Statistical Methods to Enhance Reproducible Microbiome Biomarker Discovery

Jing Ma

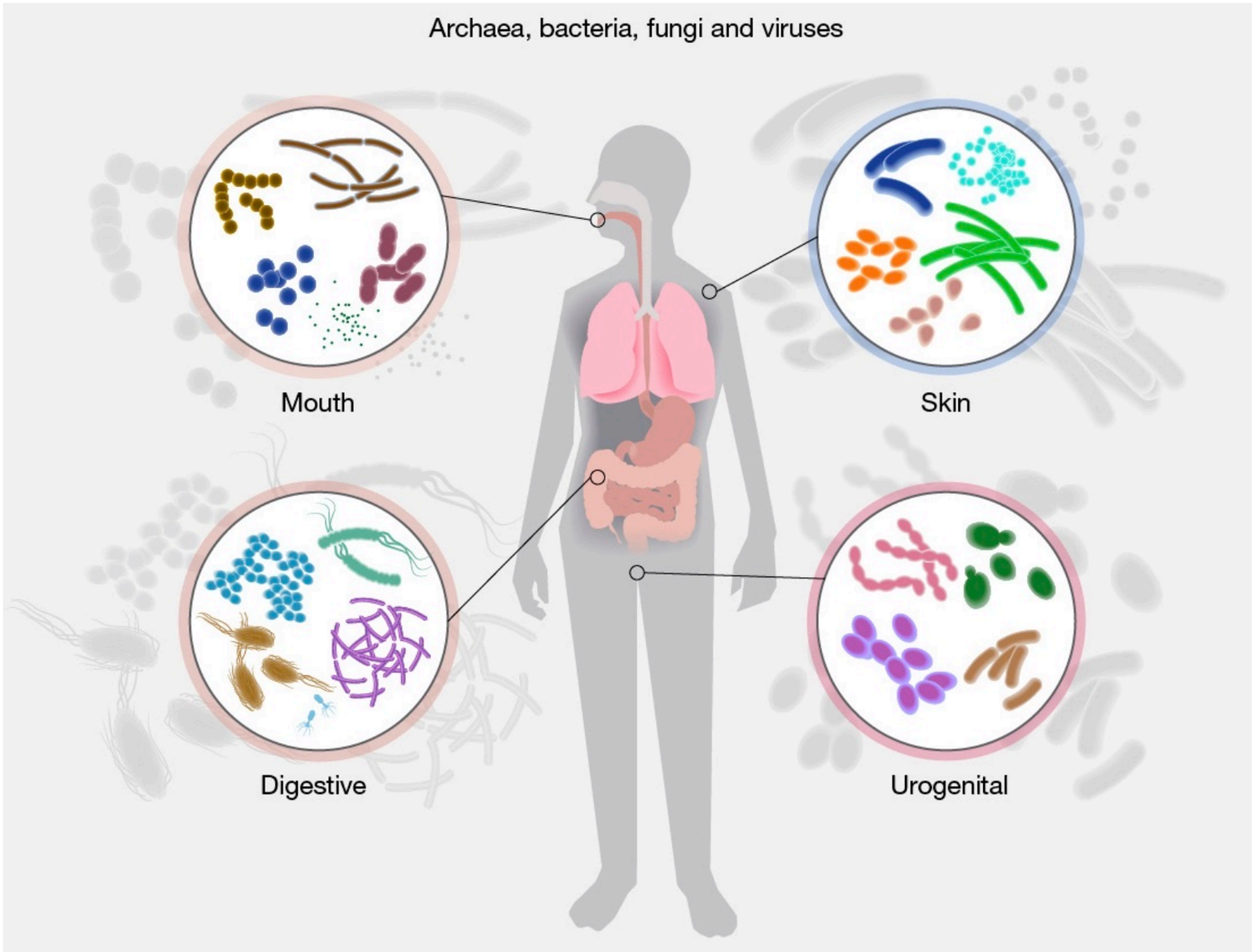
Assistant Professor, Biostatistics
Division of Public Health Sciences
Fred Hutchinson Cancer Center

25 October 2022

The Human Microbiome

Microbes + their genome

Oral health



Dermatological

Cancer, IBD, cardiovascular disease, neurological disorders

Pre-term birth, bacterial vaginosis

https://www.genome.gov/genetics-glossary/Microbiome

Microbiome and Human Cancer

Allogeneic Hematopoietic-Cell Transplantation is a curative therapy for hematologic cancers.

Complications such as graft-versus-host disease (GVHD) remain a major cause of illness and death.

ORIGINAL ARTICLE

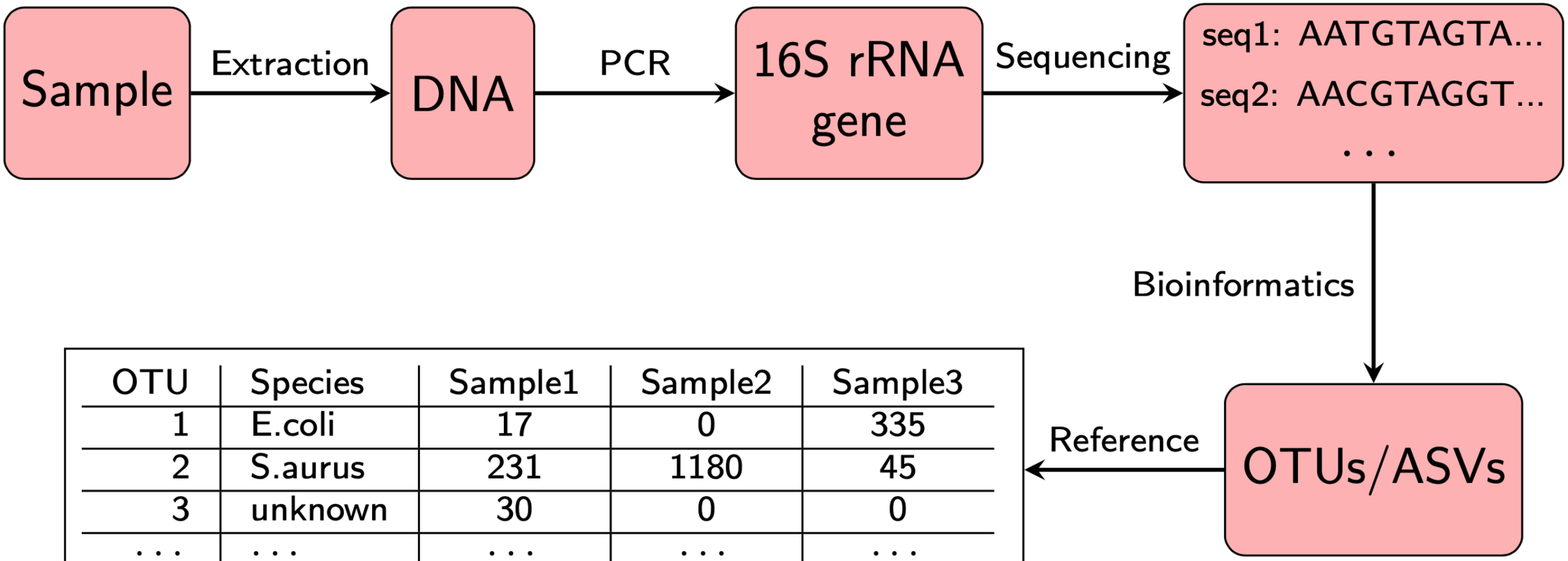
Microbiota as Predictor of Mortality in Allogeneic Hematopoietic-Cell Transplantation

Jonathan U. Peled, M.D., Ph.D., Antonio L.C. Gomes, Ph.D., Sean M. Devlin, Ph.D., Eric R. Littmann, B.A., Ying Taur, M.D., Anthony D. Sung, M.D., Daniela Weber, M.D., Daigo Hashimoto, M.D., Ph.D., Ann E. Slingerland, B.S., John B. Slingerland, B.S., Molly Maloy, M.S., Annelie G. Clurman, B.A., et al.

