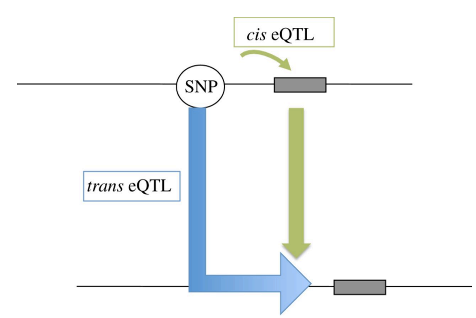
Apply Causality Tests



Build Significant Network



Validation



Challenge: Causal Inference

- Vector Autoregression (VAR)
  - Granger Causality: X → Y if including past values of X helps to predict Y
    - Fast, effective, flexible lags

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

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

$$H_A : \beta_i \neq 0 \text{ for some } i$$

Yao 2015

Challenge: High Dimension

- Fit all causes simultaneously and regularize.

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

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

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

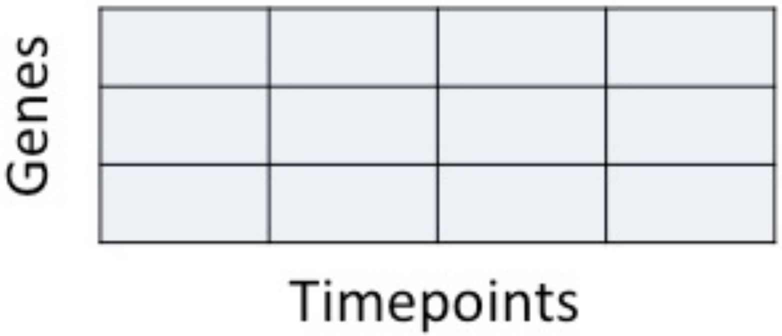
$H_0 : \beta_i^g = 0$ for given $g \in G$.

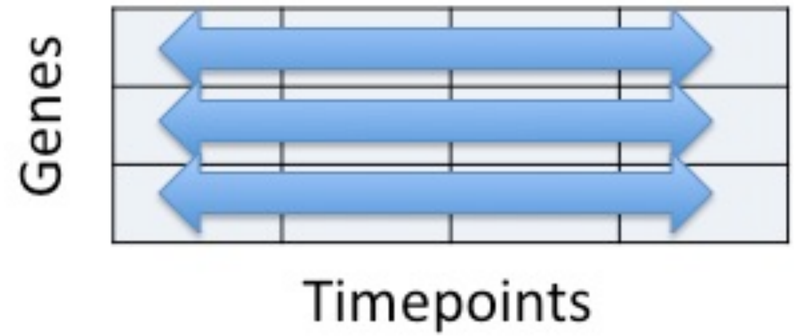
$H_A : \beta_i^g \neq 0$ for some given $g \in G$

Challenge: Statistical Significance

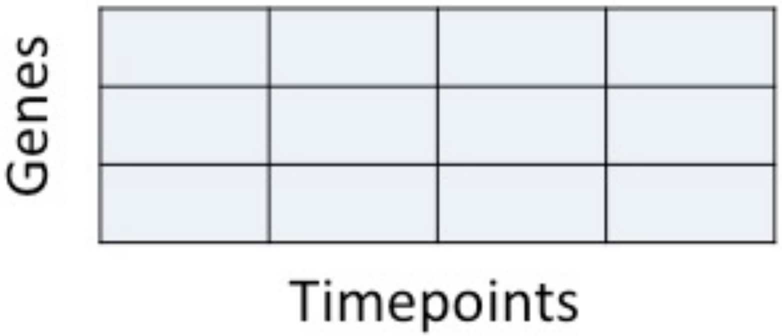
- Statistical Test is undefined for high dimension
- FDR Control is difficult for p-values
  - Around 10^8 tests
- Solution:
  - Use shuffled data as null
  - Use Coefficients instead of p-values

