Fred Hutch Cancer Center

# Network-based Integration of Microbiome and Metabolomic Data

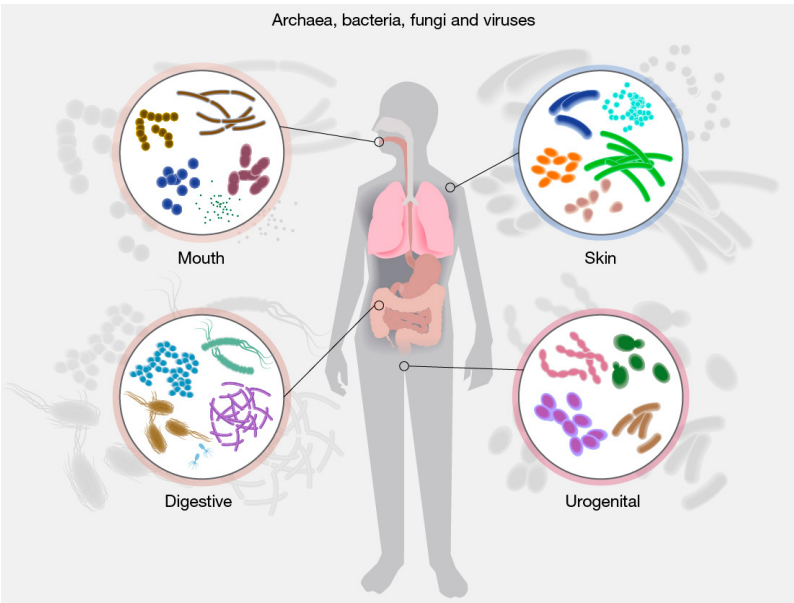
Banff, 17 July 2025

Jing Ma

# The human microbiome

Fred Hutch Cancer Center

All of the microbes and their genome, mostly bacteria




- More microbial cells than our somatic cells
- More microbial genes than our human genome
- Compositions vary within a person and between individuals
- Highly dynamic yet robust
- Association with many diseases

# Microbiome data

Fred Hutch Cancer Center

| OTU | Species | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|---|
| 1 | E.coli | 17 | 0 | 335 |
| 2 | S.aurus | 231 | 1180 | 45 |
| 3 | unknown | 30 | 0 | 0 |
| ... | ... | ... | ... | ... |

Table 1: A typical microbiome contingency table

Features of microbiome data

- high-dimensional

# Microbiome data

Fred Hutch Cancer Center

| OTU | Species | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|---|
| 1 | E.coli | 17 | 0 | 335 |
| 2 | S.aurus | 231 | 1180 | 45 |
| 3 | unknown | 30 | 0 | 0 |
| ... | ... | ... | ... | ... |

Table 1: A typical microbiome contingency table

Features of microbiome data

- high-dimensional
- sparse: lots of zeros $\rightarrow$ filtering + pseudocount replacement

# Microbiome data

Fred Hutch Cancer Center

| OTU | Species | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|---|
| 1 | E.coli | 17 | 0 | 335 |
| 2 | S.aurus | 231 | 1180 | 45 |
| 3 | unknown | 30 | 0 | 0 |
| ... | ... | ... | ... | ... |

Table 1: A typical microbiome contingency table

Features of microbiome data

- high-dimensional
- sparse: lots of zeros $\rightarrow$ filtering + pseudocount replacement
- compositional: only relative abundances are meaningful $\rightarrow$ normalization

# Metabolomics

Fred Hutch Cancer Center

Metabolomics is the study of small molecules, known as metabolites, within a biological system.

# Metabolomics

Fred Hutch Cancer Center

Metabolomics is the study of small molecules, known as metabolites, within a biological system.

Metabolites are the building blocks, intermediates, or end products of metabolism.

# Metabolomics

Fred Hutch Cancer Center

Metabolomics is the study of small molecules, known as metabolites, within a biological system.

Metabolites are the building blocks, intermediates, or end products of metabolism.

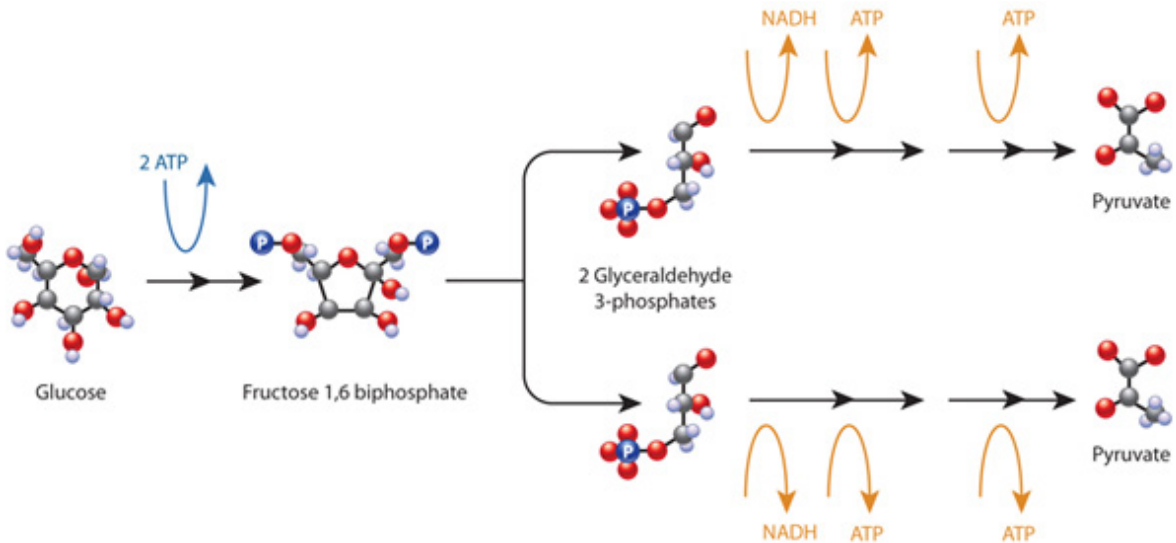



Figure 1: Glycolysis: energy is used to convert glucose to a 6 carbon form. Thereafter, energy is generated to create two molecules of pyruvate. Credit to Nature Education.