Scientific Question

Which bacterial species are associated with poor outcome (e.g., transplant survival)?

Data: Marker-Gene Sequencing

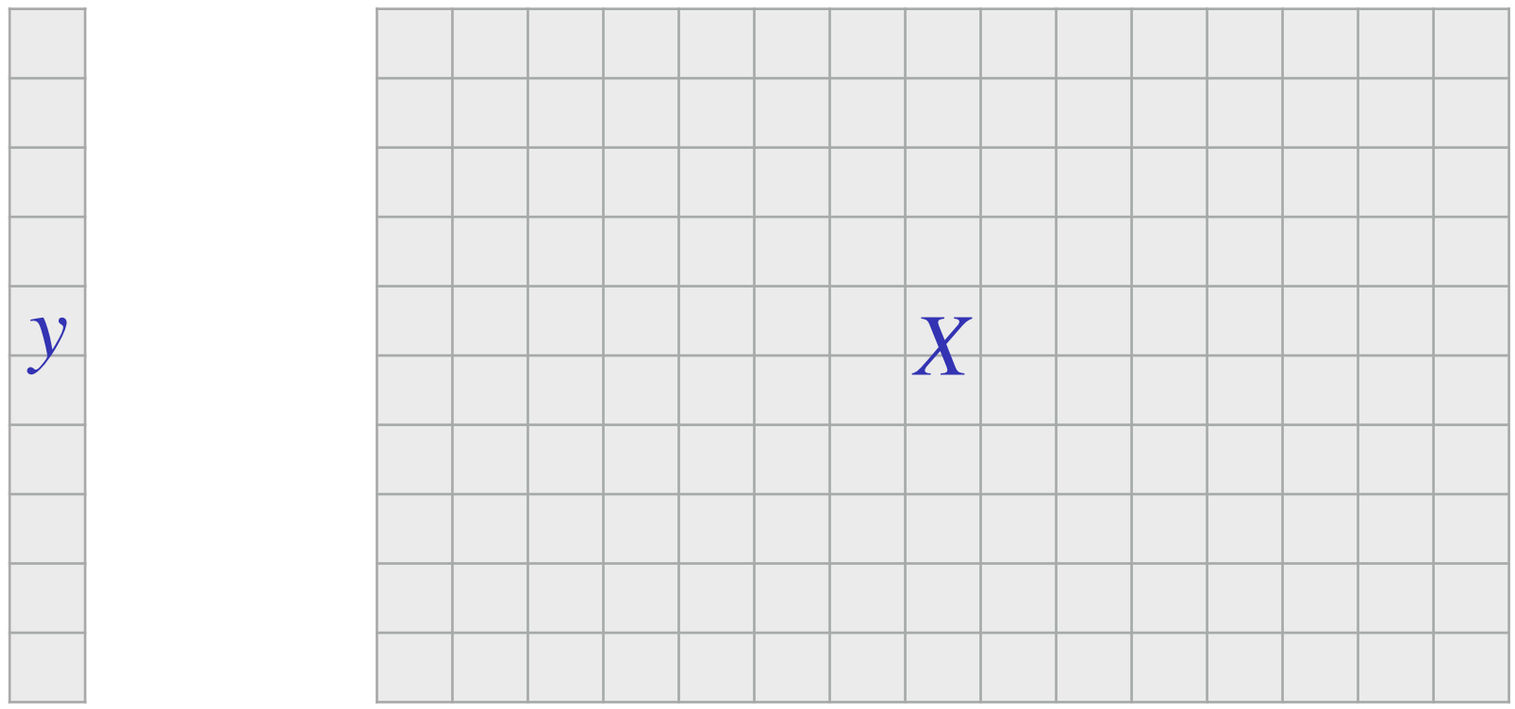


OTU	Species	Sample1	Sample2	Sample3
1	E.coli	17	0	335
2	S.aurus	231	1180	45
3	unknown	30	0	0
...	...	...	...	...