Challenge: Statistical Significance

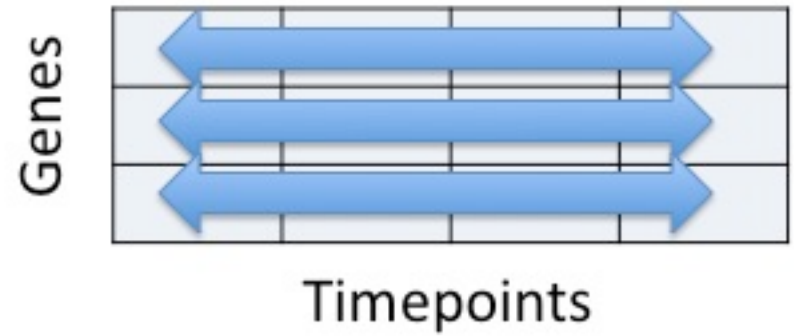




Challenge: Statistical Significance

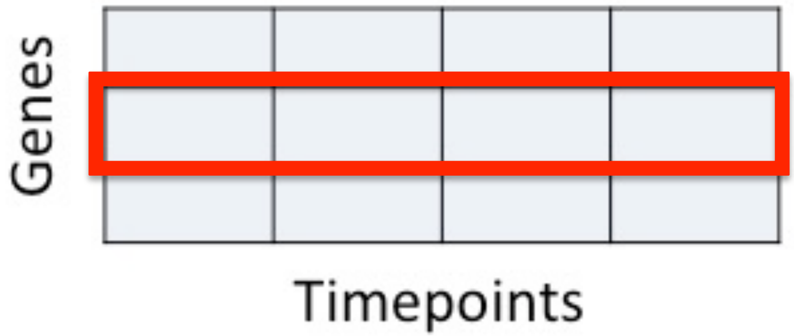


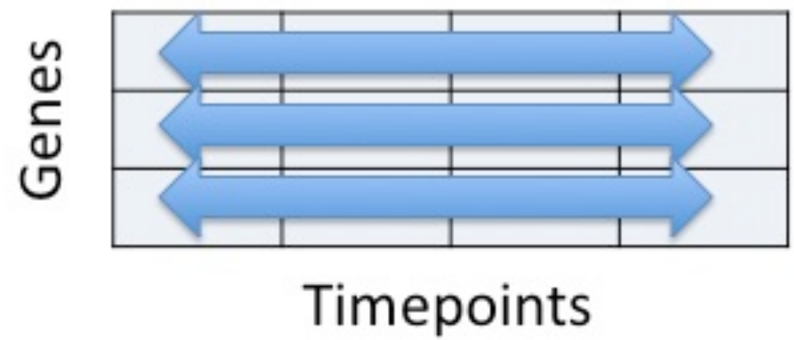
$$Y_t = \sum_{i=1}^{k} \alpha_i Y_{t-i} + \sum_{g \in G} \sum_{i=1}^{k} \beta_i^g X_{t-i}^g + \varepsilon_t$$



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

Challenge: Statistical Significance

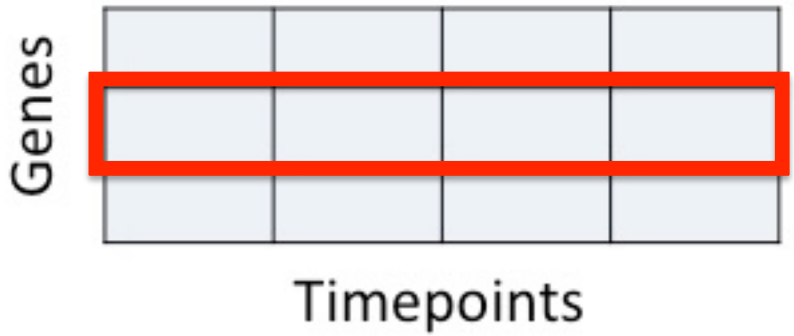


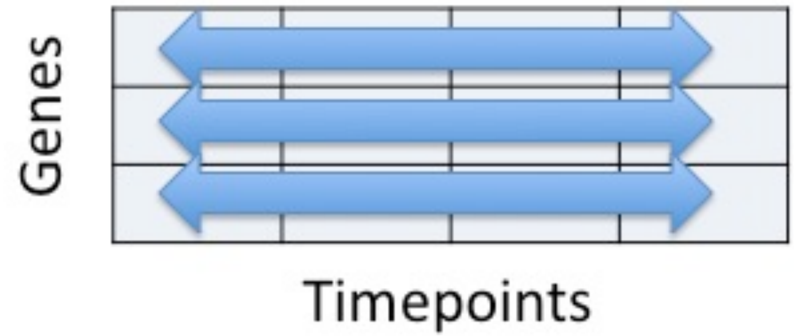


$$\boxed{Y_t} = \sum_{i=1}^{k} \alpha_i Y_{t-i} + \sum_{g \in G} \sum_{i=1}^{k} \beta_i^g X_{t-i}^g + \varepsilon_t$$

