Fred Hutch
Cancer Center

Statistical Data Integration

Principled Methods for Analyzing Multi-view Microbiome Data

8 June 2023

Jing Ma
Assistant Professor of Biostatistics
Division of Public Health Sciences

Research Overview

Fred Hutch Cancer Center

Network Biology

- Metabolic networks
- Microbial networks
- Comorbidity networks

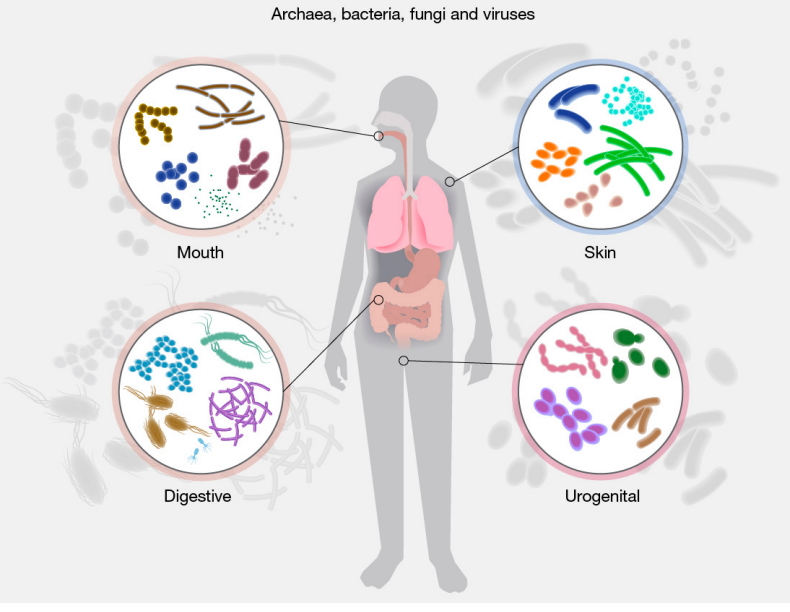
Microbiome

- Network analysis
- High-dimensional inference
- Integration with other Omics

The Human Microbiome

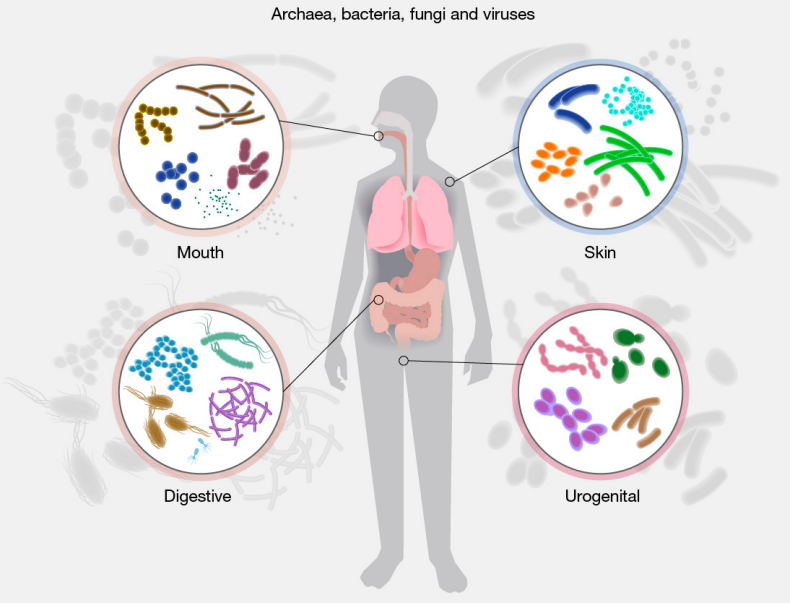
Fred Hutch Cancer Center

All of the microbes and their genome, mostly bacteria



The Human Microbiome

All of the microbes and their genome, mostly bacteria

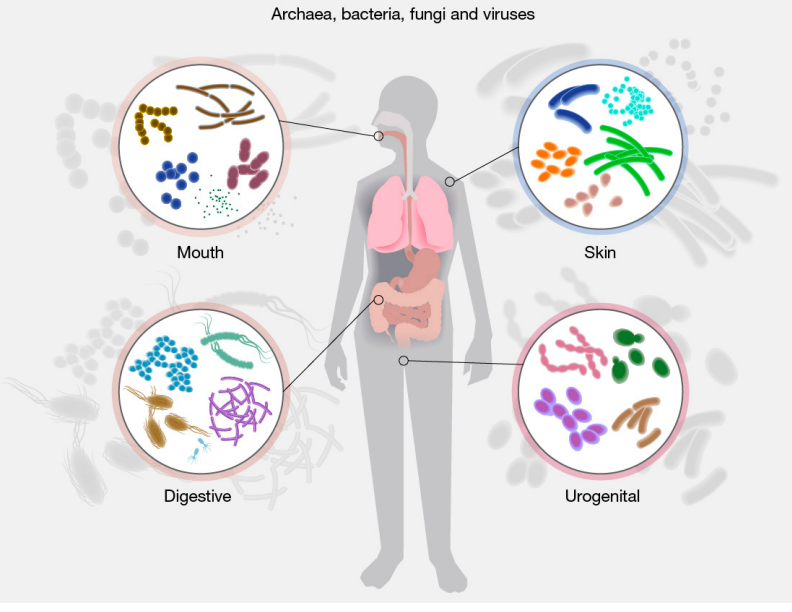


- More microbial cells than our somatic cells

The Human Microbiome

Fred Hutch Cancer Center

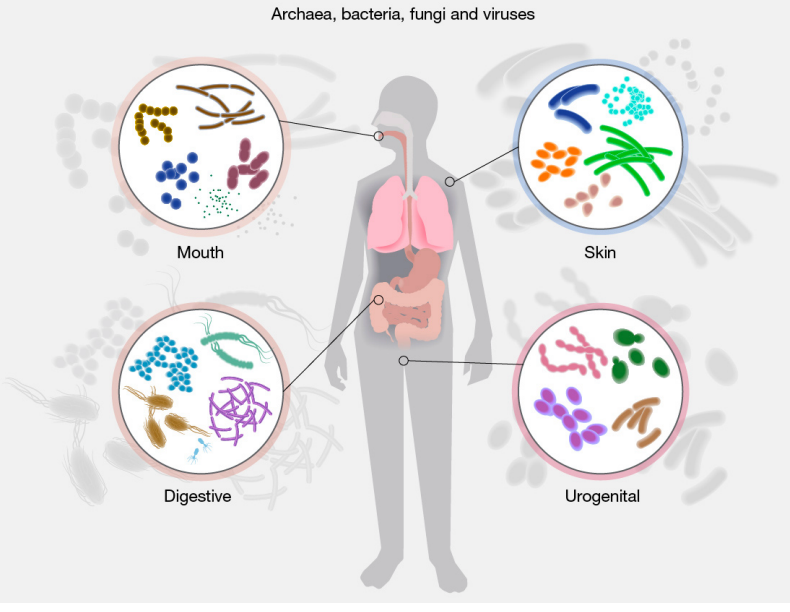
All of the microbes and their genome, mostly bacteria



- More microbial cells than our somatic cells
- More microbial genes than our human genome

The Human Microbiome

All of the microbes and their genome, mostly bacteria

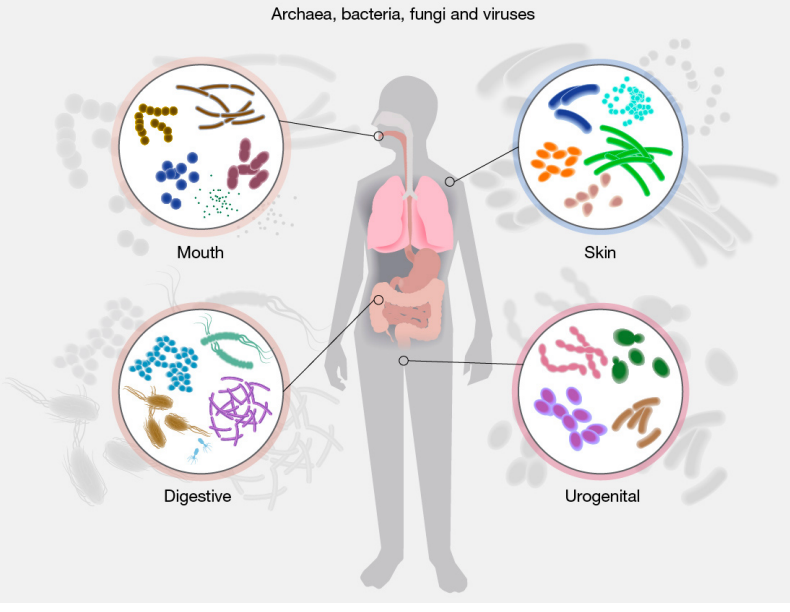


- More microbial cells than our somatic cells
- More microbial genes than our human genome
- Compositions vary within a person and between individuals

The Human Microbiome

Fred Hutch Cancer Center

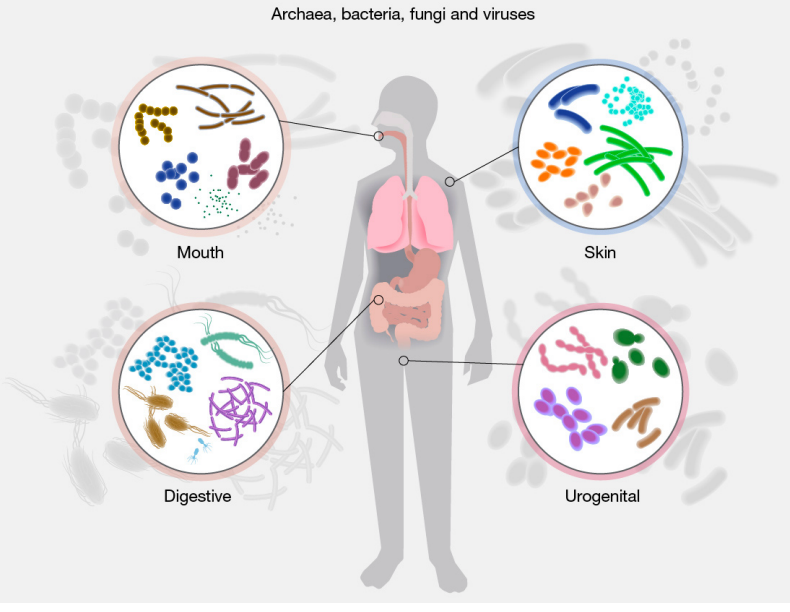
All of the microbes and their genome, mostly bacteria



- 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

The Human Microbiome

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
  - Obesity
  - Inflammatory bowel disease
  - Cancer
  - Neurological disorders
  - etc.

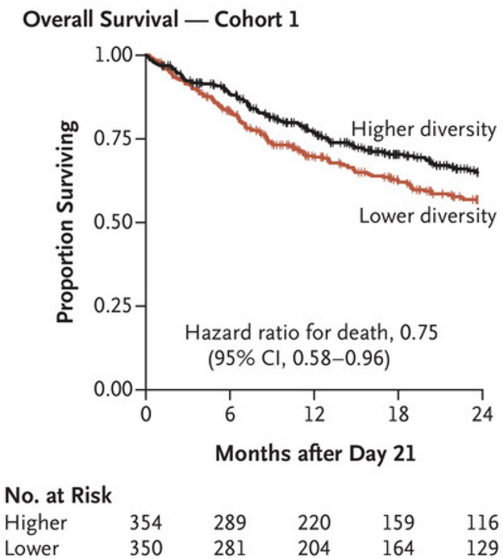
Microbiome and Cancer

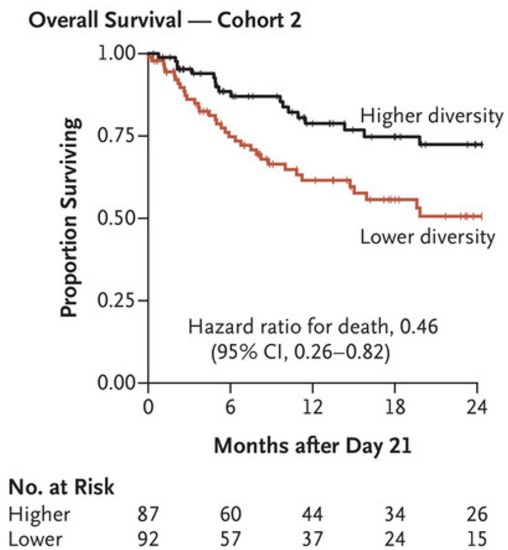
- Allogeneic Hematopoietic-Cell Transplantation (allo-HCT) is a curative therapy for hematologic cancers, but complications such as graft-versus-host disease (GVHD) remain a major cause of illness and death.

[1] Peled et al., NEJM. 20'

Microbiome and Cancer

- Allogeneic Hematopoietic-Cell Transplantation (allo-HCT) is a curative therapy for hematologic cancers, but complications such as graft-versus-host disease (GVHD) remain a major cause of illness and death.
- Lower diversity predicts poor overall survival[1].

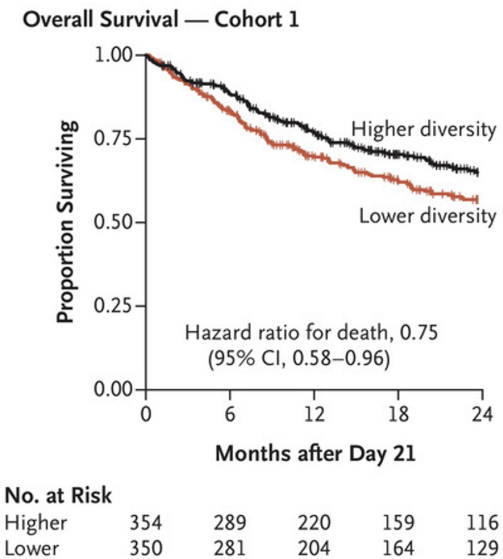


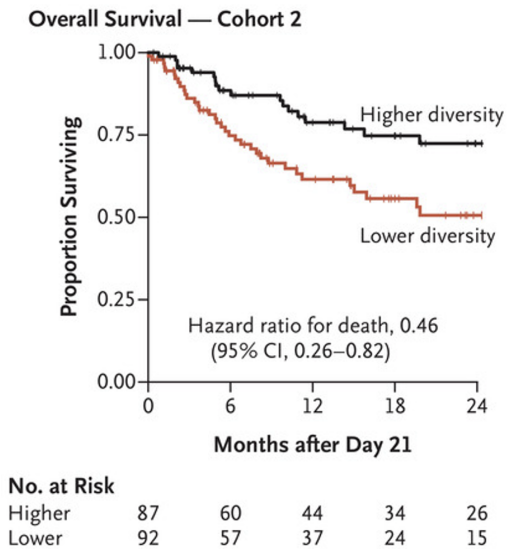


[1] Peled et al., NEJM. 20'

Microbiome and Cancer

- Allogeneic Hematopoietic-Cell Transplantation (allo-HCT) is a curative therapy for hematologic cancers, but complications such as graft-versus-host disease (GVHD) remain a major cause of illness and death.
- Lower diversity predicts poor overall survival[1].





- Interventions to restore integrity to the intestinal microbiota?

