Batch-effect Assessment on Networks Using GTEx

Matt's Lab Meeting

Yun Zhang

3-16-2016

Batch-effect on gene networks

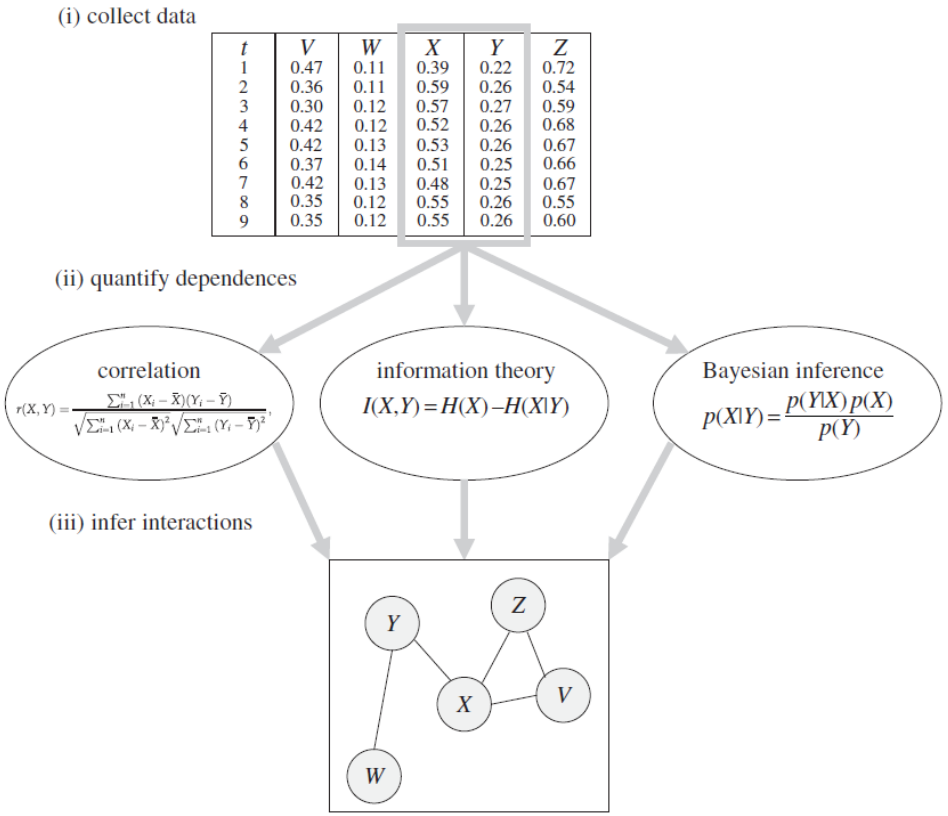
- Technical effects: when, where
- Biological effects: types of death (if used **ventilator**), **gender**, ethnicity
- Continuous effect: composition of cellular types

- How much of the reported findings is due to batch-effects?
- Does an edge in a gene network really represent a biological mechanism? Or is it just due to similar composition of cellular types?

GTEx: data preprocessing

- GTEx: Genotype Tissue Expression data
- Focused on the **lung** tissue samples:
  - 133 samples
- RNA-seq
- Variance-stabilizing transformation on the count data
- Filtered out low expression and small variation genes
  - Filtered data: 175 genes

Interaction networks: three main strategies



(i) collect data

t	*V*	*W*	*X*	*Y*	*Z*
1	0.47	0.11	0.39	0.22	0.72
2	0.36	0.11	0.59	0.26	0.54
3	0.30	0.12	0.57	0.27	0.59
4	0.42	0.12	0.52	0.26	0.68
5	0.42	0.13	0.53	0.26	0.67
6	0.37	0.14	0.51	0.25	0.66
7	0.42	0.13	0.48	0.25	0.67
8	0.35	0.12	0.55	0.26	0.55
9	0.35	0.12	0.55	0.26	0.60

(ii) quantify dependences

correlation

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}},$$

information theory

$$I(X,Y) = H(X) - H(X|Y)$$

Bayesian inference

$$p(X|Y) = \frac{p(Y|X)\,p(X)}{p(Y)}$$

(iii) infer interactions

Figure 1. Approaches for inferring interaction networks. Schematic of the process of inferring a network structure from data, showing three approaches for measuring dependence among variables: correlation-based, information theoretic and Bayesian.

Figure credit: Villaverde, A. F., & Banga, J. R. (2014). Reverse engineering and identification in systems biology: strategies, perspectives and challenges.*Journal of The Royal Society Interface*, *11*(91), 20130505.