# Metabolomic data

Fred Hutch Cancer Center

| Compound | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| Glucose | 42,062,493 | 46,507,270 | 48,849,105 |
| Glutamic acid | 1,027,679 | 1,317,161 | 2,527,070 |
| Propionate | 3,487 | 6,262 | 9,188 |
| ... | ... | ... | ... |

Table 2: Peak intensities of compounds across samples

Features of metabolomic data

- high-dimensional

# Metabolomic data

Fred Hutch Cancer Center

| Compound | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| Glucose | 42,062,493 | 46,507,270 | 48,849,105 |
| Glutamic acid | 1,027,679 | 1,317,161 | 2,527,070 |
| Propionate | 3,487 | 6,262 | 9,188 |
| ... | ... | ... | ... |

Table 2: Peak intensities of compounds across samples

Features of metabolomic data

- high-dimensional
- sparse: lots of zeros $\rightarrow$ filtering + imputation

# Metabolomic data

Fred Hutch Cancer Center

| Compound | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| Glucose | 42,062,493 | 46,507,270 | 48,849,105 |
| Glutamic acid | 1,027,679 | 1,317,161 | 2,527,070 |
| Propionate | 3,487 | 6,262 | 9,188 |
| ... | ... | ... | ... |

Table 2: Peak intensities of compounds across samples

Features of metabolomic data

- high-dimensional
- sparse: lots of zeros $\rightarrow$ filtering + imputation
- high variance $\rightarrow$ log transformation

# Microbial metabolites

Fred Hutch Cancer Center

Microbial metabolites play an important role in host immune system.

# Microbial metabolites

Fred Hutch Cancer Center

Microbial metabolites play an important role in host immune system.

- SCFA: metabolites that are produced by bacteria from dietary components

# Microbial metabolites

Fred Hutch Cancer Center

Microbial metabolites play an important role in host immune system.

- SCFA: metabolites that are produced by bacteria from dietary components
- bile acids: metabolites that are produced by the host and biochemically modified by gut bacteria

# Microbial metabolites

Fred Hutch Cancer Center

Microbial metabolites play an important role in host immune system.

- SCFA: metabolites that are produced by bacteria from dietary components
- bile acids: metabolites that are produced by the host and biochemically modified by gut bacteria
- ATP: metabolites that are synthesized de novo by gut microbes

# Detecting microbial metabolites

Fred Hutch Cancer Center

Knowledge gap:

- most of the bacterial metabolites remain unidentified.
- many known metabolites have yet to be functionally characterized.

[1] Morton et al. 19'. Nat Methods; Reiman et al. 21'. PLOS Comp Bio
[2] Quinn-Bohmann et al. 25'. Nat Microbiol

# Detecting microbial metabolites

Fred Hutch Cancer Center

Knowledge gap:

- most of the bacterial metabolites remain unidentified.
- many known metabolites have yet to be functionally characterized.

Current methods for learning microbial–metabolite interactions:

- correlation networks
- machine learning models[1]
- mechanistic models[2]

[1] Morton et al. 19'. Nat Methods; Reiman et al. 21'. PLOS Comp Bio
[2] Quinn-Bohmann et al. 25'. Nat Microbiol

# Microbe–metabolite networks

Let $A \in \{0,1\}^{p_1 \times p_2}$ denote the latent network between $p_1$ microbes and $p_2$ metabolites.

In this bipartite network, the nodes are microbes/metabolites and the edges represent associations between microbes and metabolites.

# Microbe–metabolite networks

Let $A \in \{0,1\}^{p_1 \times p_2}$ denote the latent network between $p_1$ microbes and $p_2$ metabolites.

In this bipartite network, the nodes are microbes/metabolites and the edges represent associations between microbes and metabolites.

Inference for $A$ amounts to testing:

$$H_{0,i,j} : A_{i,j} = 0 \quad \text{versus.} \quad H_{1,i,j} : A_{i,j} \neq 0,$$

for all $1 \leq i \leq p_1, 1 \leq j \leq p_2$.

# False discovery rate

FDR provides a way of quantifying the statistical significance of multiple hypothesis tests.

| | Not significant | Significant | Total |
|---|---|---|---|
| Null is true | $N_{00}$ | $N_{10}$ | $m_0$ |
| Alternative is true | $N_{01}$ | $N_{11}$ | $m_1$ |
| Total | S | R | m |

Table 3: Classification of tested hypothesis

$$\text{FDR} = E\left(\frac{N_{10}}{R \vee 1}\right), \quad \text{mFDR} = \frac{E(N_{10})}{E(R)}$$

# FDR control

The Benjamini & Hochberg (BH) procedure[3]

- Choose a desired significance level $\alpha \in (0, 1)$.
- Sort the p-values in increasing order: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
- Find the largest index $k$ such that:

$$p_{(k)} \leq \frac{k}{m}\alpha$$

- Reject all hypotheses with p-values $p_i \leq p_{(k)}$.

[3] Benjamini & Hochberg. (95') JRSSB

# Local false discovery rate

Let $X = (x_i) \in \mathbb{R}^{p_1 p_2}$ denote the observations (e.g., $z$-scores). Efron et al.[4] studied the mixture model

$$f(x) = \pi_0 f_0(x) + \pi_1 f_1(x),$$

for multiple testing, where

- $f_0$: null distribution (e.g., $\mathcal{N}(0, 1)$)
- $f_1$: the alternative distribution
- $\pi_0, \pi_1 = 1 - \pi_0$: prior probabilities

[4] Efron et al. (01') JASA
[5] Sun and Cai. (07') JASA

# Local false discovery rate

Let $X = (x_i) \in \mathbb{R}^{p_1 p_2}$ denote the observations (e.g., $z$-scores). Efron et al.[4] studied the mixture model

$$f(x) = \pi_0 f_0(x) + \pi_1 f_1(x),$$

for multiple testing, where

- $f_0$: null distribution (e.g., $\mathcal{N}(0,1)$)
- $f_1$: the alternative distribution
- $\pi_0, \pi_1 = 1 - \pi_0$: prior probabilities

The Empirical Bayes local false discovery rate[5] is:

$$\text{lFDR}(x) = \frac{\pi_0 f_0(x)}{f(x)}.$$

Reject hypotheses with $\text{lFDR}(x) < \alpha$.

[4] Efron et al. (01') JASA
[5] Sun and Cai. (07') JASA

# Incorporating structures

More power can be achieved by exploiting the local dependence structure, e.g., Hidden Markov models[6].