[1] Peled et al., NEJM. 20'

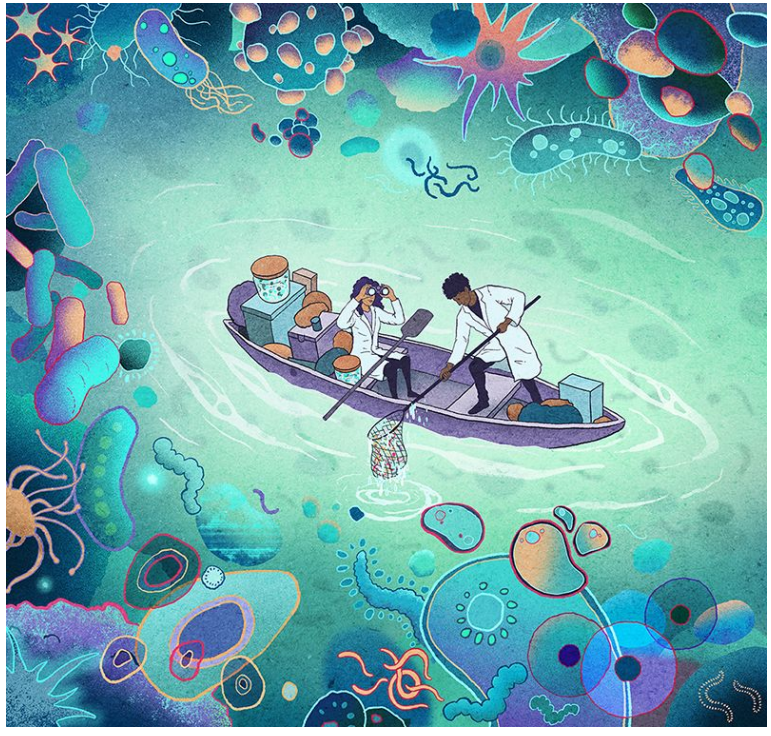
Scientific Questions

Fred Hutch Cancer Center

1. Which bacterial species are associated with poor outcome (e.g., GVHD status)?

Scientific Questions

Fred Hutch Cancer Center



1. Which bacterial species are associated with poor outcome (e.g., GVHD status)?
2. What is the mechanism underlying the association between bacterial species and clinical outcomes?

Microbiome Data

Fred Hutch Cancer Center

OTU	Species	Sample1	Sample2	Sample3
1	E.coli	17	0	335
2	S.aurus	231	1180	45
3	unknown	30	0	0
...	...	...	...	...

Table 1: A typical microbiome contingency table

Microbiome Data

Fred Hutch Cancer Center

OTU	Species	Sample1	Sample2	Sample3
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: # of species $>$ # of samples

Microbiome Data

Fred Hutch Cancer Center

OTU	Species	Sample1	Sample2	Sample3
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: # of species $>$ # of samples
- structured: species and/or samples are correlated

Microbiome Data

Fred Hutch Cancer Center

OTU	Species	Sample1	Sample2	Sample3
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: # of species $>$ # of samples
- structured: species and/or samples are correlated
- sparse: lots of zeros

Microbiome Data

Fred Hutch Cancer Center

OTU	Species	Sample1	Sample2	Sample3
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: # of species $>$ # of samples
- structured: species and/or samples are correlated
- sparse: lots of zeros
- compositional: only relative abundances are meaningful

Part I

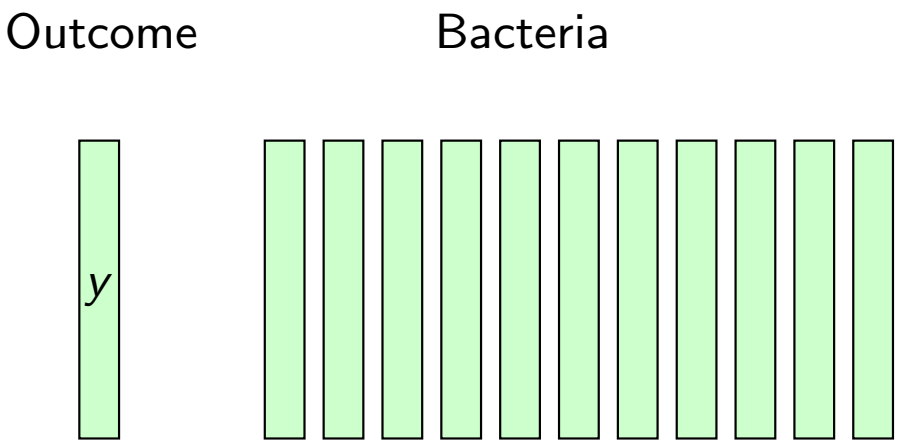
Fred Hutch
Cancer Center

Network-based Variable Selection via GMDR

Univariate Analysis

Fred Hutch Cancer Center

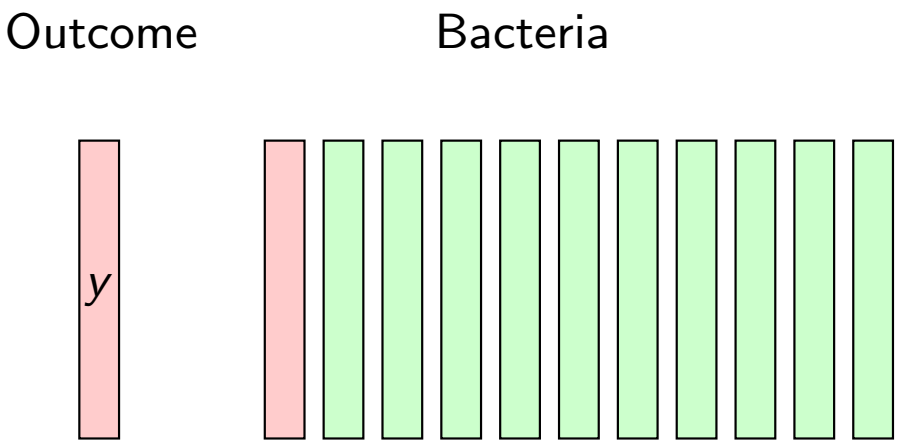
Perform univariate test of each species with respect to an outcome



Univariate Analysis

Fred Hutch Cancer Center

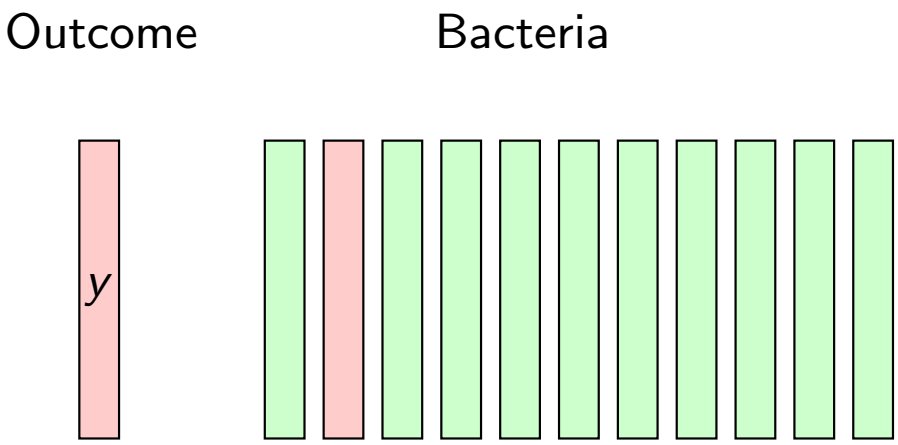
Perform univariate test of each species with respect to an outcome



Univariate Analysis

Fred Hutch Cancer Center

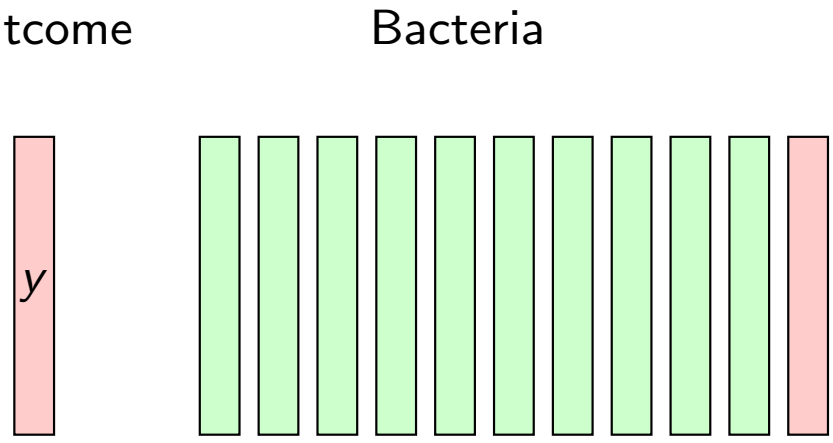
Perform univariate test of each species with respect to an outcome



Univariate Analysis

Fred Hutch Cancer Center

Perform univariate test of each species with respect to an outcome



Univariate Analysis

Fred Hutch Cancer Center

Output:

- p-value for testing each bacterial species (after correcting for multiple comparisons)

[2] Paulson et al. Nat Meth. 13'
[3] Martin et al. AoAS. 20'
[4] Ling et al. Microbiome. 21'

Univariate Analysis

Fred Hutch Cancer Center

Output:

- p-value for testing each bacterial species (after correcting for multiple comparisons)

Variations:

- different normalization methods

[2] Paulson et al. Nat Meth. 13'
[3] Martin et al. AoAS. 20'
[4] Ling et al. Microbiome. 21'

Univariate Analysis

Fred Hutch Cancer Center

Output:

- p-value for testing each bacterial species (after correcting for multiple comparisons)

Variations:

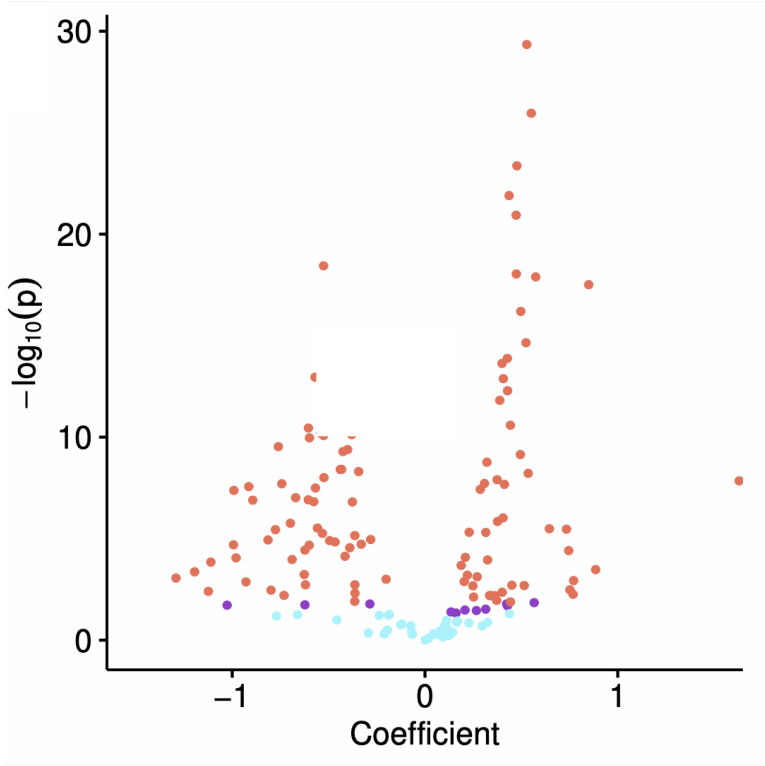
- different normalization methods
- different noise models for bacterial abundances (e.g., zero-inflated log normal[2], beta-binomial[3], zero-inflated quantile regression[4], etc.)

[2] Paulson et al. Nat Meth. 13'
[3] Martin et al. AoAS. 20'
[4] Ling et al. Microbiome. 21'

Application to Yatsunenko 12'

Which bacteria are associated with age?

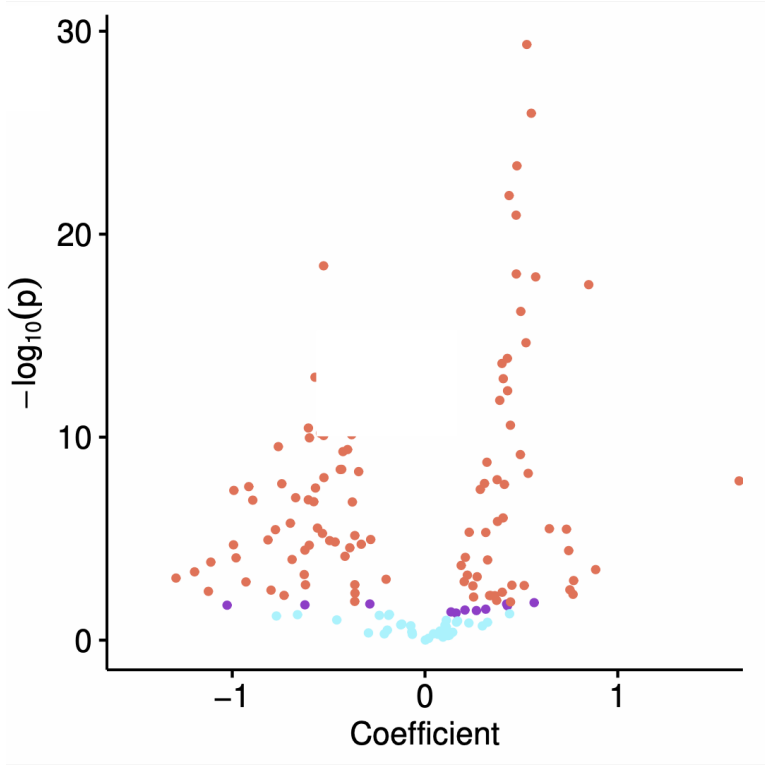
- $n = 100$
- $p = 149$ after pre-processing
- Apply univariate Spearman rank correlation test between log transformed abundance and age



Application to Yatsunenko 12'

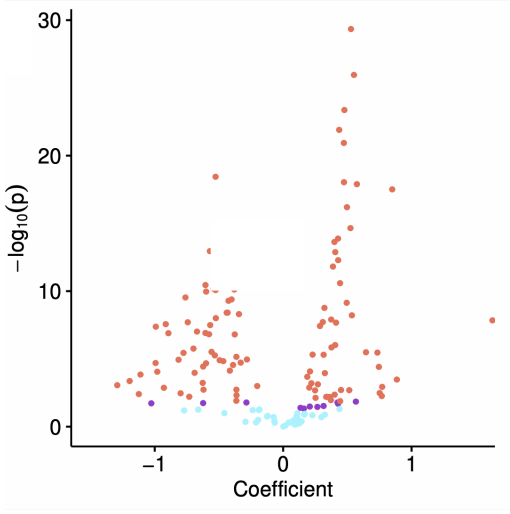
Which bacteria are associated with age?

- $n = 100$
- $p = 149$ after pre-processing
- Apply univariate Spearman rank correlation test between log transformed abundance and age

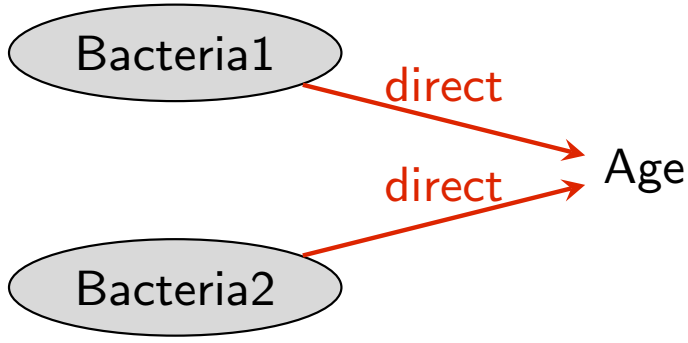


☹ Which bacteria should I target?