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

Challenge: Statistical Significance





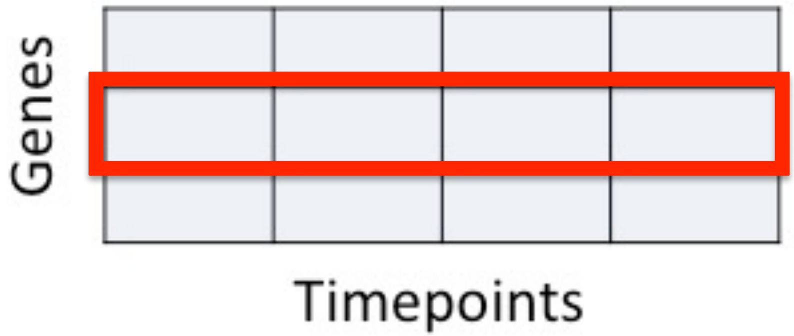
$$\boxed{Y_t} = \sum_{i=1}^{k} \alpha_i Y_{t-i} + \sum_{g \in G} \sum_{i=1}^{k} \beta_i^g X_{t-i}^g + \varepsilon_t$$

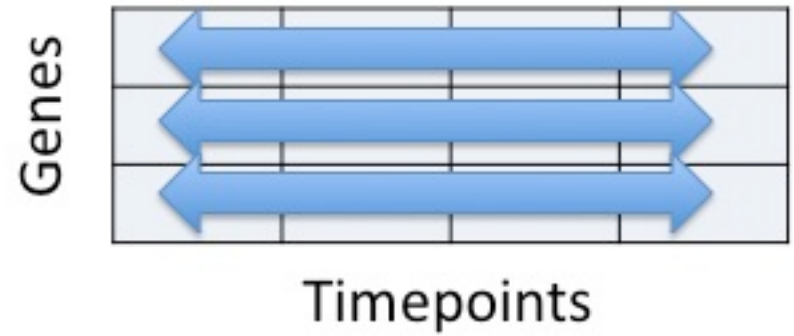
$$|\beta_i^g|$$

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

$$|\beta_i^g| = 0$$

Challenge: Statistical Significance



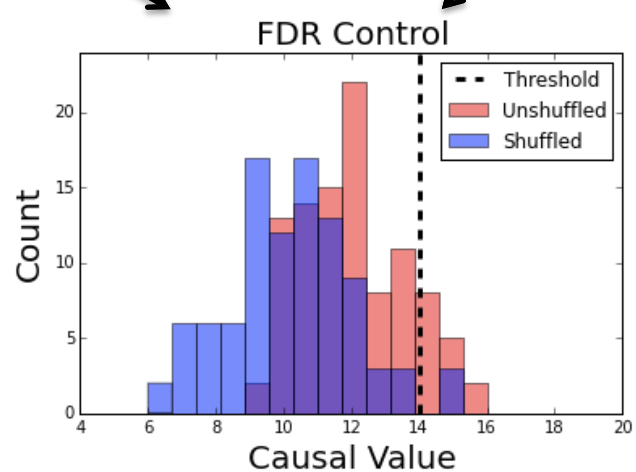


$$\boxed{Y_t} = \sum_{i=1}^{k} \alpha_i Y_{t-i} + \sum_{g \in G} \sum_{i=1}^{k} \beta_i^g X_{t-i}^g + \varepsilon_t$$

$$|\beta_i^g|$$

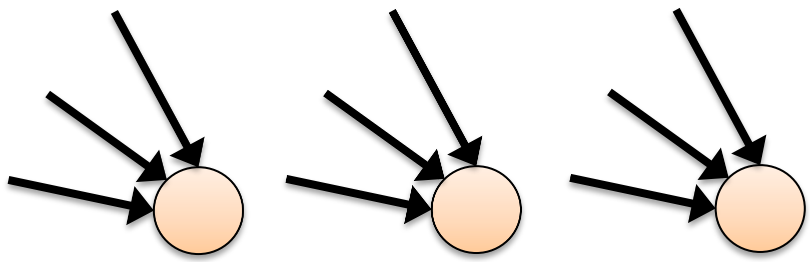
$$\boxed{Y_t} = \sum_{i=1}^{k} \alpha_i Y_{t-i} + \sum_{g \in G} \sum_{i=1}^{k} \beta_i^g X_{t-i}^g + \varepsilon_t$$