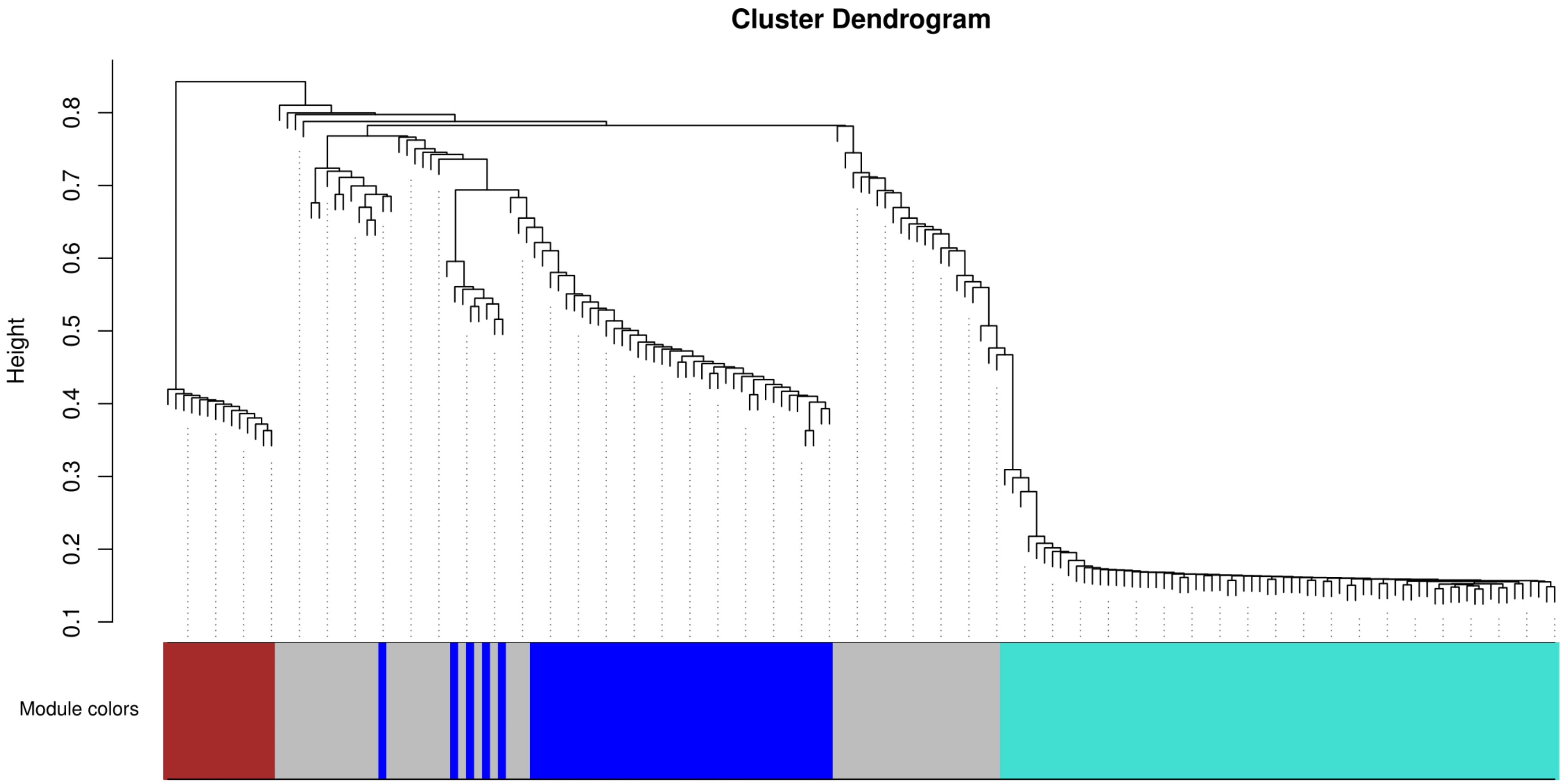
Analysis plan

- Network algorithms:
  - Correlation network: WGCNA
  - Mutual information network: ARACNE
  - Bayesian network: bnlearn
- Potential batch-effects:
  - Discrete: gender, ventilator, experiment date
  - Continuous: composition of cellular types