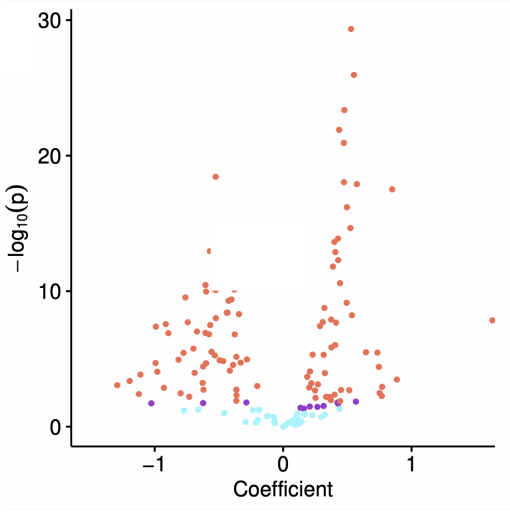
Two Possible Explanations



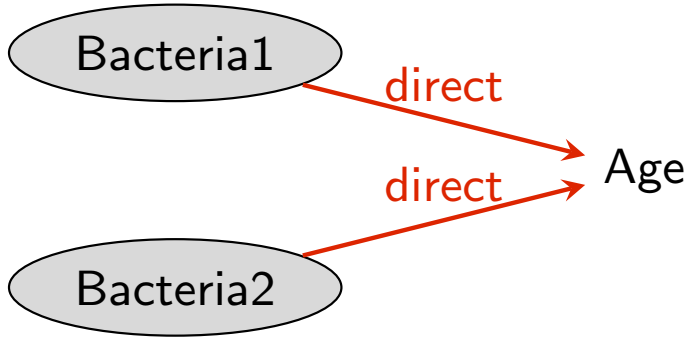
- Many bacteria are *directly* associated with age



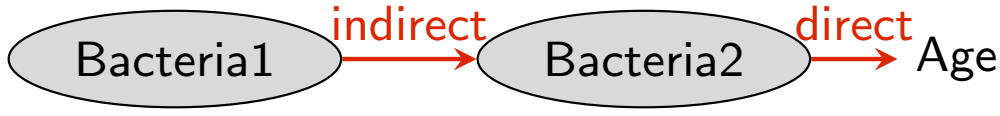
Two Possible Explanations



- Many bacteria are *directly* associated with age



- Bacteria are correlated and only a few bacteria are *directly* associated with age



Conditional Associations

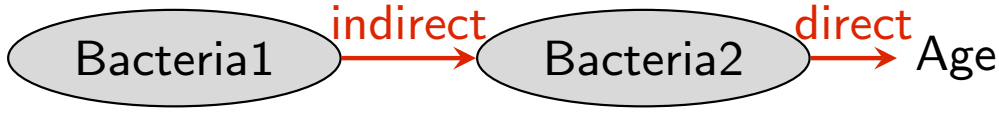


Figure 1: Conditional on Bacteria 2, Bacteria 1 is independent of Age.

Multiple linear regression

$$y_i = X_{i,1}\beta_1 + \ldots + X_{i,p}\beta_p + \text{error}_i,$$

- Coefficients β_j: conditional association between bacteria j and y
- When $p \ll n$, inference for β is straightforward by large sample theory.

Curse of Dimensionality

Fred Hutch Cancer Center

Number of unknown parameters p is usually large (e.g., $p = 149$)

Curse of Dimensionality

Fred Hutch Cancer Center

Number of unknown parameters p is usually large (e.g., $p = 149$)

- Assume sparse effects: most β_j's are zero.

  This is common, maybe reasonable, but uncheckable.

Curse of Dimensionality

Fred Hutch Cancer Center

Number of unknown parameters p is usually large (e.g., $p = 149$)

- Assume sparse effects: most β_j's are zero.
  This is common, maybe reasonable, but uncheckable.

- Use prior knowledge.
  Can boost power in biomarker discovery.

The Phylogenetic Prior

The evolutionary history of bacterial species can be summarized by a phylogenetic tree.

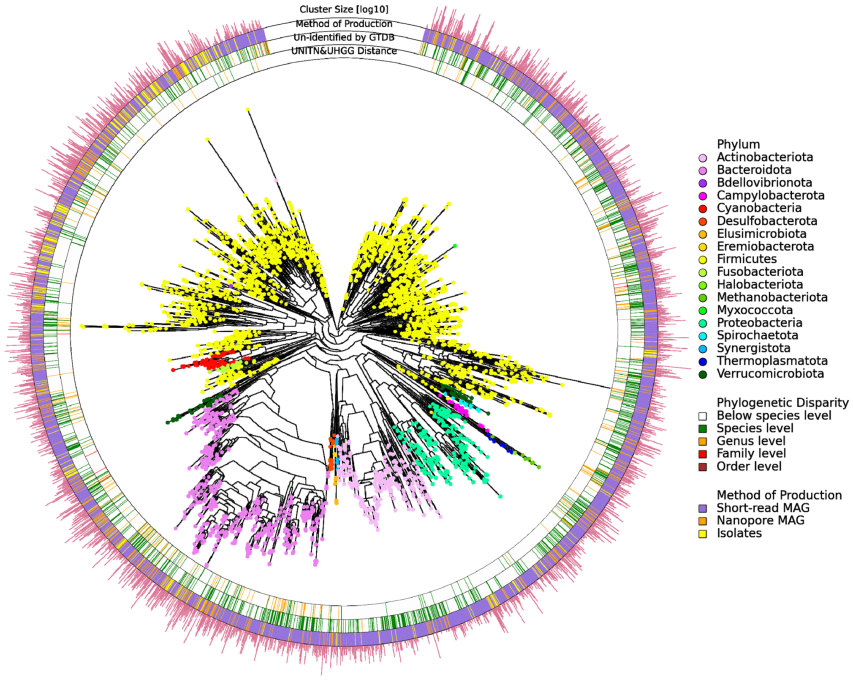


Figure 2: Leviatan et al. Nature Communications. 22'

We can construct a similarity network capturing the evolutionary relationships between bacterial species.

Are Samples Independent?

Fred Hutch Cancer Center

Most methods assume the samples are independent and identically distributed.

Is this assumption valid?

Are Samples Independent?

Most methods assume the samples are independent and identically distributed.
Is this assumption valid?

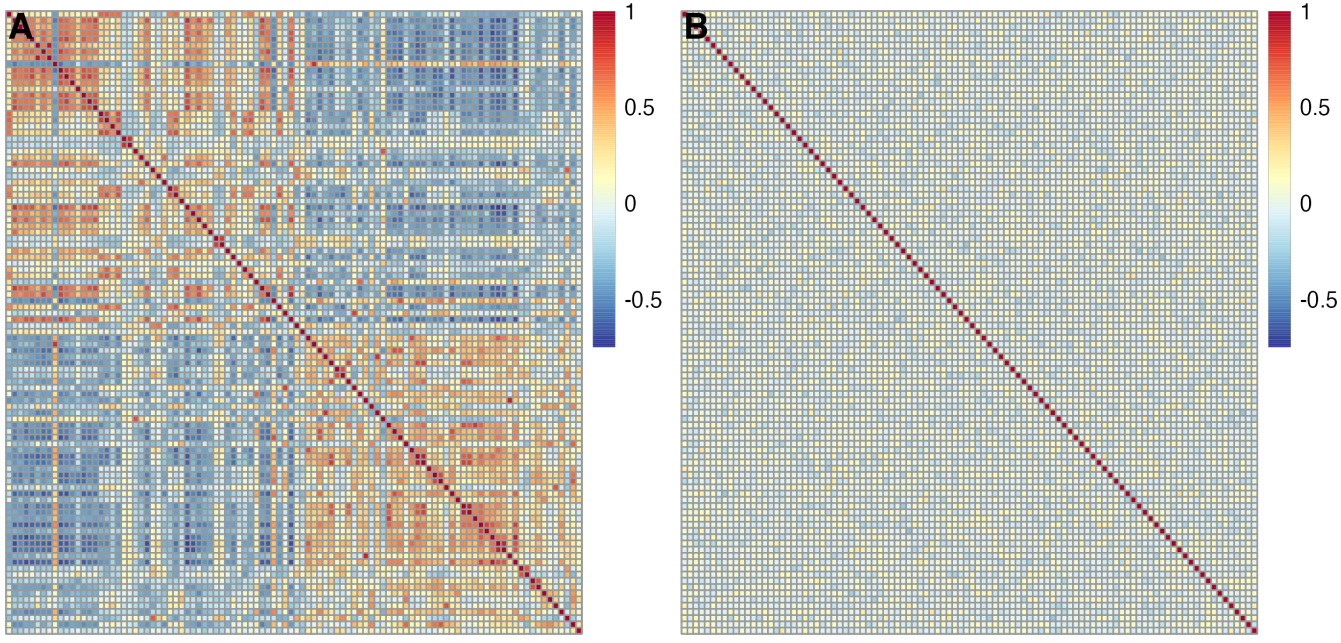


Figure 3: Correlation among (A) Yatsunenko samples; (B) independent samples

Are Samples Independent?

Most methods assume the samples are independent and identically distributed. Is this assumption valid?

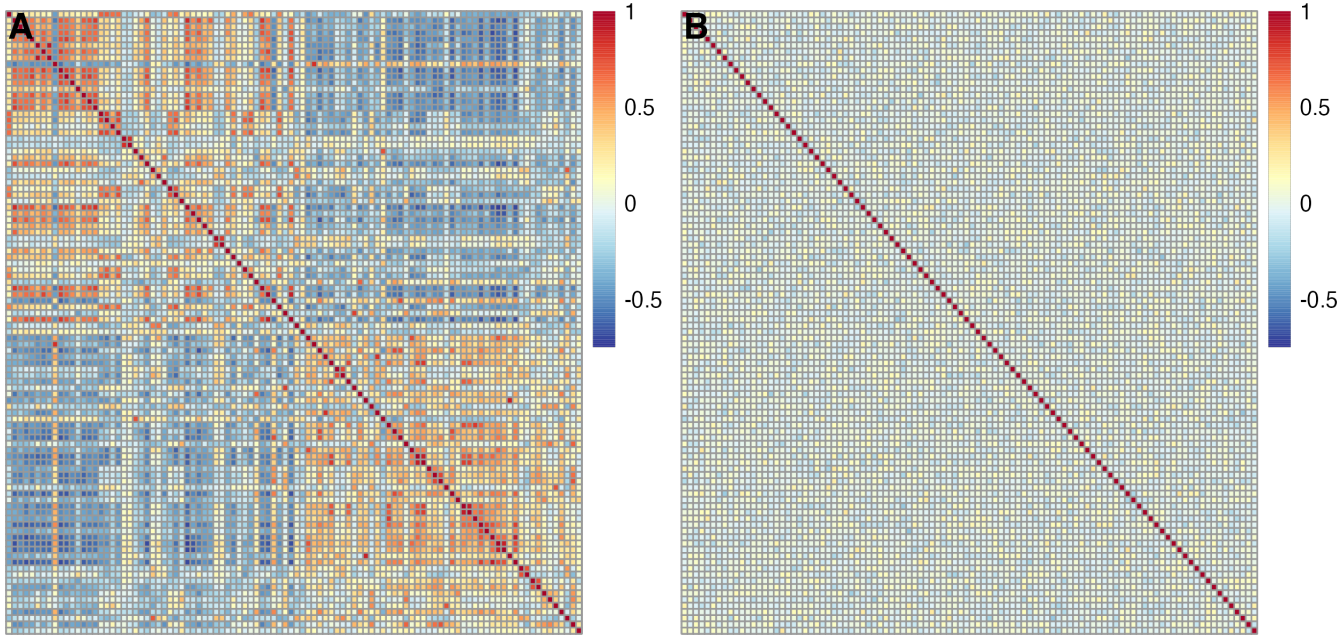


Figure 3: Correlation among (A) Yatsunenko samples; (B) independent samples

In Yatsunenko 12', subjects include individuals from the same household (twins, parent-offspring relationships). This could lead to inflated type I error if unaccounted for.

Dealing with Dependent Samples

Fred Hutch Cancer Center

Account for sample dependence analytically, e.g. via generalized least squares.

Dealing with Dependent Samples

Fred Hutch Cancer Center

Account for sample dependence analytically, e.g. via generalized least squares.

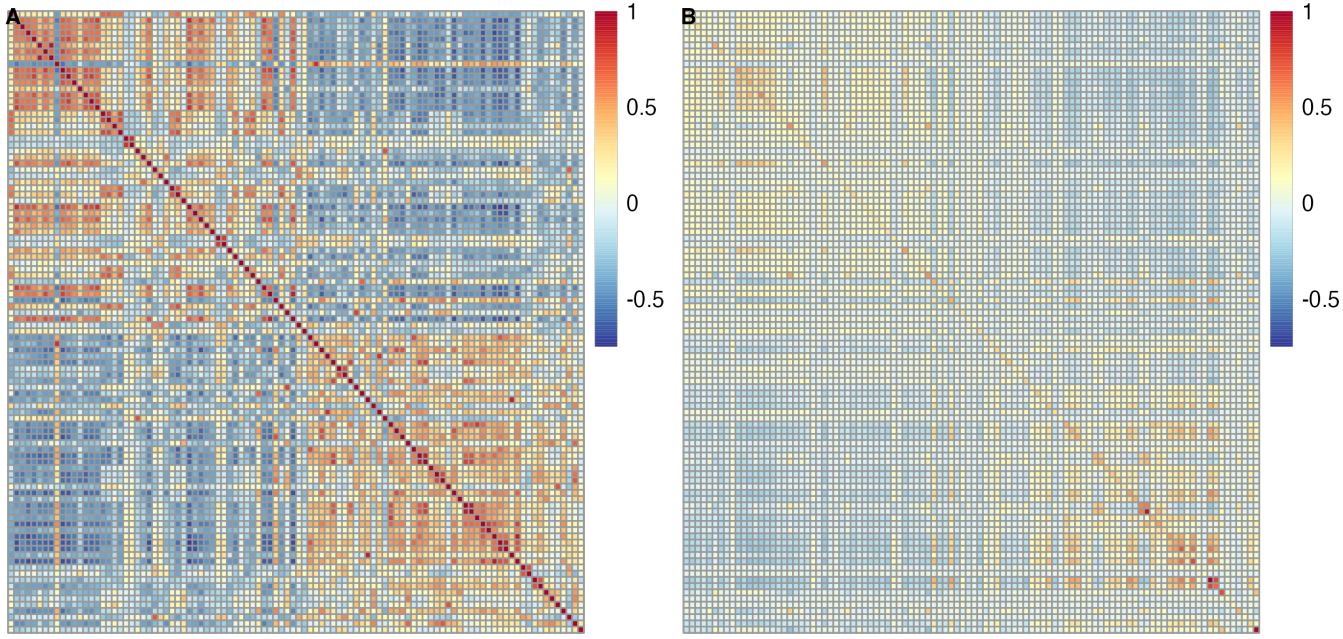


Figure 4: Sample correlation from (A) 16S abundance and (B) metagenomic pathway abundance

Dealing with Dependent Samples

Account for sample dependence analytically, e.g. via generalized least squares.

