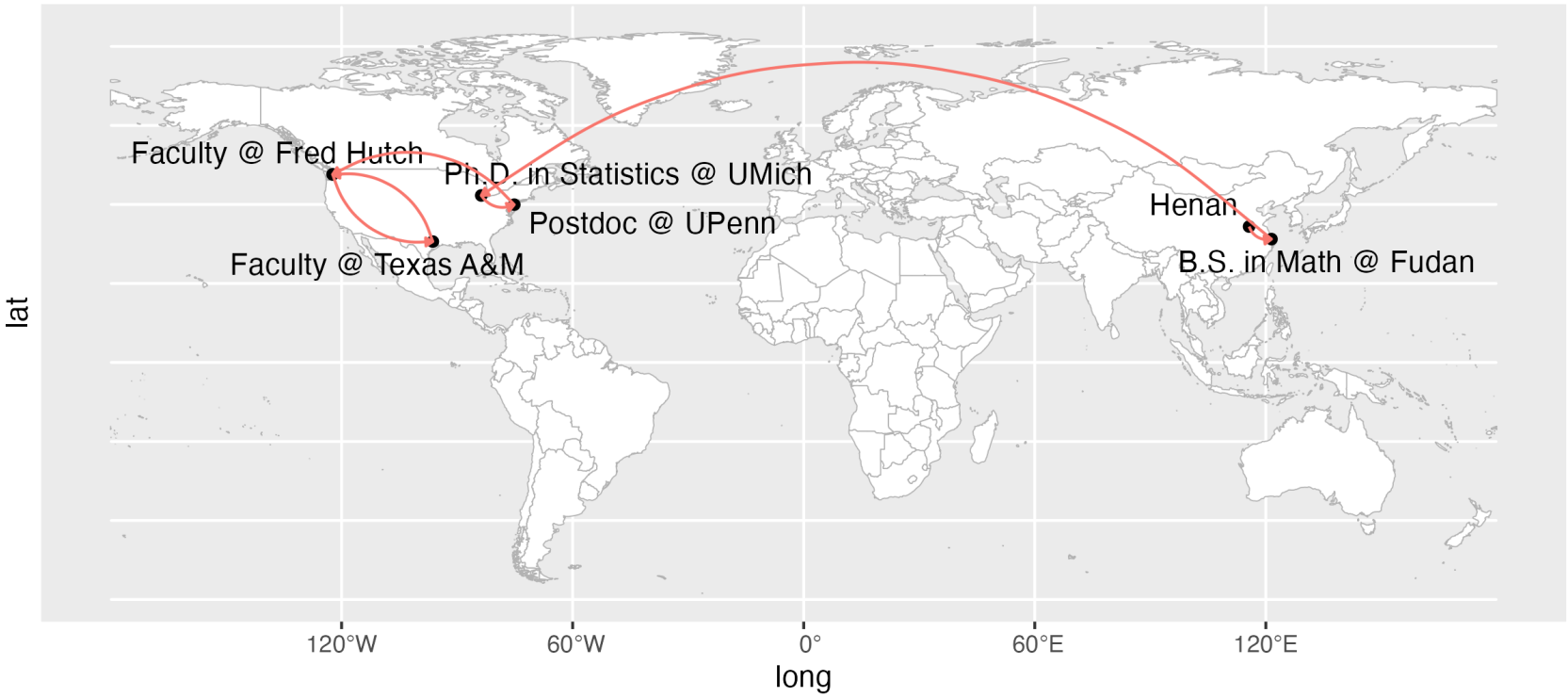
Statistical Modeling for Enhanced Microbiome Biomarker Discovery

24 April 2024

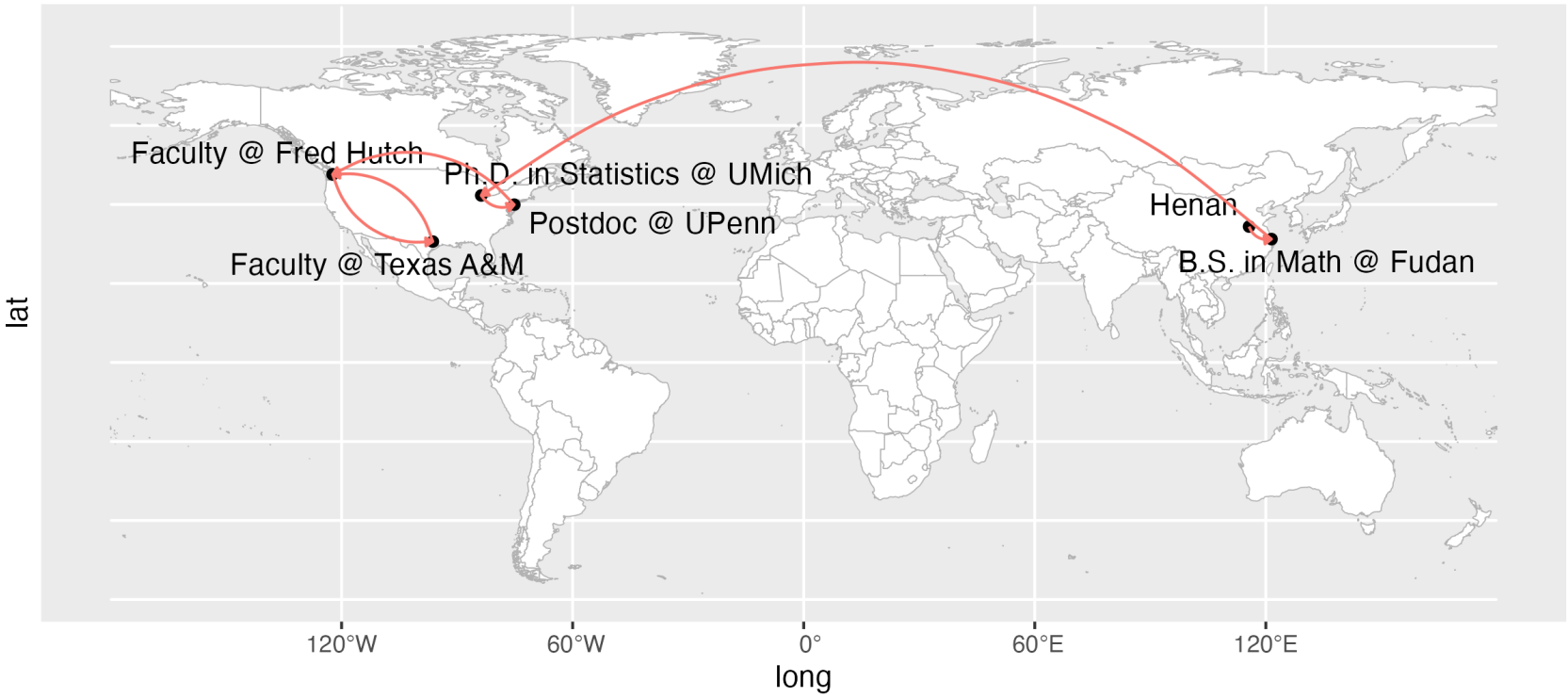
Jing Ma

Associate Professor of Biostatistics
Division of Public Health Sciences

My Journey


Faculty @ Fred Hutch
Ph.D. in Statistics @ UMich
Postdoc @ UPenn
Faculty @ Texas A&M
Henan
B.S. in Math @ Fudan
lat
120°W
60°W
0°
60°E
120°E
long

My Journey



You can read more about my story here.

My Journey

- My thesis: graphical models, high-dimensional data analysis
- My postdoctoral training: graphical models for microbiome data, high-dimensional data analysis
- Now: statistics in microbiome, neuroscience, and aging

My Job

Fred Hutch is an independent, nonprofit organization, that also serves as UW Medicine's cancer program.

Research Institutes

- Independent research
- Collaborative research
- Mentoring students

Universities

- Independent research
- Collaborative research
- Mentoring students
- Teaching

What We Do

We develop statistical methods to study the human microbiome.