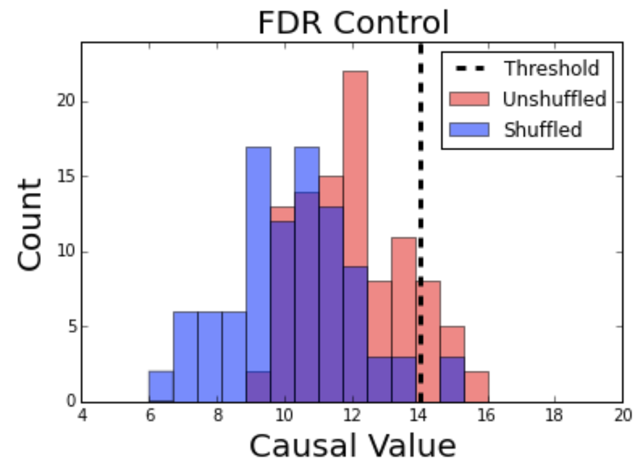
$$|\beta_i^g| = 0$$



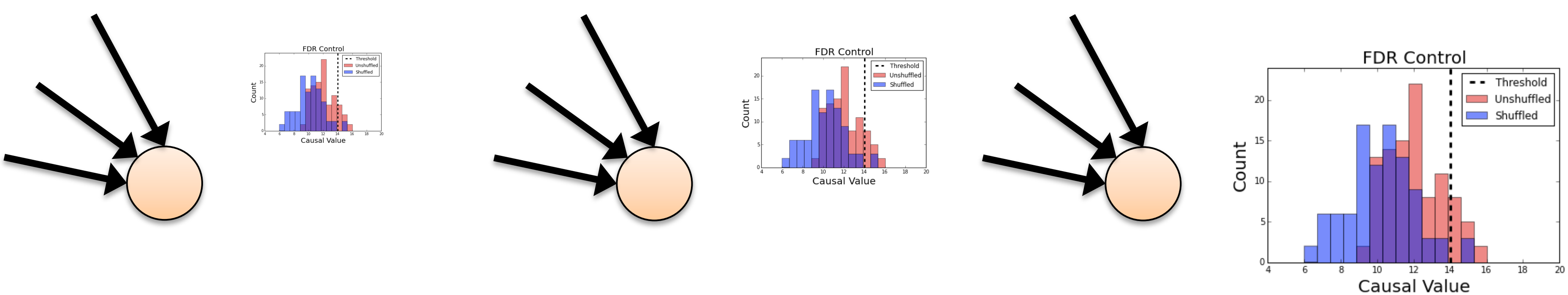
Challenge: Statistical Significance

- Global FDR



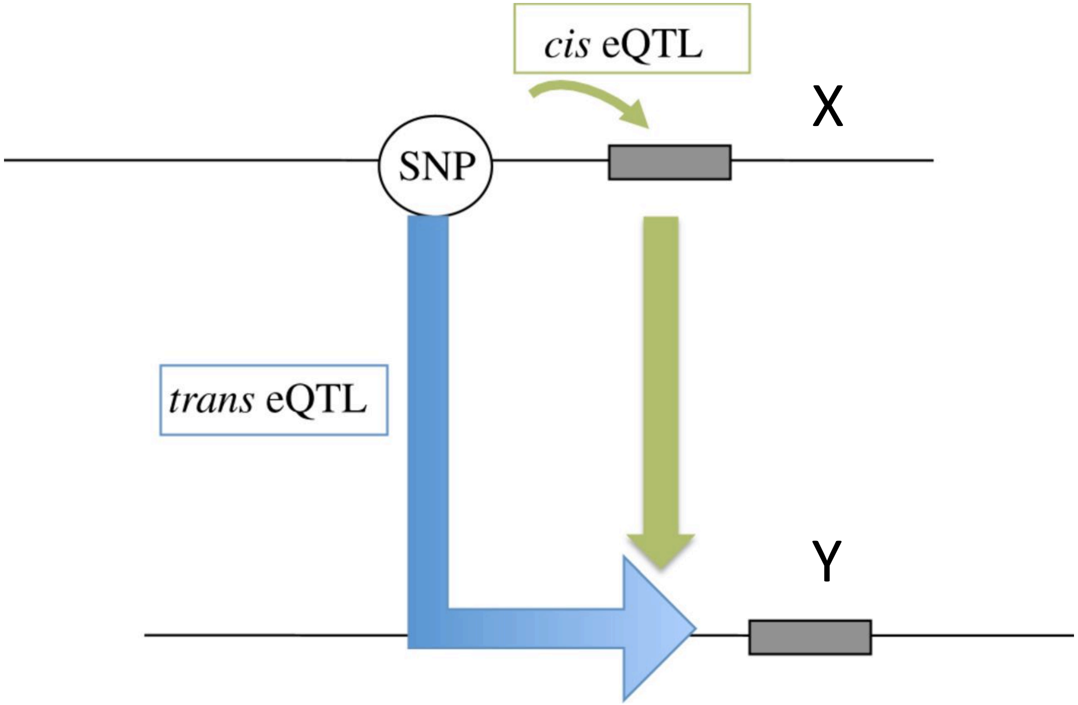


