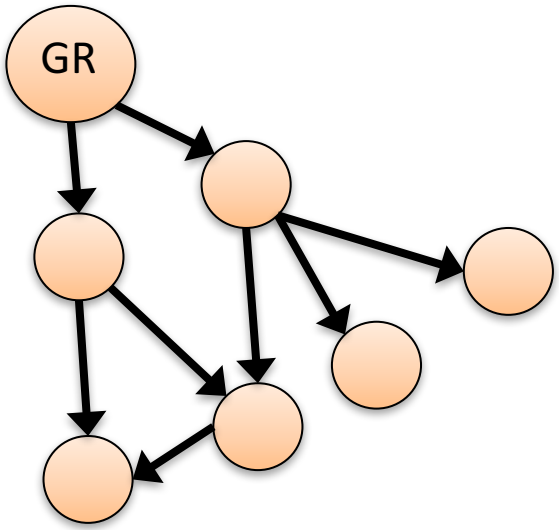
Robust Causal Network Pipeline for Gene Expression Time Series



Jonathan Lu

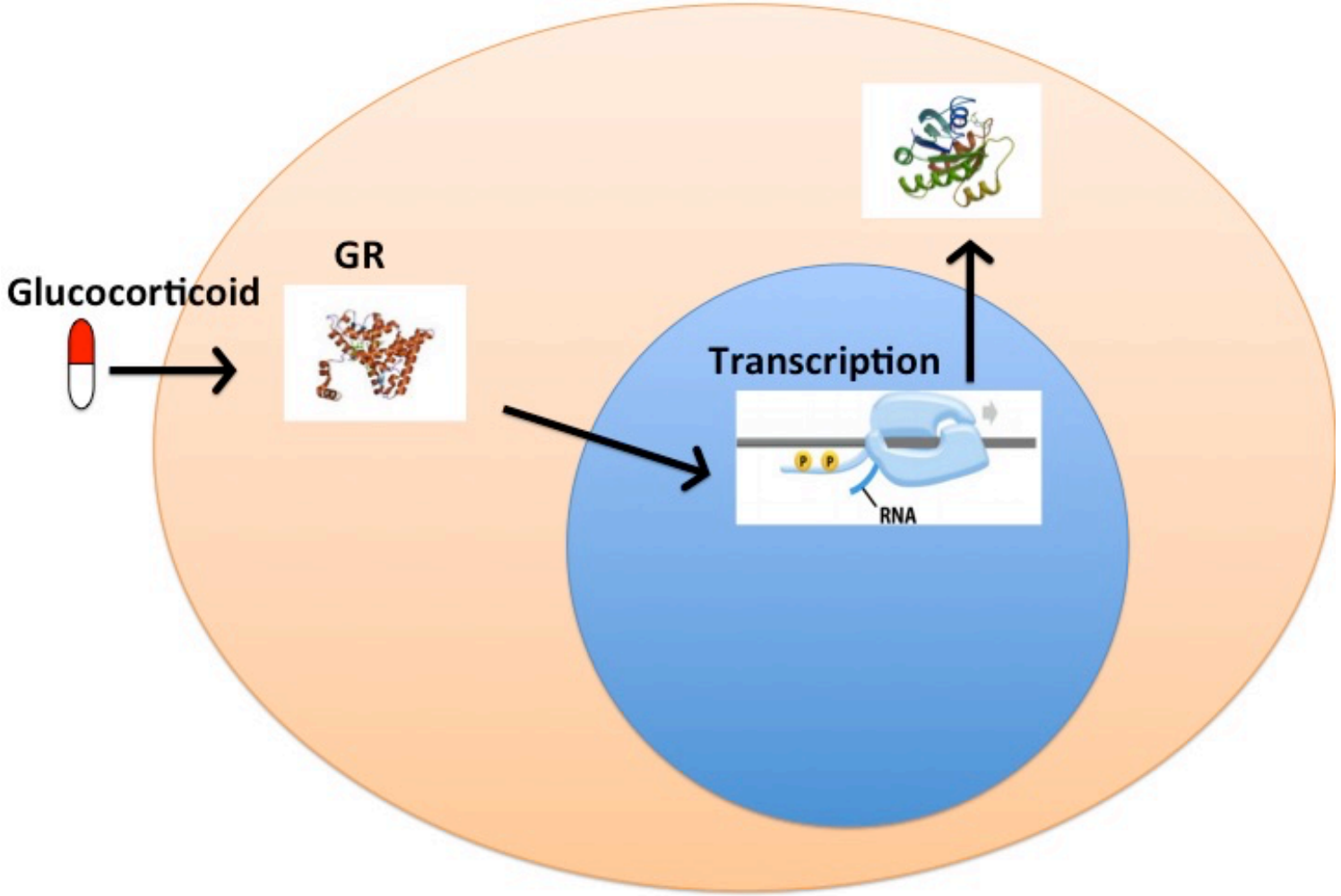
Bianca Dumitrascu, Prof. Barbara Engelhardt