… a contingency table of <u>bacterial</u> abundances

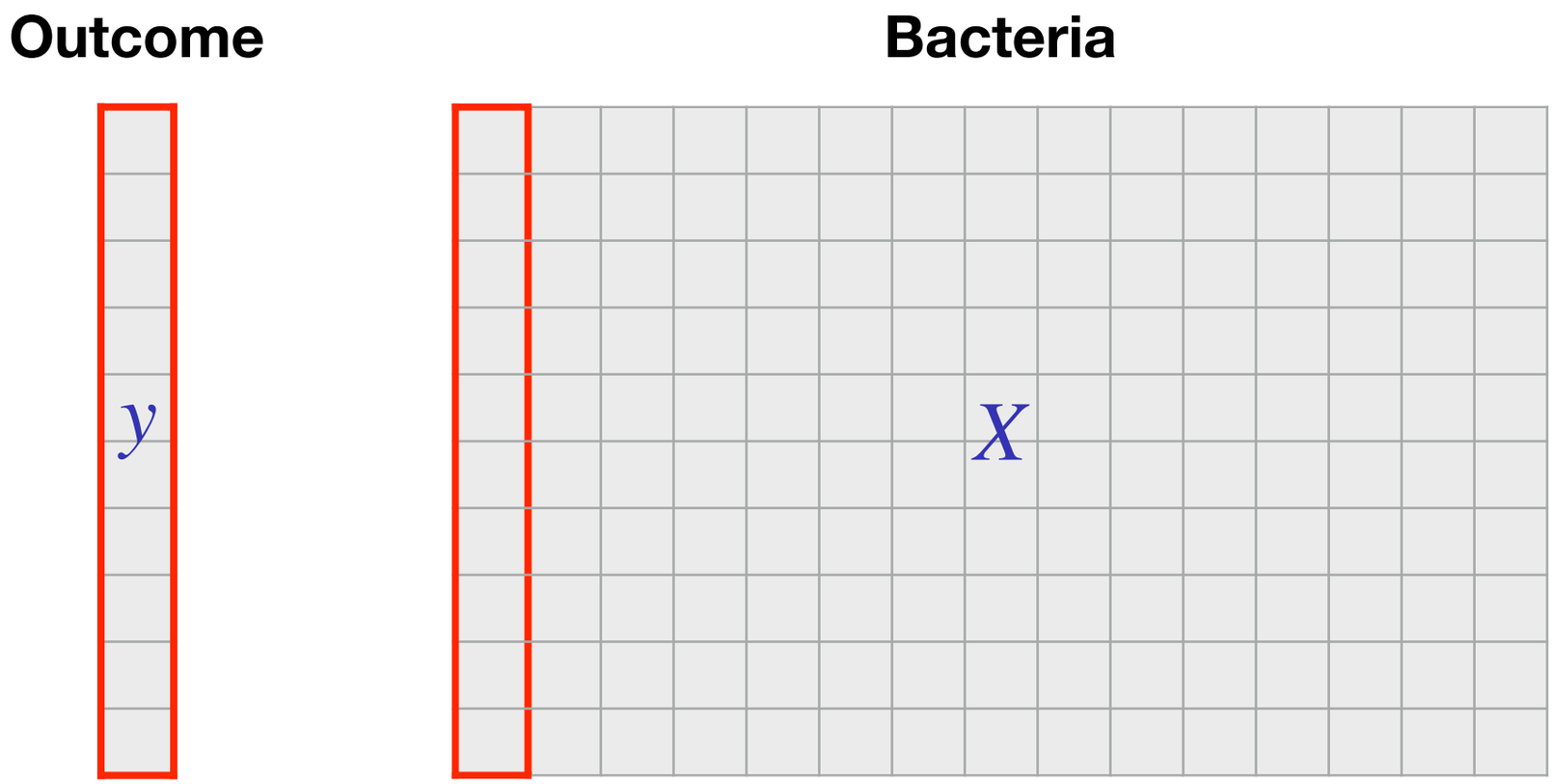
Conventional Analysis

Perform univariate test of each species with respect to an outcome



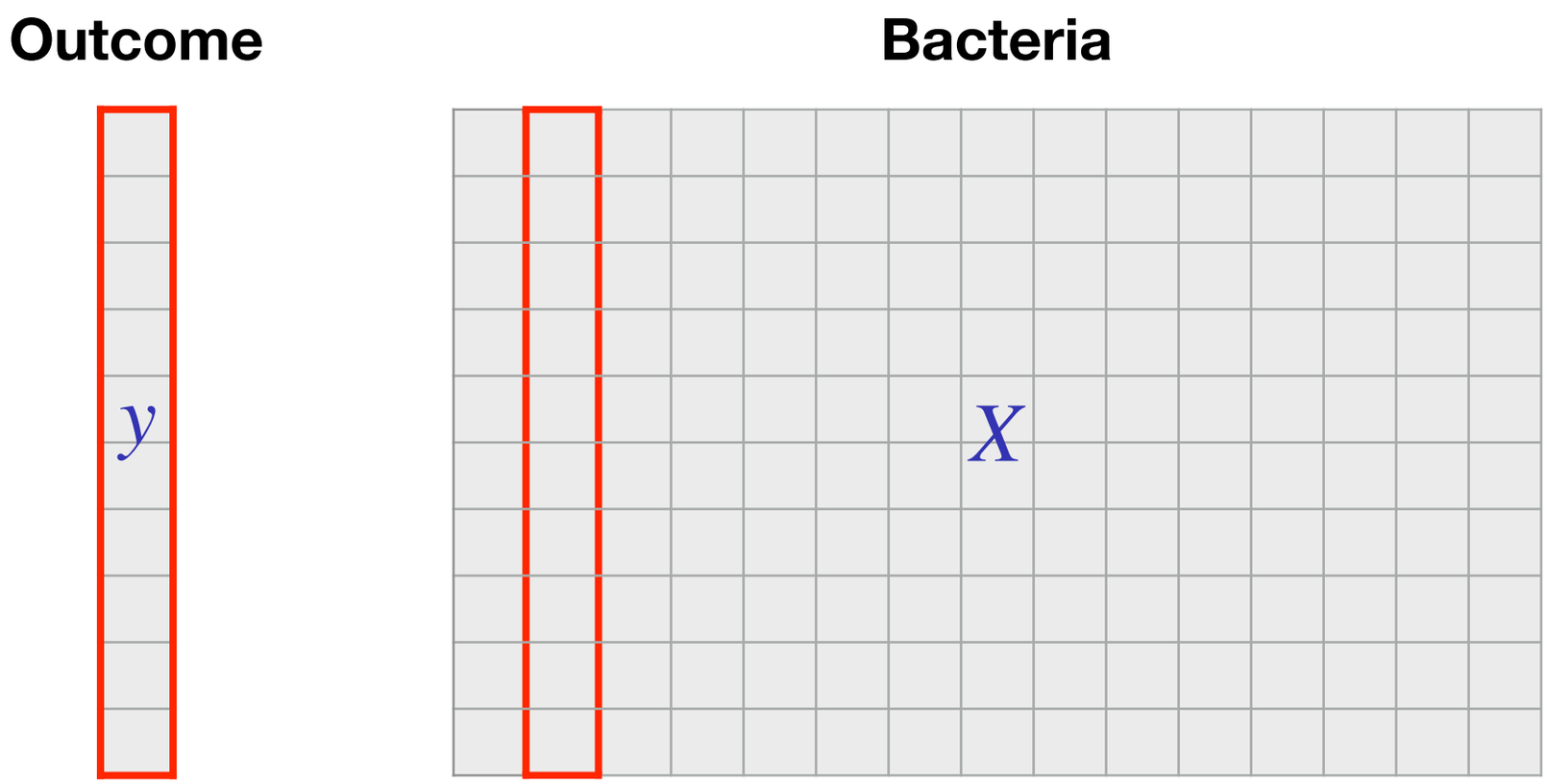
Conventional Analysis

Perform univariate test of each species with respect to an outcome



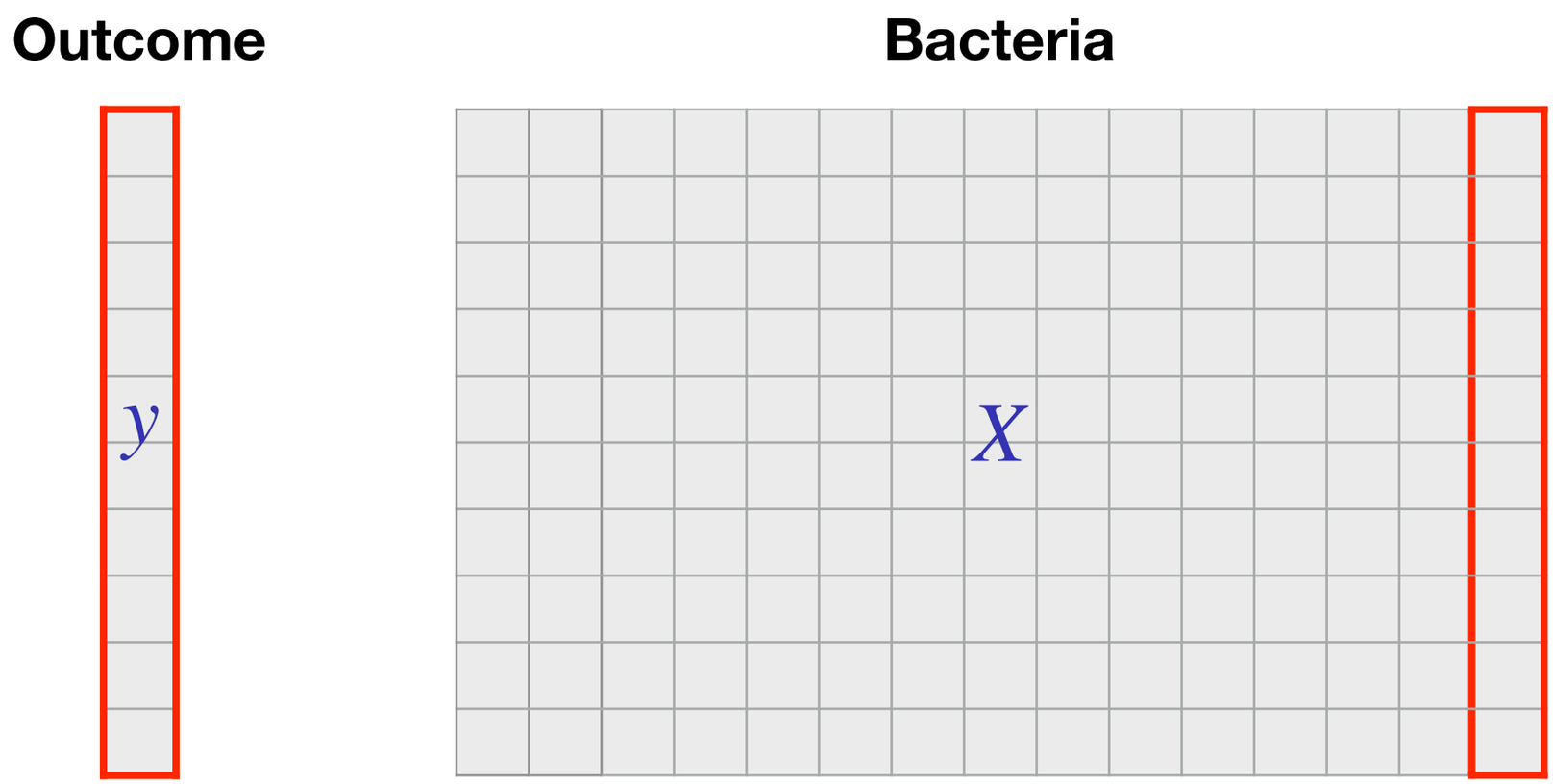
Conventional Analysis

Perform univariate test of each species with respect to an outcome



Conventional Analysis

Perform univariate test of each species with respect to an outcome

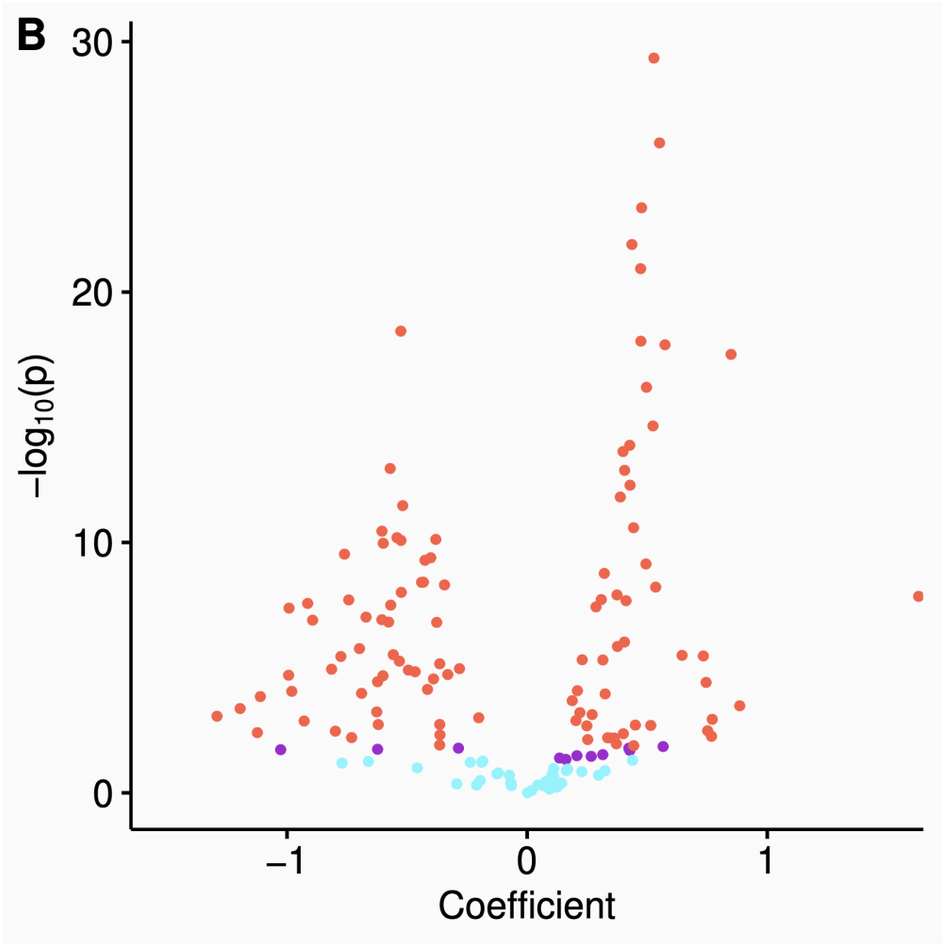


Output: a vector of p-values (after correcting for multiple comparisons)

Application to Yatsunenko 12’

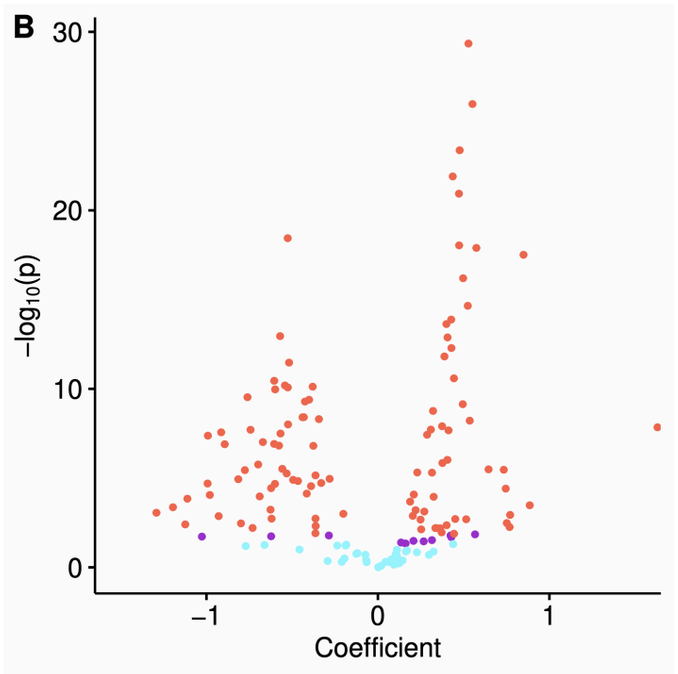
Which bacteria are associated with age?

- n=100 samples
- p=149 bacteria
- Outcome: age