- Local FDR

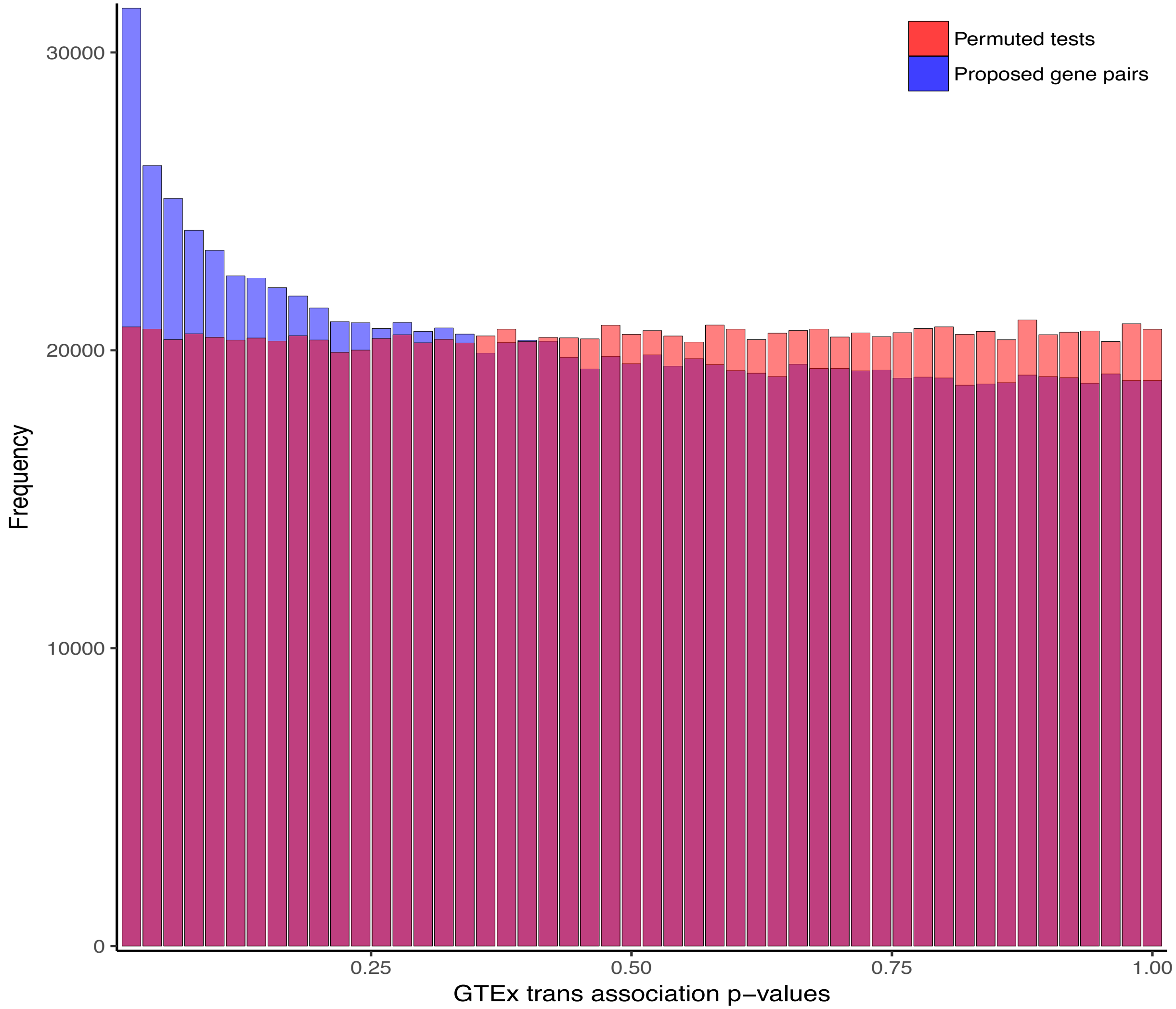


Validation

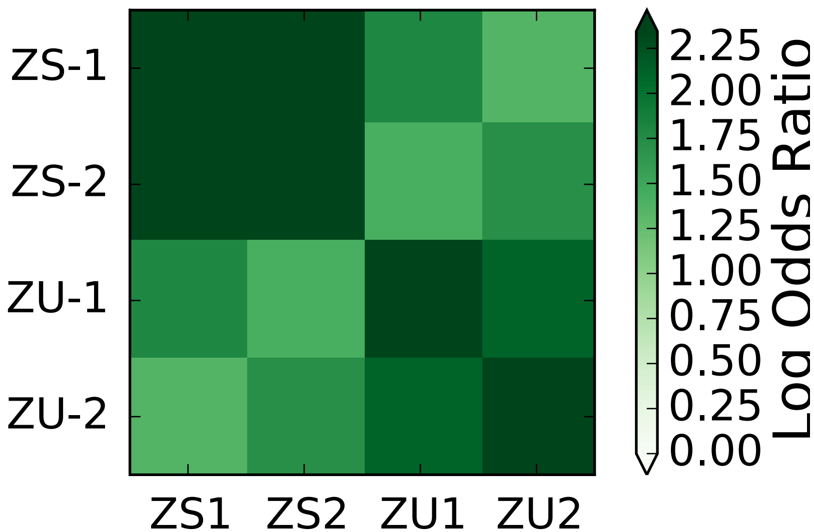
- Validate X → Y by:
  - Trans-eQTL
    - Association test in GTEx Data