4/25/2018

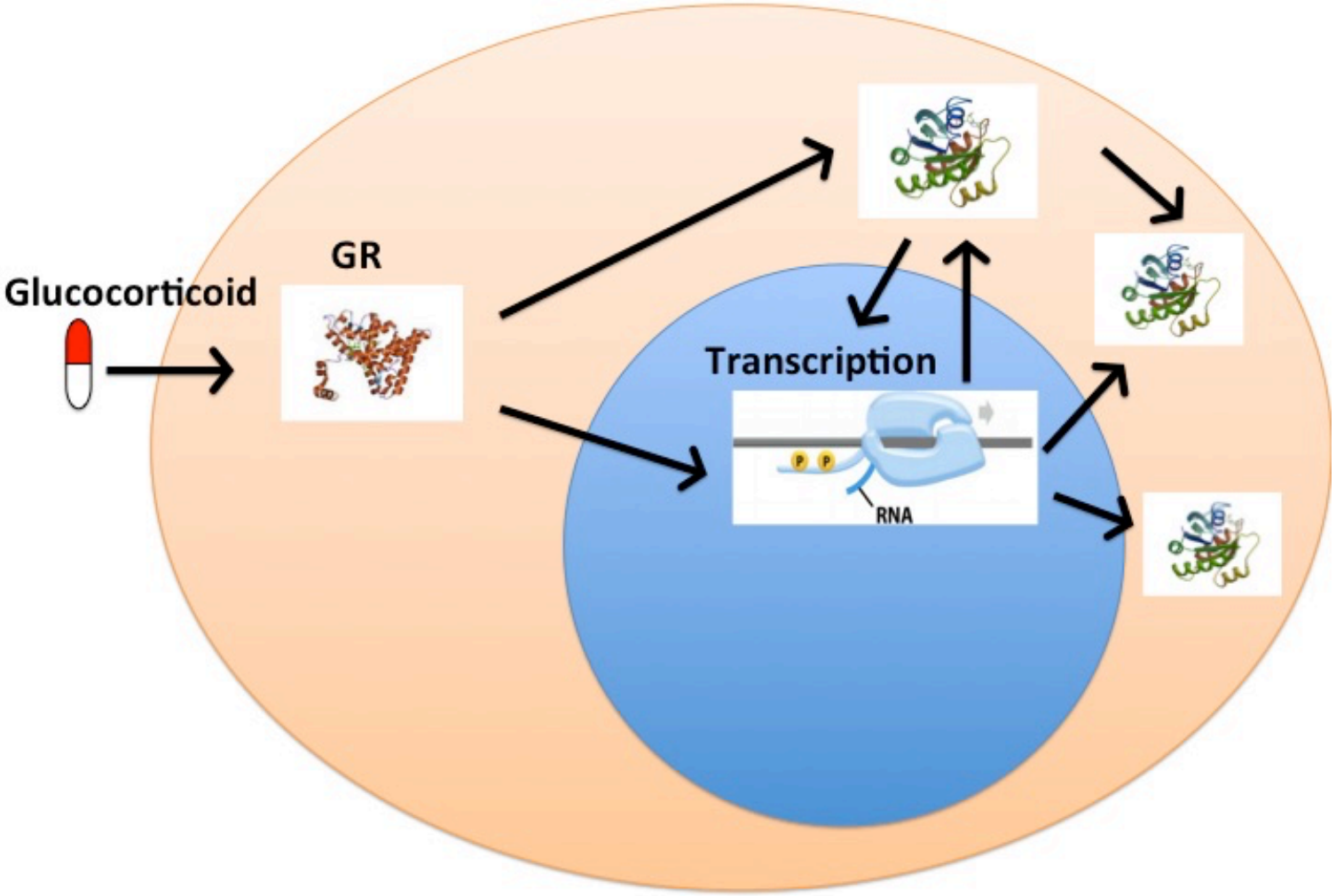
Goal: Understand Glucocorticoid Response