- Univariate spearman rank correlation test

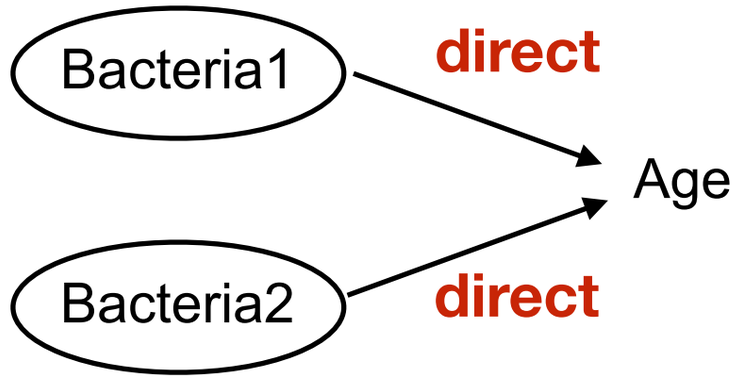


😨 **Which bacteria should I target?**

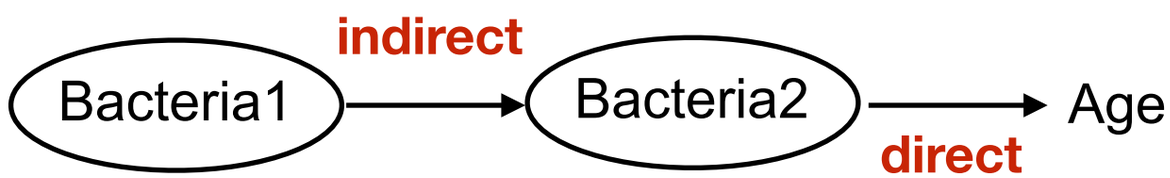
Possible Explanations



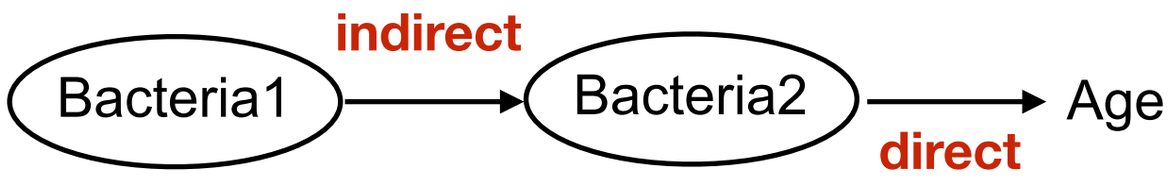
1. Many bacteria are *directly* associated with age



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



Conditional Association



Conditional on bacteria 2, bacteria 1 is independent of age.

$$y_i = X_{i,1}\beta_1 + X_{i,2}\beta_2 + \ldots + X_{i,p}\beta_p + \text{noise}_i,$$

Coefficient β_j: the conditional association between bacteria j and age

β can be solved by R commands: `lm(y~X1+X2)`

Curse of Dimensionality

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

- Reliable inference requires sample size $n \gg p$.

  This is desirable but often infeasible.