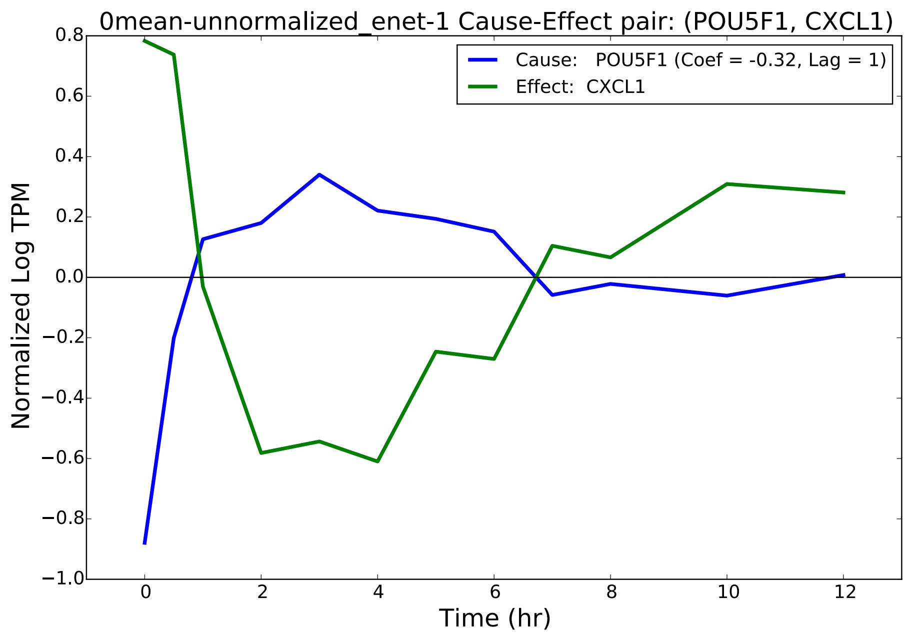


Validation


Permuted tests
Proposed gene pairs
Frequency
30000
20000
10000
0
0.25
0.50
0.75
1.00
GTEx trans association p−values