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

Glucocorticoid Transcriptional Response is Complex

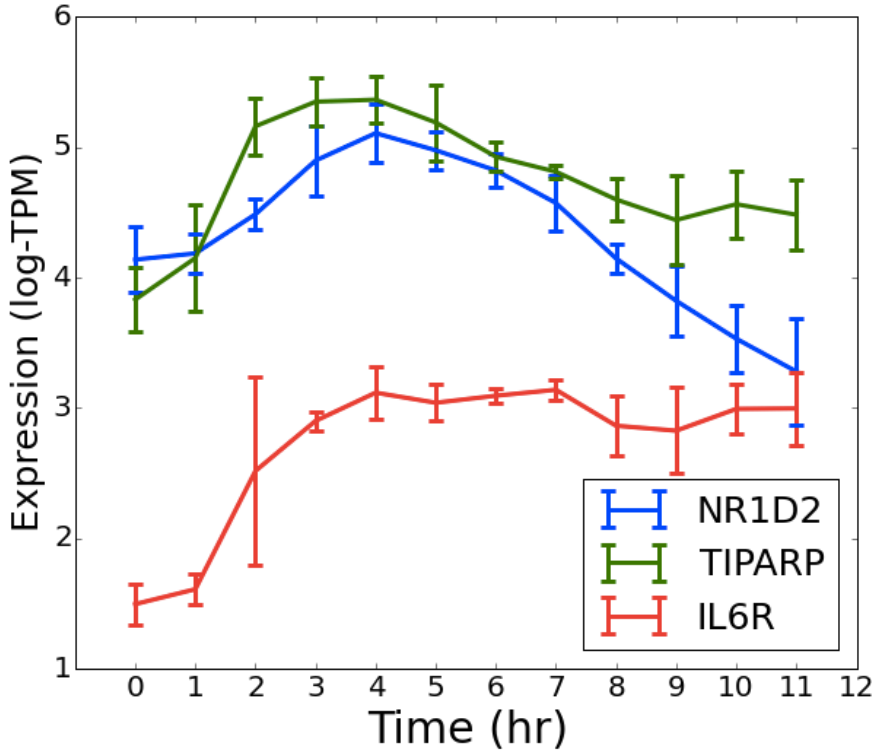


Glucocorticoid Transcriptional Response is Complex



Data

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

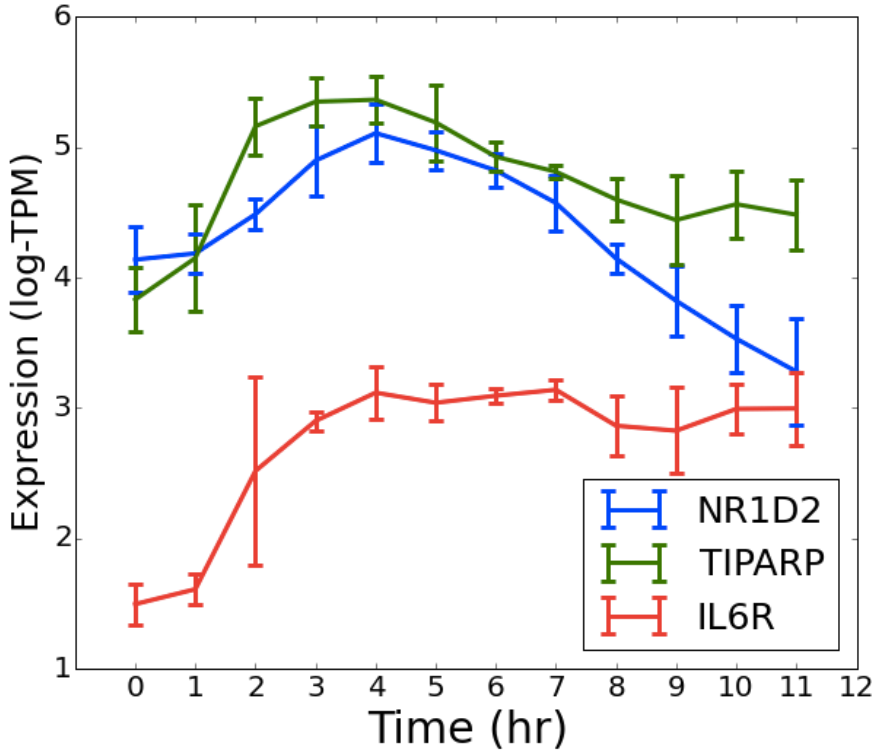


Data

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

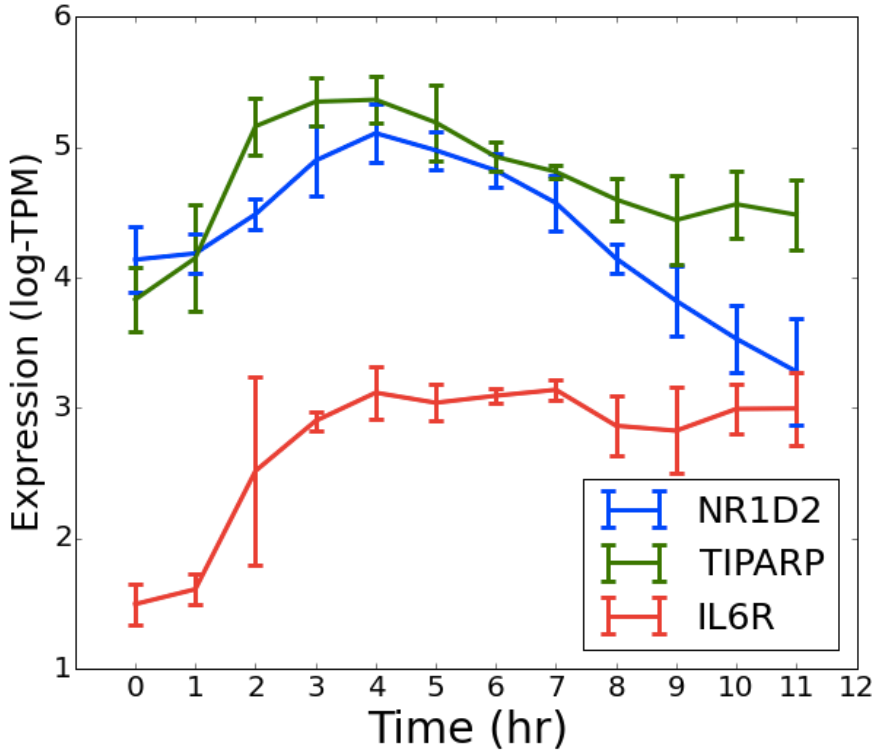
Challenges

- Causal Inference
- Statistical Significance
- Scalability



Data

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



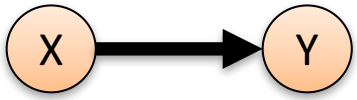
Challenges

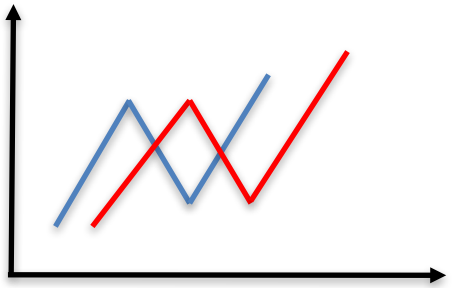
- Causal Inference
- Statistical Significance
- Scalability

Goal

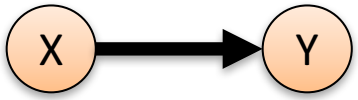
1. Build causal network pipeline to overcome challenges
2. Validate method on community benchmarks, real data

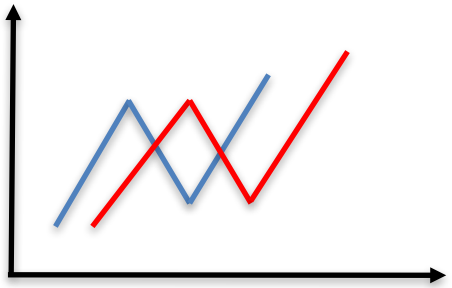
What is a causal edge?



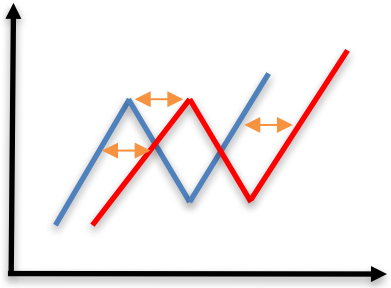


What is a causal edge?



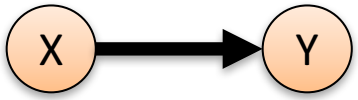


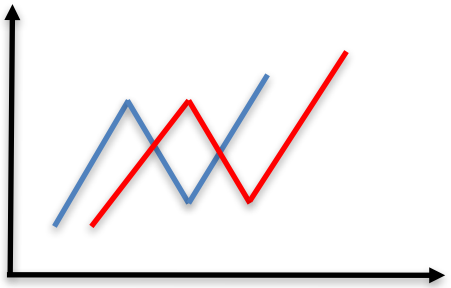
Mutual Information



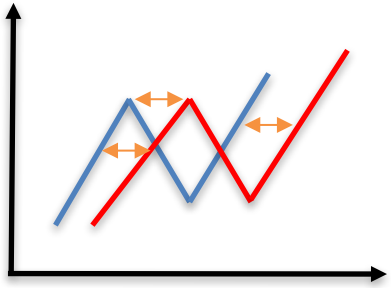
$$I^k(X,Y) = -\sum_{t=k}^{T} P(X_{t-k}, Y_t) \log \frac{P(X_{t-k}, Y_t)}{P(X_{t-k})\, P(Y_t)}$$

What is a causal edge?