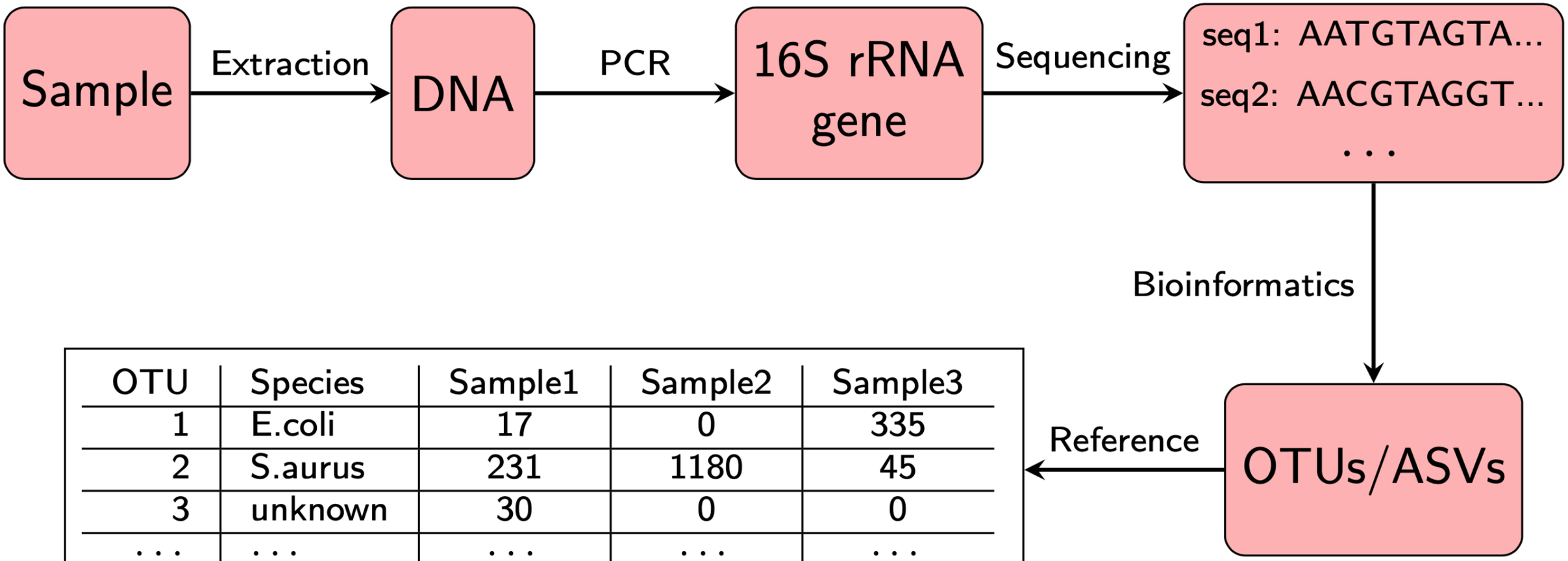
- Assume sparsity: most β_j's are zero.

  This is common, maybe reasonable, but uncheckable.

- Use prior knowledge.

  Well-curated priors are often available a priori.

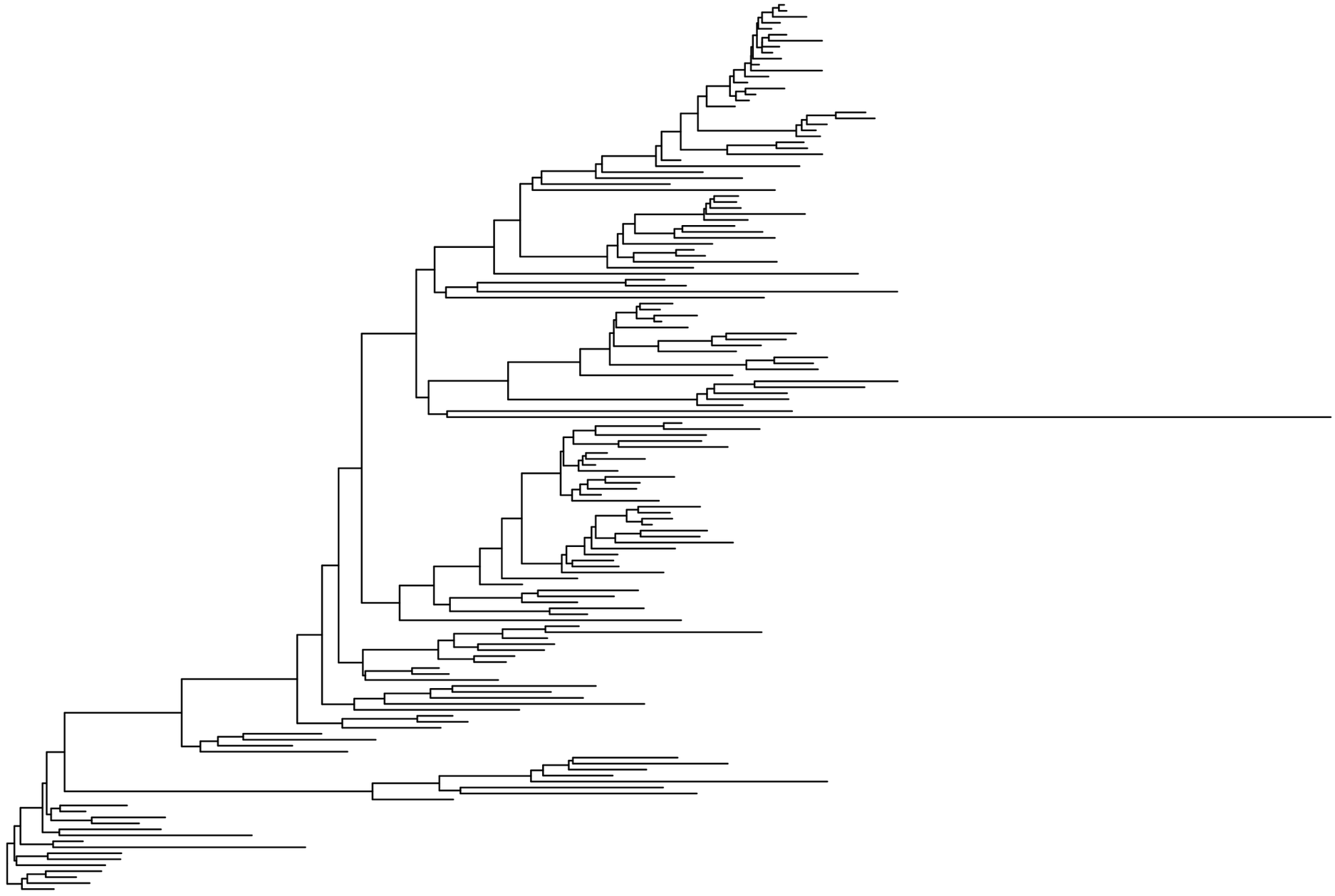
Prior Knowledge



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

Genetic sequences define similarity between bacteria.

The Phylogenetic Prior



We can construct a phylogenetic tree where each tip represents a bacteria from genetic sequences.