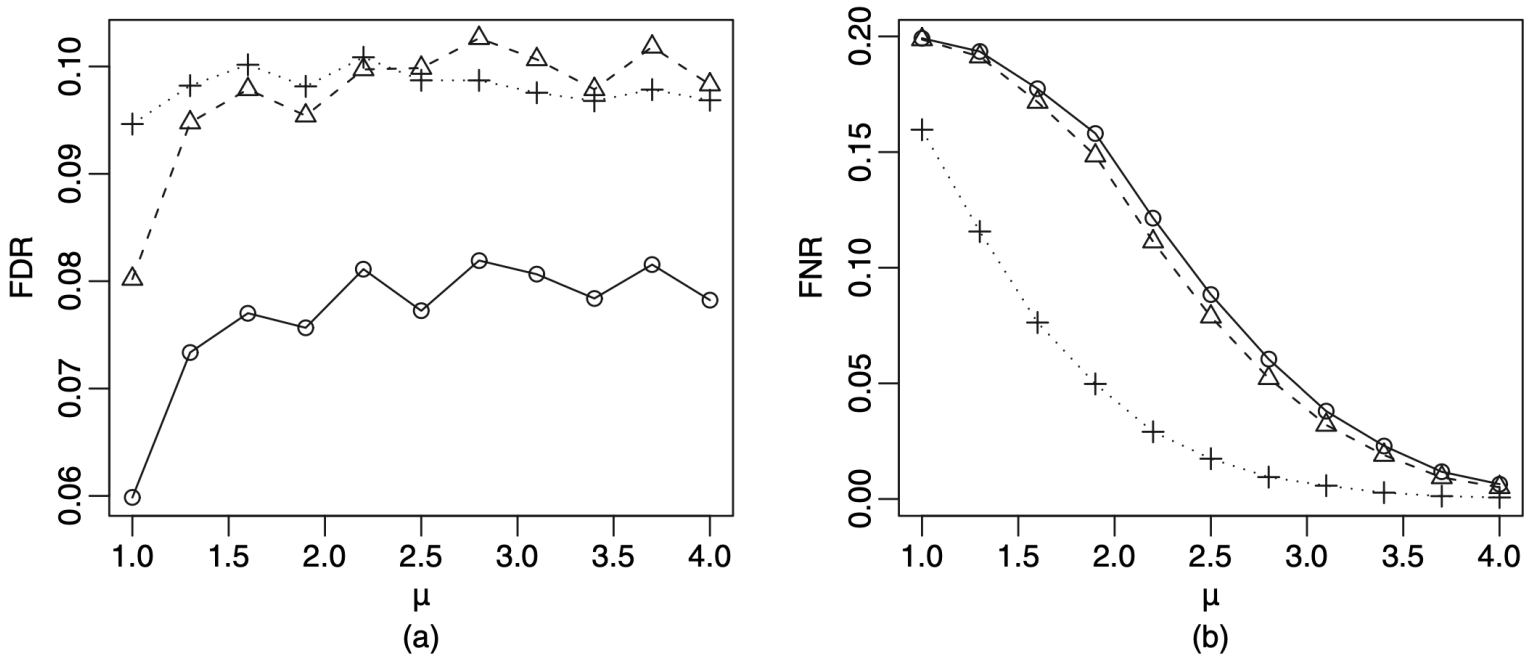



**Fig. 1.** Comparison of BH (○), AP (△) and OR (+) in an HMM (the FDR level is set at 0.10): (a) FDR *versus* $\mu$; (b) FNR *versus* $\mu$

[6]Sun and Cai (09') JRSSB

# Knowledge gap

- Nearly all existing works on large scale multiple testing require vector inputs and are not optimized for matrix-valued observations.

[7] Cai and Liu (16') JASA

# Knowledge gap

- Nearly all existing works on large scale multiple testing require vector inputs and are not optimized for matrix-valued observations.
- Methods for correlation testing exist[7], but they do not take into account the structural information in the data.

[7] Cai and Liu (16') JASA

# Knowledge gap

- Nearly all existing works on large scale multiple testing require vector inputs and are not optimized for matrix-valued observations.
- Methods for correlation testing exist[7], but they do not take into account the structural information in the data.

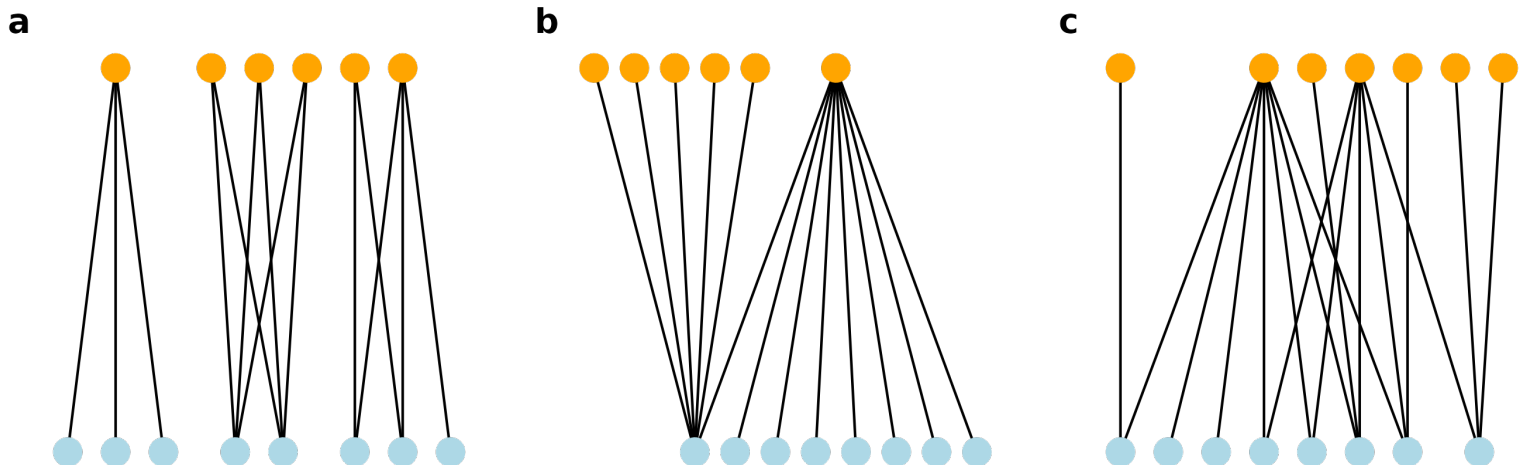



Figure 2: Topology of interest: (a) three biclusters, (b) fully nested graph, and (c) preferential attachment.

[7] Cai and Liu (16') JASA

# Latent graph model

Let $X = (x_{i,j}) \in \mathbb{R}^{p_1 \times p_2}$ denote the matrix-valued observations (e.g., $z$-scores).

# Latent graph model

Let $X = (x_{i,j}) \in \mathbb{R}^{p_1 \times p_2}$ denote the matrix-valued observations (e.g., $z$-scores).