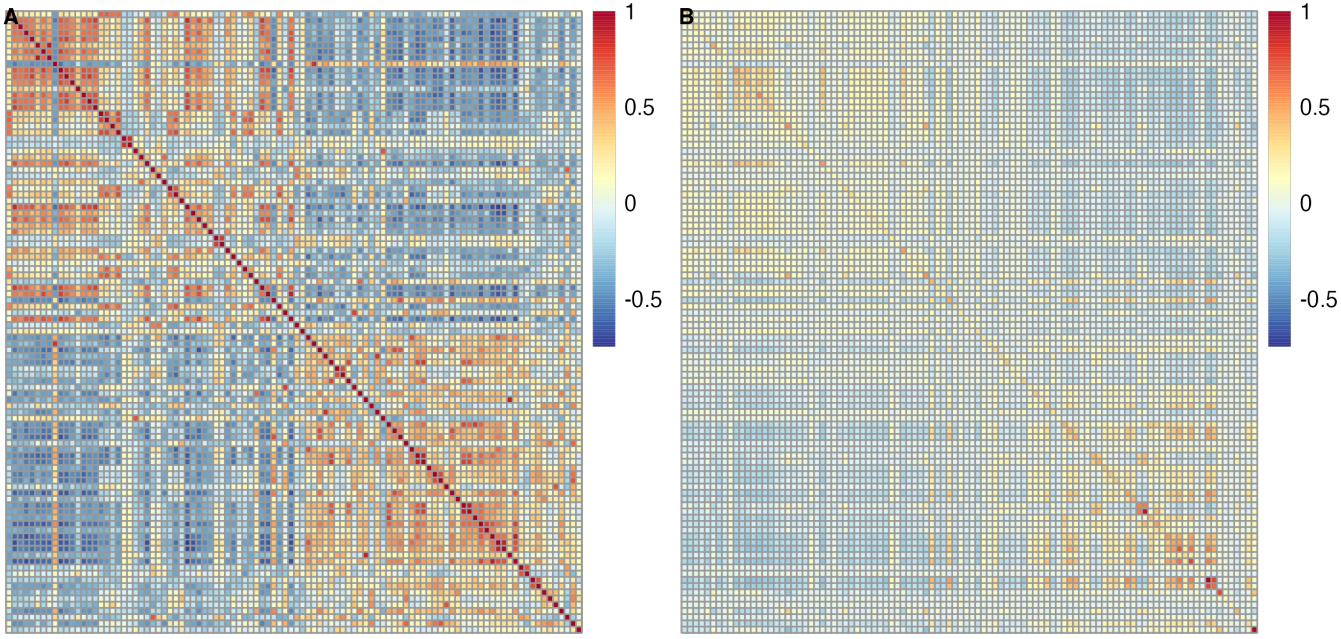


Figure 4: Sample correlation from (A) 16S abundance and (B) metagenomic pathway abundance

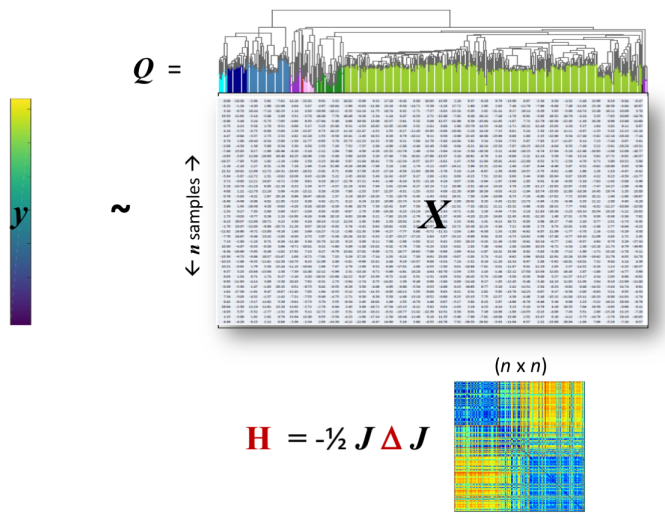
Sample correlation from metagenomic data provides prior knowledge on the dependence structure among samples.

Generalized Matrix Decomposition Regression

$$y_i = X_{i,1}\beta_1 + \ldots + X_{i,p}\beta_p + \text{error}_i,$$

subject to the constraints that

- the coefficients β are smooth with respect to a variable similarity network Q
- the error covariance is smooth with respect to the inverse of a sample similarity network H



GMDR Algorithm

Fred Hutch Cancer Center

- Perform GMD[5] on X

$$\min_{U,S,V} \|X - USV'\|_{H,Q}$$

where $\|A\|_{H,Q} = \text{trace}(A'HAQ)$.

[5] Allen et al. JASA. 14'
[6] Wang et al. AoAS. 23'

GMDR Algorithm

Fred Hutch Cancer Center

- Perform GMD[5] on X

$$\min_{U,S,V} \|X - USV'\|_{H,Q}$$

  where $\|A\|_{H,Q} = \text{trace}(A'HAQ)$.

- Regress y on GMD components with index set $\mathcal{I}$

$$\hat{\gamma}(\mathcal{I}) = \arg\min_{\gamma} \|y - [US]_{\mathcal{I}}\gamma\|_H^2$$

[5] Allen et al. JASA. 14'

[6] Wang et al. AoAS. 23'

GMDR Algorithm

Fred Hutch Cancer Center

- Perform GMD[5] on X

$$\min_{U,S,V} \|X - USV'\|_{H,Q}$$

  where $\|A\|_{H,Q} = \text{trace}(A'HAQ)$.

- Regress y on GMD components with index set $\mathcal{I}$

$$\hat{\gamma}(\mathcal{I}) = \arg\min_{\gamma} \|y - [US]_{\mathcal{I}}\gamma\|_H^2$$

- Calculate

$$\hat{\beta}_{GMDR}(\mathcal{I}) = [QV]_{\mathcal{I}}\hat{\gamma}(\mathcal{I})$$

[5] Allen et al. JASA. 14'
[6] Wang et al. AoAS. 23'

GMDR Algorithm

Fred Hutch Cancer Center

- Perform GMD[5] on X

$$\min_{U,S,V} \|X - USV'\|_{H,Q}$$

  where $\|A\|_{H,Q} = \text{trace}(A'HAQ)$.

- Regress y on GMD components with index set $\mathcal{I}$

$$\hat{\gamma}(\mathcal{I}) = \arg\min_{\gamma} \|y - [US]_{\mathcal{I}}\gamma\|_H^2$$

- Calculate

$$\hat{\beta}_{GMDR}(\mathcal{I}) = [QV]_{\mathcal{I}}\hat{\gamma}(\mathcal{I})$$

- Perform GMD inference (GMDI[6]) $H_{0,j} : \beta_j = 0$ for $j = 1, \ldots, p$.

[5] Allen et al. JASA. 14'

[6] Wang et al. AoAS. 23'

Remarks

Fred Hutch Cancer Center

GMDR can be viewed as generalization of principal component regression or ridge regression to two-way structured data

- Discrete: use selected GMD components
- Continuous: shrink all GMD components by variation explained

Robust GMDR and GMDI

Fred Hutch Cancer Center

What if prior information is misspecified? Suppose Q is correct but H is misspecified.

Robust GMDR and GMDI

What if prior information is misspecified? Suppose Q is correct but H is misspecified.

1. Test the association between H^{-1} and $H_x = XX'$:

$$\text{KRV}(H_x, H) = \frac{\text{tr}(\widetilde{H}_x \widetilde{H})}{\sqrt{\text{tr}(\widetilde{H}_x^2)\text{tr}(\widetilde{H}^2)}}$$

where $\widetilde{H}_x = JH_xJ$ and $\widetilde{H} = JH^{-1}J$. J is the double-centering matrix.

Robust GMDR and GMDI

What if prior information is misspecified? Suppose Q is correct but H is misspecified.

1. Test the association between H^{-1} and $H_x = XX'$:

$$\text{KRV}(H_x, H) = \frac{\text{tr}(\widetilde{H}_x \widetilde{H})}{\sqrt{\text{tr}(\widetilde{H}_x^2)\text{tr}(\widetilde{H}^2)}}$$

   where $\widetilde{H}_x = JH_xJ$ and $\widetilde{H} = JH^{-1}J$. J is the double-centering matrix.

2. For partially informative structures, use a likelihood criterion to weight the prior against a uninformative baseline.

Robust GMDR and GMDI

Fred Hutch Cancer Center

What if prior information is misspecified? Suppose Q is correct but H is misspecified.

1. Test the association between H^{-1} and $H_x = XX'$:

$$\text{KRV}(H_x, H) = \frac{\text{tr}(\widetilde{H}_x \widetilde{H})}{\sqrt{\text{tr}(\widetilde{H}_x^2)\text{tr}(\widetilde{H}^2)}}$$

   where $\widetilde{H}_x = JH_xJ$ and $\widetilde{H} = JH^{-1}J$. J is the double-centering matrix.

2. For partially informative structures, use a likelihood criterion to weight the prior against a uninformative baseline.

   Let $H(\tau) = \tau H + (1-\tau)I_n$ for $\tau \in (0,1)$. Then for some $c_1, c_2 > 0$

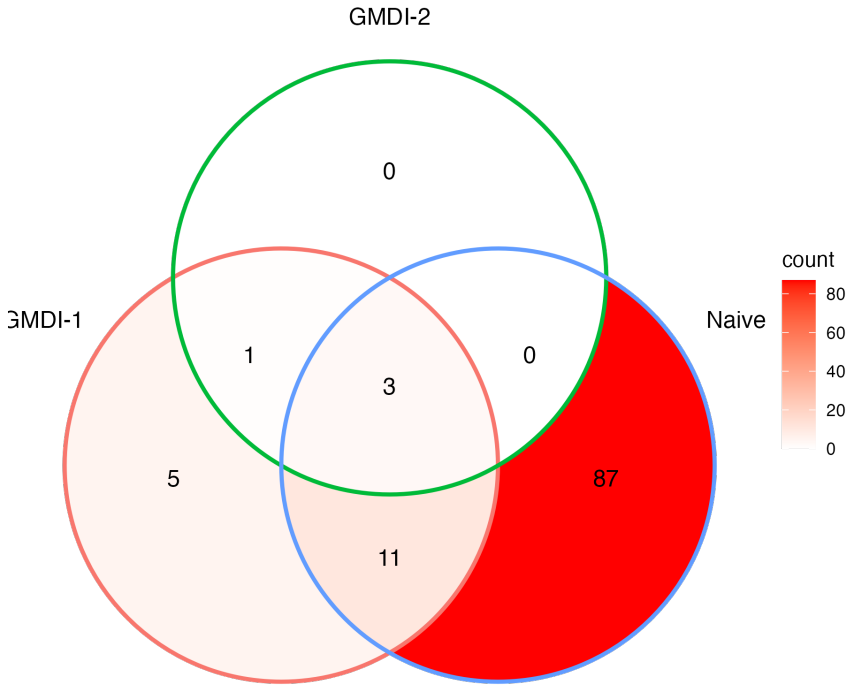
$$\begin{aligned} y &= X\beta + \epsilon \\ \beta &\sim N(0, c_1 Q) \\ \epsilon &\sim N(0, c_2\{H(\tau)\}^{-1}) \end{aligned}$$

Revisit Yatsunenko 12'

Fred Hutch Cancer Center

Which bacteria are associated with age?

- $n = 100$
- $p = 149$
- FDR $= 0.1$
- Results from robust GMDI



Visualization

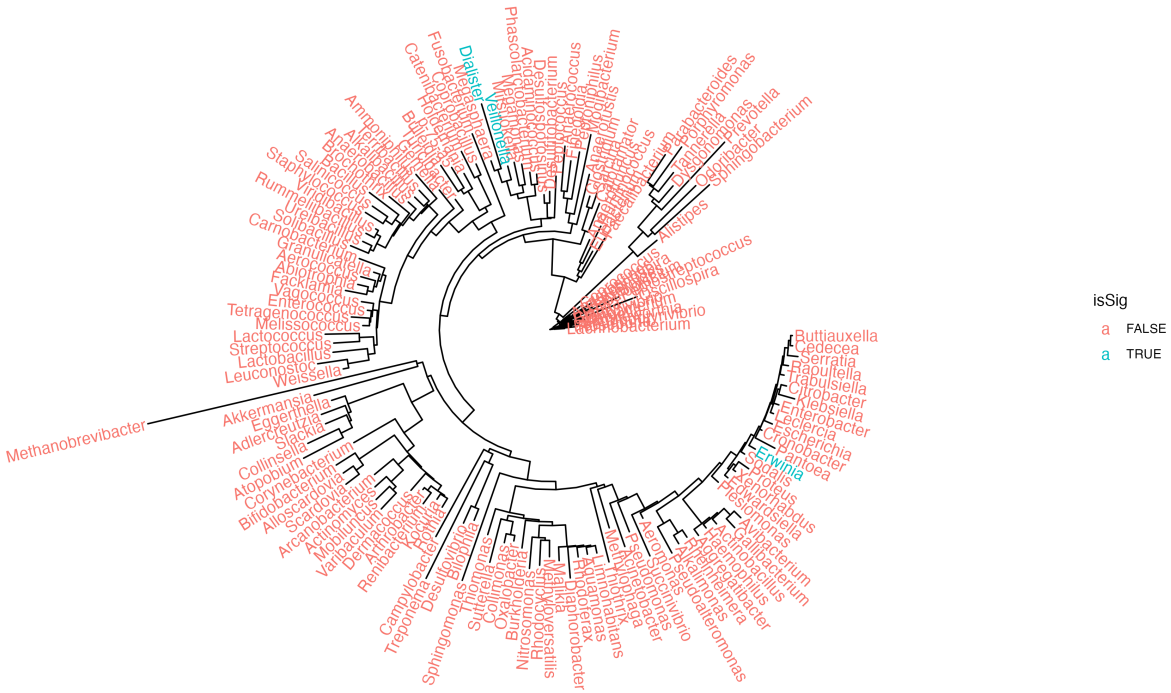


Figure 5: *Dialister* and *Veillonella* are phylogenetically close.

- Dialister has been shown to play a role in age-related diseases, such as obesity and diabetes[7].
- Veillonella is a signature of infant (4-month old) microbiome and breast feeding[8].

[7] Xu et al., 20'; Gurung et al., 20'
[8] Backhed et al., 15'

Summary of Part I

Fred Hutch Cancer Center

Variable selection with high-dimensional and structured microbiome data is a hard problem!

[9] Wang et al. AoAS. 23'

Summary of Part I

Fred Hutch Cancer Center

Variable selection with high-dimensional and structured microbiome data is a hard problem!

We provide GMDR and GMDI[9] which

[9] Wang et al. AoAS. 23'

Summary of Part I

Fred Hutch Cancer Center

Variable selection with high-dimensional and structured microbiome data is a hard problem!

We provide GMDR and GMDI[9] which

- identify features that are conditionally associated with an outcome

[9] Wang et al. AoAS. 23'

Summary of Part I

Fred Hutch Cancer Center

Variable selection with high-dimensional and structured microbiome data is a hard problem!

We provide GMDR and GMDI[9] which

- identify features that are conditionally associated with an outcome
- incorporate prior knowledge to alleviate model complexity and produce biologically meaningful results

[9] Wang et al. AoAS. 23'

Summary of Part I

Fred Hutch Cancer Center

Variable selection with high-dimensional and structured microbiome data is a hard problem!

We provide GMDR and GMDI[9] which

- identify features that are conditionally associated with an outcome
- incorporate prior knowledge to alleviate model complexity and produce biologically meaningful results
  - Compositional constraints can be imposed by a degenerate Q.

[9] Wang et al. AoAS. 23'

Summary of Part I

Fred Hutch Cancer Center

Variable selection with high-dimensional and structured microbiome data is a hard problem!

We provide GMDR and GMDI[9] which

- identify features that are conditionally associated with an outcome
- incorporate prior knowledge to alleviate model complexity and produce biologically meaningful results
  - Compositional constraints can be imposed by a degenerate Q.
- can handle mis-specification in priors.

[9] Wang et al. AoAS. 23'

Summary of Part I

Fred Hutch Cancer Center

Variable selection with high-dimensional and structured microbiome data is a hard problem!