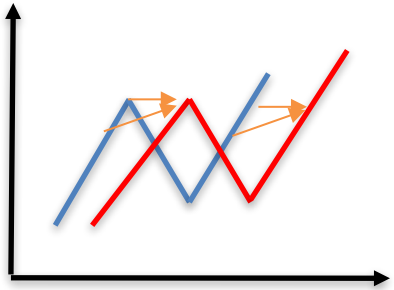




Mutual Information



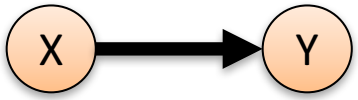
Vector Autoregression

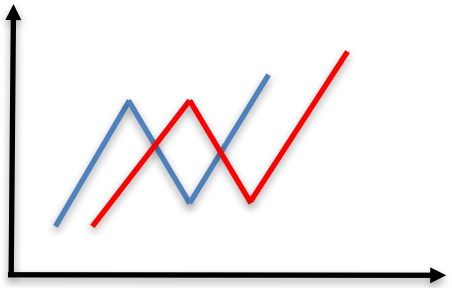


$$I^k(X,Y) = -\sum_{t=k}^{T} P(X_{t-k},Y_t)\log\frac{P(X_{t-k},Y_t)}{P(X_{t-k})\,P(Y_t)}$$

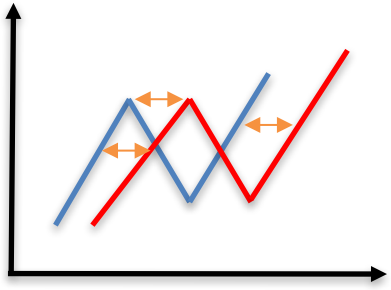
$$Y_t = \sum_{i=1}^{K} a_i Y_{t-i} + \sum_{i=1}^{K} b_i X_{t-i} + \varepsilon_t$$

What is a causal edge?

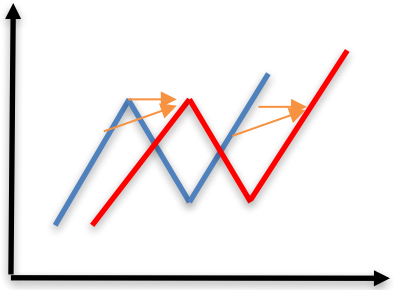




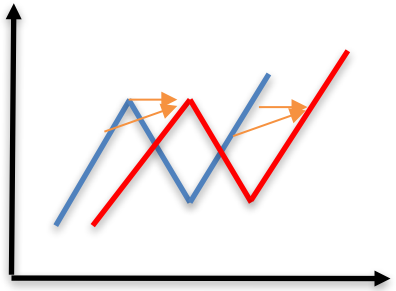
Mutual Information

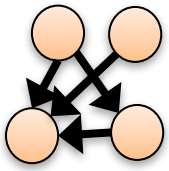


Vector Autoregression



Dynamic Bayesian Network



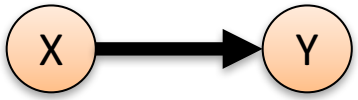


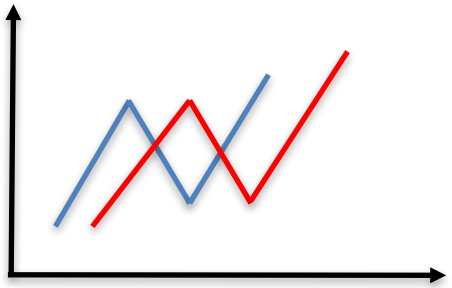
$$I^k(X,Y) = -\sum_{t=k}^{T} P(X_{t-k},Y_t)\log\frac{P(X_{t-k},Y_t)}{P(X_{t-k})\,P(Y_t)}$$

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

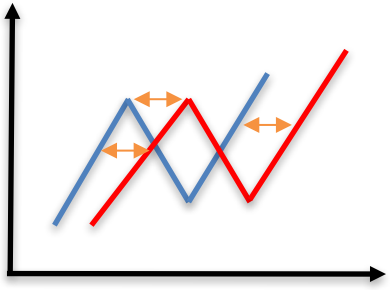
$$P(X_1 \ldots X_1^n \ldots X_T^n) = P(X_1)\prod_{t=2}^{T}\prod_{i=1}^{n} P(X_t^i \mid pa(X_t^i))$$

What is a causal edge?

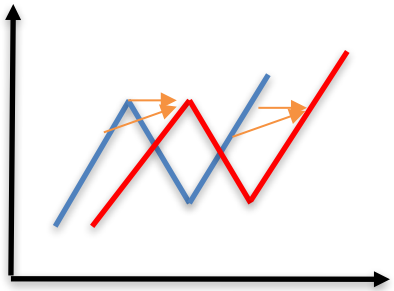




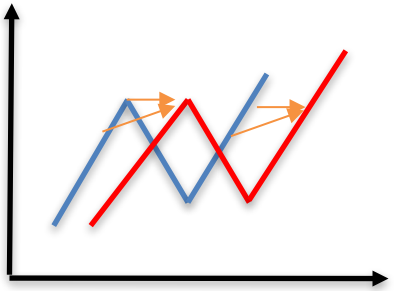
Mutual Information

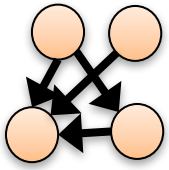


Vector Autoregression

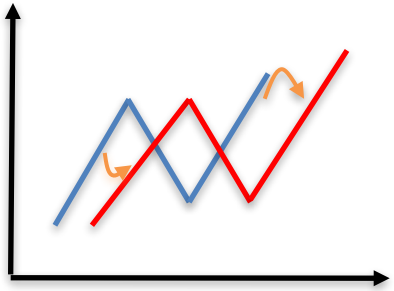


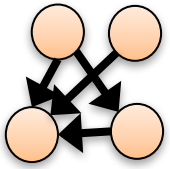
Dynamic Bayesian Network





Gaussian Process





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

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

$$P(X_1 \dots X_1^n \dots X_T^n) = P(X_1)\prod_{t=2}^{T}\prod_{i=1}^{n} P(X_t^i \mid pa(X_t^i))$$

$$X_t = f\big(pa(X_t)\big), f \sim GP(m, \mathrm{k})$$

Previous Work

Feature	Mutual Information [1]	Vector Autoregression[2]	Dynamic Bayesian Network[3]	Gaussian Process[4]
Effective	~	✗	~	✓
Scalable	✓	✓	✗	✗
Statistical Significance	✗	~	~	✗

Note: we only discuss existing methods. For example, it is possible for a future GP method to be developed that, e.g. does statistical significance.