We model $X$ using a latent bipartite stochastic block model (biSBM), defined with respect to row clustering $Z_1 = (Z_{i,1})$ and column clustering $Z_2 = (Z_{j,2})$. For $i = 1, \ldots, p_1$ and $j = 1, \ldots, p_2$:

$$\begin{aligned} Z_{i,1} &\sim \text{Multi}(1, \boldsymbol{\alpha}_1), \\ Z_{j,2} &\sim \text{Multi}(1, \boldsymbol{\alpha}_2), \\ A_{i,j} \mid Z_1, Z_2 &\sim \text{Bern}(\pi_{Z_{i,1}, Z_{j,2}}), \\ x_{i,j} \mid A_{i,j}, Z_1, Z_2 &\sim A_{i,j} g_{\nu_{Z_{i,1}, Z_{j,2}}} + (1 - A_{i,j}) g_{0,\nu_0}. \end{aligned} \tag{1}$$

Model parameters are $\boldsymbol{\theta} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\pi}, \boldsymbol{\nu}, \boldsymbol{\nu}_0)$.

# Multiple testing procedure

Ideally, we need

$$P(A_{i,j} = 0 \mid X)$$

to control the IFDR. However, they are intractable in our context.

# Multiple testing procedure

Fred Hutch Cancer Center

Ideally, we need

$$P(A_{i,j} = 0 \mid X)$$

to control the IFDR. However, they are intractable in our context.

Instead, we use the *structured $\ell$-value*

$$P(A_{i,j} = 0 \mid X, Z_1, Z_2).$$

Reject the hypotheses if the $\ell$-value is small, where the threshold is chosen to control mFDR.

# Multiple testing procedure

Ideally, we need

$$P(A_{i,j} = 0 \mid X)$$

to control the IFDR. However, they are intractable in our context.

Instead, we use the *structured $\ell$-value*

$$P(A_{i,j} = 0 \mid X, Z_1, Z_2).$$

Reject the hypotheses if the $\ell$-value is small, where the threshold is chosen to control mFDR.

The *structured $\ell$ values* provide much more information than a single observation $x_{i,j}$ and will considerably help to make the final decision.

# Identifiability

Fred Hutch Cancer Center

The *Gaussian* noisy biSBM is identifiable under the constraint that all elements of $\{(0, \sigma_0), (\mu_{q,l}, \sigma_{q,l}), 1 \leq q \leq B_1, 1 \leq l \leq B_2\}$ are distinct.

# Identifiability

Fred Hutch Cancer Center

The *Gaussian* noisy biSBM is identifiable under the constraint that all elements of $\{(0, \sigma_0), (\mu_{q,l}, \sigma_{q,l}), 1 \leq q \leq B_1, 1 \leq l \leq B_2\}$ are distinct.

In general, the requirement of all elements being distinct is not necessary, as the model is also identifiable if there is a single alternative distribution such that $\sigma_{q,l} = \sigma$ and $\mu_{q,l} = \mu$.

# Parameter estimation

Need to use the EM algorithm due to the latent variables $A, Z_1, Z_2$. Let $Q$ denote a probability distribution of the latent variables.

$$\log \mathcal{L}(X;\boldsymbol{\theta}) = \underbrace{E_Q[\log \mathcal{L}(X, A, Z_1, Z_2; \boldsymbol{\theta})] + \mathcal{H}(Q)}_{ELBO} + KL(Q \| P_{A,Z_1,Z_2|X;\boldsymbol{\theta}}), \tag{2}$$

1. Initialize $\boldsymbol{\theta}^{(0)}$.
2. E-step: evaluate the expectation in (2) with respect to $Q = P_{A,Z_1,Z_2|X;\boldsymbol{\theta}^{(t)}}$.
   - Mean-field approximation
3. M-step: update $\boldsymbol{\theta}^{(t+1)}$ by maximizing the ELBO.
4. Iterate between E- and M-step until convergence.

# Selecting the number of clusters

Fred Hutch Cancer Center

The numbers of blocks $B_1, B_2$ are unknown.

# Selecting the number of clusters

Fred Hutch Cancer Center

The numbers of blocks $B_1, B_2$ are unknown.

We use the integrated classification likelihood (ICL) criterion to select the optimal $B_1$ and $B_2$, allowing them to be different.

# Simulations

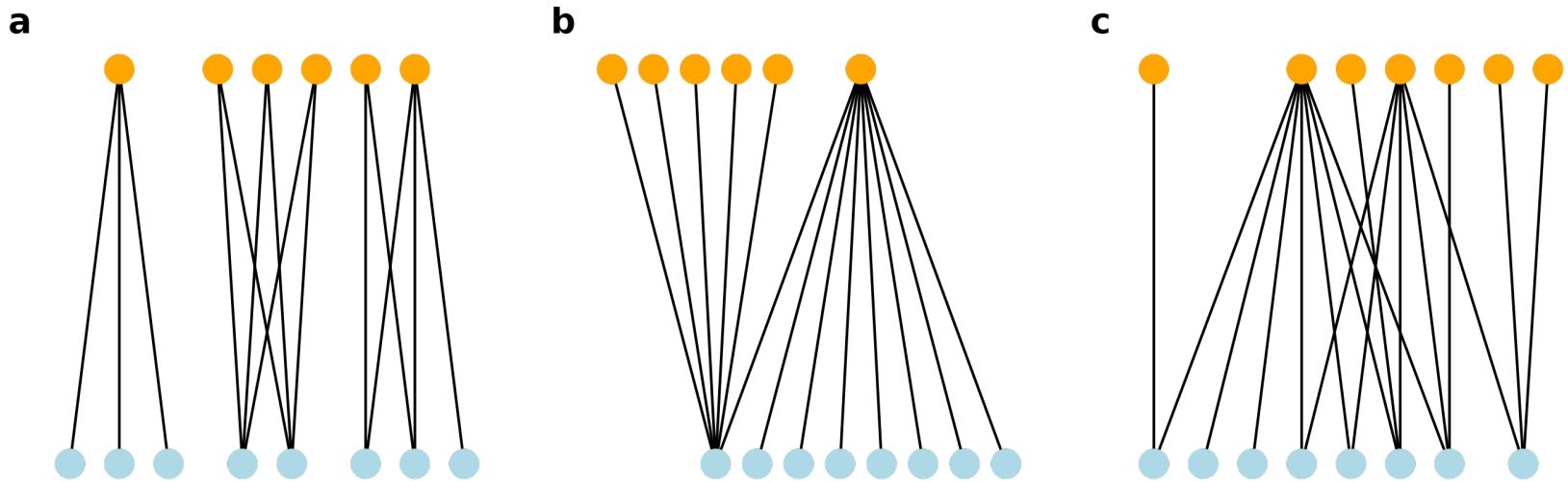



Figure 3: Illustrations of the latent bipartite network used in simulations: (a) three biclusters, (b) fully nested graph, and (c) preferential attachment

(a) Modular structures

(b) Nested graph in ecology: generalist vs specialist

(c) Preferential attachment: the rich gets richer

# Simulation parameters

- $p_1 = 150, p_2 = 200$
- $\mathcal{N}(0,1)$ versus $\mathcal{N}(2,1)$
- Compare the average performance over 100 simulations
- Both the new and the SC procedures[8] were implemented assuming known null density.