Results





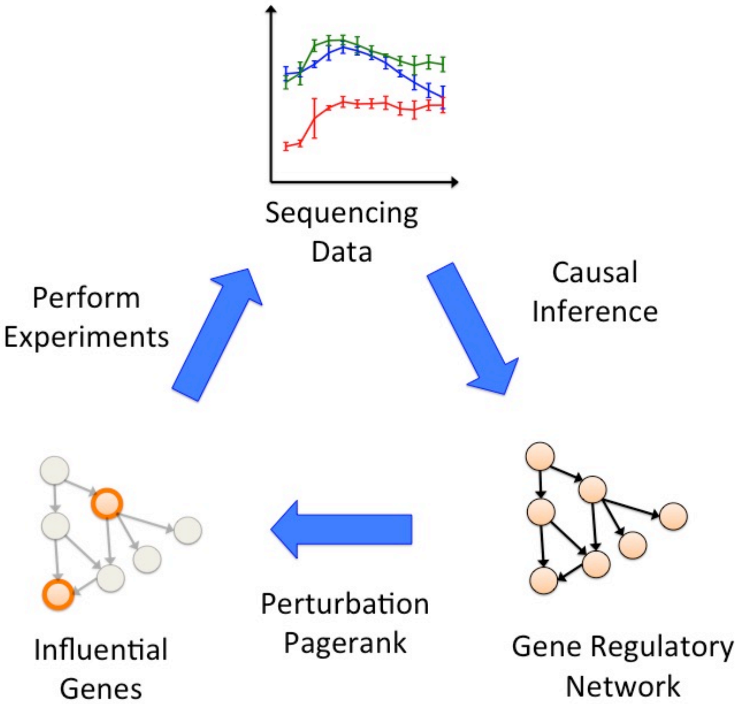
- ZS, ZU: Normalization Types
- 1, 2: Lag Numbers

Conclusion

1. We use Vector Autoregression to build causal networks from gene expression time series.
2. We address challenges of dimensionality and statistical significance.
3. Our causal networks are robust, validating on external data and uncovering strong signals.

Future work

1. Iteratively suggest new experiments and refine networks
   – Perturbation Pagerank
2. Learn causal relations under different perturbations



Acknowledgments

Engelhardt Lab (Princeton)

Bianca Dumitrascu

Brian Jo

Barbara Engelhardt

Ari, Derek, Allison, Greg, Izzy, ...

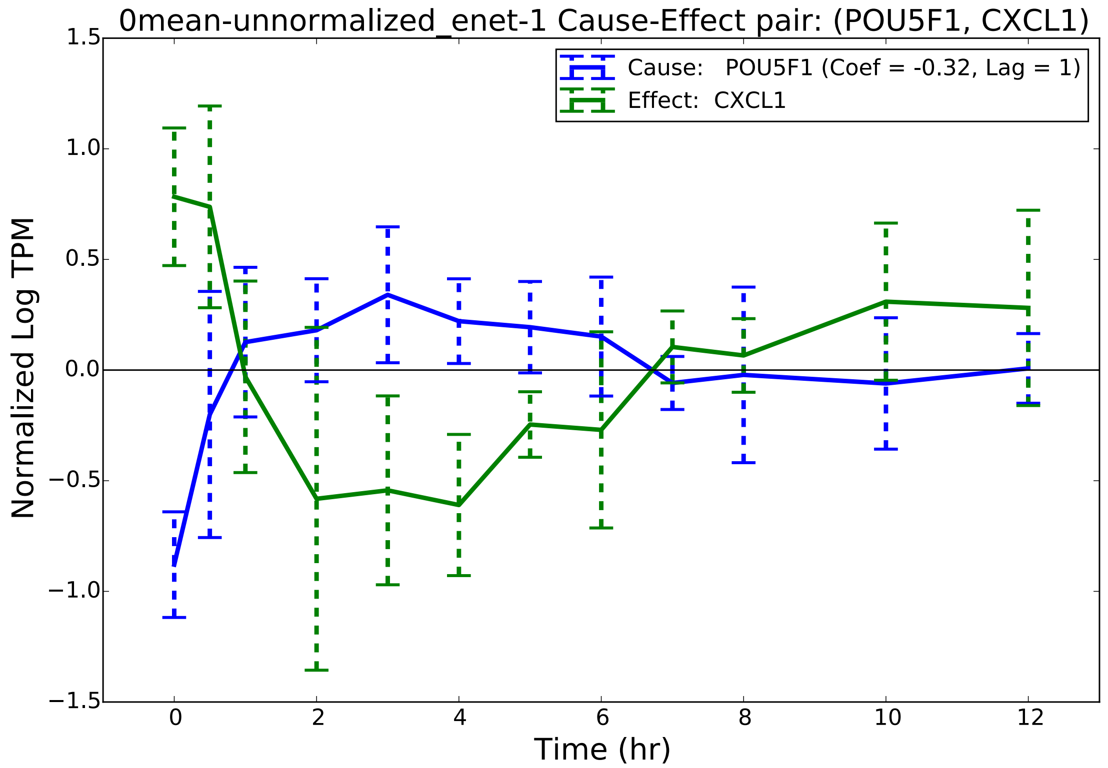
Reddy Lab (Duke)