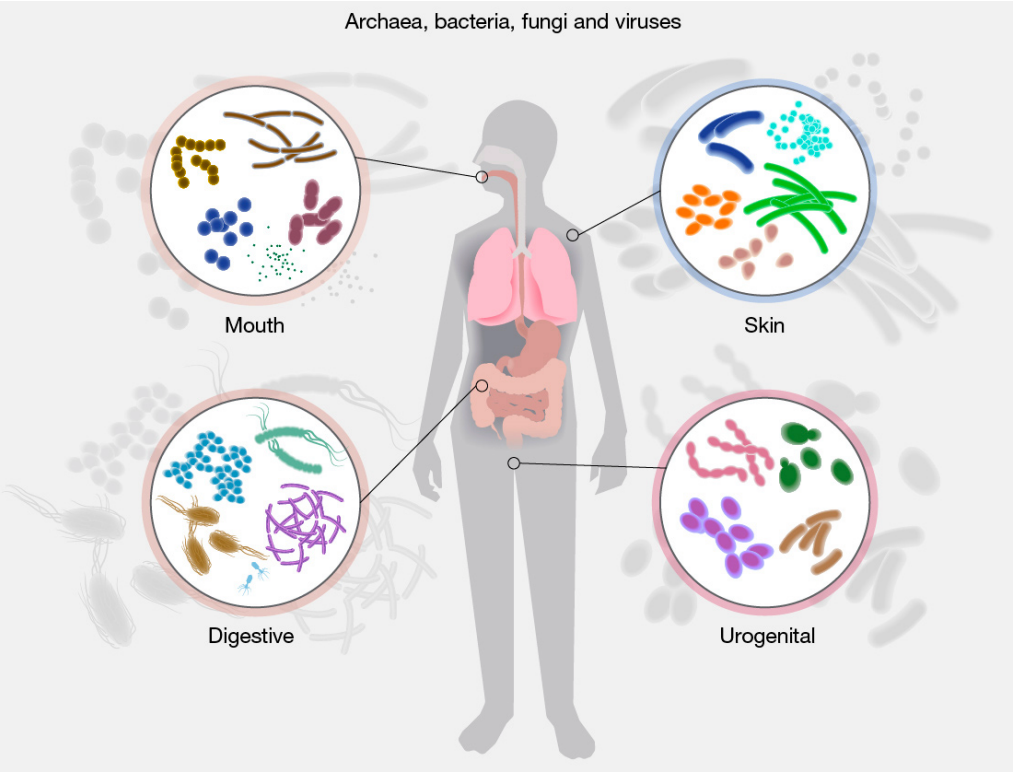


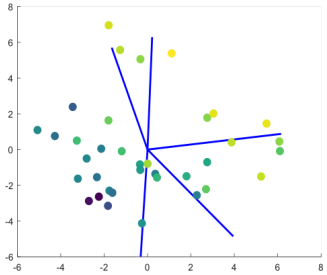
Figure 1: Composition of the human microbiome varies by body sites. They play important roles in human health and have been associated with many diseases. Most of the microbes are bacteria and live in human gut.

What We Do

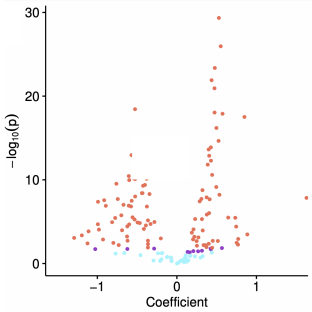
Preprocessing (Bioinformatics)

Community-level Analysis

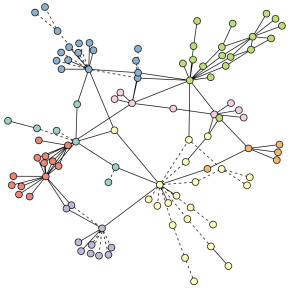
Visualization



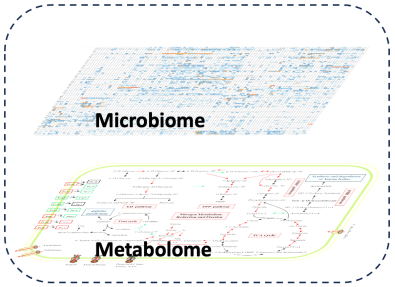
Feature Selection



Network Analysis

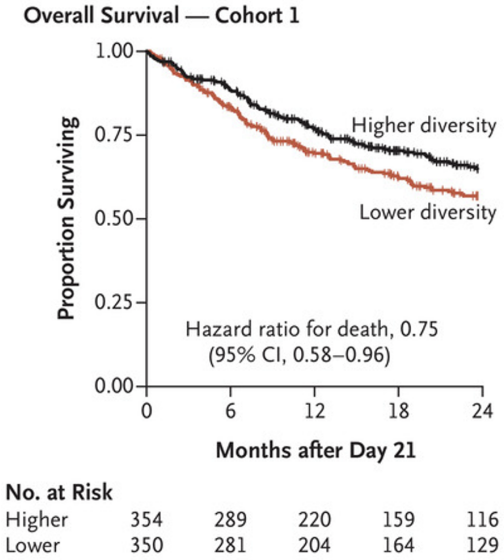


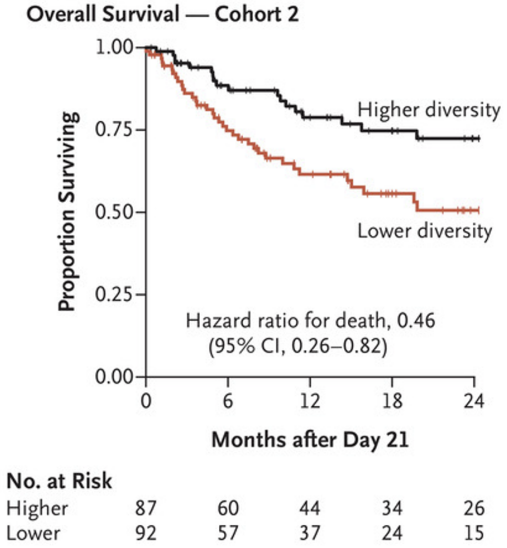
Integration



Microbiome and Cancer Treatment

- 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 Question

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

Microbiome Data

	Sample1	Sample2	Sample3
ASV1	17	0	335
ASV2	231	1180	45
ASV3	30	0	0
...	...	...	...
Age	25	48	65
Disease	yes	no	yes

Table 1: A typical microbiome contingency table

Microbiome Data

	Sample1	Sample2	Sample3
ASV1	17	0	335
ASV2	231	1180	45
ASV3	30	0	0
...	...	...	...
Age	25	48	65
Disease	yes	no	yes

Table 1: A typical microbiome contingency table

Analytical challenges

- high-dimensional: # of taxa > # of samples

Microbiome Data

	Sample1	Sample2	Sample3
ASV1	17	0	335
ASV2	231	1180	45
ASV3	30	0	0
...	...	...	...
Age	25	48	65
Disease	yes	no	yes

Table 1: A typical microbiome contingency table

Analytical challenges

- high-dimensional: # of taxa > # of samples
- structured: taxa and/or samples are correlated

Microbiome Data

	Sample1	Sample2	Sample3
ASV1	17	0	335
ASV2	231	1180	45
ASV3	30	0	0
...	...	...	...
Age	25	48	65
Disease	yes	no	yes