[8] Sun and Cai (07') JASA

# Simulation results

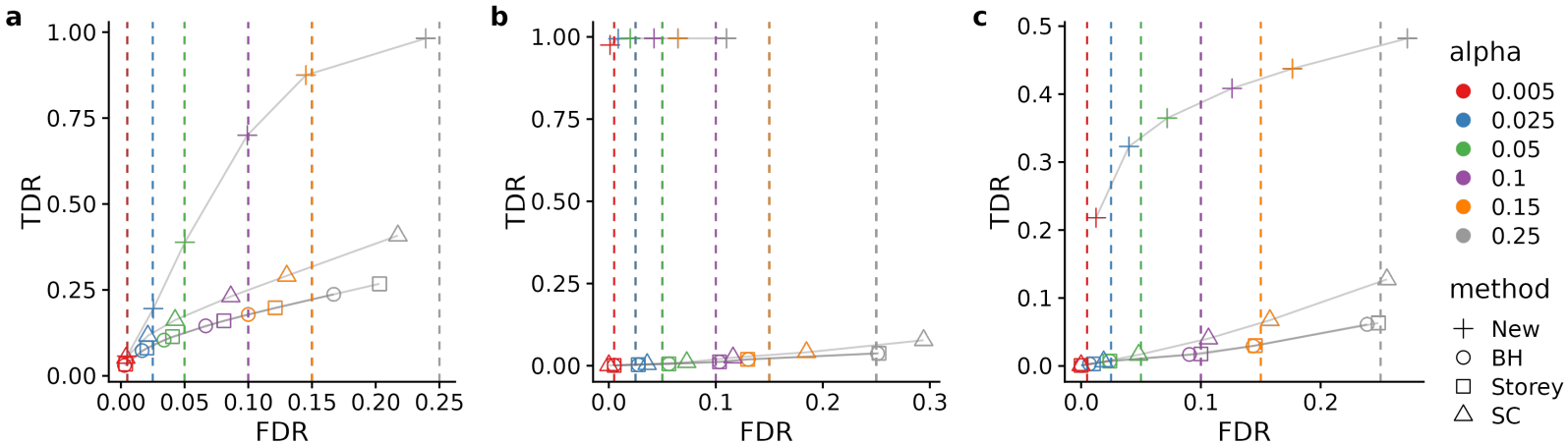



Figure 4: Plot of the empirical $(\mathrm{FDR}, \mathrm{TDR})$ as a function of the nominal level $\alpha$ for the new procedure, BH, Storey's $q$-value, and the SC procedure. Dashed lines indicate the nominal level $\alpha$.

# Simulation results

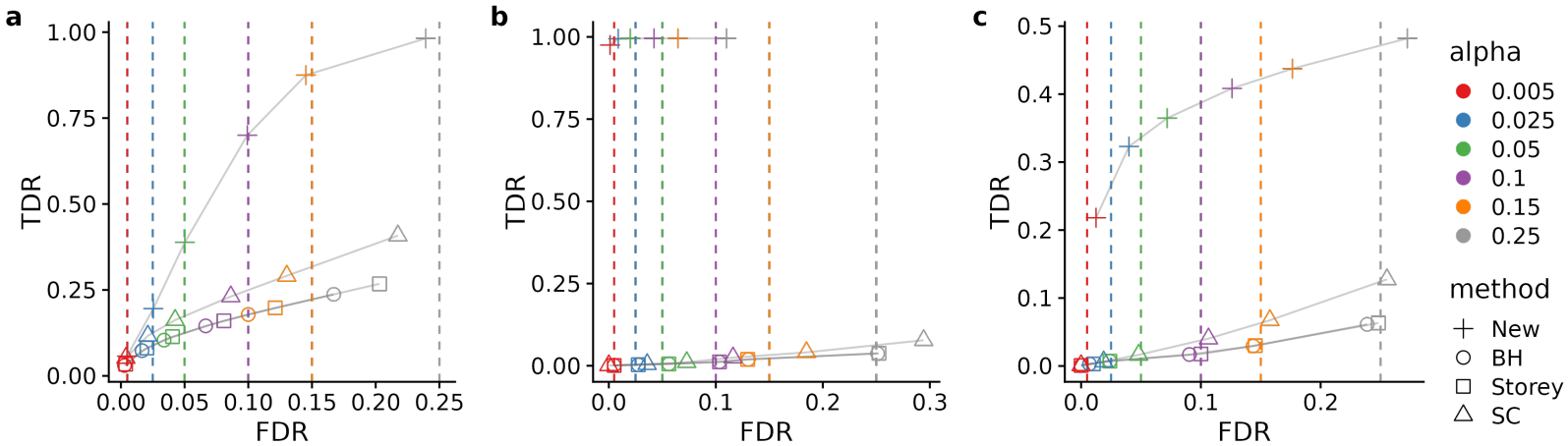



Figure 4: Plot of the empirical $(\mathrm{FDR}, \mathrm{TDR})$ as a function of the nominal level $\alpha$ for the new procedure, BH, Storey's $q$-value, and the SC procedure. Dashed lines indicate the nominal level $\alpha$.

(a) ICL selected a model with three biclusters correctly in 90% of the simulations.