1 *Meyer 2007, Zoppoli 2010*
2 *Opgen-Rhein 2007, Yao 2015*
3 *Hartemink 2001, Young 2013*
4 *Penfold 2015, Penfold 2012*

Previous Work

Feature	Mutual Information [1]	Vector Autoregression[2]	Dynamic Bayesian Network[3]	Gaussian Process[4]	Our Work (Vector Autoregression)
Effective	~	✗	~	✓	✓
Scalable	✓	✓	✗	✗	✓
Statistical Significance	✗	~	~	✗	✓

Note: we only discuss existing methods. For example, it is possible for a future GP method to be developed that, e.g. does statistical significance

1 *Meyer 2007, Zoppoli 2010*
2 *Opgen-Rhein 2007, Yao 2015*
3 *Hartemink 2001, Young 2013*
4 *Penfold 2015, Penfold 2012*

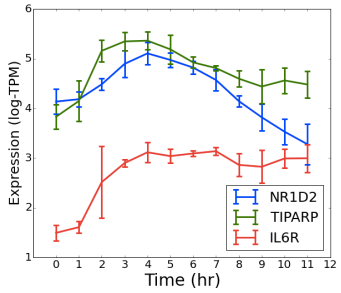
Approach

BETS: **B**ootstrap **E**lastic net regression from **T**ime **S**eries

	Our Work
Effective	Elastic Net, Bootstrap Stability Selection
Scalability	Massive Parallelization
Statistical Significance	Statistical Null and False Discovery Control from Permuted Data

Pipeline Workflow

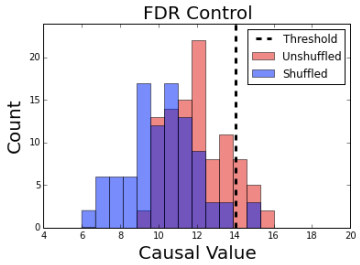
Preprocess Data



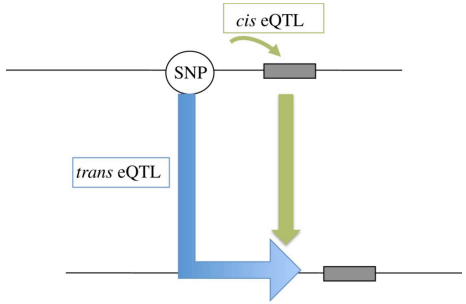
Apply Causality Tests

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

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, interpretable

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

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

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

Yao 2015

Challenge: High Dimension

- Fit all causes simultaneously and regularize.

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

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

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

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

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

Challenge: High Dimension

- Fit all causes simultaneously and regularize.

