Statistical Methods for Association Analysis of Microbiome Data

Hongjiao Liu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:
Michael C. Wu, Chair
Wei Sun
Ting Ye

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Statistical Methods for Association Analysis of Microbiome Data

Hongjiao Liu

Chair of the Supervisory Committee:
Michael C. Wu
Department of Biostatistics

The human microbiome is an integral component of the human body. High-throughput sequencing techniques have provided detailed information on abundance and phylogeny of individual taxa in the human microbiome. A variety of association studies based on microbiome data has emerged in recent years, revealing important relationships among microbial features as well as between the microbiome and host health. Challenges specific to microbiome data, such as high-dimensionality and sparsity, call for novel statistical approaches. Meanwhile, common practical needs in association analyses, such as covariate adjustment and analysis of clustered data, can be extended to microbiome data. Here we present four projects on novel statistical methods for association analyses of microbiome data.

In Project 1, we propose a powerful kernel-based approach for microbiome genome-wide association studies (GWASs), where we evaluate the covariate-adjusted association between groups of genetic variants at the gene level and the overall microbiome composition at the community level. In Project 2, we develop a kernel-based multivariate independence test for clustered data and apply the test to evaluate the association between the overall microbiome composition and a multivariate trait based on longitudinal data. In Project 3, we propose a multivariate approach to construct microbial association networks, where we develop a conditional independence test to assess the pairwise association between multivariate microbial features, such as bacterial genera composed of multiple species. In Project 4, we propose a

novel approach for one-sample Mendelian randomization with a microbial exposure, which allows us to evaluate the causal effect of individual microbial taxa on a continuous health outcome with an improved power.

TABLE OF CONTENTS

Page

List of Figures iv

List of Tables ix

Chapter 1: Introduction 1
1.1 Genetic Association Analysis for Microbiome Phenotypes 2
1.2 Multivariate Independence Testing for Clustered Data 3
1.3 Multivariate Approach to Microbial Network Construction 5
1.4 Mendelian Randomization with a Microbial Exposure 6

Chapter 2: Kernel-based Genetic Association Analysis for Microbiome Phenotypes 8
2.1 Introduction 8
2.2 Methods 11
2.3 Results 22
2.4 Discussion 31

Chapter 3: A Kernel-based Test of Independence for Cluster-correlated Data . . . 40
3.1 Introduction 40
3.2 Background 43
3.3 HSIC for Cluster-correlated Data 45
3.4 Simulation Studies 49
3.5 Application to Real Data 55
3.6 Discussion 57

Chapter 4: Multivariate Conditional Independence Testing for Vaginal Microbial Network Construction 59
4.1 Introduction 59
4.2 Data and Problem Description 62

4.3 Methods . . . 64
4.4 Simulation Studies . . . 68
4.5 Data Application . . . 73
4.6 Discussion . . . 82

Chapter 5: Mendelian Randomization with a Microbial Exposure . . . 84
5.1 Introduction . . . 84
5.2 Methods . . . 87
5.3 Simulation Studies . . . 94
5.4 Data Application . . . 99
5.5 Discussion . . . 107

Chapter 6: Discussion . . . 109
6.1 Summary . . . 109
6.2 Future Work . . . 110

Bibliography . . . 112

Appendix A: Supplementary Materials for Chapter 2 . . . 135
A.1 Derivation of Covariate-adjusted KRV Coefficient . . . 135
A.2 Taxon-level Microbiome GWAS of the HCHS/SOL Study . . . 138
A.3 Analyses to Assess the Robustness of the *IL23R-C1orf141* Signal . . . 139
A.4 Supplementary Tables and Figures . . . 141

Appendix B: Supplementary Materials for Chapter 3 . . . 150
B.1 Preliminary Results . . . 150
B.2 Proof of Theorem 3.3.2 . . . 161
B.3 Proof of Theorem 3.3.3 . . . 165
B.4 Proof of Proposition 3.3.4 . . . 167
B.5 Proofs of the Lemmas . . . 170
B.6 Additional Simulations . . . 174
B.7 Additional Details on Implementation . . . 179
B.8 MsFLASH Study . . . 181

Appendix C: Supplementary Materials for Chapter 4 183
C.1 Connection between CRV and SEcov 183
C.2 Proof of Theorem 4.3.2 184
C.3 Additional Details on Simulation 186
C.4 Additional Details on Real Data Application 189

LIST OF FIGURES

Figure Number Page

2.1 Illustration of covariate-adjusted KRV for microbiome genome-wide association studies. 12

2.2 P-value QQ-plots from the first-stage gene-level analysis of the HCHS/SOL data. Each panel corresponds to a QQ-plot based on a distinct microbiome kernel. In the adjusted KRV, the top 5 PCs of genome-wide genetic variability were adjusted. $\lambda_{GC,0.1}$ represents the genomic inflation factor evaluated at the upper 10th percentile. 36