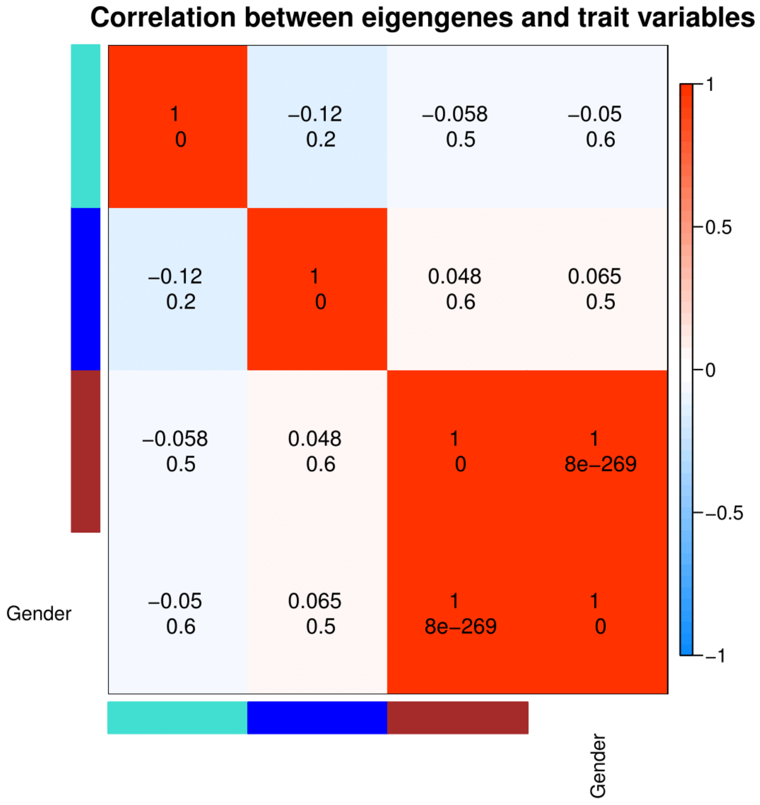
Correlation network

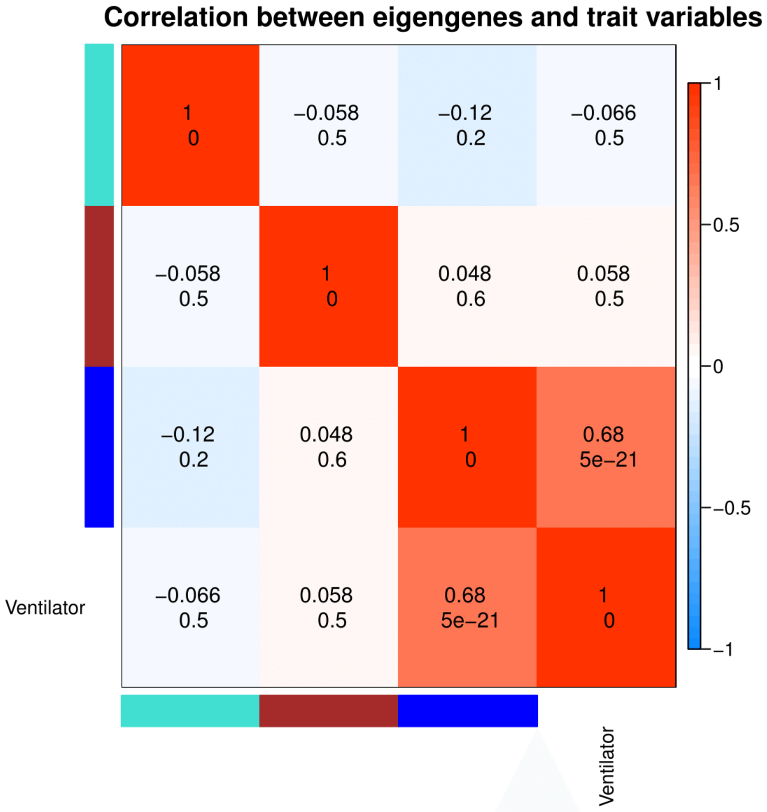
- WGCNA: Weighted Gene Co-expression Network Analysis
  - Weight of an edge between two genes is based on correlation measures
  - Gene modules and their corresponding eigengenes
- Investigate potential batch-effects
  - Gender
  - Ventilator
- Adjust known/unknown batch-effects using existing methods
  - ComBat – for known effects
  - Surrogate Variable Analysis (SVA) – for unknown effects

Gene modules identified by WGCNA

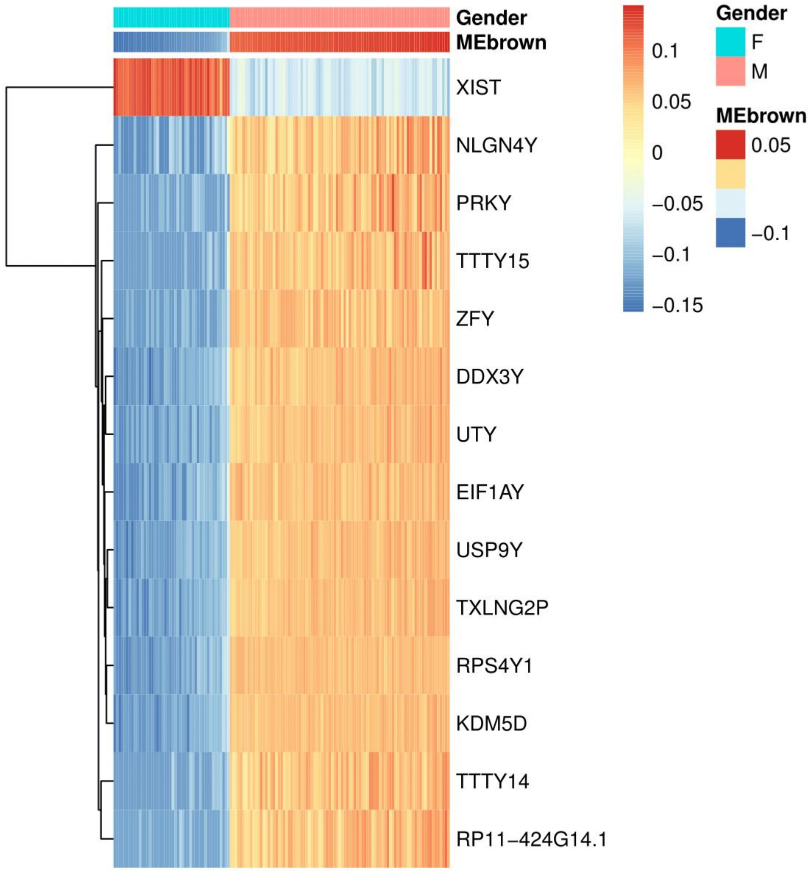


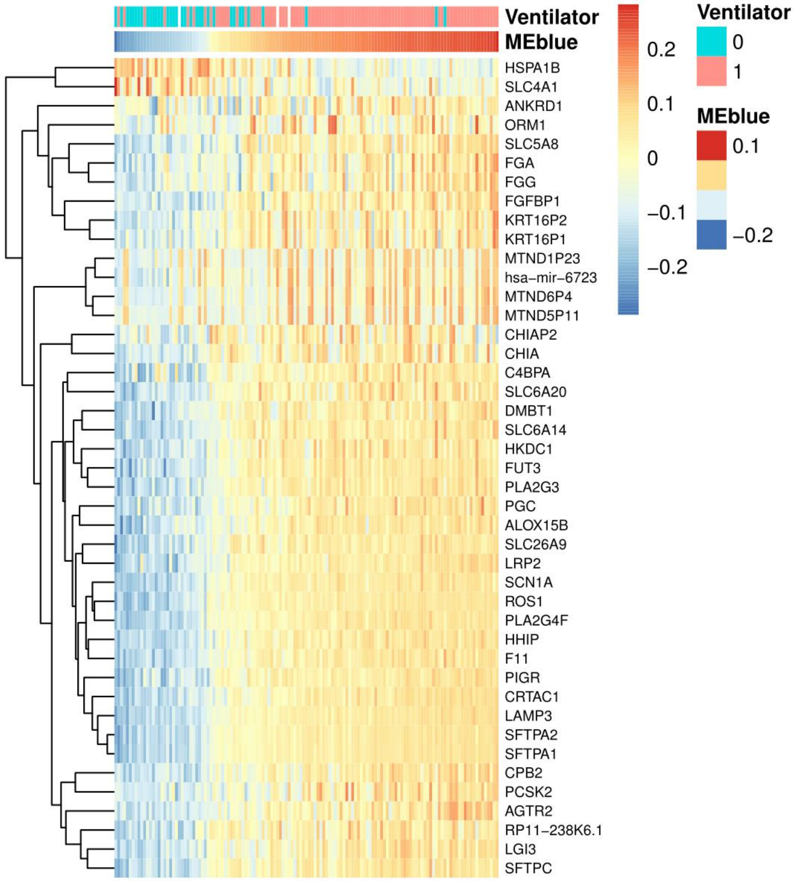
Suspects of gene modules induced by batch-effects