# Simulation results

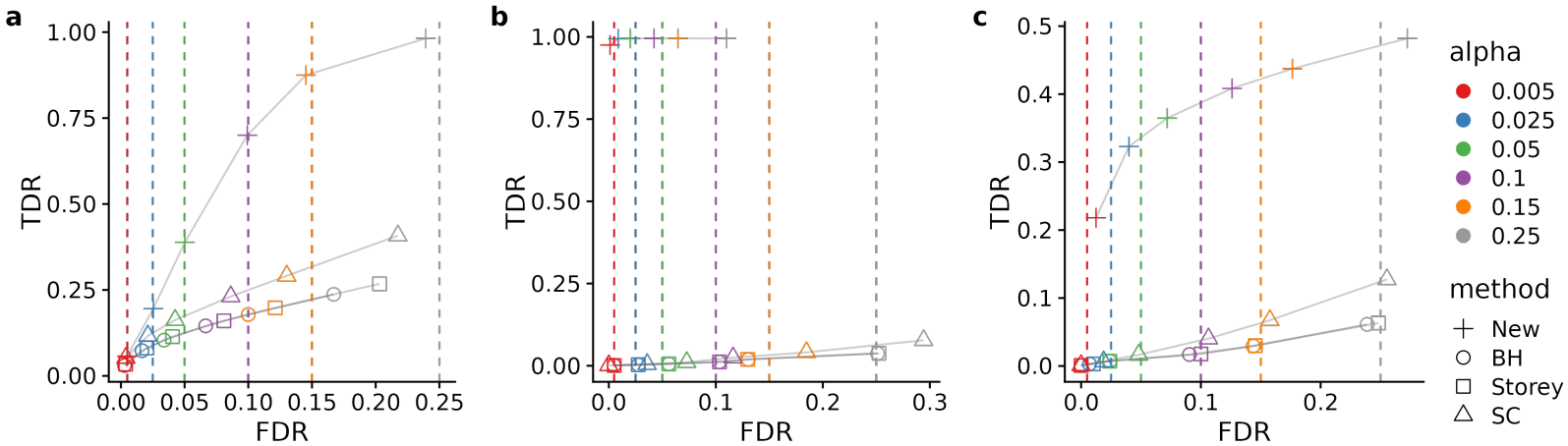



Figure 4: Plot of the empirical $(\mathrm{FDR}, \mathrm{TDR})$ as a function of the nominal level $\alpha$ for the new procedure, BH, Storey's $q$-value, and the SC procedure. Dashed lines indicate the nominal level $\alpha$.

(a) ICL selected a model with three biclusters correctly in 90% of the simulations.

(b) ICL selected two biclusters in 83% of the simulations.

# Simulation results

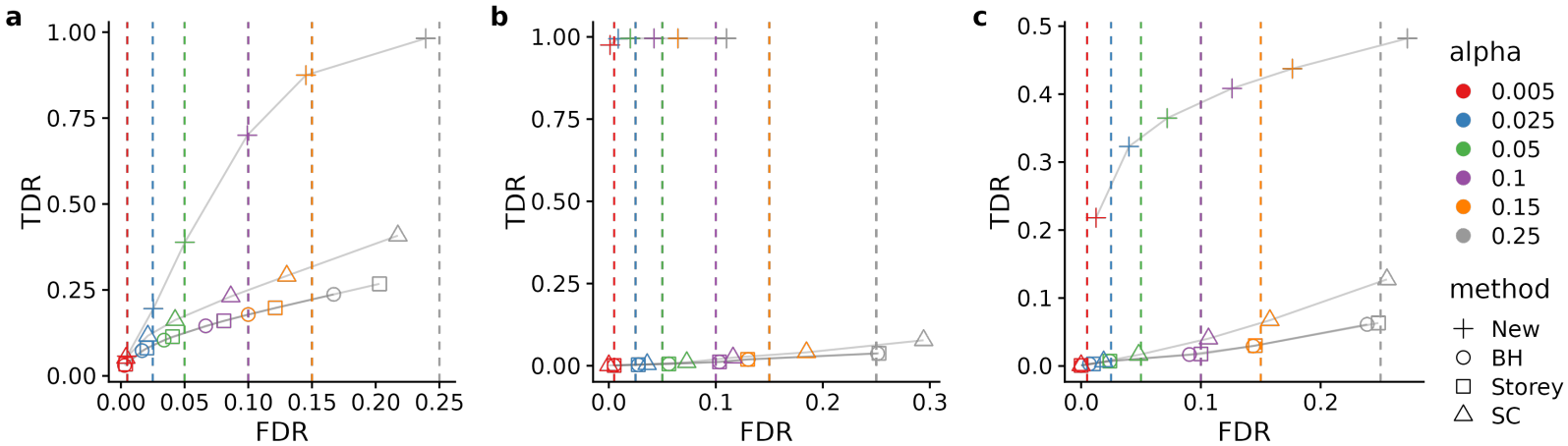



Figure 4: Plot of the empirical $(\mathrm{FDR}, \mathrm{TDR})$ as a function of the nominal level $\alpha$ for the new procedure, BH, Storey's $q$-value, and the SC procedure. Dashed lines indicate the nominal level $\alpha$.

(a) ICL selected a model with three biclusters correctly in 90% of the simulations.

(b) ICL selected two biclusters in 83% of the simulations.

(c) ICL mostly found three clusters in type I vertex and one cluster in type II vertex.

# Bacterial vaginosis

- BV is the most common vaginal condition, affecting an estimated 30% of women at any given time[9].
- BV is associated with increased transmission of HIV and STIs as well as increased risk of preterm labour.
- Diagnosis relies on microscopy to identify BV-like bacteria by morphology alone (Nugent Scoring).
- The pathogenesis of BV remains unclear.

[9] McMillan et al. (15') Scientific Reports

# Data

- $Y_1 \in R^{n \times p_1}$: relative abundances from 49 genera obtained from 16S; centered log ratio transformed while keep zeros unchanged.

# Data

- $Y_1 \in R^{n \times p_1}$: relative abundances from 49 genera obtained from 16S; centered log ratio transformed while keep zeros unchanged.
- $Y_2 \in R^{n \times p_2}$: concentrations of 128 metabolites from GC-MS; log transformed

# Data

- $Y_1 \in R^{n \times p_1}$: relative abundances from 49 genera obtained from 16S; centered log ratio transformed while keep zeros unchanged.
- $Y_2 \in R^{n \times p_2}$: concentrations of 128 metabolites from GC-MS; log transformed
- Outcome: normal ($n = 79$) vs BV ($n = 45$) status defined by nugent score

# Data

- $Y_1 \in R^{n \times p_1}$: relative abundances from 49 genera obtained from 16S; centered log ratio transformed while keep zeros unchanged.
- $Y_2 \in R^{n \times p_2}$: concentrations of 128 metabolites from GC-MS; log transformed
- Outcome: normal ($n = 79$) vs BV ($n = 45$) status defined by nugent score
- Aim is to understand the microbial functional changes during BV.