We provide GMDR and GMDI[9] which

- identify features that are conditionally associated with an outcome
- incorporate prior knowledge to alleviate model complexity and produce biologically meaningful results
  - Compositional constraints can be imposed by a degenerate Q.
- can handle mis-specification in priors.

Limitations

[9] Wang et al. AoAS. 23'

Summary of Part I

Fred Hutch Cancer Center

Variable selection with high-dimensional and structured microbiome data is a hard problem!

We provide GMDR and GMDI[9] which

- identify features that are conditionally associated with an outcome
- incorporate prior knowledge to alleviate model complexity and produce biologically meaningful results
  - Compositional constraints can be imposed by a degenerate Q.
- can handle mis-specification in priors.

Limitations

- Data are observational, so need to be cautious about making any causal interpretations.

[9] Wang et al. AoAS. 23'

Summary of Part I

Fred Hutch Cancer Center

Variable selection with high-dimensional and structured microbiome data is a hard problem!

We provide GMDR and GMDI[9] which

- identify features that are conditionally associated with an outcome
- incorporate prior knowledge to alleviate model complexity and produce biologically meaningful results
  - Compositional constraints can be imposed by a degenerate Q.
- can handle mis-specification in priors.

Limitations

- Data are observational, so need to be cautious about making any causal interpretations.
- Zeros in data are not directly accounted for.

[9] Wang et al. AoAS. 23'

Part II

Fred Hutch Cancer Center

1. Which bacterial species are associated with poor outcome (e.g., GVHD status)?
2. What is the mechanism underlying the association between bacterial species and clinical outcomes?

Immune Regulation by Microbial Metabolites

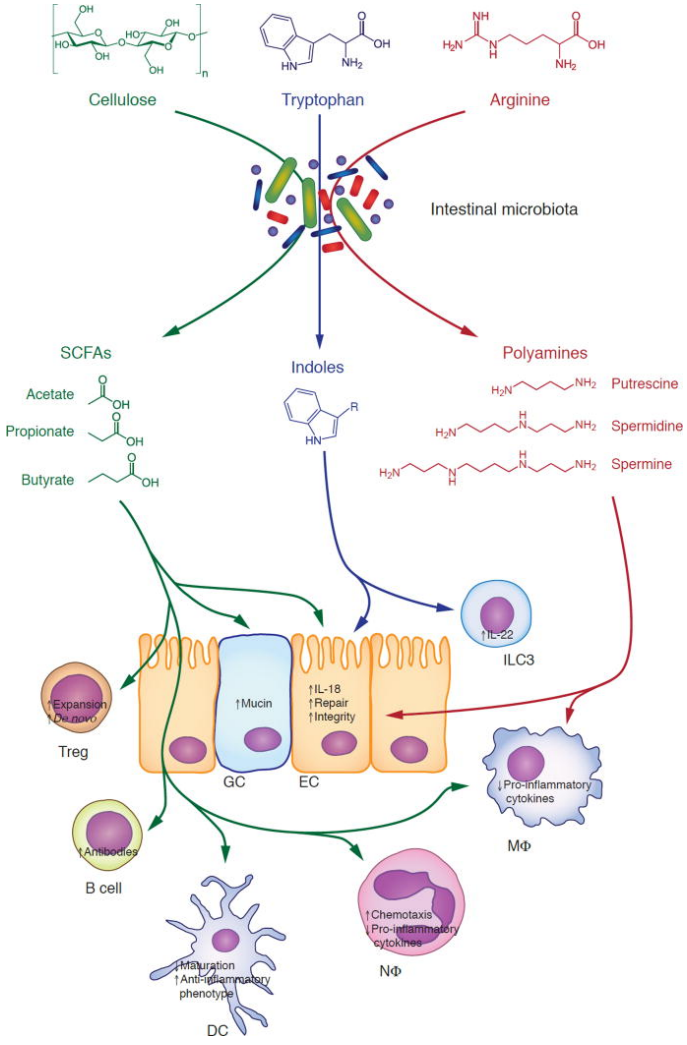


Figure 6: Postler and Ghosh. Cell Metabolism, 17'

- Gut bacteria convert dietary nutrients to immunomodulatory metabolites.

Immune Regulation by Microbial Metabolites

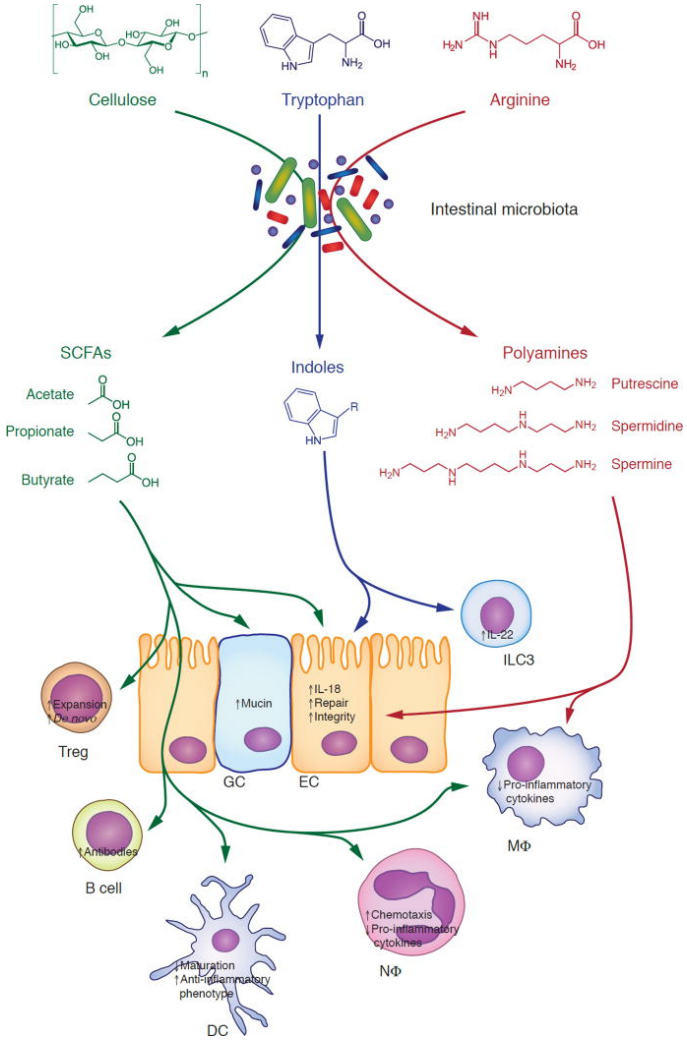


Figure 6: Postler and Ghosh. Cell Metabolism, 17'

- Gut bacteria convert dietary nutrients to immunomodulatory metabolites.
- Scientific Question: *Which bacteria produce which metabolites and eventually lead to changes in clinical outcomes?*

Immune Regulation by Microbial Metabolites

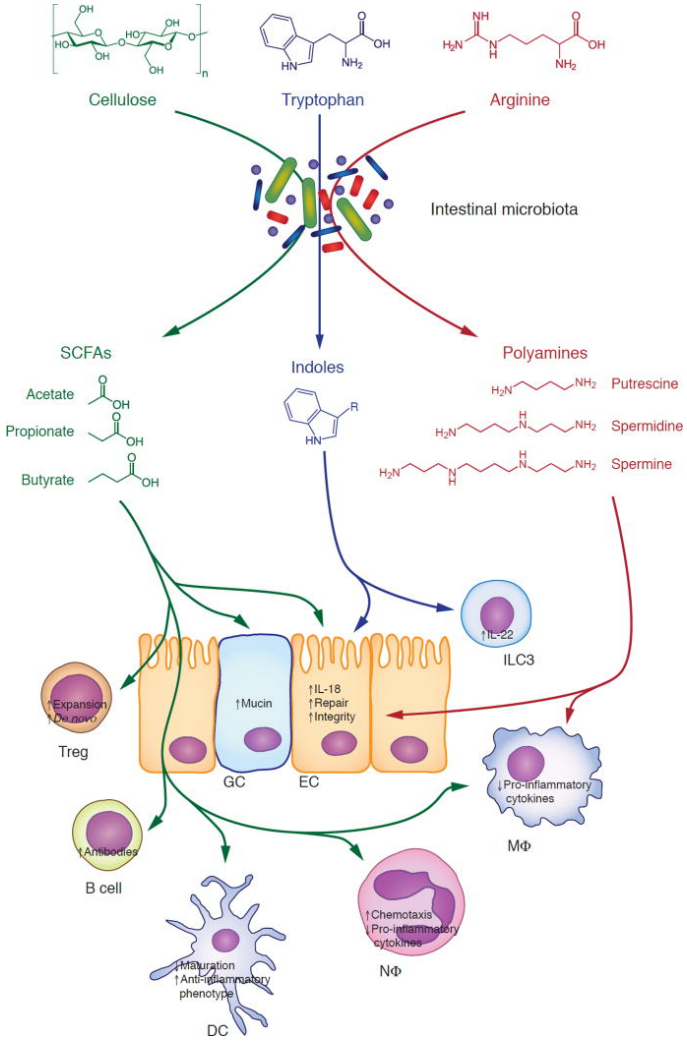


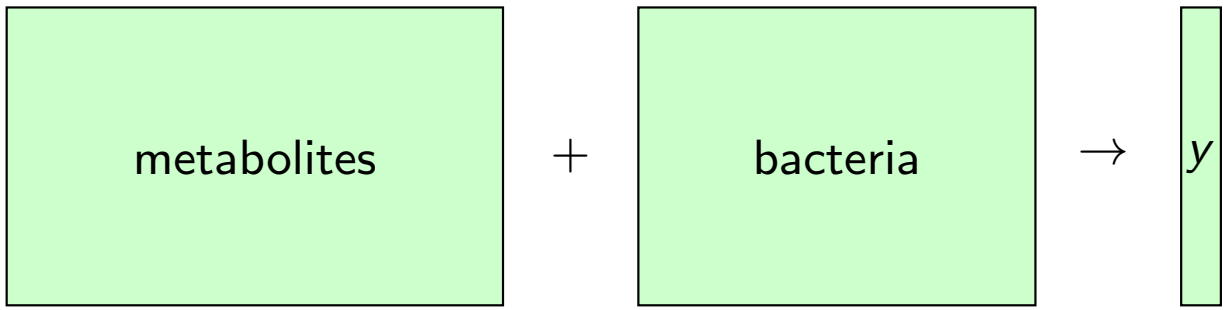
Figure 6: Postler and Ghosh. Cell Metabolism, 17'

- Gut bacteria convert dietary nutrients to immunomodulatory metabolites.
- Scientific Question: *Which bacteria produce which metabolites and eventually lead to changes in clinical outcomes?*
- Challenges: Impractical to experimentally validate all combinations in complex microbial communities.

Multi-view Regression

Fred Hutch Cancer Center

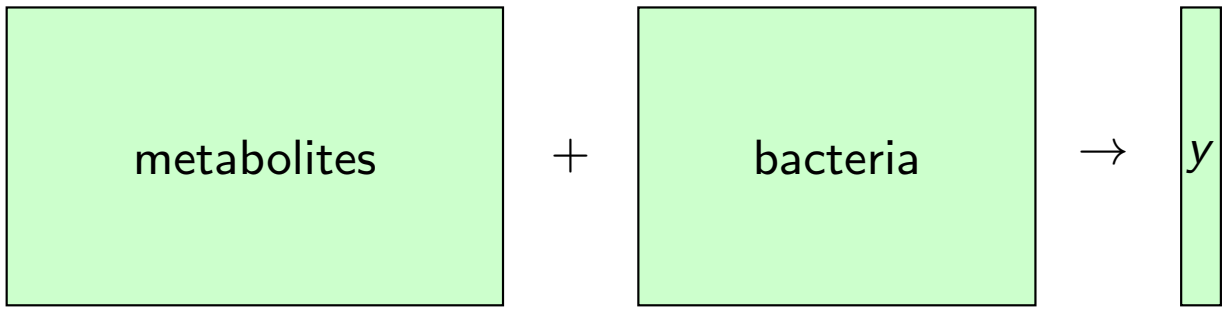
Objective: identify coherent patterns across datasets that are associated with a clinical outcome (y).



[10] Ding et al. PNAS. 22'
[11] Li et al. JASA. 22'
[12] Kakade and Foster. COLT. 07'

Multi-view Regression

Objective: identify coherent patterns across datasets that are associated with a clinical outcome (y).



Multi-view regression is a common statistical problem with applications in genomics[10], neuroscience[11], and machine learning[12].

[10] Ding et al. PNAS. 22'
[11] Li et al. JASA. 22'
[12] Kakade and Foster. COLT. 07'

Current Methods and Limitations

Fred Hutch Cancer Center

Current Methods

- Canonical correlation (e.g., CVR[13], DIABLO[14])

[13] Luo et al. Biostatistics. 16'
[14] Singh et al. Bioinformatics. 19'
[15] Palzer et al. CSDA. 22'
[16] Zhu et al. Biostatistics. 16'

Current Methods and Limitations

Fred Hutch Cancer Center

Current Methods

- Canonical correlation (e.g., CVR[13], DIABLO[14])
- Joint matrix decomposition (e.g., sJIVE[15])

[13] Luo et al. Biostatistics. 16'
[14] Singh et al. Bioinformatics. 19'
[15] Palzer et al. CSDA. 22'
[16] Zhu et al. Biostatistics. 16'

Current Methods and Limitations

Fred Hutch Cancer Center

Current Methods

- Canonical correlation (e.g., CVR[13], DIABLO[14])
- Joint matrix decomposition (e.g., sJIVE[15])
- Hierarchical integration (e.g., LRM[16])

[13] Luo et al. Biostatistics. 16'
[14] Singh et al. Bioinformatics. 19'
[15] Palzer et al. CSDA. 22'
[16] Zhu et al. Biostatistics. 16'

Current Methods and Limitations

Fred Hutch Cancer Center

Current Methods

- Canonical correlation (e.g., CVR[13], DIABLO[14])
- Joint matrix decomposition (e.g., sJIVE[15])
- Hierarchical integration (e.g., LRM[16])

Limitations: lack of uncertainty quantification

[13] Luo et al. Biostatistics. 16'
[14] Singh et al. Bioinformatics. 19'
[15] Palzer et al. CSDA. 22'
[16] Zhu et al. Biostatistics. 16'

Current Methods and Limitations

Fred Hutch Cancer Center

Current Methods

- Canonical correlation (e.g., CVR[13], DIABLO[14])
- Joint matrix decomposition (e.g., sJIVE[15])
- Hierarchical integration (e.g., LRM[16])

Limitations: lack of uncertainty quantification

- Focus on discrimination or prediction of clinical outcome as opposed to inference

[13] Luo et al. Biostatistics. 16'
[14] Singh et al. Bioinformatics. 19'
[15] Palzer et al. CSDA. 22'
[16] Zhu et al. Biostatistics. 16'

Current Methods and Limitations

Fred Hutch Cancer Center

Current Methods

- Canonical correlation (e.g., CVR[13], DIABLO[14])
- Joint matrix decomposition (e.g., sJIVE[15])
- Hierarchical integration (e.g., LRM[16])

Limitations: lack of uncertainty quantification

- Focus on discrimination or prediction of clinical outcome as opposed to inference
- Rank selection is based on maximizing prediction performance

[13] Luo et al. Biostatistics. 16'
[14] Singh et al. Bioinformatics. 19'
[15] Palzer et al. CSDA. 22'
[16] Zhu et al. Biostatistics. 16'