Correlation between eigengenes and trait variables

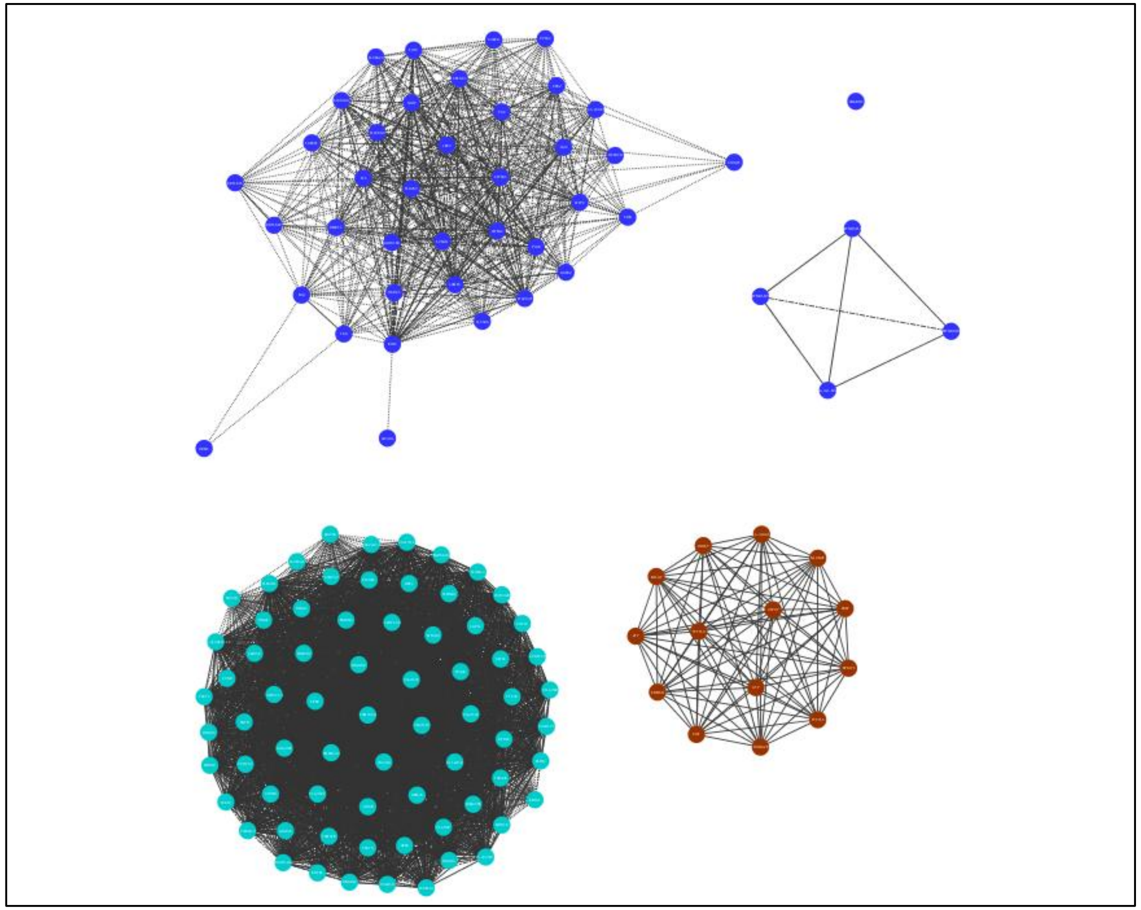

Correlation between eigengenes and trait variables

Gene expression profile of suspicious gene modules



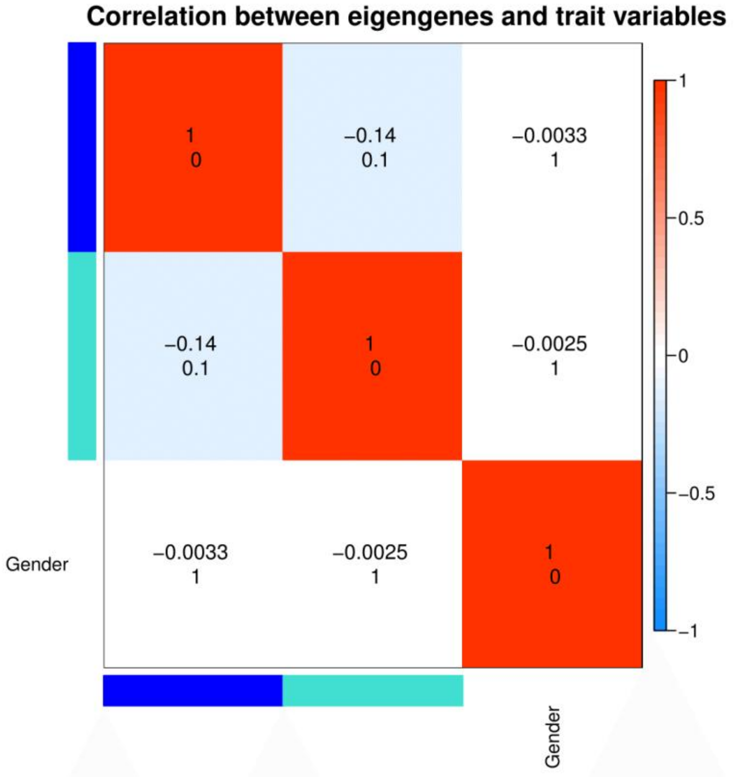


WGCNA network with gene modules

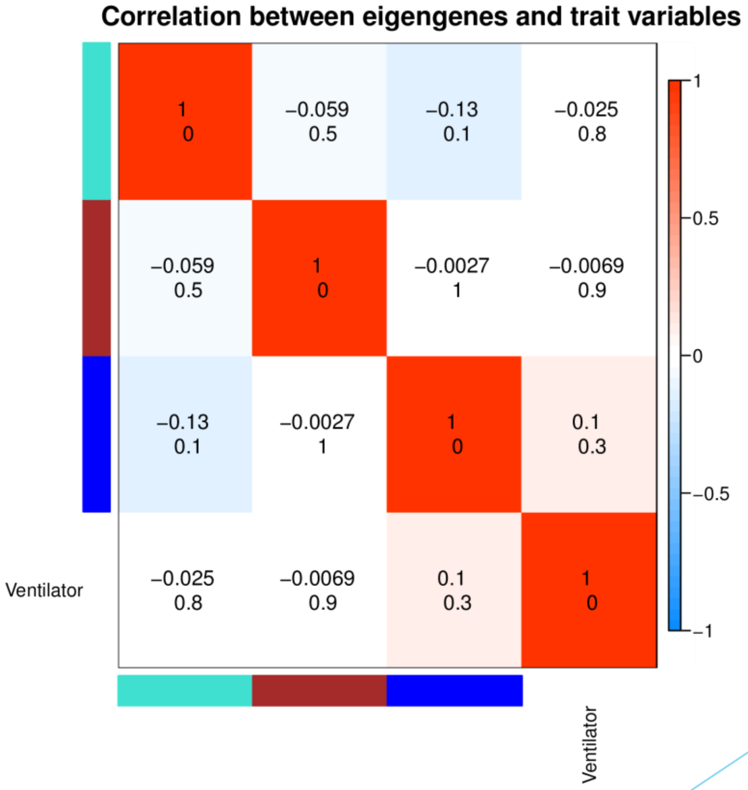