Ian McDowell

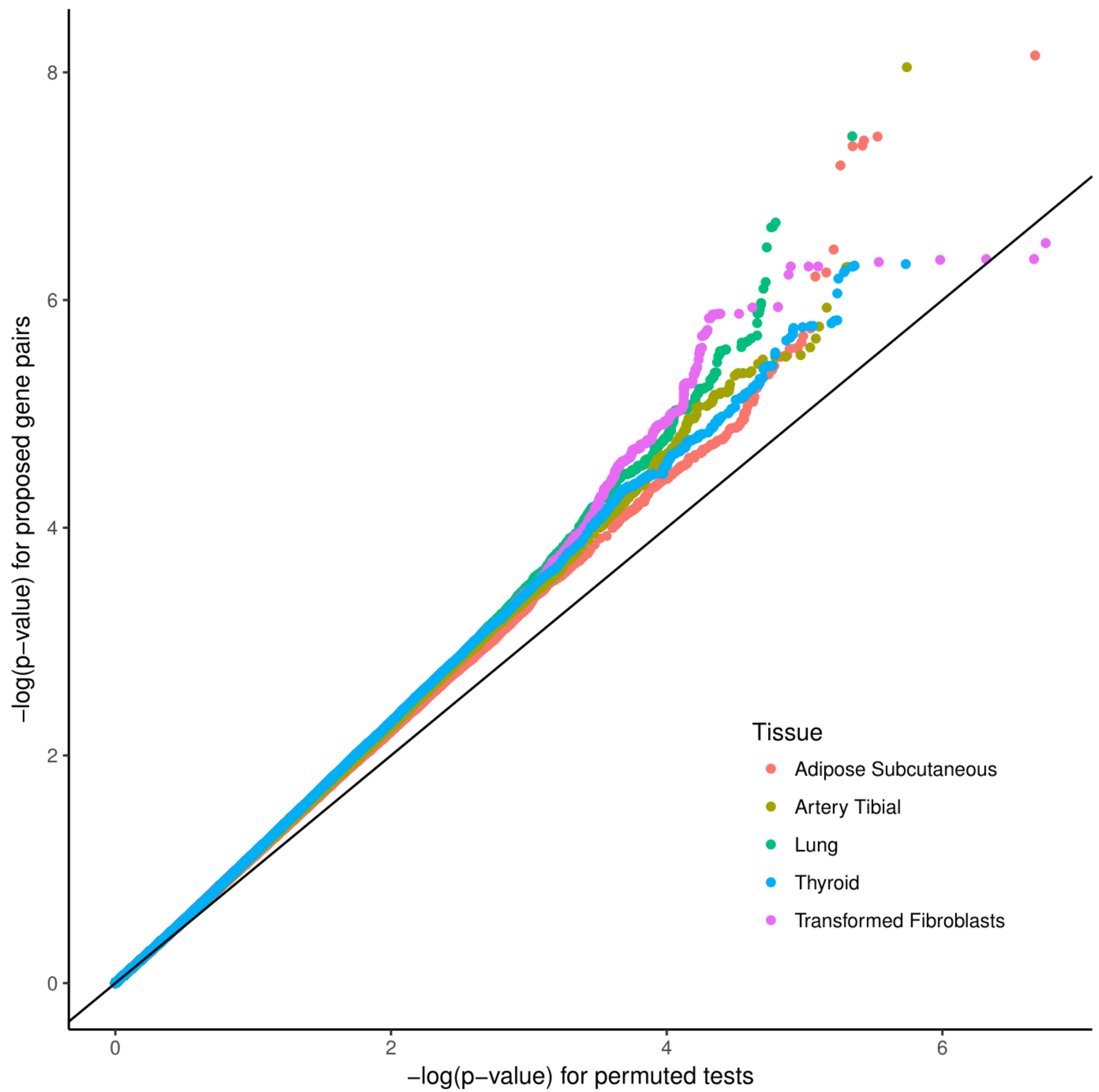
Tim Reddy

Data collection Team

Extra


0mean-unnormalized_enet-1 Cause-Effect pair: (POU5F1, CXCL1)
Cause: POU5F1 (Coef = -0.32, Lag = 1)
Effect: CXCL1
Normalized Log TPM
Time (hr)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
2
4
6
8
10
12

Validation



−log(p−value) for proposed gene pairs
−log(p−value) for permuted tests
0
2
4
6
8
Tissue
Adipose Subcutaneous
Artery Tibial
Lung
Thyroid
Transformed Fibroblasts