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

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

$$f_{\text{LASSO}}(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_1$$

$$f_{\text{RIDGE}}(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_2^2$$

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

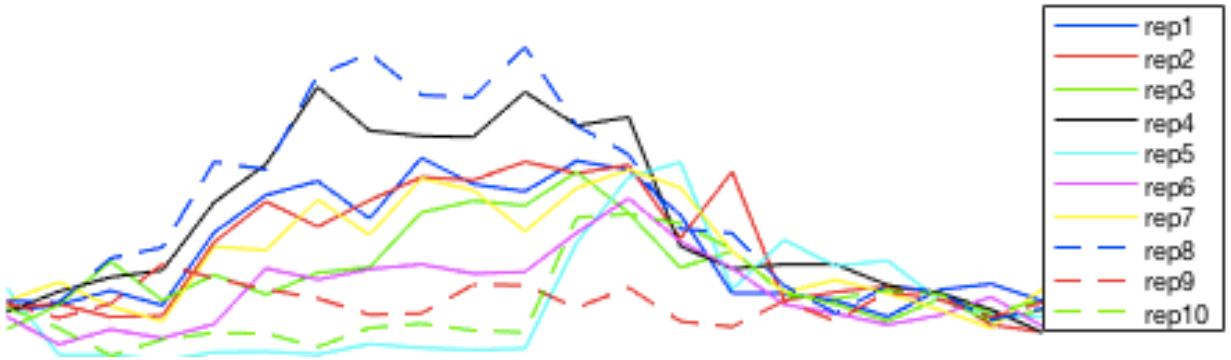
Both sparsity & correlated genes

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

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

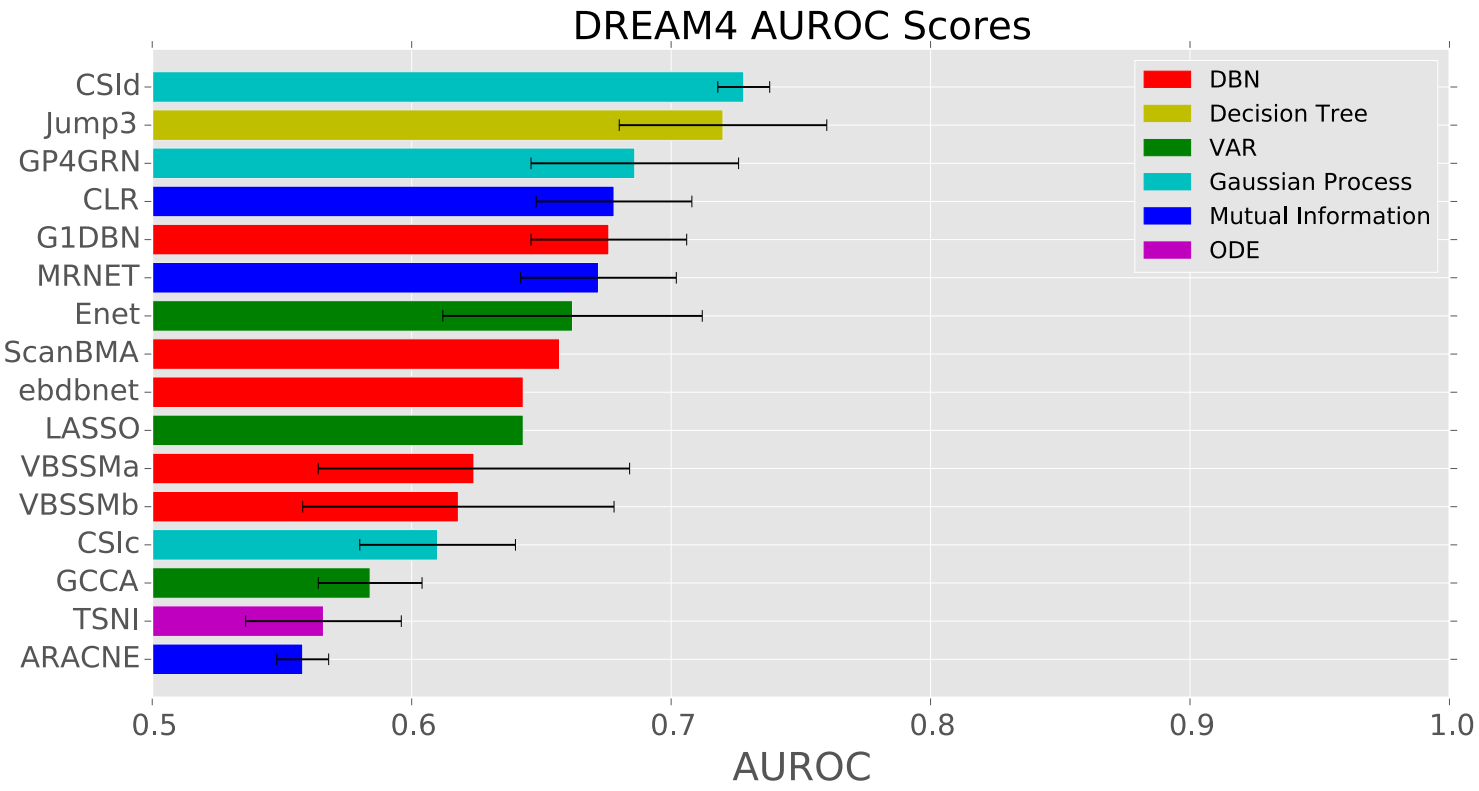
Evaluation

- DREAM4 Network Inference Challenge
- 100 genes, 21 timepoint time series, 10 replicates