Edge visibility:
weight > 0.5

Batch-Effect Adjustment Using ComBat (One Batch-Effect A Time)

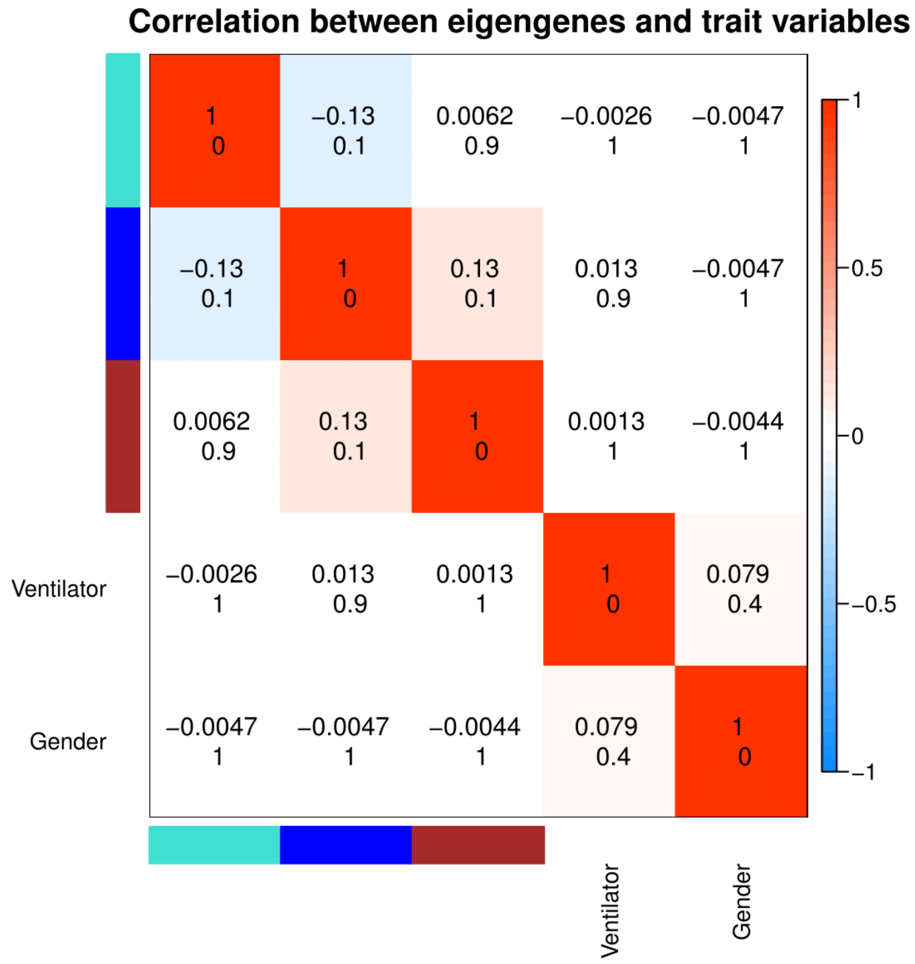


Correlation between eigengenes and trait variables



Correlation between eigengenes and trait variables

Batch-Effect Adjustment Using ComBat (Combined Batch-Effects)