Graph-constrained Regression

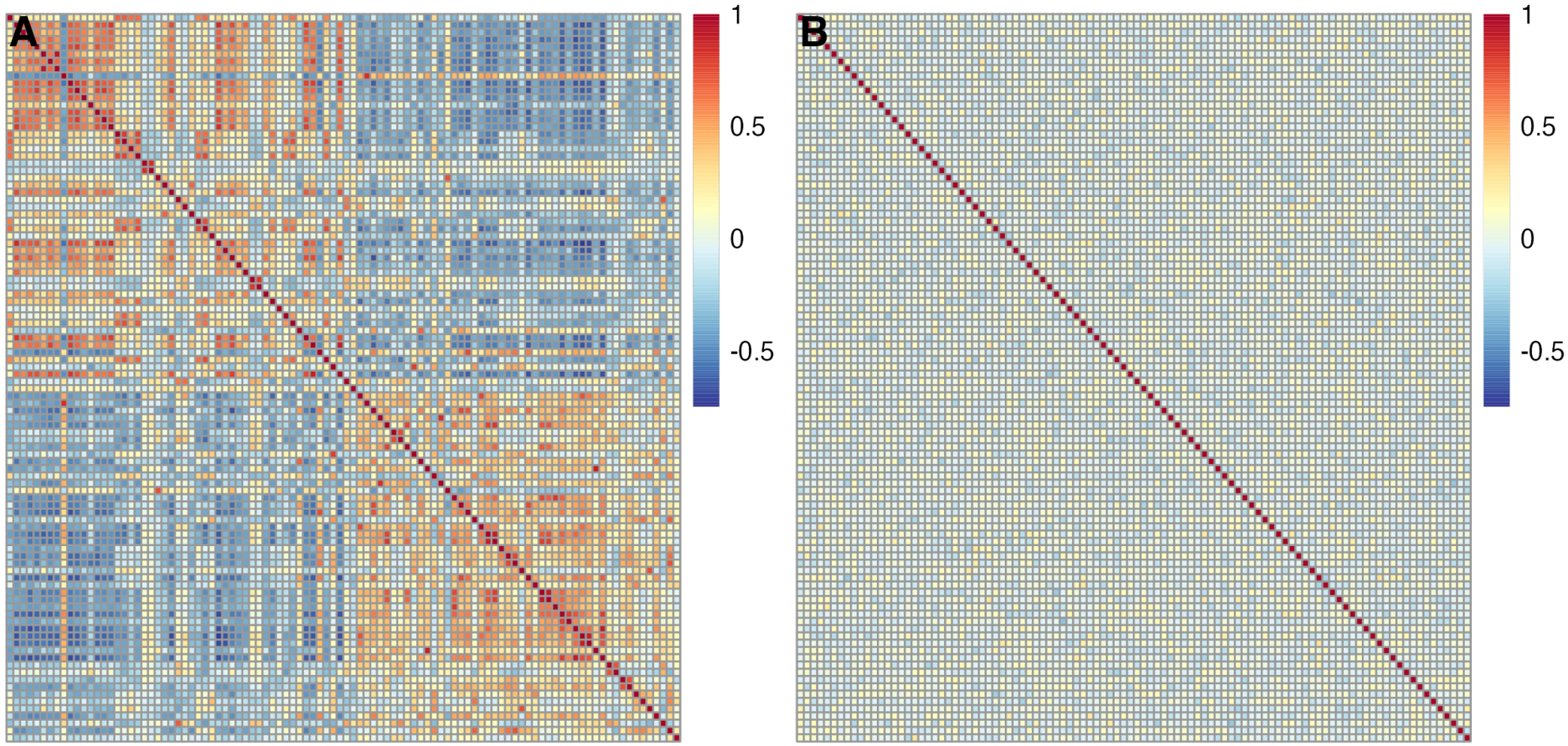
$$y_i = X_{i,1}\beta_1 + X_{i,2}\beta_2 + \ldots + X_{i,p}\beta_p + \text{noise}_i,$$

Subject to the constraint that

- the coefficients β are smooth with respect to the phylogenetic tree

Are Samples Independent?

But wait! So far we have assumed the samples are independent and identically distributed. Is this assumption valid?



Correlation among (A) Yatsunenko samples; (B) independent samples

Samples Are Dependent

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

- Effective sample size $\leq n$
- May lead to inflated type I error

Dealing with Sample Dependence

- Study design: recruit diverse subjects that are less dependent

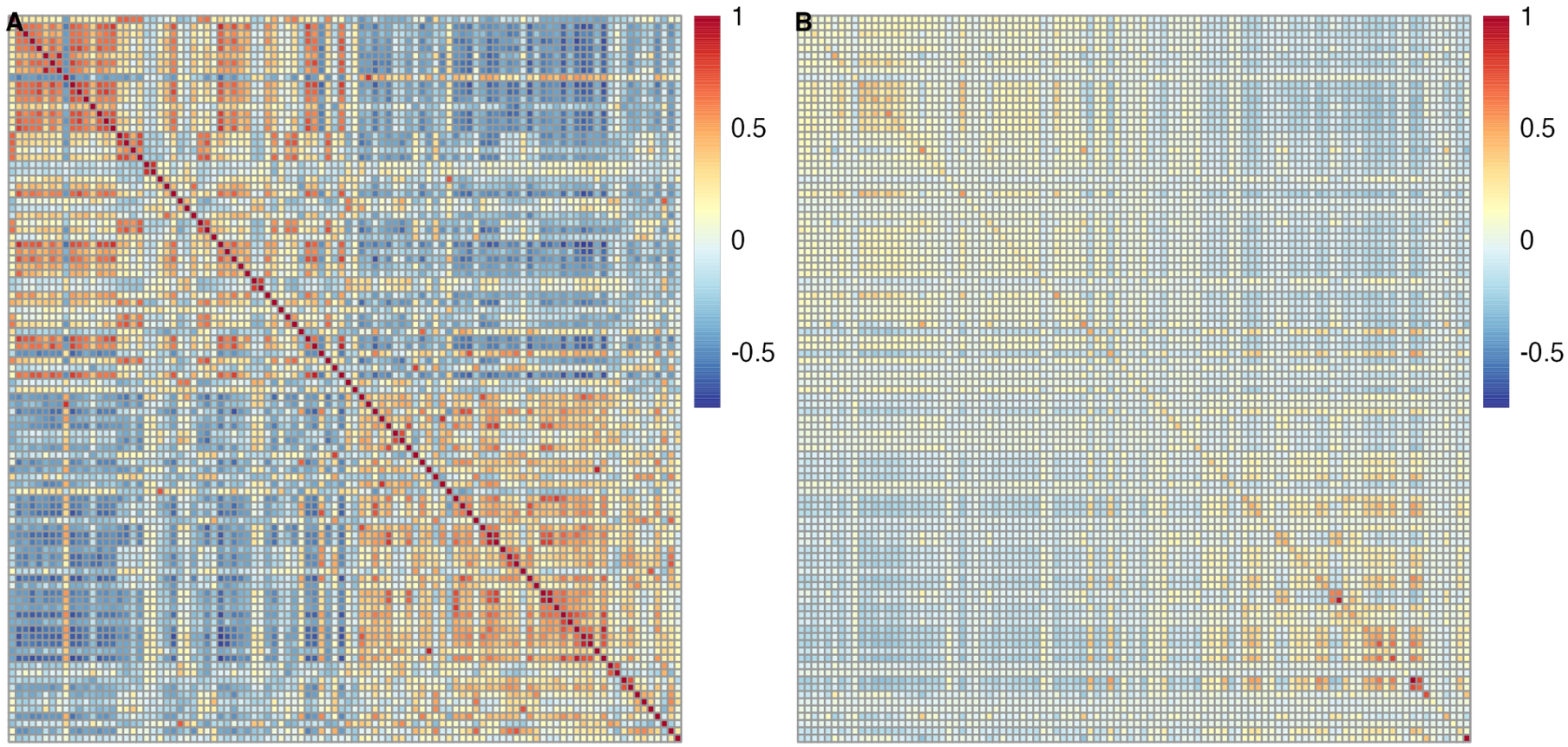
This is desirable but maybe costly.

- Account for sample dependence analytically

This requires prior knowledge.