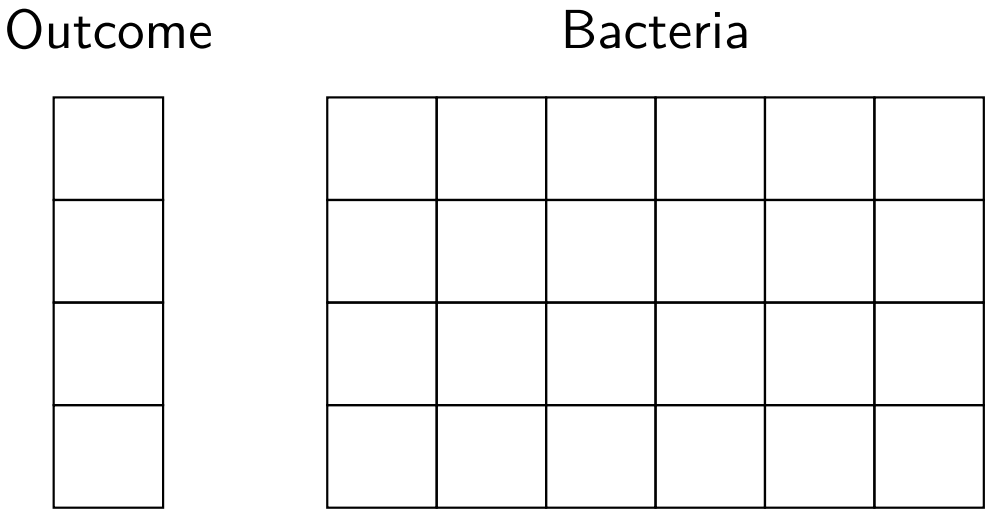
Table 1: A typical microbiome contingency table

Analytical challenges

- high-dimensional: # of taxa > # of samples
- structured: taxa and/or samples are correlated
- etc.

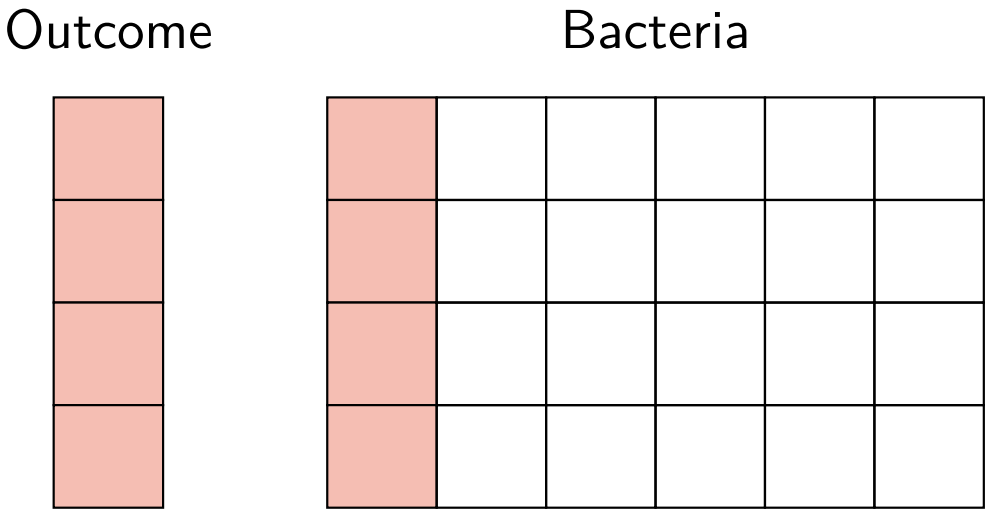
Univariate Feature Selection

Perform univariate test of each species with respect to an outcome



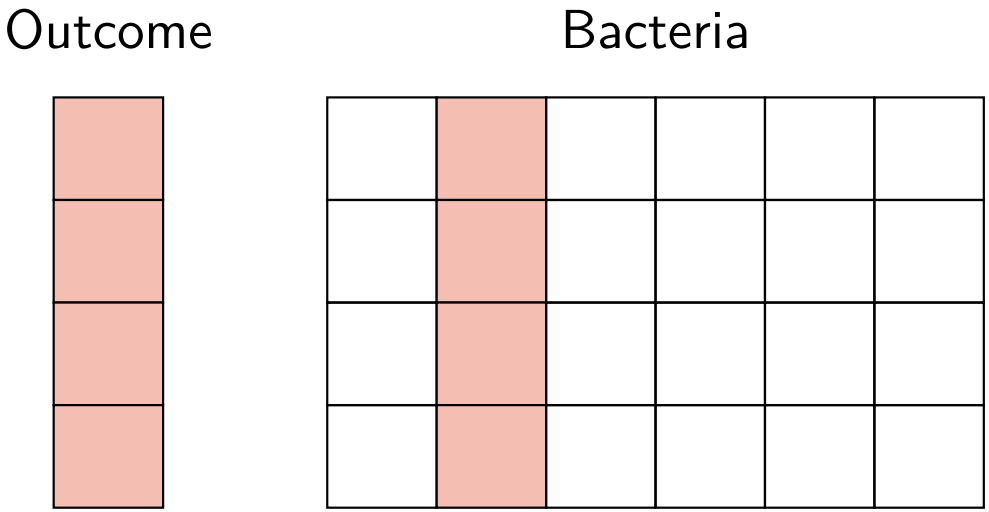
Univariate Feature Selection

Perform univariate test of each species with respect to an outcome



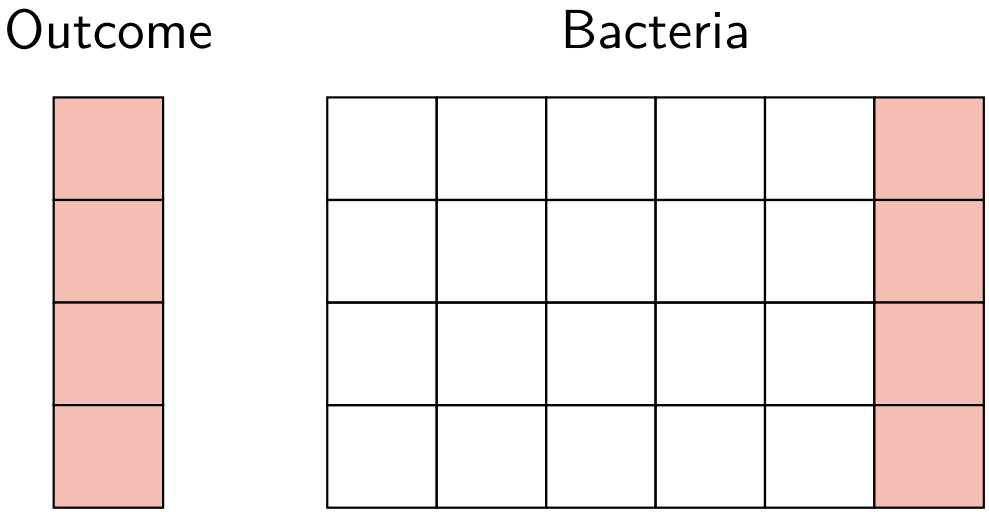
Univariate Feature Selection

Perform univariate test of each species with respect to an outcome



Univariate Feature Selection

Perform univariate test of each species with respect to an outcome



Univariate Feature Selection

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 Feature Selection

Output:

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

Options

- different normalization methods

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

Univariate Feature Selection

Fred Hutch Cancer Center

Output:

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

Options

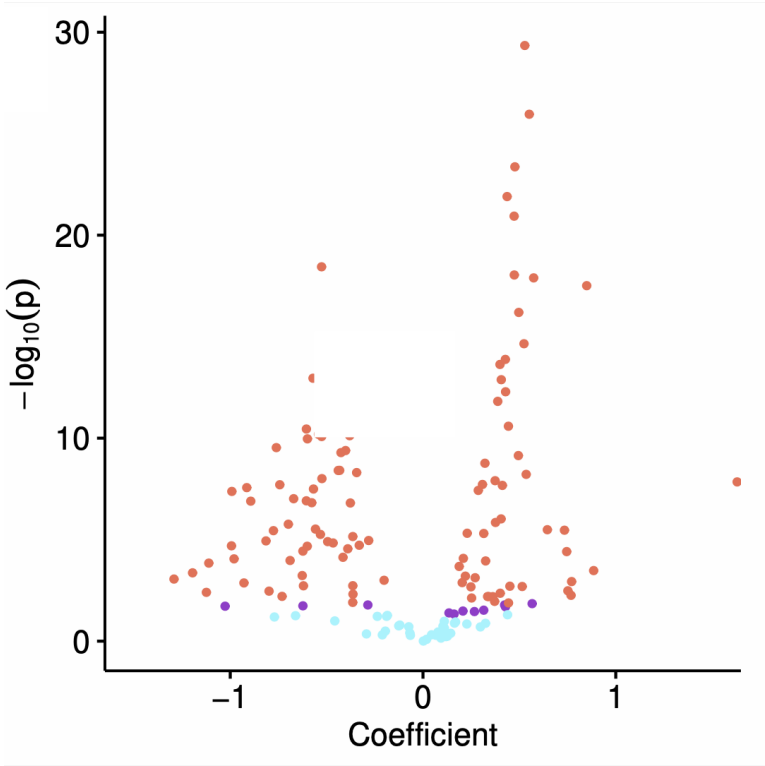
- different normalization methods
- different noise models for bacterial abundances (e.g., zero-inflated log normal[2], beta-binomial[3], zero-inflated quantile regression[4], etc.)

[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?

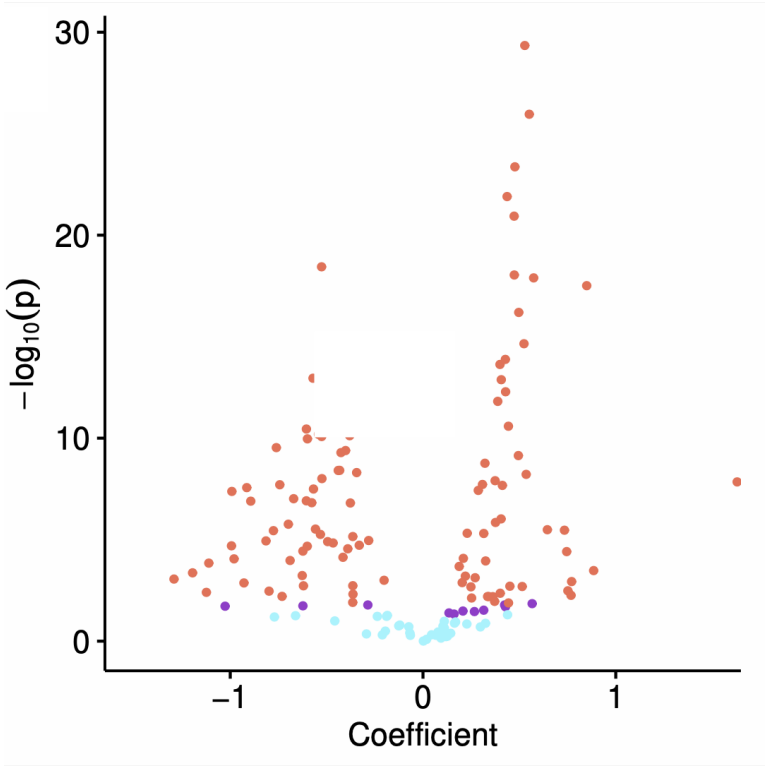
- $n = 100$
- $p = 149$
- Apply univariate Spearman rank correlation test between log transformed abundance and age



Application to Yatsunenko 12'

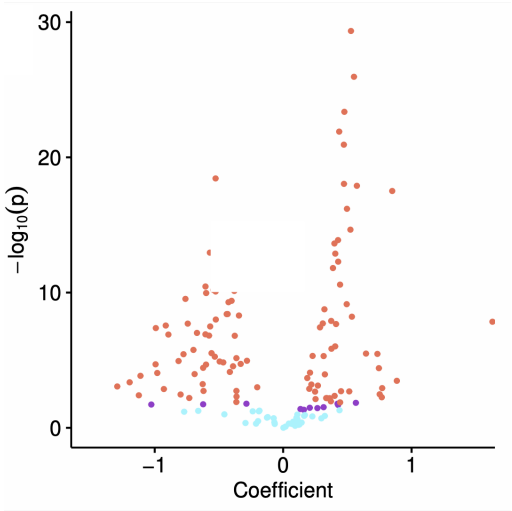
Which bacteria are associated with age?

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

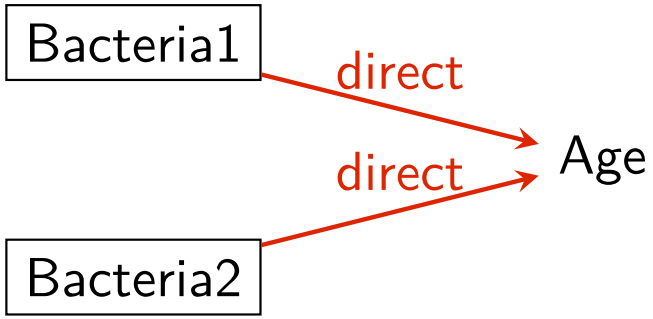


☹ Which bacteria should I target?

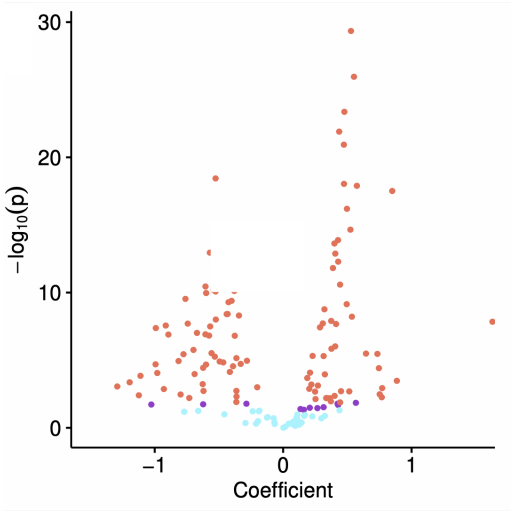
Two Possible Explanations



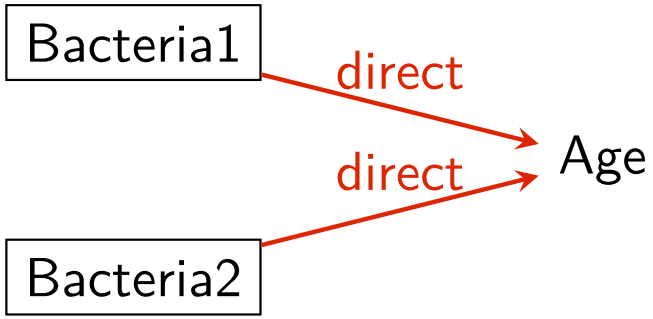
1. Many bacteria are *directly* associated with age