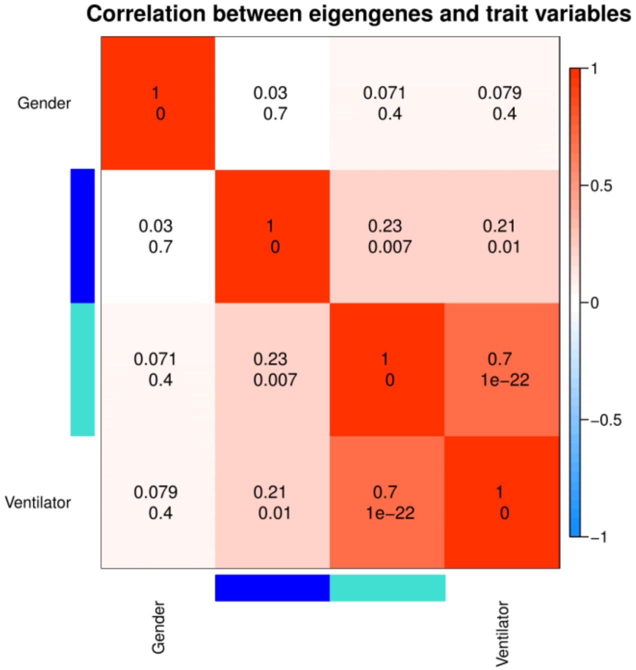


Correlation between eigengenes and trait variables

	🟩	🟦	🟫	Ventilator	Gender
🟩	1 0	−0.13 0.1	0.0062 0.9	−0.0026 1	−0.0047 1
🟦	−0.13 0.1	1 0	0.13 0.1	0.013 0.9	−0.0047 1
🟫	0.0062 0.9	0.13 0.1	1 0	0.0013 1	−0.0044 1
Ventilator	−0.0026 1	0.013 0.9	0.0013 1	1 0	0.079 0.4
Gender	−0.0047 1	−0.0047 1	−0.0044 1	0.079 0.4	1 0

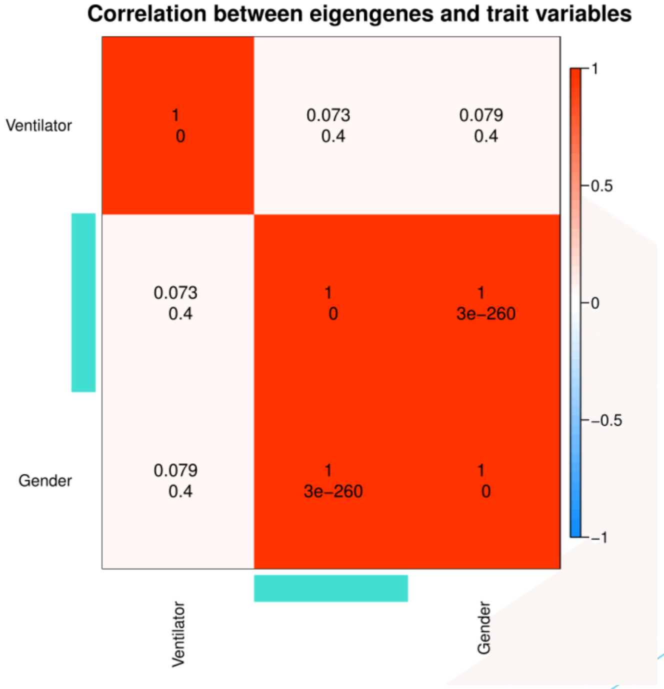
Batch-effect adjustment using SVA (Bad Example)

Adjust for gender



Correlation between eigengenes and trait variables

Adjust for ventilator



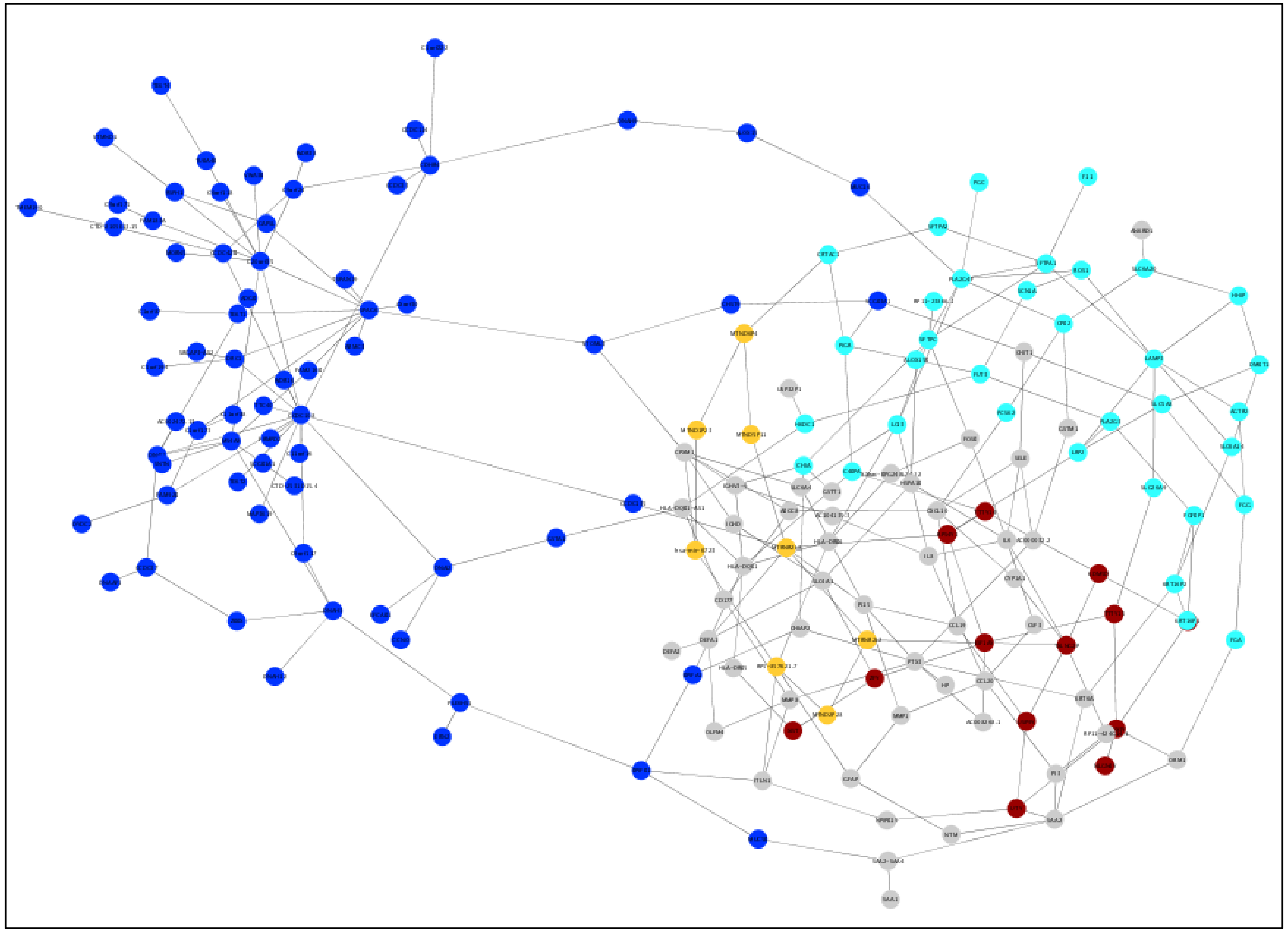
Correlation between eigengenes and trait variables