# Test statistics

Let $\hat{Y}_1$ and $\hat{Y}_2$ denote, respectively, the standardized data. The sample correlation is defined by

$$\hat{\rho}_{i,j} = \frac{1}{n} \sum_{k=1}^{n} \hat{Y}_{1,k,i} \hat{Y}_{2,k,j}.$$

[10] Cai and Liu (16') JASA

# Test statistics

Let $\hat{Y}_1$ and $\hat{Y}_2$ denote, respectively, the standardized data. The sample correlation is defined by

$$\hat{\rho}_{i,j} = \frac{1}{n}\sum_{k=1}^{n} \hat{Y}_{1,k,i}\,\hat{Y}_{2,k,j}.$$

Let

$$s_{i,j} = \frac{1}{n}\sum_{i=1}^{n}(2\hat{Y}_{1,k,i}\,\hat{Y}_{2,k,j} - \hat{\rho}_{k,i}\,\hat{Y}_{1,k,i} - \hat{\rho}_{k,i}\,\hat{Y}_{2,k,j})^2.$$

[10] Cai and Liu (16') JASA

# Test statistics

Let $\hat{Y}_1$ and $\hat{Y}_2$ denote, respectively, the standardized data. The sample correlation is defined by

$$\hat{\rho}_{i,j} = \frac{1}{n}\sum_{k=1}^{n} \hat{Y}_{1,k,i}\,\hat{Y}_{2,k,j}.$$

Let

$$s_{i,j} = \frac{1}{n}\sum_{i=1}^{n}(2\hat{Y}_{1,k,i}\,\hat{Y}_{2,k,j} - \hat{\rho}_{k,i}\,\hat{Y}_{1,k,i} - \hat{\rho}_{k,i}\,\hat{Y}_{2,k,j})^2.$$

The test statistic

$$x_{i,j} = \frac{2\hat{\rho}_{i,j}}{\sqrt{s_{i,j}/n}} \to \mathcal{N}(0,1),$$

under finite fourth moment condition[10].

[10] Cai and Liu (16') JASA

# Test statistics

Two-sample inference comparing BV to normal patients:

$$x_{i,j} = \frac{2(\hat{\rho}_{i,j}^{(1)} - \hat{\rho}_{i,j}^{(2)})}{\sqrt{s_{i,j}^{(1)}/n_1 + s_{i,j}^{(2)}/n_2}}$$

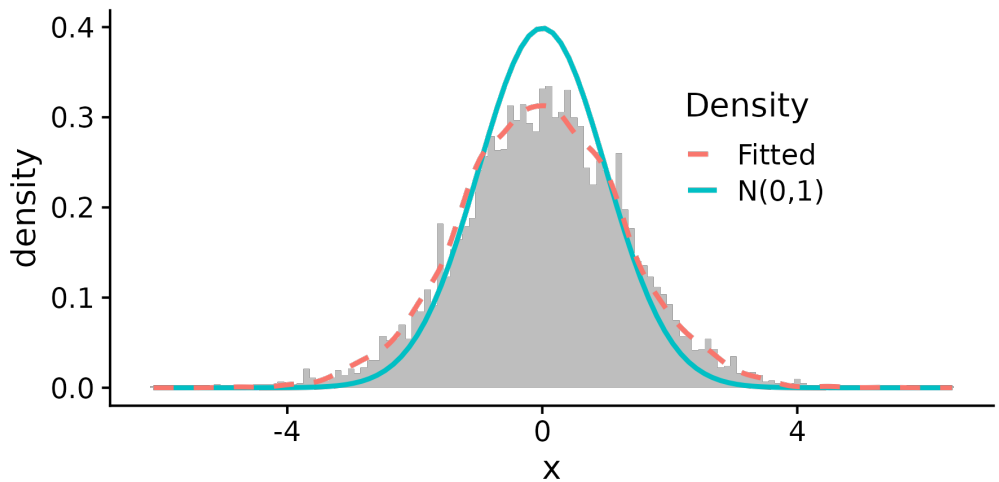



Figure 5: Histogram of observed $z$-scores compared to the standard normal and the estimated marginal distribution by the proposed approach.

# Results

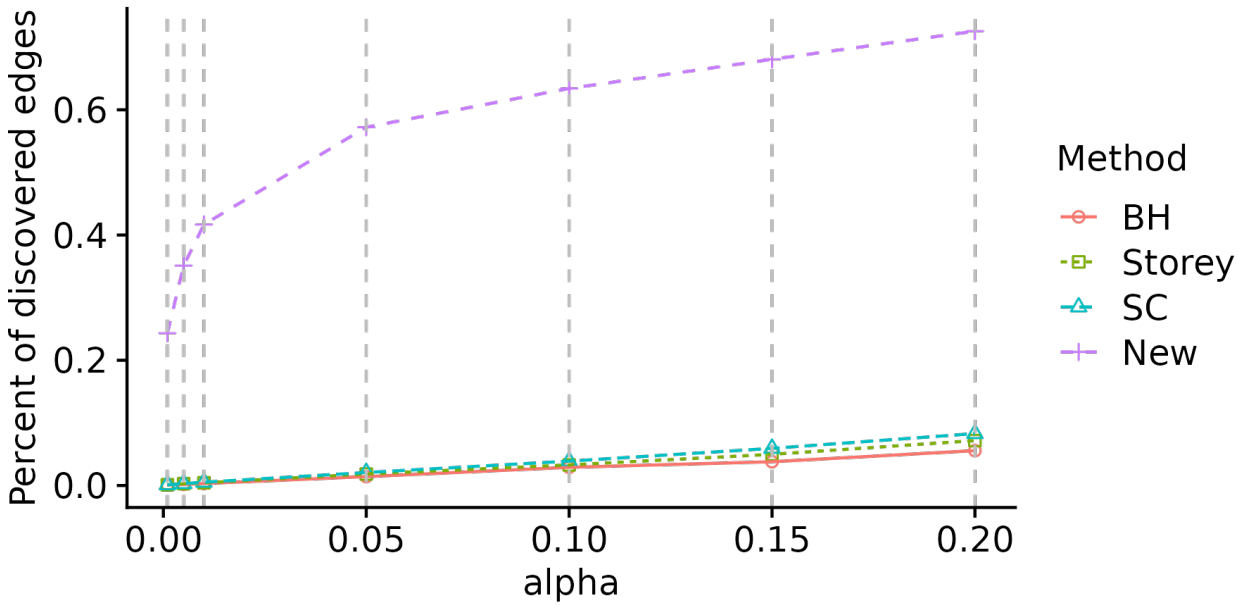



Figure 6: Percent of rejected edges as a function of the significance level $\alpha$ for the different procedures.

# Results

The new procedure groups microbes and metabolites with similar association patterns into biclusters.