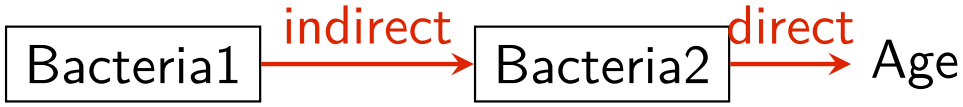
Two Possible Explanations



1. Many bacteria are *directly* associated with age



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



Conditional Associations

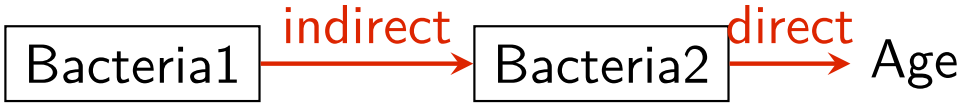


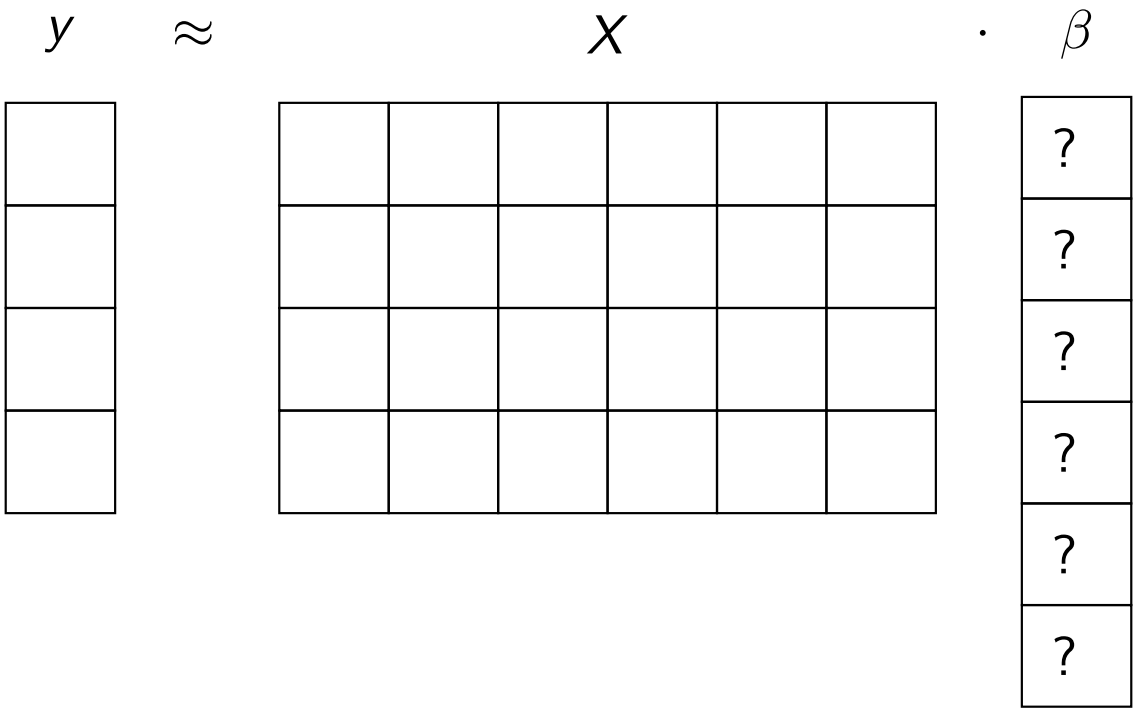
Figure 2: Conditional on Bacteria 2, Bacteria 1 is independent of Age.

Multiple linear regression

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

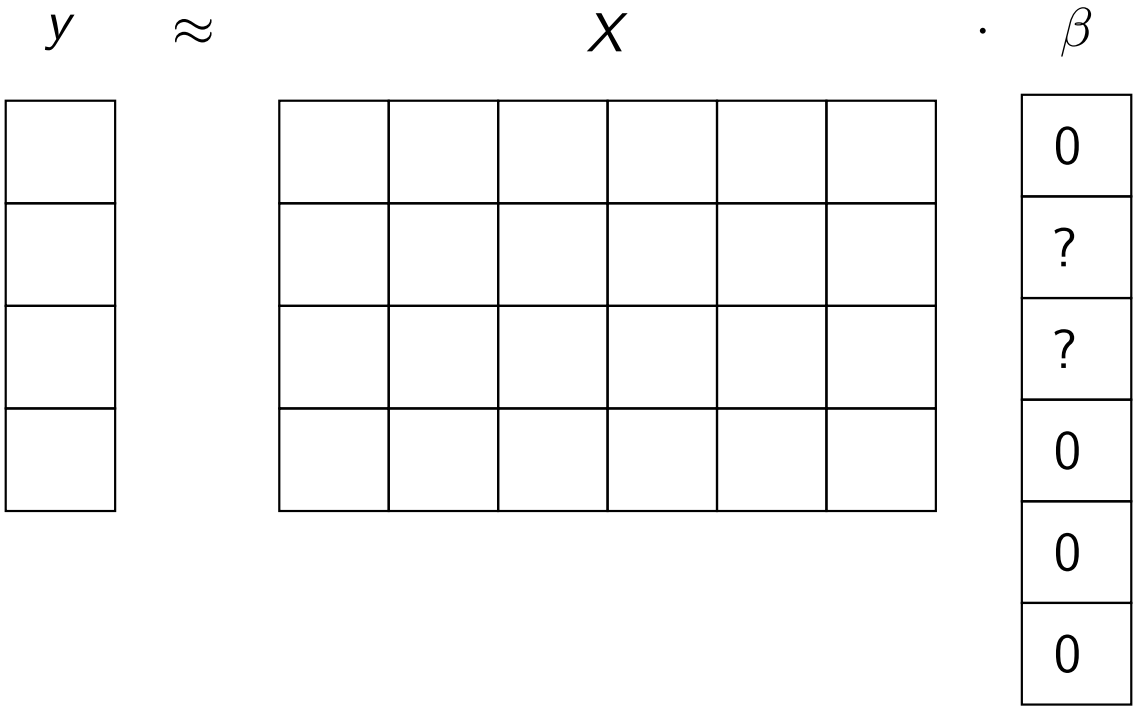
- Coefficients β_j: conditional association between bacteria j and y
- When $p \ll n$, inference for β is straightforward by large sample theory.

Curse of Dimensionality



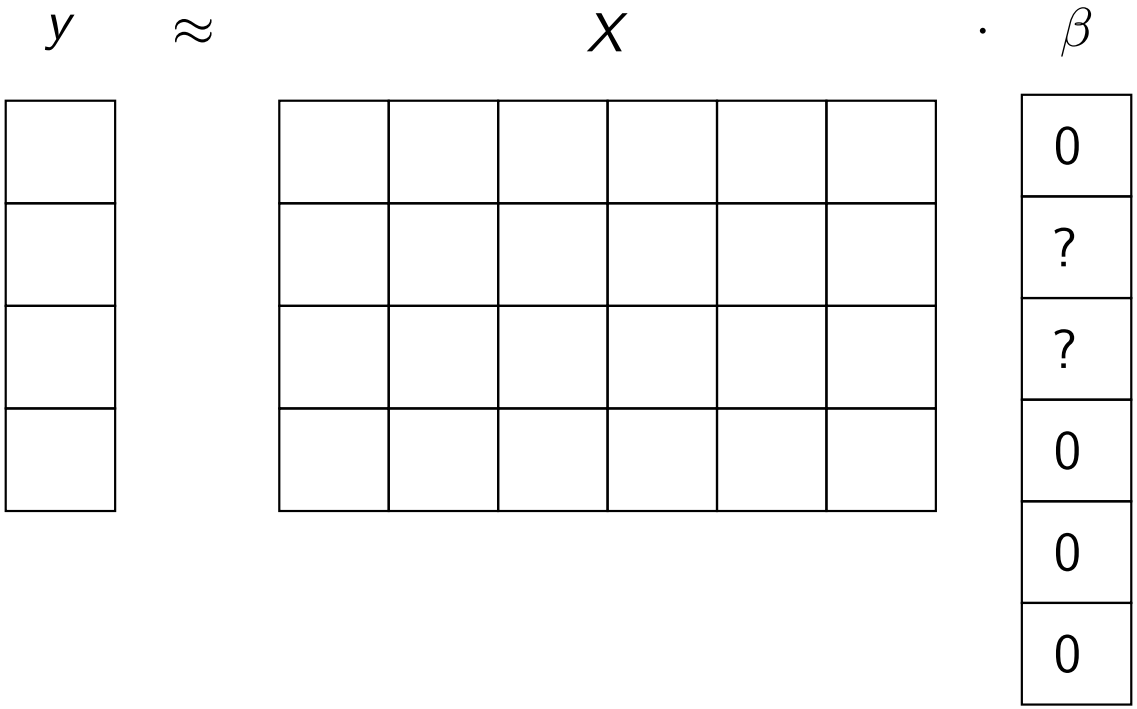
If the number of unknown parameters exceeds the sample size, we have identifiability issue!