Mutual information network

- ARACNE: Algorithm for the Reconstruction of Accurate Cellular Networks
  - Mutual information (MI) measures the degree of statistical dependency between two variables
  - Weight of an edge between two genes is based on MI
- Investigate potential batch-effects
  - 4 suspicious gene clusters due to potential batch-effects

Gene cluster	Potential batch-effect
Cluster 1	Bronchial epithelium (spatial: which part of the lung) (changes in cellular composition)
Cluster 2	Gender
Cluster 3	Sequencing date
Cluster 4	Ventilator

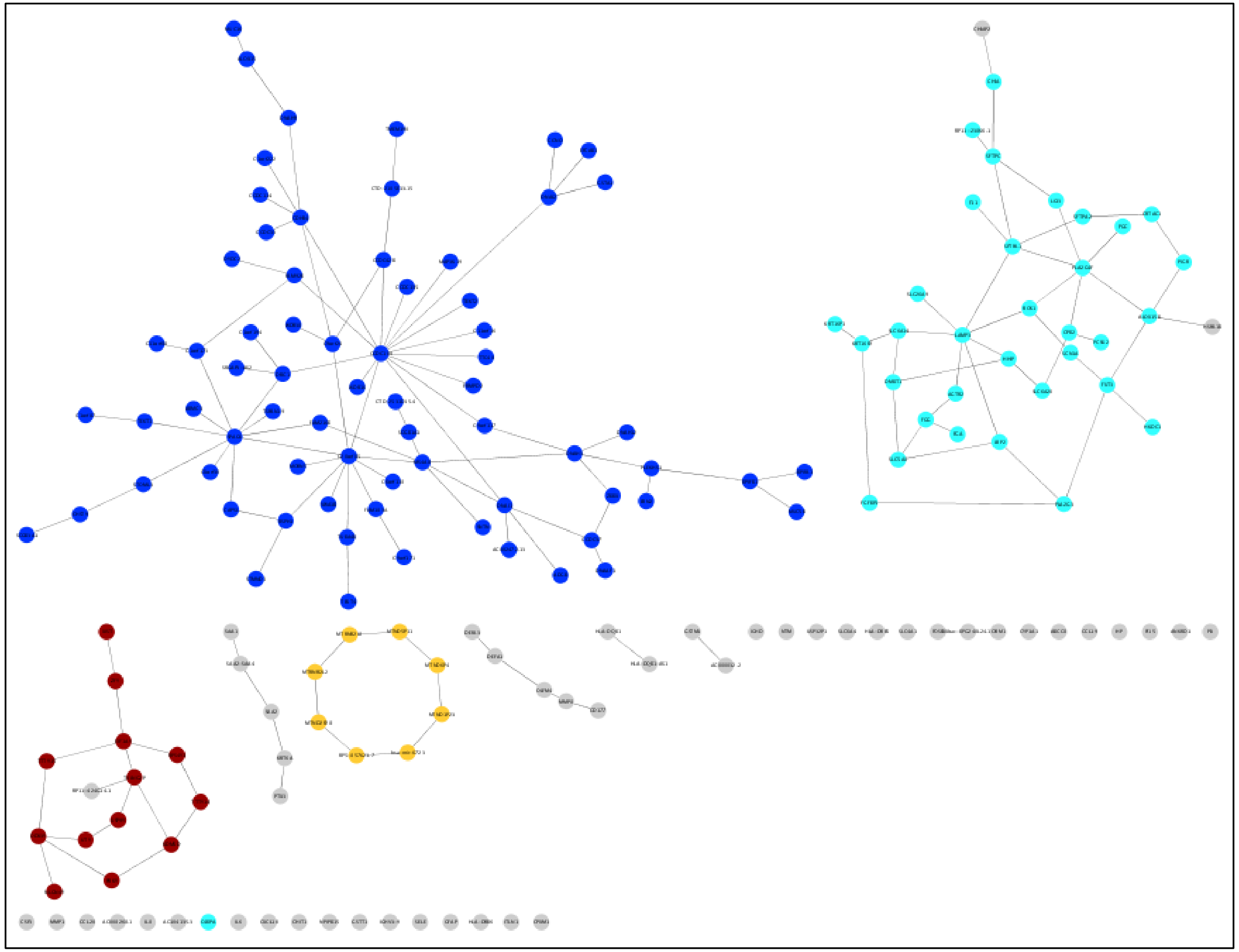
ARACNE network with suspicious gene clusters