Current Methods and Limitations

Fred Hutch Cancer Center

Current Methods

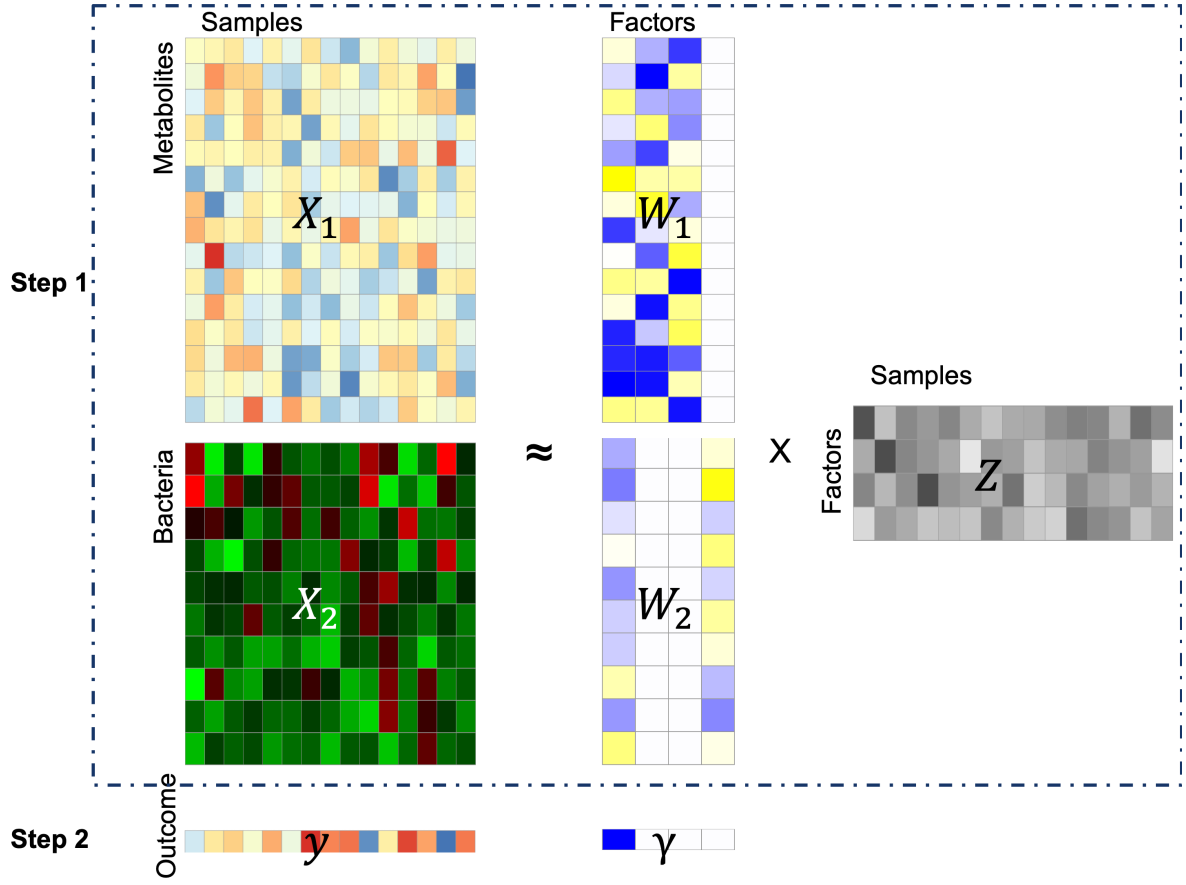
- Canonical correlation (e.g., CVR[13], DIABLO[14])
- Joint matrix decomposition (e.g., sJIVE[15])
- Hierarchical integration (e.g., LRM[16])

Limitations: lack of uncertainty quantification

- Focus on discrimination or prediction of clinical outcome as opposed to inference
- Rank selection is based on maximizing prediction performance
- Variable selection is achieved by penalization

[13] Luo et al. Biostatistics. 16'
[14] Singh et al. Bioinformatics. 19'
[15] Palzer et al. CSDA. 22'
[16] Zhu et al. Biostatistics. 16'

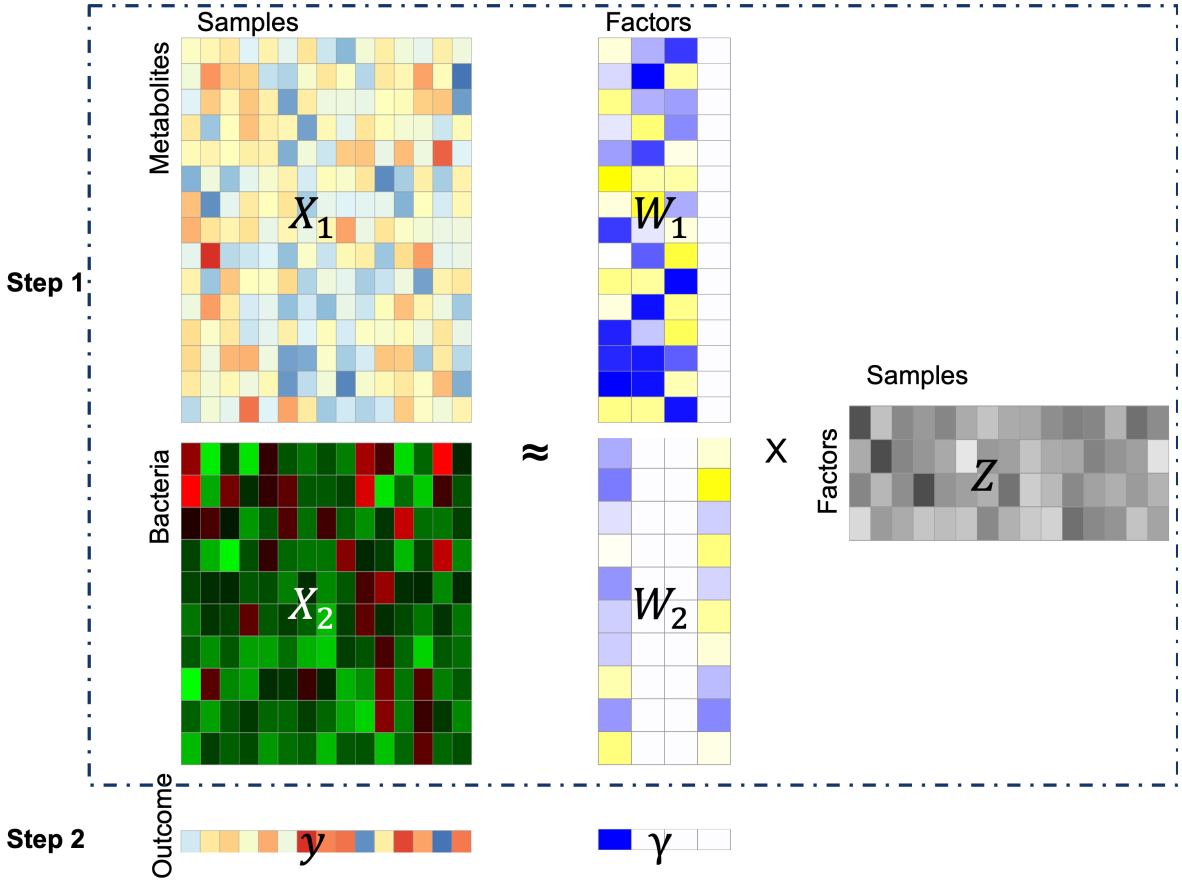
Our Framework: Group Factor Regression



GFAR

1. decomposes two data matrices (X_1, X_2) into a set of factors Z and loading matrices (W_1, W_2)

Our Framework: Group Factor Regression



GFAR

1. decomposes two data matrices (X_1, X_2) into a set of factors Z and loading matrices (W_1, W_2)
2. performs association analysis between latent factors and y

Group Factor Regression

Fred Hutch Cancer Center

$$\begin{bmatrix} X_{1,i} \\ X_{2,i} \end{bmatrix} = \begin{bmatrix} W_1^c & W_1^s & 0 \\ W_2^c & 0 & W_2^s \end{bmatrix} \begin{bmatrix} Z_i^c \\ Z_{1,i}^s \\ Z_{2,i}^s \end{bmatrix} + \begin{bmatrix} e_{1,i} \\ e_{2,i} \end{bmatrix}$$

$$y_i = \gamma_c' Z_i^c + \gamma_1' Z_{1,i}^s + \gamma_2' Z_{2,i}^s + \epsilon_i$$

Group Factor Regression

Fred Hutch Cancer Center

$$\begin{bmatrix} X_{1,i} \\ X_{2,i} \end{bmatrix} = \begin{bmatrix} W_1^c & W_1^s & 0 \\ W_2^c & 0 & W_2^s \end{bmatrix} \begin{bmatrix} Z_i^c \\ Z_{1,i}^s \\ Z_{2,i}^s \end{bmatrix} + \begin{bmatrix} e_{1,i} \\ e_{2,i} \end{bmatrix}$$

$$y_i = \gamma_c' Z_i^c + \gamma_1' Z_{1,i}^s + \gamma_2' Z_{2,i}^s + \epsilon_i$$

- W_1^c and W_2^c are low-rank loading matrices for *joint variation*.

Group Factor Regression

Fred Hutch Cancer Center

$$\begin{bmatrix} X_{1,i} \\ X_{2,i} \end{bmatrix} = \begin{bmatrix} W_1^c & W_1^s & 0 \\ W_2^c & 0 & W_2^s \end{bmatrix} \begin{bmatrix} Z_i^c \\ Z_{1,i}^s \\ Z_{2,i}^s \end{bmatrix} + \begin{bmatrix} e_{1,i} \\ e_{2,i} \end{bmatrix}$$

$$y_i = \gamma_c' Z_i^c + \gamma_1' Z_{1,i}^s + \gamma_2' Z_{2,i}^s + \epsilon_i$$

- W_1^c and W_2^c are low-rank loading matrices for *joint variation*.
- W_1^s and W_2^s are low-rank loading matrices for *separate variation*.

Group Factor Regression

Fred Hutch Cancer Center

$$\begin{bmatrix} X_{1,i} \\ X_{2,i} \end{bmatrix} = \begin{bmatrix} W_1^c & W_1^s & 0 \\ W_2^c & 0 & W_2^s \end{bmatrix} \begin{bmatrix} Z_i^c \\ Z_{1,i}^s \\ Z_{2,i}^s \end{bmatrix} + \begin{bmatrix} e_{1,i} \\ e_{2,i} \end{bmatrix}$$

$$y_i = \gamma_c' Z_i^c + \gamma_1' Z_{1,i}^s + \gamma_2' Z_{2,i}^s + \epsilon_i$$

- W_1^c and W_2^c are low-rank loading matrices for *joint variation*.
- W_1^s and W_2^s are low-rank loading matrices for *separate variation*.
- Identifiability

$$\mathrm{E} \begin{bmatrix} Z_i^c \\ Z_{1,i}^s \\ Z_{2,i}^s \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathrm{Var} \begin{bmatrix} Z_i^c \\ Z_{1,i}^s \\ Z_{2,i}^s \end{bmatrix} = \begin{bmatrix} I_{k_c} & 0 & 0 \\ 0 & I_{k_1^s} & \Phi \\ 0 & \Phi & I_{k_2^s} \end{bmatrix}$$

Group Factor Regression

Fred Hutch Cancer Center

$$\begin{bmatrix} X_{1,i} \\ X_{2,i} \end{bmatrix} = \begin{bmatrix} W_1^c & W_1^s & 0 \\ W_2^c & 0 & W_2^s \end{bmatrix} \begin{bmatrix} Z_i^c \\ Z_{1,i}^s \\ Z_{2,i}^s \end{bmatrix} + \begin{bmatrix} e_{1,i} \\ e_{2,i} \end{bmatrix}$$

$$y_i = \gamma_c' Z_i^c + \gamma_1' Z_{1,i}^s + \gamma_2' Z_{2,i}^s + \epsilon_i$$

- W_1^c and W_2^c are low-rank loading matrices for *joint variation*.
- W_1^s and W_2^s are low-rank loading matrices for *separate variation*.
- Identifiability

$$\mathrm{E} \begin{bmatrix} Z_i^c \\ Z_{1,i}^s \\ Z_{2,i}^s \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathrm{Var} \begin{bmatrix} Z_i^c \\ Z_{1,i}^s \\ Z_{2,i}^s \end{bmatrix} = \begin{bmatrix} I_{k_c} & 0 & 0 \\ 0 & I_{k_1^s} & \Phi \\ 0 & \Phi & I_{k_2^s} \end{bmatrix}$$

- Error terms

$$\mathrm{E} \begin{bmatrix} e_{1,i} \\ e_{2,i} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mathrm{Var} \begin{bmatrix} e_{1,i} \\ e_{2,i} \end{bmatrix} = \begin{bmatrix} \Psi_1 & 0 \\ 0 & \Psi_2 \end{bmatrix}, \quad \epsilon_i \sim N(0, \sigma^2)$$

Scientific Questions

Fred Hutch Cancer Center

1. Is *joint variation* shared across datasets associated with clinical outcome?

Scientific Questions

Fred Hutch Cancer Center

1. Is *joint variation* shared across datasets associated with clinical outcome?
    - Test $\gamma_c = 0$

Scientific Questions

Fred Hutch Cancer Center

1. Is *joint variation* shared across datasets associated with clinical outcome?
   - Test $\gamma_c = 0$
   - Joint variation is likely due to the underlying biological mechanism.

Scientific Questions

Fred Hutch Cancer Center

1. Is *joint variation* shared across datasets associated with clinical outcome?
    - Test $\gamma_c = 0$
    - Joint variation is likely due to the underlying biological mechanism.
    - If NOT, there is perhaps little benefit in predicting outcome with joint variation.

Scientific Questions

Fred Hutch Cancer Center

1. Is *joint variation* shared across datasets associated with clinical outcome?
    - Test $\gamma_c = 0$
    - Joint variation is likely due to the underlying biological mechanism.
    - If NOT, there is perhaps little benefit in predicting outcome with joint variation.
    - If YES, one might want to know the variables contributing to this association as these variables may define the underlying 'causal' pathway.

Scientific Questions

Fred Hutch Cancer Center

1. Is *joint variation* shared across datasets associated with clinical outcome?
   - Test $\gamma_c = 0$
   - Joint variation is likely due to the underlying biological mechanism.
   - If NOT, there is perhaps little benefit in predicting outcome with joint variation.
   - If YES, one might want to know the variables contributing to this association as these variables may define the underlying 'causal' pathway.
2. Is *separate variation* associated with clinical outcome?

Scientific Questions

Fred Hutch Cancer Center

1. Is *joint variation* shared across datasets associated with clinical outcome?
   - Test $\gamma_c = 0$
   - Joint variation is likely due to the underlying biological mechanism.
   - If NOT, there is perhaps little benefit in predicting outcome with joint variation.
   - If YES, one might want to know the variables contributing to this association as these variables may define the underlying 'causal' pathway.

2. Is *separate variation* associated with clinical outcome?
   - Test $\gamma_1 = 0$ and $\gamma_2 = 0$

Scientific Questions

Fred Hutch Cancer Center

1. Is *joint variation* shared across datasets associated with clinical outcome?
   - Test $\gamma_c = 0$
   - Joint variation is likely due to the underlying biological mechanism.
   - If NOT, there is perhaps little benefit in predicting outcome with joint variation.
   - If YES, one might want to know the variables contributing to this association as these variables may define the underlying 'causal' pathway.

2. Is *separate variation* associated with clinical outcome?
   - Test $\gamma_1 = 0$ and $\gamma_2 = 0$
   - Separate variation is likely due to technical effects, host factors, etc.