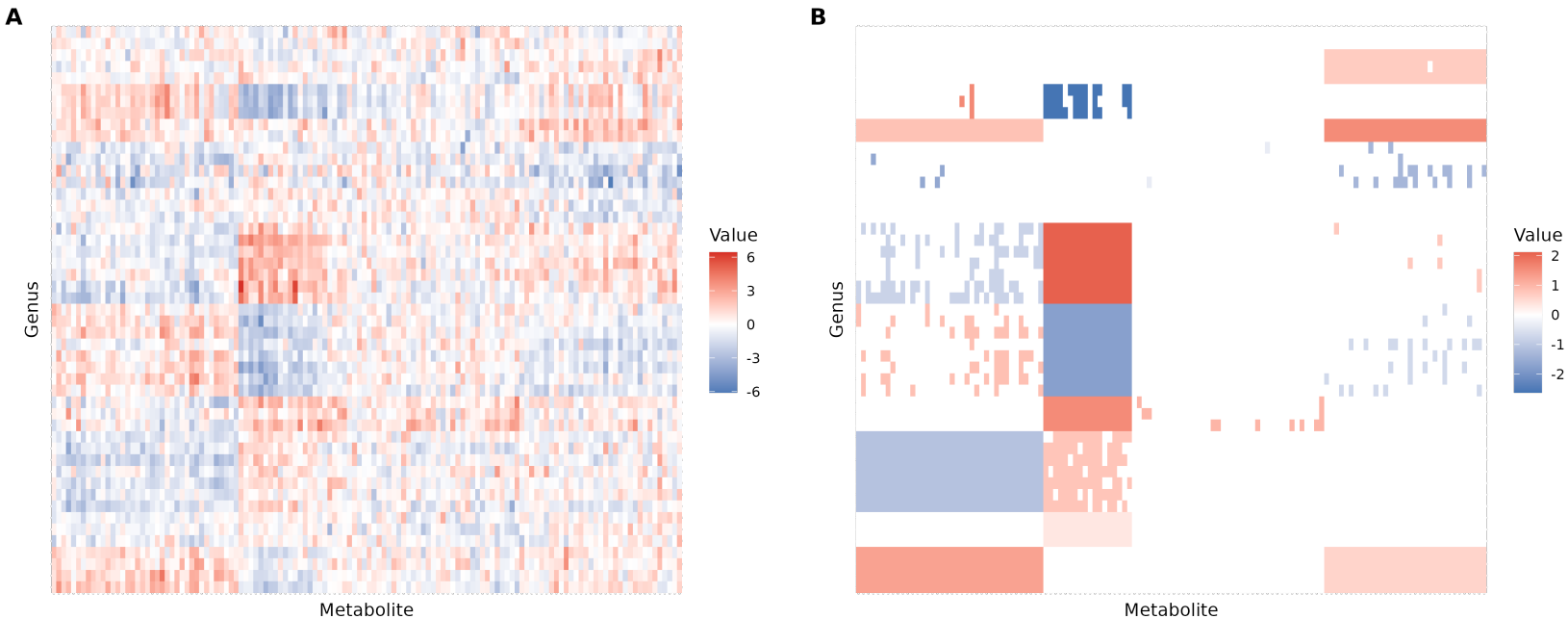



Figure 7: Heat map of the data (A) compared to the estimated graph by the proposed approach at $\alpha = 0.1\%$ (B). Rows and columns are ordered by the inferred clustering.

# Results

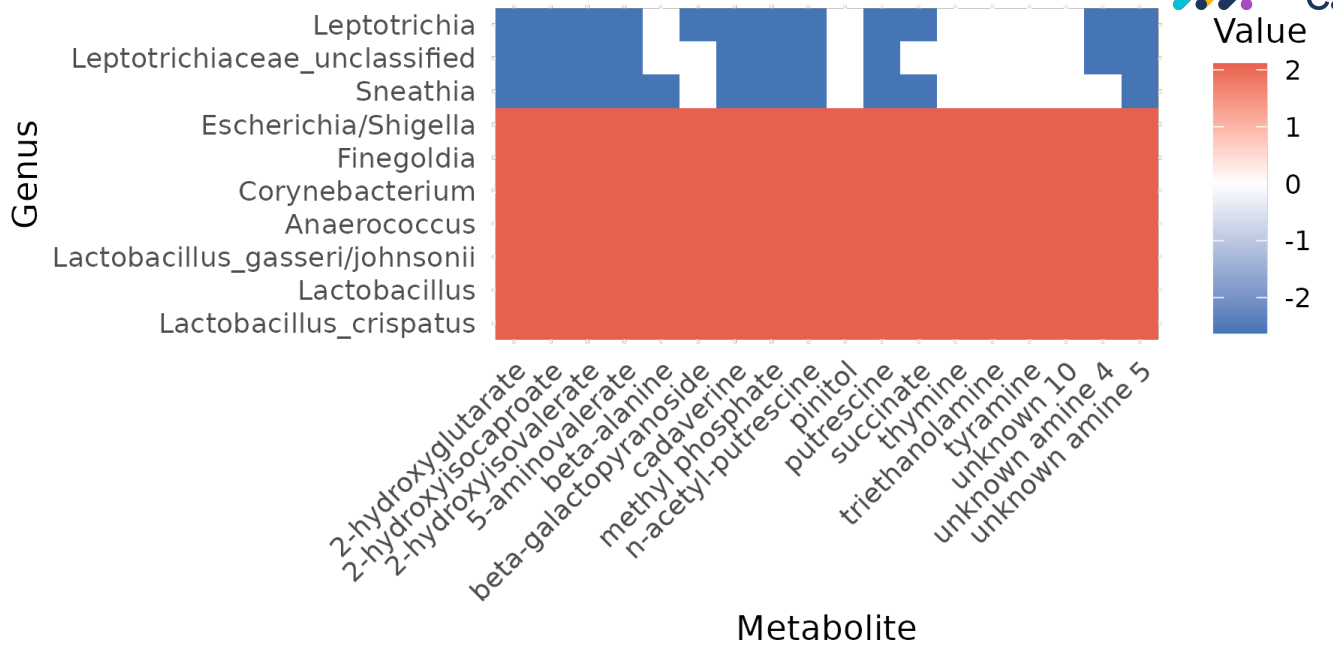



Figure 8: A zoom in view of two biclusters with the largest mean difference.

- Top bicluster consists of *Leptotrichia* and *Sneathia* which are emerging pathogens implicated in BV. Association of these genera with metabolites is higher in BV patients.
- Bottom bicluster consists of *Lactobacillus* species, important for keeping a healthy vaginal microbiome. Association of these genera with metabolites is higher in normal individuals.
- Shed light on uncharacterized metabolites through "Guilt by Association".

# Summary

Fred Hutch Cancer Center

metaMint

- preprint: arXiv.2506.12275
- Software: https://github.com/drjingma/metaMint

Future directions

- Degree heterogeneity is not accounted for.
- Computation: the method is slower than existing methods. Alternative model estimation and/or selection is helpful.

NIH National Institute of General Medical Sciences