Curse of Dimensionality



We can solve this problem if an oracle tells us that only a small number of coefficients are nonzero!

Curse of Dimensionality



We can solve this problem if an oracle tells us that only a small number of coefficients are nonzero!

In reality, we do not know which coefficients are nonzero, nor do we know if the coefficients are sparse!

Use Prior to Improve Power

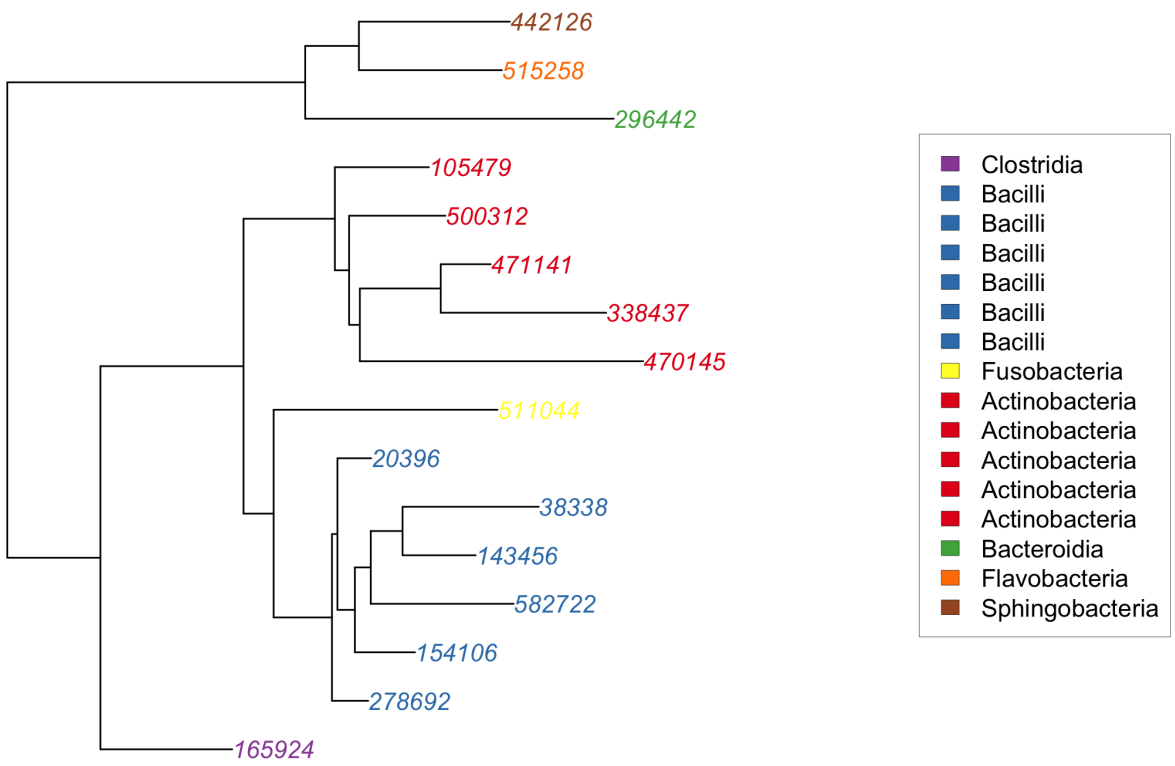
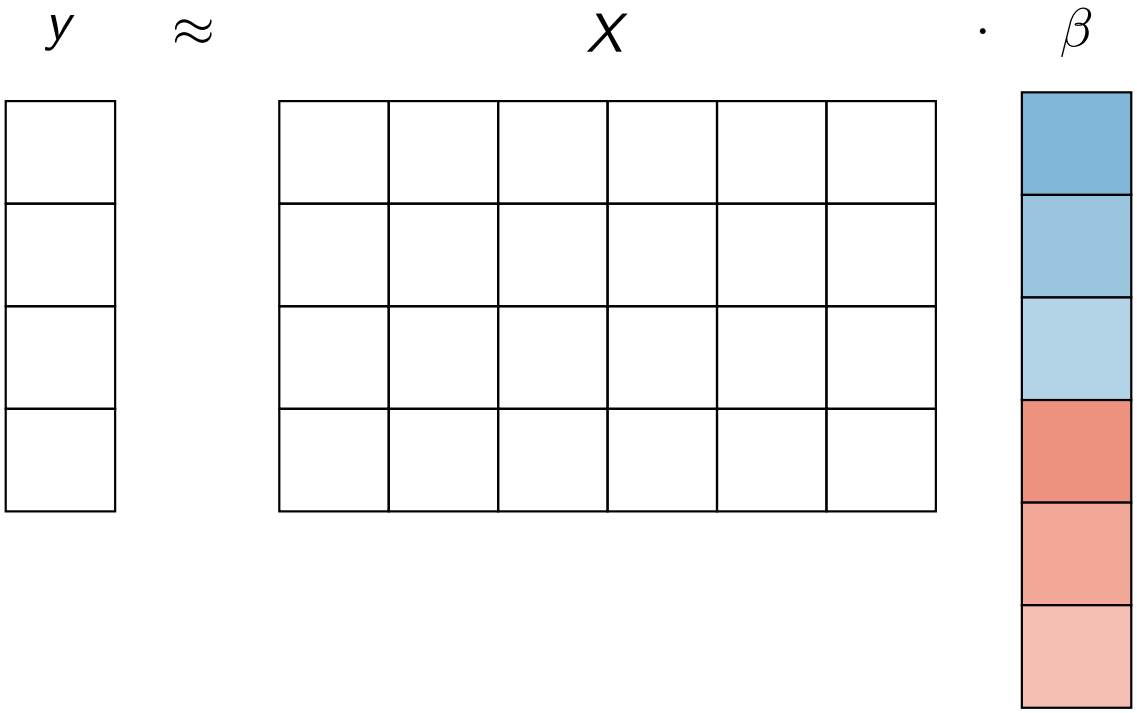


Figure 3: A simple phylogenetic tree

Bacteria closer on the tree are more similar in their DNA content and have similar effects on the outcome (a reasonable assumption).

Use Prior to Improve Power



Assume coefficients are smooth with respect to the prior!

Are Observations Independent?

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

Are Observations Independent?