Prior Knowledge

In Yatsunenko 12', additional data on the same 100 individuals were collected via shotgun metagenomics sequencing



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

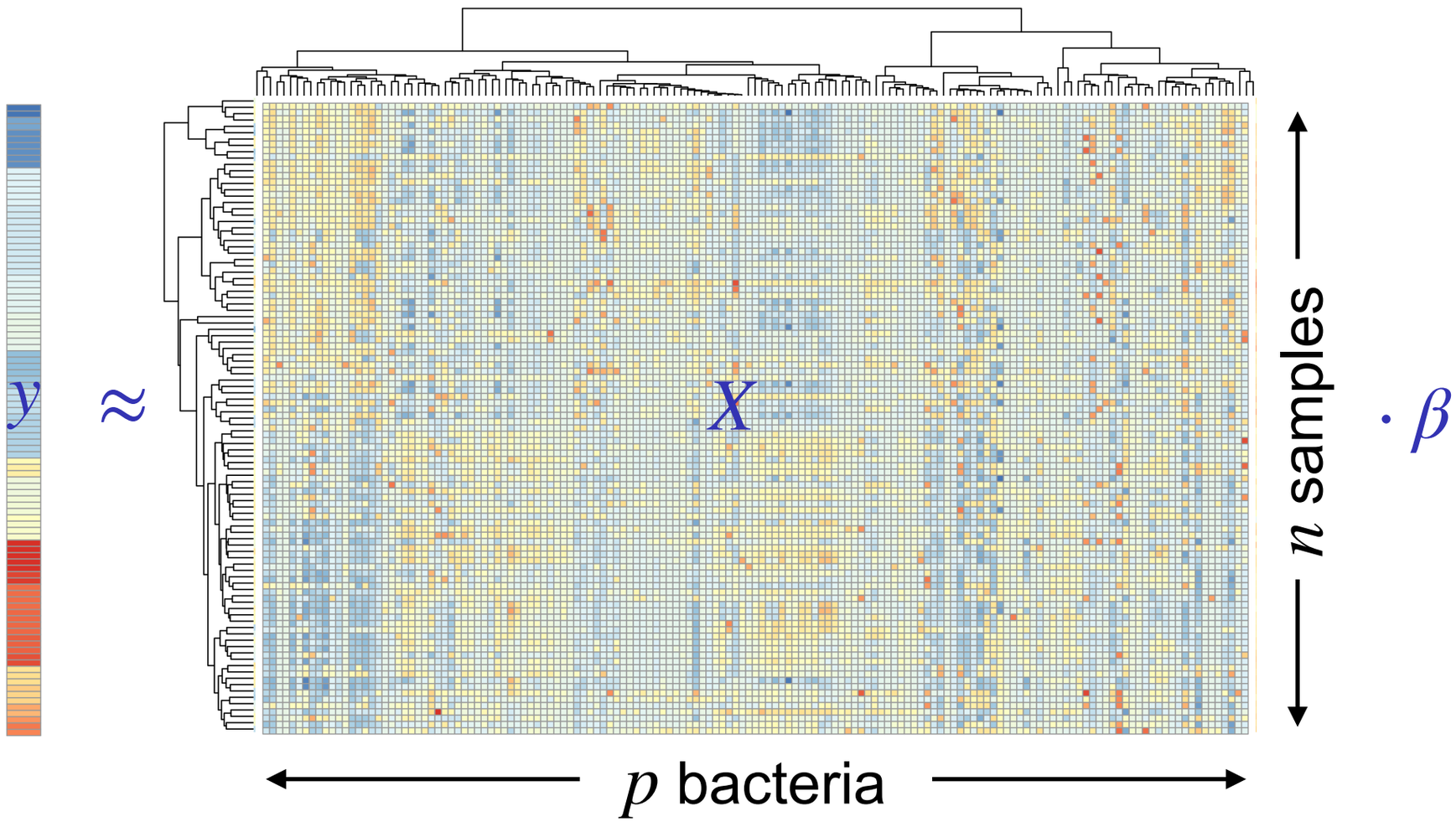
$$y_i = X_{i,1}\beta_1 + X_{i,2}\beta_2 + \ldots + X_{i,p}\beta_p + \text{noise}_i,$$

Subject to the constraint that

- the coefficients β are smooth with respect to the phylogenetic tree
- the noise covariance is a linear function of a pre-specified sample similarity kernel H and the identity matrix

Output: p-value for testing the null $\beta_j = 0$ for $j = 1,\ldots,p$.

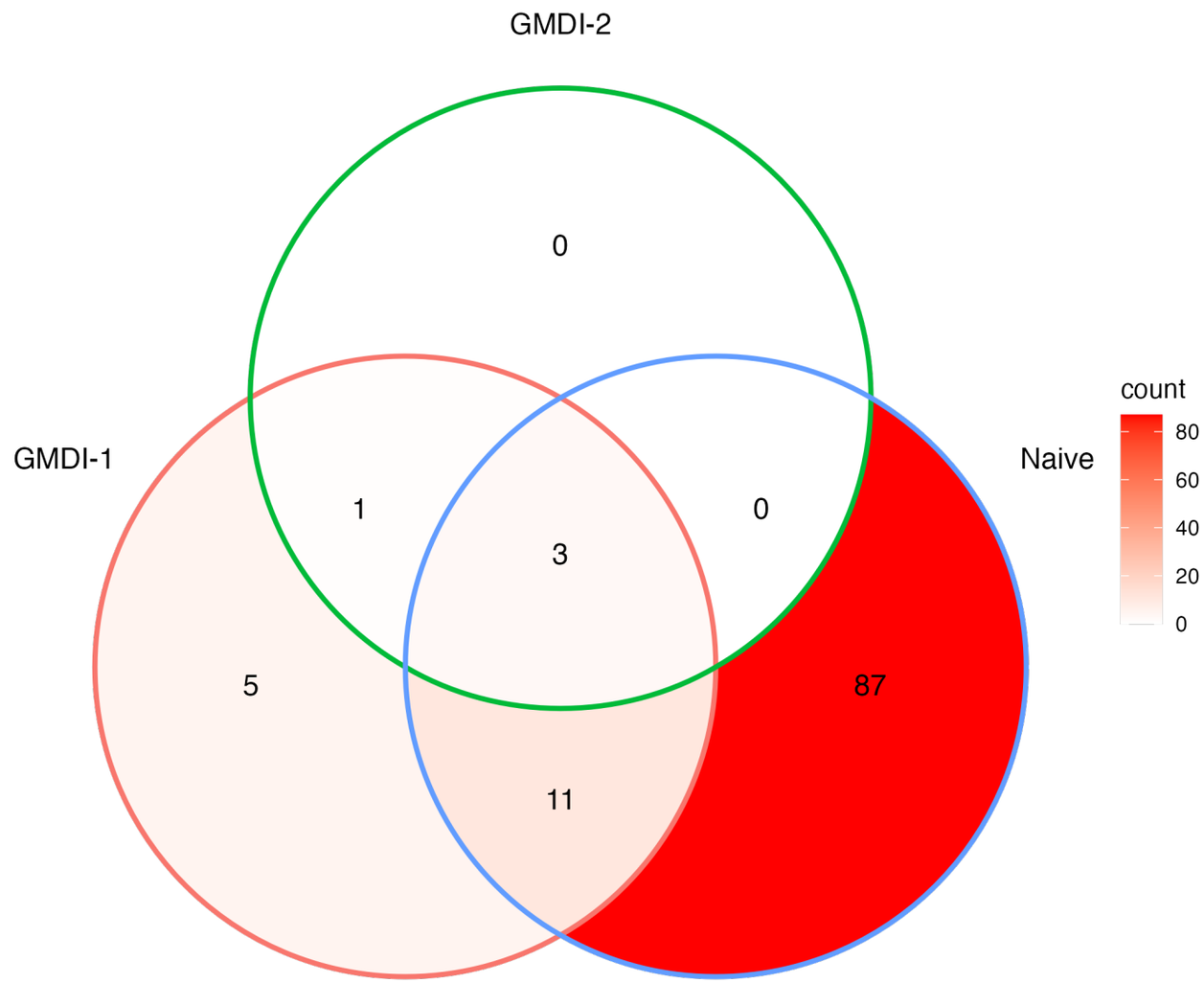
Generalized Matrix Decomposition Regression



Application to Yatsunenko 12’