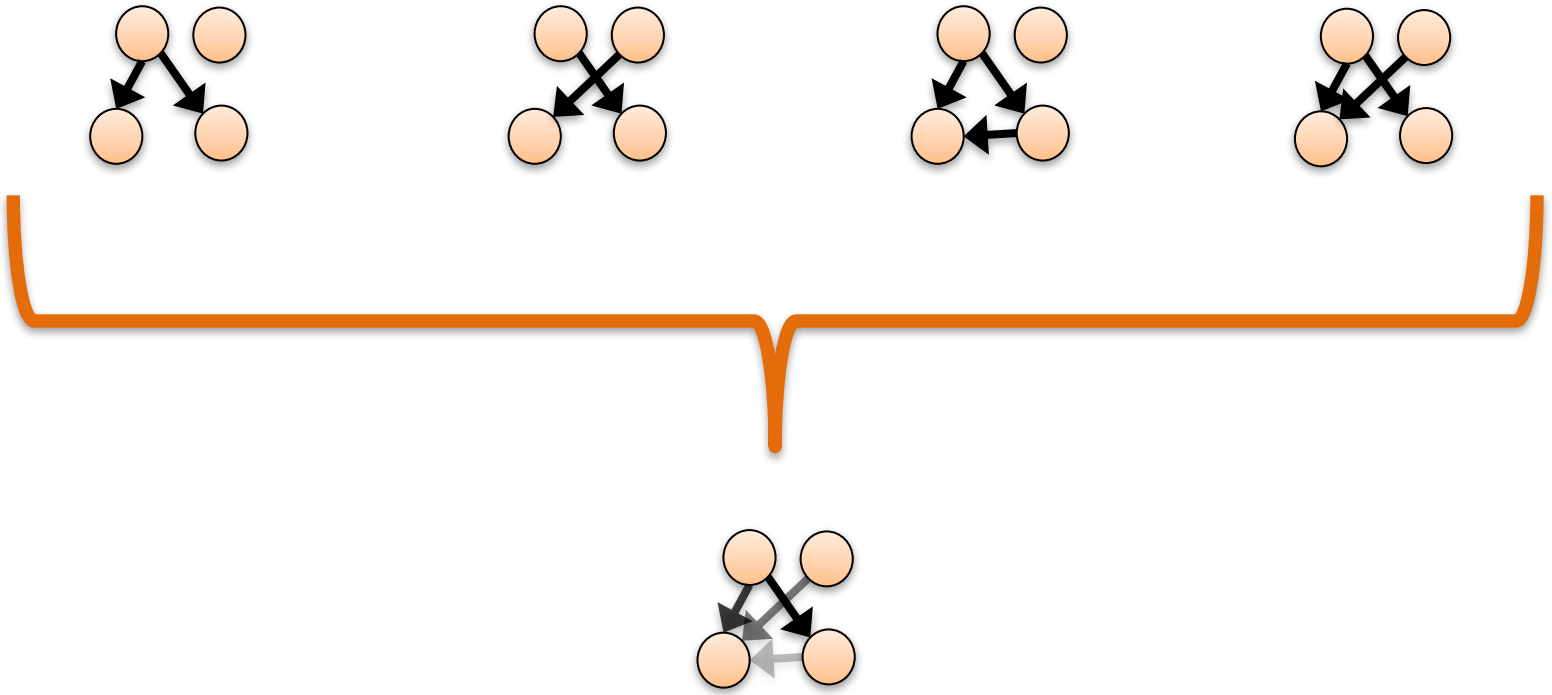
Performance

- Enet: Rank by coefficient
- 7^{th}/16, but best of VAR



Challenge: Robustness

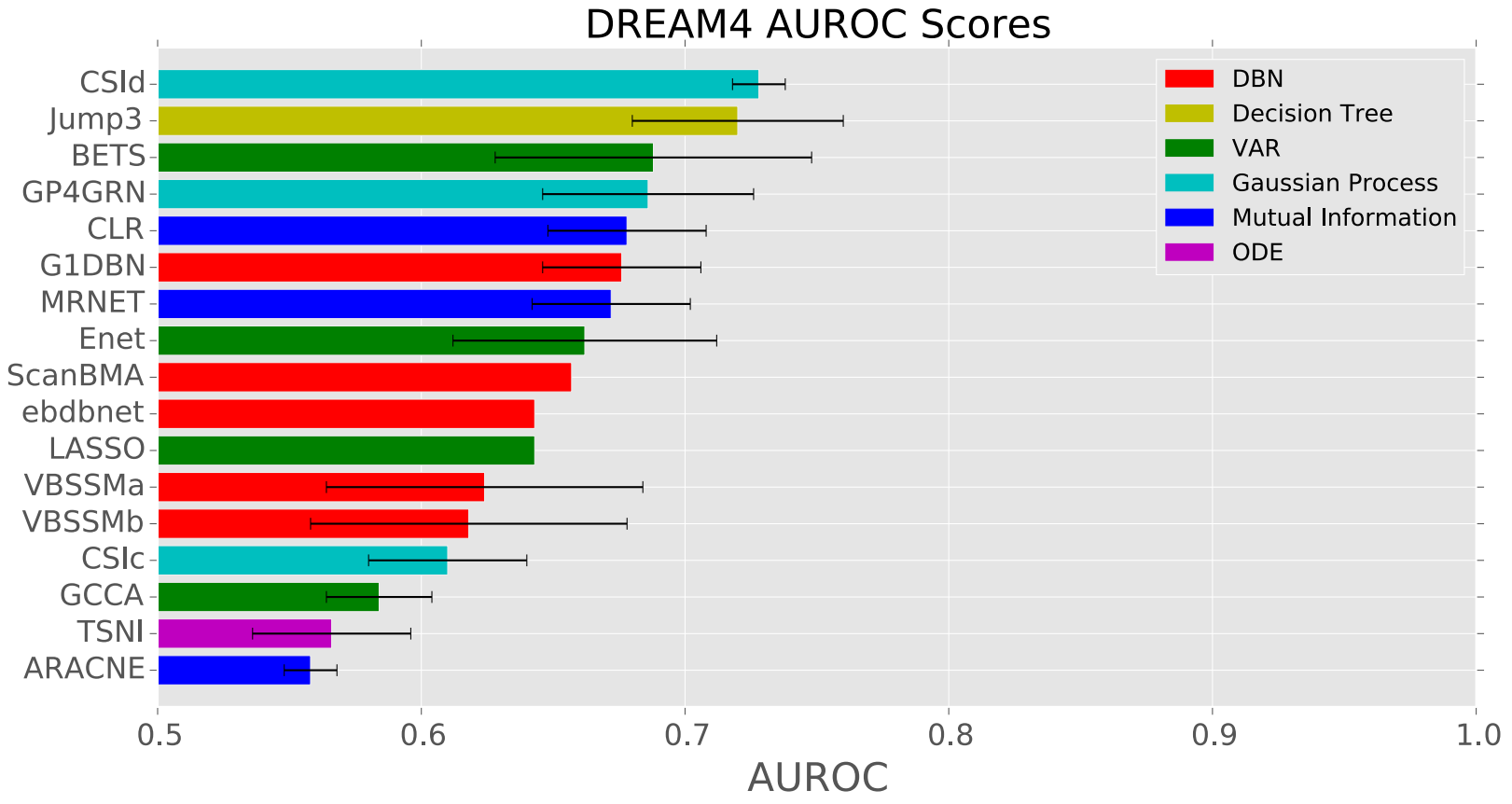
- How sensitive are inferred edges?
- Bootstrap Frequency:
  - Infer from 1000 Samples with replacement



Buhlmann 2014

Performance

- BETS: Rank by bootstrap frequency
- Huge improvement! 3rd/17



Challenge: Scalability

- Enet: 3000 fits
  - 40 hrs (~1 min/fit)*
- BETS: 1000 networks x 3000 fits each
  - 214 days!
- Solution: Massive Parallelization
  - 28,000 jobs on Della cluster
  - Complete in 4 days!

* = Very high variance, depending on the hyperparameter used

Timing

Method	Parallelized?	DREAM: 100 gene Elapsed Time	GGR: 2768 gene Real Time
CSI	Yes	9.2 hr	3 days per gene was insufficient, now 7 days per gene
Jump3	No	45 hr	Failed to Complete
BETS	Yes	4.8 hr	4 days

Evaluation

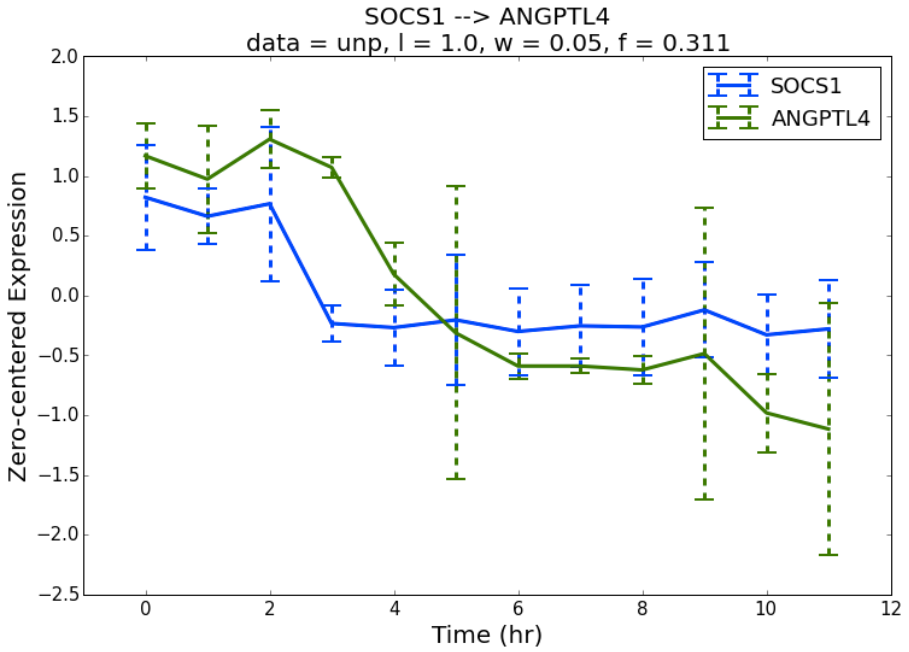
- Apply to GR
  - 31000 edges, FDR 0.2
- Held-out Dataset: Over-expression
  - Gene TF is biologically set to a higher level
  - Consider edges: TF → G
  - Compute G's fold-change between over-expression, original
  - Edge = logit(FC)?