Scientific Questions

Fred Hutch Cancer Center

1. Is *joint variation* shared across datasets associated with clinical outcome?
   - Test $\gamma_c = 0$
   - Joint variation is likely due to the underlying biological mechanism.
   - If NOT, there is perhaps little benefit in predicting outcome with joint variation.
   - If YES, one might want to know the variables contributing to this association as these variables may define the underlying 'causal' pathway.

2. Is *separate variation* associated with clinical outcome?
   - Test $\gamma_1 = 0$ and $\gamma_2 = 0$
   - Separate variation is likely due to technical effects, host factors, etc.
   - In case of microbiome + metabolome, is the gap between the metabolic potential and observed metabolic activity associated with outcome?

Scientific Questions

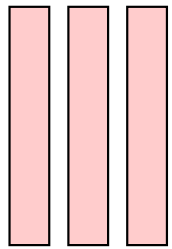
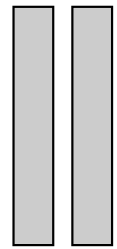
Fred Hutch Cancer Center

1. Is *joint variation* shared across datasets associated with clinical outcome?
   - Test $\gamma_c = 0$
   - Joint variation is likely due to the underlying biological mechanism.
   - If NOT, there is perhaps little benefit in predicting outcome with joint variation.
   - If YES, one might want to know the variables contributing to this association as these variables may define the underlying 'causal' pathway.

2. Is *separate variation* associated with clinical outcome?
   - Test $\gamma_1 = 0$ and $\gamma_2 = 0$
   - Separate variation is likely due to technical effects, host factors, etc.
   - In case of microbiome + metabolome, is the gap between the metabolic potential and observed metabolic activity associated with outcome?
   - Correlation-based methods do not uncover separate variation.

GFAR Algorithm

Step 1(a): Low-rank approximation to each dataset[17]

$\widehat{U}_1 =$  $\widehat{U}_2 =$ 

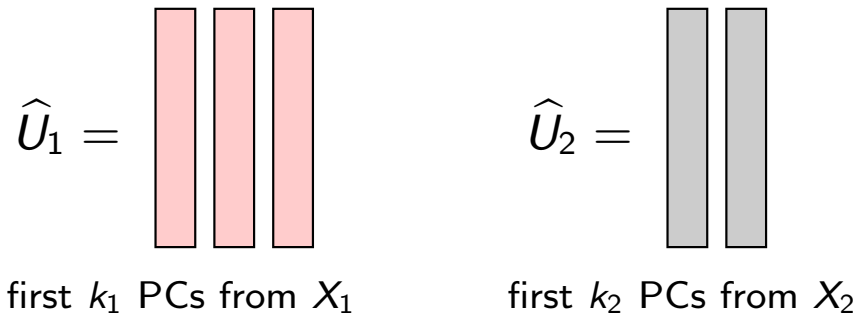
first k_1 PCs from X_1 first k_2 PCs from X_2

[17] Individual rank is estimated by Gavish and Donoho. IEEE, 14'

[18] Joint rank is estimated by Andreou et al. Econometrica, 19'.

GFAR Algorithm

Step 1(a): Low-rank approximation to each dataset[17]



Step 1(b): Canonical correlation analysis on $(\widehat{U}_1, \widehat{U}_2)$ to identify joint factor

- Perform SVD

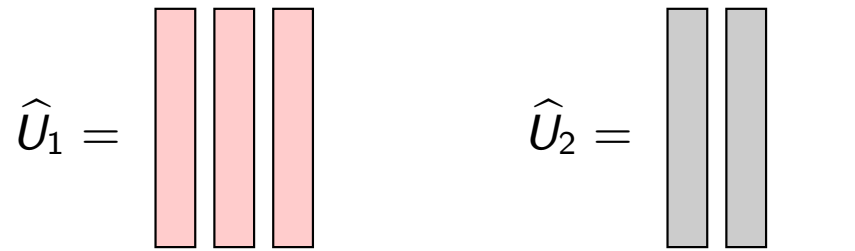
$$\widehat{U}_1' \widehat{U}_2 = 0.9 R_1 Q_1' + 0.1 R_2 Q_2'$$

[17] Individual rank is estimated by Gavish and Donoho. IEEE, 14'

[18] Joint rank is estimated by Andreou et al. Econometrica, 19'.

GFAR Algorithm

Step 1(a): Low-rank approximation to each dataset[17]



Step 1(b): Canonical correlation analysis on $(\widehat{U}_1, \widehat{U}_2)$ to identify joint factor

- Perform SVD

$$\widehat{U}_1' \widehat{U}_2 = 0.9 R_1 Q_1' + 0.1 R_2 Q_2'$$

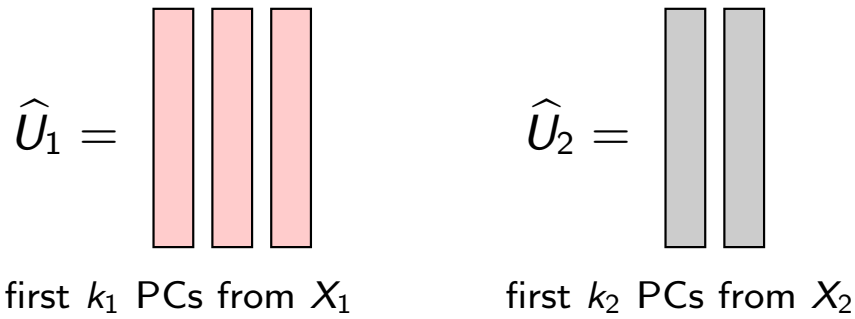
- Estimate joint factor by $\widehat{U}_1 R_{1:k_c}$ or $\widehat{U}_2 Q_{1:k_c}$, where $k_c = 1$ is the rank[18] of Z^c.

[17] Individual rank is estimated by Gavish and Donoho. IEEE, 14'

[18] Joint rank is estimated by Andreou et al. Econometrica, 19'.

GFAR Algorithm

Step 1(a): Low-rank approximation to each dataset[17]



Step 1(b): Canonical correlation analysis on $(\widehat{U}_1, \widehat{U}_2)$ to identify joint factor

- Perform SVD
$$\widehat{U}_1' \widehat{U}_2 = 0.9 R_1 Q_1' + 0.1 R_2 Q_2'$$
- Estimate joint factor by $\widehat{U}_1 R_{1:k_c}$ or $\widehat{U}_2 Q_{1:k_c}$, where $k_c = 1$ is the rank[18] of Z^c.
- Estimate individual factors by projecting out the joint factor in each dataset.

[17] Individual rank is estimated by Gavish and Donoho. IEEE, 14'

[18] Joint rank is estimated by Andreou et al. Econometrica, 19'.

GFAR Algorithm

Step 2: Inference

- Association analysis by testing $H_0 : \gamma_c = 0$ with

$$T = n\frac{y'(I_n - \widehat{P}_0)y - y'(I_n - \widehat{P}_1)y}{y'(I_n - \widehat{P}_0)y} \quad \xrightarrow{\text{null}} \quad \chi^2_{k_c}$$

GFAR Algorithm

Fred Hutch Cancer Center

Step 2: Inference

- Association analysis by testing $H_0 : \gamma_c = 0$ with

$$T = n\frac{y'(I_n - \widehat{P}_0)y - y'(I_n - \widehat{P}_1)y}{y'(I_n - \widehat{P}_0)y} \quad \xrightarrow{\text{null}} \quad \chi^2_{k_c}$$

Subspace recovery error in $\widehat{P}_0$ and $\widehat{P}_1$ is accounted for.

GFAR Algorithm

Fred Hutch Cancer Center

Step 2: Inference

- Association analysis by testing $H_0 : \gamma_c = 0$ with

$$T = n\frac{y'(I_n - \widehat{P}_0)y - y'(I_n - \widehat{P}_1)y}{y'(I_n - \widehat{P}_0)y} \xrightarrow{\text{null}} \chi^2_{k_c}$$

  Subspace recovery error in $\widehat{P}_0$ and $\widehat{P}_1$ is accounted for.

- Variable selection by thresholding joint loadings

$$\widehat{W}_1^c = \frac{1}{n}X_1'\widehat{Z}^c, \quad \widehat{W}_2^c = \frac{1}{n}X_2'\widehat{Z}^c$$

  or regression coefficients

$$\hat{\beta}_1 = \frac{1}{n}\widehat{W}_1^c(\widehat{Z}^c)'y, \quad \hat{\beta}_2 = \frac{1}{n}\widehat{W}_2^c(\widehat{Z}^c)'y$$

Relation to sJIVE

Fred Hutch Cancer Center

- GFAR studies association between joint/separate variation and outcome using a score test.

Relation to sJIVE

Fred Hutch Cancer Center

- GFAR studies association between joint/separate variation and outcome using a score test.
- GFAR algorithm is decomposition-based while sJIVE algorithm is iterative.

Relation to sJIVE

Fred Hutch Cancer Center

- GFAR studies association between joint/separate variation and outcome using a score test.
- GFAR algorithm is decomposition-based while sJIVE algorithm is iterative.
- GFAR selects joint rank by testing significance of canonical correlation coefficients while sJIVE uses cross-validation.

Application to BactoCARB Study

Fred Hutch Cancer Center

A randomized crossover study of 80 subjects aimed at the effects of high/low glycemic load on a variety of biomarkers.

- $X_1 = 144$ metabolites & bile acids (blood)
- $X_2 = 134$ microbial genus abundances (stool)
- $y =$ enterolactone 'ENL' (urine)

Application to BactoCARB Study

Fred Hutch Cancer Center

A randomized crossover study of 80 subjects aimed at the effects of high/low glycemic load on a variety of biomarkers.

- $X_1 = 144$ metabolites & bile acids (blood)
- $X_2 = 134$ microbial genus abundances (stool)
- $y =$ enterolactone 'ENL' (urine)

Hypothesis: Some metabolites are of microbial origin and these microbial metabolites are responsible for ENL.

BactoCARB: Association Analysis

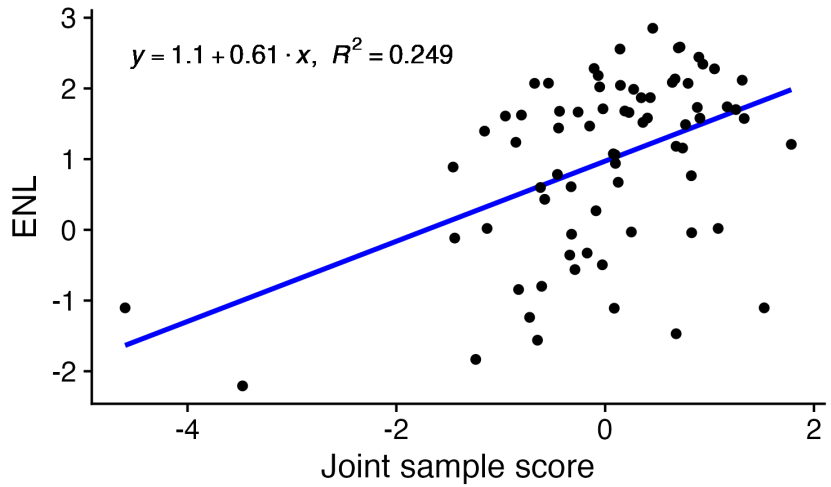
Fred Hutch Cancer Center

Ranks: $\hat{k}_c = 1$, $\hat{k}_1^s = 15$ and $\hat{k}_2^s = 7$

BactoCARB: Association Analysis

Ranks: $\hat{k}_c = 1$, $\hat{k}_1^s = 15$ and $\hat{k}_2^s = 7$

Joint variation is significantly associated with ENL: p-value $< 10^{-4}$.

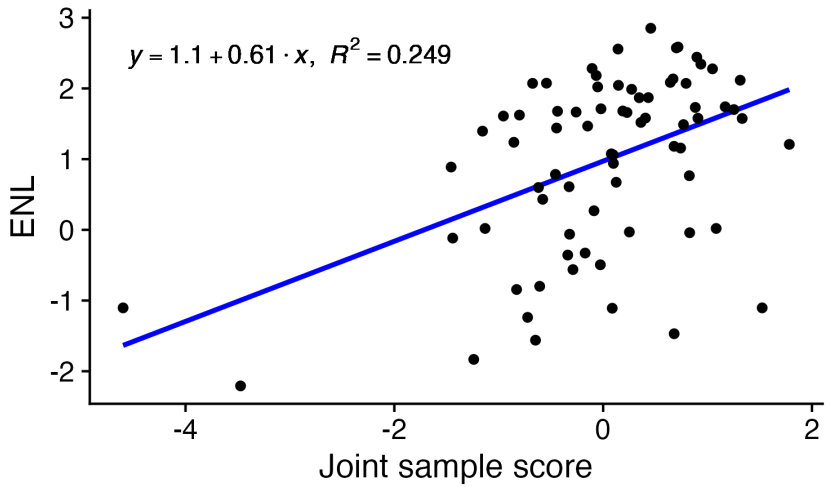


BactoCARB: Association Analysis

Fred Hutch Cancer Center

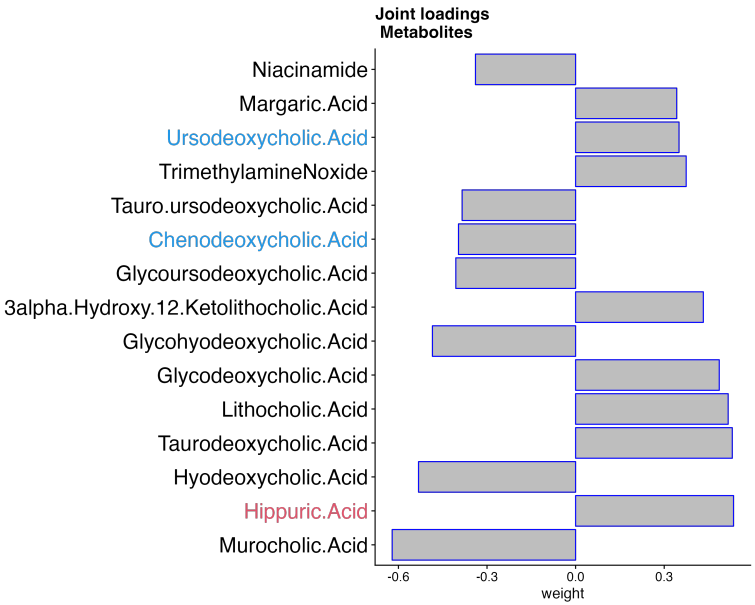
Ranks: $\hat{k}_c = 1$, $\hat{k}_1^s = 15$ and $\hat{k}_2^s = 7$

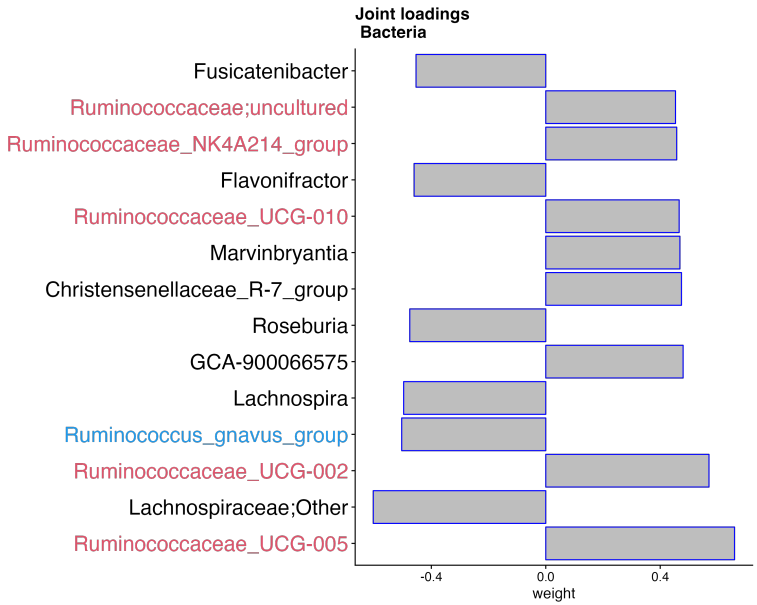
Joint variation is significantly associated with ENL: p-value $< 10^{-4}$.