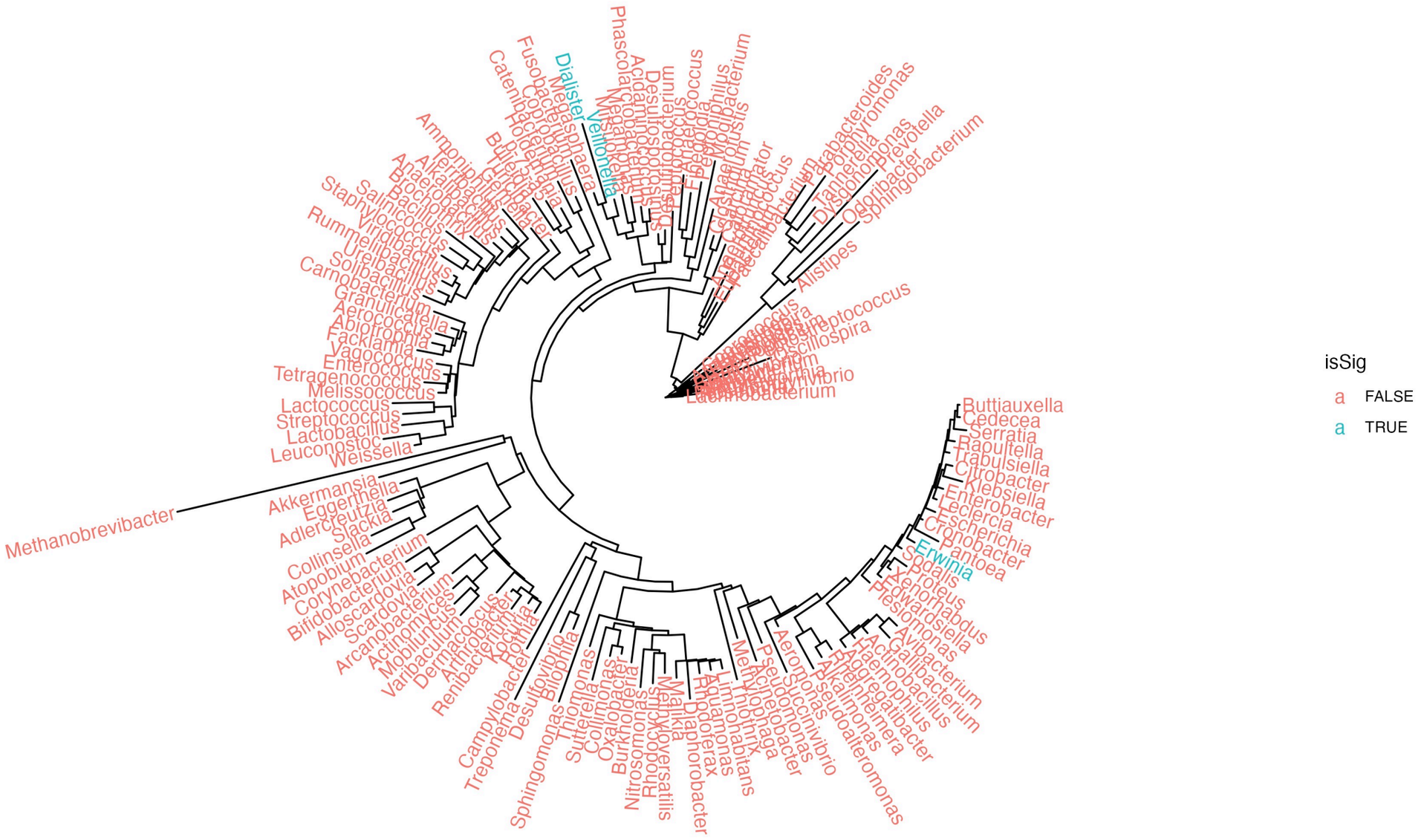
Which bacteria are associated with age?

- n=100 samples
- p=149 bacteria
- Outcome: age
- FDR = 0.1



Dialister, Erwinia, Veillonella

Visualization



Dialister and *Veillonella* are phylogenetically close.

Conclusion

- A good study design is the most important step towards reproducibility
- Always check model assumptions
- Multiple regression identifies features that are *conditionally* associated with an outcome
- Incorporating prior knowledge alleviates model complexity and yields biologically meaningful results

  Domain expertise is needed!

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

References

| Editor's Pick | Methods and Protocols | 17 December 2019

The Generalized Matrix Decomposition Biplot and Its Application to Microbiome Data

Authors: Yue Wang, Timothy W. Randolph, Ali Shojaie, Jing Ma | AUTHORS INFO & AFFILIATIONS

DOI: https://doi.org/10.1128/mSystems.00504-19 • Check for updates

[Submitted on 16 Apr 2021]

Generalized Matrix Decomposition Regression: Estimation and Inference for Two-way Structured Data

Yue Wang, Ali Shojaie, Timothy W. Randolph, Jing Ma

This paper studies high-dimensional regression with two-way structured data. To estimate the high-dimensional coefficient vector, we propose the generalized matrix decomposition regression (GMDR) to efficiently leverage any auxiliary information on row and column structures. The GMDR extends the principal component regression (PCR) to two-way structured data, but unlike PCR, the GMDR selects the components that are most predictive of the outcome, leading to more accurate prediction. For inference on regression coefficients of individual variables, we propose the generalized matrix decomposition inference (GMDI), a general high-dimensional inferential framework for a large family of estimators that include the proposed GMDR estimator. GMDI provides more flexibility for modeling relevant auxiliary row and column structures. As a result, GMDI does not require the true regression coefficients to be sparse; it also allows dependent and heteroscedastic observations. We study the theoretical properties of GMDI in terms of both the type-I error rate and power and demonstrate the effectiveness of GMDR and GMDI on simulation studies and an application to human microbiome data.

Data & Code Availability

Published: 09 May 2012

Human gut microbiome viewed across age and geography

Tanya Yatsunenko, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N. Baldassano, Andrey P. Anokhin, Andrew C. Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J. Gregory Caporaso, Catherine A. Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight & Jeffrey I. Gordon

Nature **486**, 222–227 (2012) | Cite this article

Installation

Currently, the only way to install the package is from GitHub.

```
install.package("devtools")
devtools::install_github("pknight24/KPR")
```

Data processing

The demonstrate the KPR model, we use data from the study by (Yatsunenko et al, 2012), which is included in the package. The `yatsunenko` object is a list containing raw microbe abundance data, patristic distances between genera, and the country of origin and age of each subject. We will fit a kernel penalized regression model with the microbial abundances as the penalized variables, and subject age as the outcome.

In order to incorporate the patristic distance matrix into the regression model, we first need to convert it to a similarity kernel using a function provided by the `KPR` package.

```
library(KPR)
data(yatsunenko)

age <- yatsunenko$age
counts <- yatsunenko$raw.counts
patristic <- yatsunenko$patristic
```

Acknowledgement

Ma Lab

Kristyn Pantoja (Novartis)

Yue Wang (ASU)

Ilias Moysidis (CERTH)

Fang Nan

Xinyi Xie

Collaborators

Parker Knight

Tim Randolph

Ali Shojaie

David Jones

Amanda Phipps

Kate Markey

Robert Kaplan

Dog Aging Project

Funding

R01 GM145772

PAM/MRI Pilot

TDS Pilot

Microbiome and Human Cancer

Article | Published: 11 March 2020

Microbiome analyses of blood and tissues suggest cancer diagnostic approach

Gregory D. Poore, Evguenia Kopylova, Qiyun Zhu, Carolina Carpenter, Serena Fraraccio, Stephen Wandro, Tomasz Kosciolek, Stefan Janssen, Jessica Metcalf, Se Jin Song, Jad Kanbar, Sandrine Miller-Montgomery, Robert Heaton, Rana Mckay, Sandip Pravin Patel, Austin D. Swafford & Rob Knight ✉

Nature **579**, 567–574 (2020) | Cite this article

Micronoma™