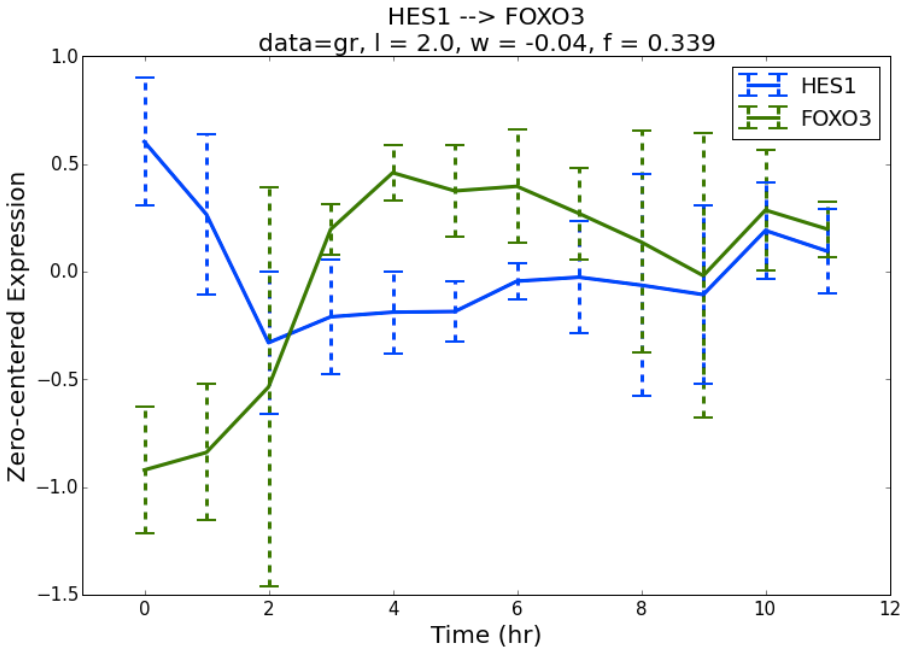
Validation Results

- Is-Positive-Edge ~ logit(log2 fold-change)
  - Pos = logit(-0.6848* log2FC + -3.7622)
  - **Log2fc p-value: 0.000186**
- Is-Negative-Edge ~ logit(log2 fold-change)
  - Neg = logit(0.4176 * log2FC – 3.9617)
  - Log2FC p-value: 0.165
- Is-Edge ~ logit(abs-log2 fold-change)
  - Edge = logit(0.3718* abs-log2FC - 3.2250)
  - Abs-log2FC p-value: **0.0964**

Interesting Edges?

- Search Space: 31000
- Metric: Bootstrapped coefficient with variance over time





Conclusion

1. We develop a novel method based on VAR to build causal networks from gene expression time series.
2. We address challenges of causal inference, statistical significance, and scalability.
3. We test our method extensively against other methods and data types.

Acknowledgments

Engelhardt Lab (Princeton)

Bianca Dumitrascu

Brian Jo

Barbara Engelhardt

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

Reddy Lab (Duke)

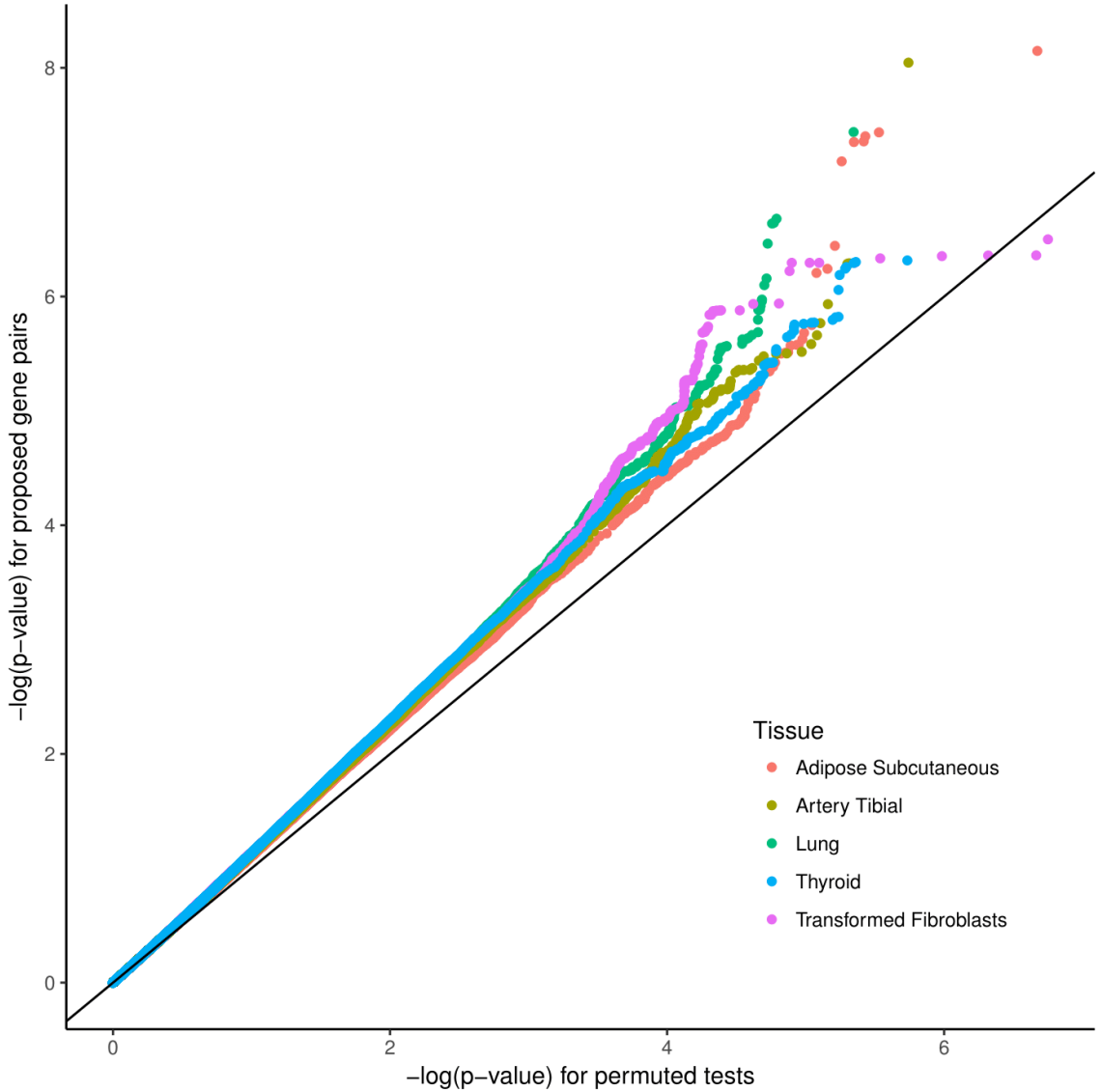
Ian McDowell

Tim Reddy

Data collection Team

Extra

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