Separate variation is NOT significantly associated with ENL: p-values 0.40 and 0.67.

BactoCARB: Active Variables

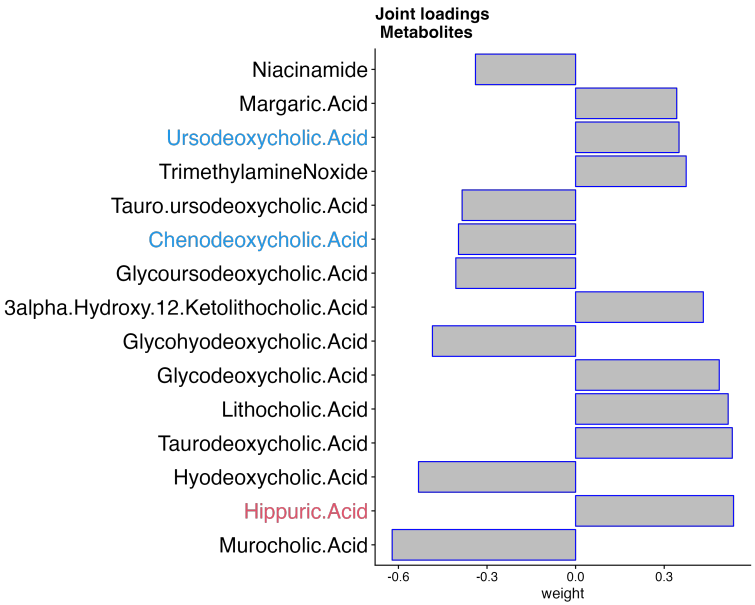


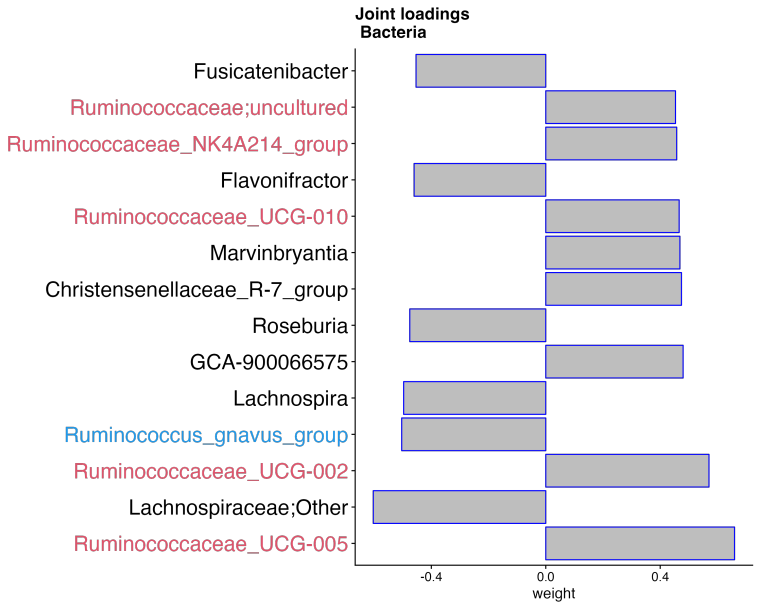


- Hippuric.Acid is microbially derived and has known association with ENL[19].

[19] Miles et al. Food Funct. 18'
[20] Pallister et al. *Hippurate as a metabolomic marker of gut microbiome diversity...* Sci Rep. 17'
[21] Lee et al. JLR. 13'
[22] Crost et al. FEMS Microbiology Reviews. 23'

BactoCARB: Active Variables

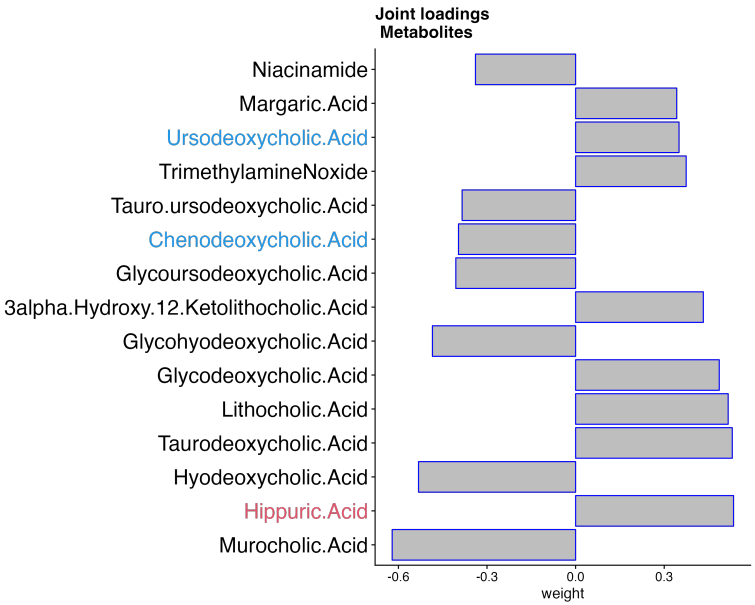


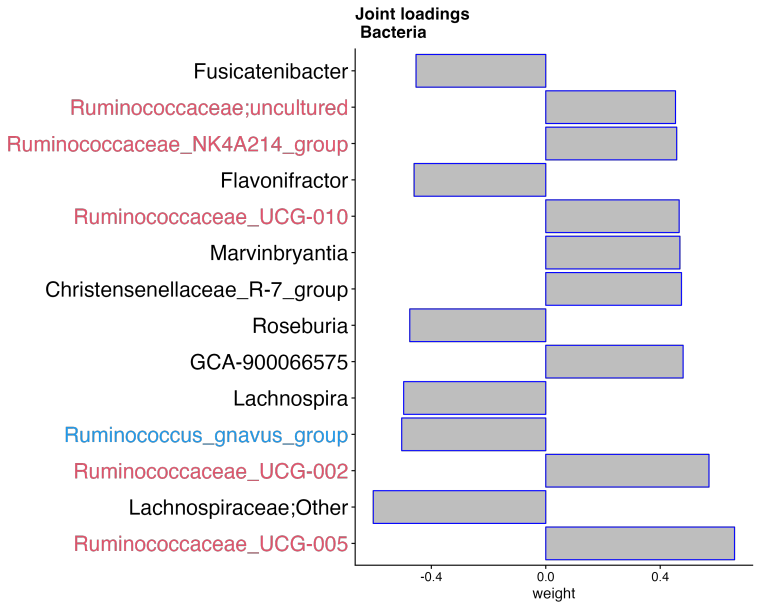


- Hippuric.Acid is microbially derived and has known association with ENL[19].
- *Ruminococcaceae* family is associated with Hippurate[20].

[19] Miles et al. Food Funct. 18'
[20] Pallister et al. *Hippurate as a metabolomic marker of gut microbiome diversity...* Sci Rep. 17'
[21] Lee et al. JLR. 13'
[22] Crost et al. FEMS Microbiology Reviews. 23'

BactoCARB: Active Variables





- Hippuric.Acid is microbially derived and has known association with ENL[19].
- *Ruminococcaceae* family is associated with Hippurate[20].
- *Ruminococcus_gnavus_group* helps convert Ursodeoxycholic.Acid into Chenodeoxycholic.Acid[21] and is associated with inflammatory diseases[22].

[19] Miles et al. Food Funct. 18'
[20] Pallister et al. *Hippurate as a metabolomic marker of gut microbiome diversity...* Sci Rep. 17'
[21] Lee et al. JLR. 13'
[22] Crost et al. FEMS Microbiology Reviews. 23'

Summary of Part II

Fred Hutch Cancer Center

Multi-view regression is a common statistical problem.

Summary of Part II

Fred Hutch Cancer Center

Multi-view regression is a common statistical problem.

We provide GFAR which

- performs association analysis between shared/joint variation and a clinical outcome.

Summary of Part II

Fred Hutch Cancer Center

Multi-view regression is a common statistical problem.

We provide GFAR which

- performs association analysis between shared/joint variation and a clinical outcome.
- identifies variables in each dataset that are jointly associated with a clinical outcome $\rightarrow$ active microbial metabolic pathway?

Summary of Part II

Fred Hutch Cancer Center

Multi-view regression is a common statistical problem.

We provide GFAR which

- performs association analysis between shared/joint variation and a clinical outcome.
- identifies variables in each dataset that are jointly associated with a clinical outcome $\rightarrow$ active microbial metabolic pathway?

Work in progress

- Variable selection accounting for uncertainty in estimation of factor loadings.

Summary of Part II

Fred Hutch Cancer Center

Multi-view regression is a common statistical problem.

We provide GFAR which

- performs association analysis between shared/joint variation and a clinical outcome.
- identifies variables in each dataset that are jointly associated with a clinical outcome $\rightarrow$ active microbial metabolic pathway?

Work in progress

- Variable selection accounting for uncertainty in estimation of factor loadings.
- Extension to three or more datasets allowing partially shared joint variation.

Future Goals

Fred Hutch Cancer Center

Network Biology

- Metabolic networks
- Microbial networks
- Comorbidity networks
- Brain connectivity networks

Microbiome

- Network analysis
- High-dimensional inference
- Integration with other Omics
- Cancer prevention
- Gut-brain association

Other Areas

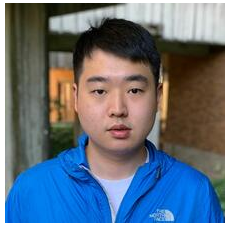
New Opportunities!

Acknowledgement

Ma Lab








Yue Wang, Ilias Moysidis, Kristyn Pantoja, Xinyi Xie, Wenjie Guan, Antoinette Fang

Collaborators





Tim Randolph, Ali Shojaie, David Jones, Kate Markey, Robert Kaplan

Funding

NIH R01 GM145772, FHCC TDS IRC Pilot

Biostatistics Program

GMD Inference

Fred Hutch Cancer Center

- Find the estimation bias

$$B_j = \mathrm{E}[\hat{\beta}_j^{\mathcal{W}} - \beta_j] = (QV\mathcal{W}V'\beta)_j - \beta_j$$

- Correct the estimation bias via an initial estimator $\hat{\beta}^{init} = D\tilde{\beta}(\lambda)$ where

$$Q = D\Delta D', \quad \tilde{\beta}(\lambda) = \arg\min_{\beta}\{\|y - XD\beta\|_H^2 + \lambda\|\Delta^{-1/2}\beta\|_1\}.$$

- Obtain inference for $H_{0,j} : \beta_j = 0$ using de-biased $\tilde{\beta}_j^{\mathcal{W}} = \hat{\beta}_j^{\mathcal{W}} - \widehat{B}_j$ and the p-value

$$2\left\{1 - \Phi\left((R_{jj}^w)^{1/2}\left\{|\tilde{\beta}_j^{\mathcal{W}}| - \|K_{(j,\cdot)}^w\|_\infty \left(\frac{\log p}{n}\right)^{1/2-r}\right\}\right)\right\}$$

where $R^w = \hat{\sigma}^2 QV\mathcal{W}S^{-2}V'Q$ and $K^w = (QV\mathcal{W}V' - I_p)D$.

GMDI Assumptions

Fred Hutch Cancer Center

Linear model: $y = X\beta + \epsilon$ with $\mathrm{Cov}(\epsilon \mid X) = \Psi = L_\Psi' L_\Psi$. There exists sub-Gaussian $\tilde{\epsilon}$ such that $\epsilon = L_\Psi' \tilde{\epsilon}$.

- Informative H:
$$\|L_\Psi H L_\Psi' - \sigma^2 I_n\| = o(1), \quad n \to \infty$$
- β is Q-smooth ($\tilde{\beta} = D'\beta$):
$$\|\tilde{\beta}_{S_0^c}\|_1 \leq O\{\sqrt{|S_0| \log p/n}\}, \quad |S_0| = o\{(n/\log p)^r\}, r \in (0, 1/2).$$
- A compatibility assumption w.r.t Q and H
$$0 < \underline{c} \leq \frac{\|\widetilde{X}_A v\|^2}{n\|v\|^2} \leq \bar{c} < \infty,$$
where $\widetilde{X} = H^{1/2} X Q^{1/2}$ and $|A| \geq M_1^* s_0 + 1$.

Tests for Informative Structures

Fred Hutch Cancer Center

Test for informative Q is done by KRV.

Test for informative H is done by both KRV and MiRKAT[23] because the row structure H needs to be informative for both X and y.

$$y = K(X) + \epsilon$$

for a pre-specified kernel $K = \tau H^{-1}$.

[23] Zhao et al. AJHG. 15'

Robust GMDR and GMDI

The weighting scheme can be extended to handle multiple partially informative row structures.

$$H(\tau) = \sum_{j=1}^{N-1} \tau_j H_j + (1 - \sum_{j=1}^{N-1} \tau_j) I_n$$

for $\tau = (\tau_1, \ldots, \tau_{N-1})$ satisfying $\tau_j \in (0, 1)$ and $\sum_{j=1}^{N-1} \tau_j \leq 1$.

More sophisticated strategy is needed to handle partially informative column structures Q as $\tau Q + (1 - \tau) I_p$ shares the same eigenvectors as Q.

GMDR and GMDI with Confounders

Let $Z = (z_1, \ldots, z_n)'$ denote a low-dimensional matrix of covariates. Consider

$$y = g(Z) + X\beta + \epsilon$$

where g is an $n \times 1$ vector of random effects with mean 0 and covariance $\sigma_z^2 K_z$ for some pre-specified kernel $K(\cdot, \cdot)$.

Letting $\delta = g(Z) + \epsilon$, the new model

$$y = X\beta + \delta$$

has row structure $\sigma^{-2} H(\sigma_z^2 K_z \sigma^{-2} H + I_n)^{-1}$.

GFAR: Selecting Joint Rank

Fred Hutch Cancer Center

Inference for the joint rank k_c ($\underline{k} = \min\{\hat{k}_1, \hat{k}_2\}$[24])

$$
\begin{aligned}
H(0) &= \{1 > \rho_1 \geq \cdots \geq \rho_{\underline{k}}\} \\
H(1) &= \{\rho_1 = 1 > \rho_2 \geq \cdots \geq \rho_{\underline{k}}\} \\
&\cdots \\
H(\underline{k}) &= \{\rho_1 = \cdots = \rho_{\underline{k}} = 1\}
\end{aligned}
$$

Test statistic ($p = p_1 \wedge p_2$)

$$
\tilde{\xi}(k_c) = p\sqrt{n}(0.5\mathrm{tr}(\widehat{\Sigma}_U^2))^{-1/2}\left[\sum_{\ell=1}^{k_c}\hat{\rho}_\ell - k_c + \frac{1}{p}\mathrm{tr}(\widehat{\Sigma}_U)\right]
$$

where ρ_ℓ's are the k_c largest sample canonical correlation coefficients.

Estimate k_c by

$$
\hat{k}_c = \max\{r : 1 \leq r \leq \underline{k}, \tilde{\xi}(k_c) > -c(p\sqrt{n})^\tau\}
$$

for constants $c > 0$ and $0 < \tau < 1$.

[24] $\hat{k}_1$ and $\hat{k}_2$ are consistent estimators of the rank in each dataset.