Color legend:

⬜	NA
🟦	cluster1
🟥	cluster2
🟨	cluster3
🟦	cluster4

ARACNE network with suspicious gene clusters



Color legend:

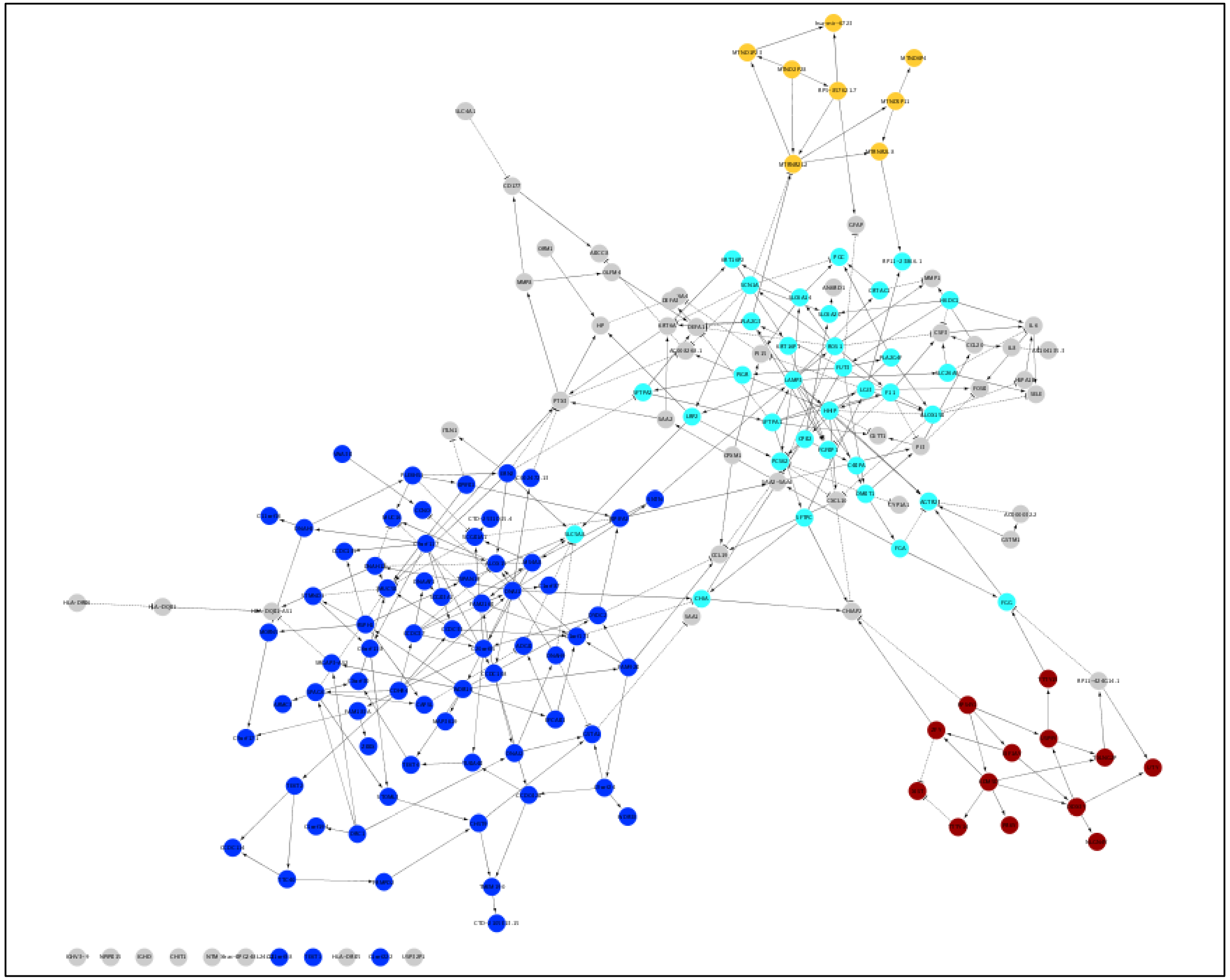
⬜	NA
🟦	cluster1
🟥	cluster2
🟨	cluster3
🟦	cluster4

Edge visibility:
weight > 0.1

Bayesian network

- Bayesian network
  - A Bayesian network is a representation of a joint probability distribution
  - Directed acyclic graph (DAG)
  - Weight of each edge is signed (+/-) from the parent gene to the child gene
- bnlearn: R package of learning Bayesian networks
  - Hill-climbing greedy search algorithm
- Investigate potential batch-effects
  - 4 suspicious gene clusters due to potential batch-effects

Bayesian network with suspicious gene clusters



Color legend:

⬜ NA

🟦 cluster1

🟥 cluster2

🟨 cluster3

🟦 cluster4

Edge visibility:
|weight| > 0.3

Conclusion

- All currently existing batch-effect correction methods focus on adjusting batch-effects at the gene level analyses, e.g. gene differential expression
- We conducted the first assessment of batch-effect on gene networks

- **Batch-effect has impact on all types of network algorithm**
- When we obtain a network, we should be cautious in interpreting the gene interactions (edges). Whether an edge is presented due to a real biological function between two genes? Edges may be presented purely due to some batch-effects.

Future work

- Simulations
  - Typical batch-effect on 2 or more groups, e.g. date, gender, etc.
  - Composition effect, e.g. % bronchial epithelium (changes in cellular composition)
- Batch-effect correction algorithms
  - ComBat + known batch variables
  - SVA + regression
  - Regression + known batch variables
  - PCA + regression