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

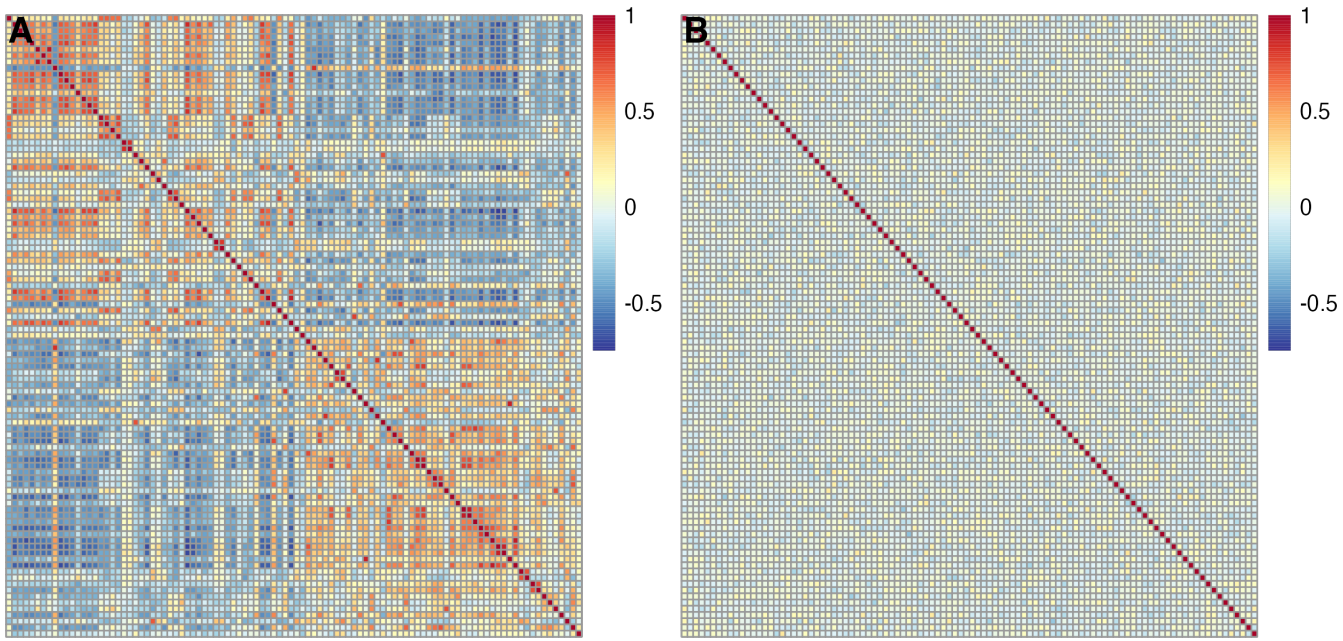


Figure 4: Correlation among (A) Yatsunenko samples; (B) independent samples

Are Observations Independent?

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

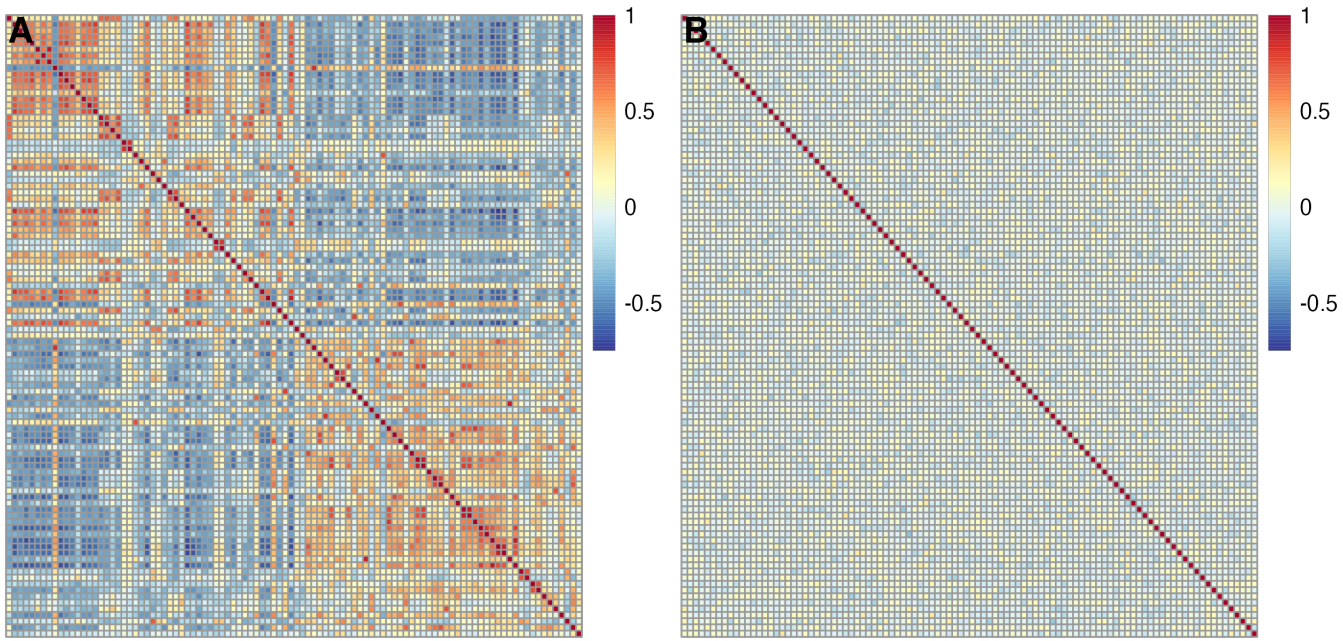


Figure 4: Correlation among (A) Yatsunenko samples; (B) independent samples

In Yatsunenko 12', subjects include individuals from the same household (twins, parent-offspring relationships).

Observations are Correlated

If a treatment works in a person's microbiome, it is more likely to work in the microbiome of their twin siblings.

It is unfair to count good outcomes in both individuals as 2 independent pieces of evidence for the treatment's effectiveness.

Doing so artificially increases the sample size, decreases the P values, and potentially results in effects being deemed significant when they should not be (a type I error).

Accounting for Correlated Observations

If an oracle tells us the correlation among observations, we can use this knowledge to de-correlate the observations. This is called generalized least squares.

Accounting for Correlated Observations

If an oracle tells us the correlation among observations, we can use this knowledge to de-correlate the observations. This is called generalized least squares.

In the absence of this oracle, we can derive a prior on sample correlation from auxiliary data.

Accounting for Correlated Observations

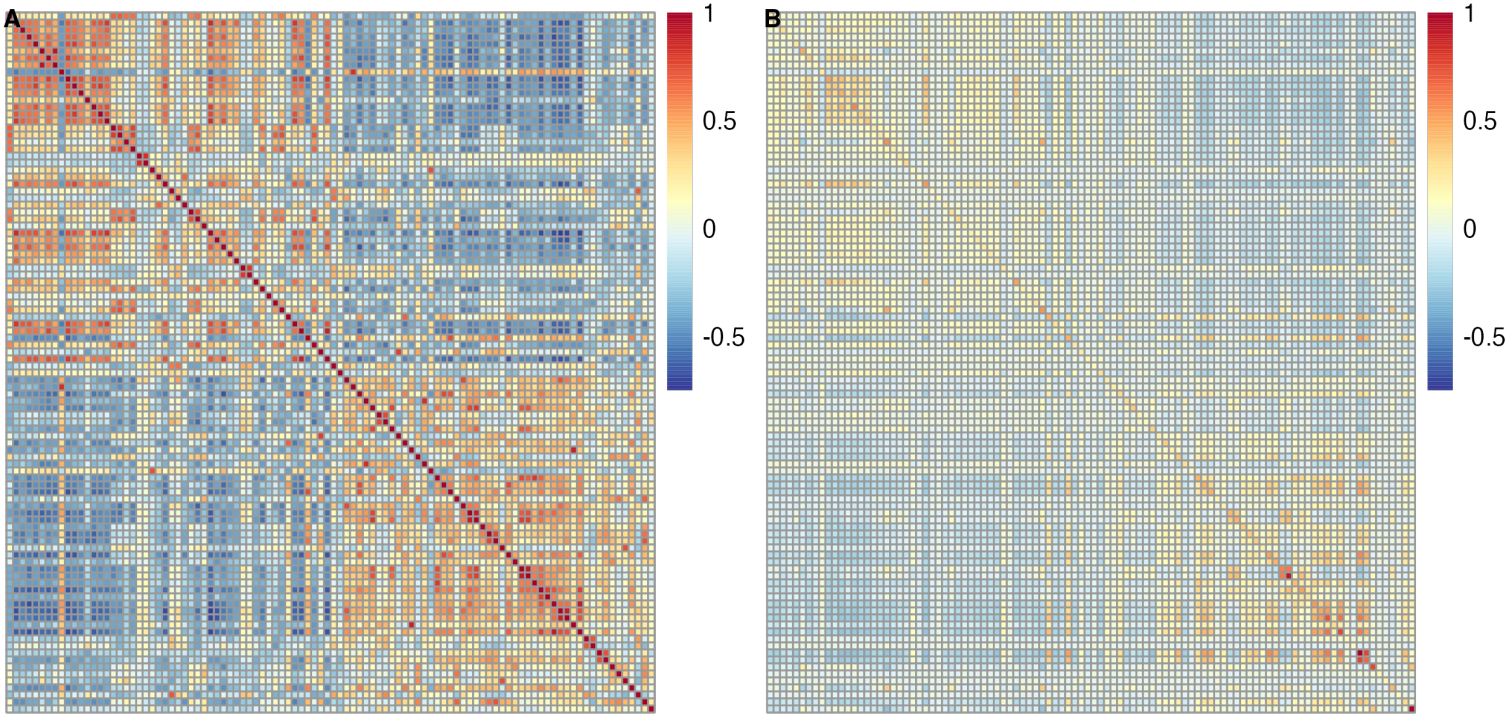


Figure 5: Sample correlation in Yatsunenko 12' from (A) 16S abundance and (B) metagenomic pathway abundance

GMDR

Generalized Matrix Decomposition Regression

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

Yue Wang

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 a sample similarity network H

Prior Misspecification

Our prior may be biased/incomplete!

Prior Misspecification

Fred Hutch Cancer Center

Our prior may be biased/incomplete!

We propose robust GMDR:

1. Test the association between given prior and observed correlations using the Kernel RV (KRV) coefficient. KRV rejects the null $\rightarrow$ prior is at least partially informative.

Prior Misspecification

Fred Hutch Cancer Center

Our prior may be biased/incomplete!

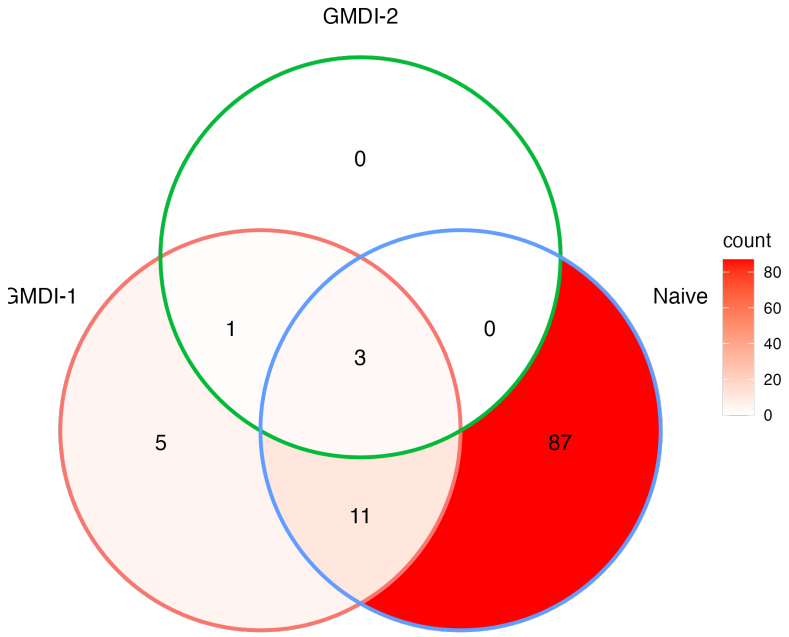
We propose robust GMDR:

1. Test the association between given prior and observed correlations using the Kernel RV (KRV) coefficient. KRV rejects the null $\rightarrow$ prior is at least partially informative.
2. For partially informative prior, use a likelihood criterion to weight the prior against an uninformative baseline.

Revisit Yatsunenko 12'

Which bacteria are associated with age?

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



GMDI-1: discrete shrinkage; GMDI-2: continuous shrinkage

Visualization

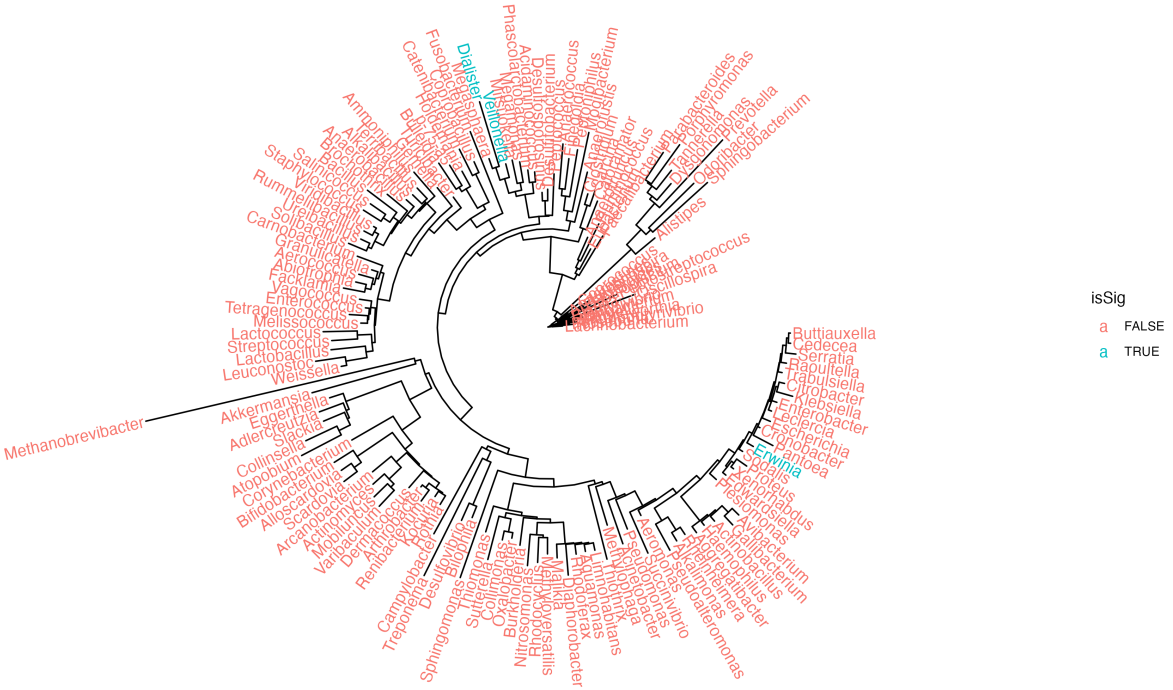


Figure 6: *Dialister* and *Veillonella* are phylogenetically close.

- Dialister has been shown to play a role in age-related diseases, such as obesity and diabetes[5].
- Veillonella is a signature of infant (4-month old) microbiome and breast feeding[6].

---

[5] Xu et al., 20'; Gurung et al., 20'

[6] Backhed et al., 15'

Current Opportunities

Network biology

- microbial networks: can we improve power by modeling the latent network structure?

Current Opportunities

Network biology

- microbial networks: can we improve power by modeling the latent network structure?
- disease association networks: how do disease co-occurrences vary with age? how to integrate data from epidemiological surveys with electronic medical records?

Current Opportunities

Network biology

- microbial networks: can we improve power by modeling the latent network structure?
- disease association networks: how do disease co-occurrences vary with age? how to integrate data from epidemiological surveys with electronic medical records?

Microbiome

- gut-brain association: how to define the association and perform valid inference?

Current Opportunities

Fred Hutch Cancer Center

Network biology

- microbial networks: can we improve power by modeling the latent network structure?
- disease association networks: how do disease co-occurrences vary with age? how to integrate data from epidemiological surveys with electronic medical records?

Microbiome

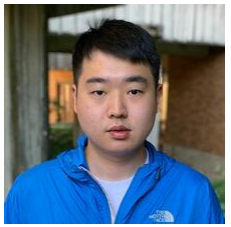
- gut-brain association: how to define the association and perform valid inference?
- microbiome and cancer: what features of the microbiome are correlated with cancer diagnosis and treatment?

Acknowledgement

Ma Lab







Yue Wang (CU), Ilias Moysidis (CERTH), Kristyn Pantoja (Novartis), Xinyi Xie, Wenjie Guan

Collaborators





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

Funding

National Institute of General Medical Sciences

FHCC TDS IRC Pilot