2.3 Manhattan plots and linkage disequilibrium (LD; R^2) heatmaps from the second-stage variant-level analysis of the HCHS/SOL data, using the PC-adjusted KRV. Each panel corresponds to a distinct gene or gene region. The Bray-Curtis kernel was used for analysis of variants in the *IL23R-C1orf141* region; the unweighted UniFrac kernel was used for analysis of variants in *ZFR* and *MTMR12*. The top 5 PCs of genome-wide genetic variability were adjusted. The red lines represent variant-level significance after Bonferroni correction ($\alpha = 8.98 \times 10^{-5}$ for variants in the *IL23R-C1orf141* region, and 1.08×10^{-4} for variants in *ZFR* and *MTMR12*). A large R^2 value indicates high LD. 37

2.4 PC2 vs. PC1 from kernel PCA on the microbiome kernel, colored by genotype of top variants from the significant genes in the HCHS/SOL study. For each variant, a 95% confidence ellipse (shown as a filled ellipse with black borders) was constructed for individuals from each genotype. The Bray-Curtis kernel was used for the top variant in the *IL23R-C1orf141* region; the unweighted UniFrac kernel was used for the top variants in *ZFR* and *MTMR12*. The percent of variance captured by each kernel PC was provided in the axis labels. Panels **B**, **D**, **F** show enlarged versions of the confidence ellipses from panels **A**, **C**, **E**. 38

2.5 Empirical power of covariate-adjusted KRV and competing methods at nominal level $\alpha = 0.05$ for different microbiome kernels under small effect sizes. Panel **A**: A single SNP affects the abundance of common OTUs. Panel **B**: A single SNP affects the abundance of OTUs from a common phylogenetic cluster. Panel **C**: A single SNP affects the abundance of rare OTUs. In each scenario, linear kernel was used for genetic data. 39

3.1 Empirical power of $\mathbf{HSIC_{cl}}$ and competing methods at nominal level $\alpha = 0.05$ under **Power Scenario 1**. The x-axis represents the proportion of variables in Y that are associated with X. The top row shows results based on the Gaussian kernel, and the bottom row shows results based on the linear kernel. 54

3.2 Empirical power of $\mathbf{HSIC_{cl}}$ and competing methods at nominal level $\alpha = 0.05$ under **Power Scenario 2**. The x-axis represents the proportion of variables in Y that are associated with X. The top row shows results based on the Gaussian kernel, and the bottom row shows results based on the linear kernel. 55

4.1 Example relationship between two microbial taxa, Genus A and Genus B, where correlations of opposite directions are present among the sub-taxa: positive association between Species A1 and Species B1; negative association between Species A2 and Species B1. 61

4.2 Empirical type I error rates (Panel **A**) and power (Panel **B**) of CRV at a significance level of 0.05 in conditional independence testing. 70

4.3 Performance of different methods for recovering a cluster-type synthetic microbial network, with false discovery rate controlled at 0.2. Conditional correlations of opposite directions are present among species within 50% of all genera. 72

4.4 ROC curves of different methods in recovering synthetic microbial networks of different topologies, averaged over 10 simulations. Conditional correlations of opposite directions are present among species within 50% of all genera. . . 74

4.5 Genus-level vaginal microbial network based on the combined sample of 647 subjects of PIN study, with false discovery rate controlled at 0.2. Panel **A**: Constructed genus-level network; the sizes of the nodes are proportional to their genus-level microbial abundance among all subjects. Panel **B**: Heatmap of three network centrality measures for each genus in the constructed network. 76

4.6 Genus-level vaginal microbial networks in White women (Panel **A**; $n = 373$) vs. Black women (Panel **B**; $n = 274$) of PIN study, with false discovery rate controlled at 0.2. The sizes of the nodes are proportional to their genus-level microbial abundance in the combined sample. 78

4.7 Heatmaps of network centrality measures for each genus in genus-level vaginal microbial networks based on White women (Panel **A**; $n = 373$) vs. Black women (Panel **B**; $n = 274$) of PIN study. 79

5.1 Graphical representation of assumptions on genetic instruments in Mendelian randomization. 88

5.2 Empirical coverage of the 95% confidence interval and empirical power of different methods under **Scenario 1**, where the microbial abundance is generated from a zero-inflated Poisson model. The dashed line on the left figure indicates nominal 95% coverage. 98

5.3 Empirical coverage of the 95% confidence interval and empirical power of different methods under **Scenario 2**, where the microbial abundance is generated from a zero-inflated negative binomial model. The dashed line on the left figure indicates nominal 95% coverage. 99

5.4 Empirical coverage of the 95% confidence interval and empirical power of different methods under **Scenario 3**, where the microbial abundance is generated from a beta-binomial model. The dashed line on the left figure indicates nominal 95% coverage. 100

5.5 Forest plot comparing the causal effect estimate and 95% CI of different IV methods from Mendelian randomization analysis of the HCHS/SOL data. . 106

A.1 Manhattan plots from the first-stage gene-level analysis of the HCHS/SOL data, using the PC-adjusted KRV. Each panel corresponds to a distinct microbiome kernel. The top 5 PCs of genome-wide genetic variability were adjusted. The red lines represent the genome-wide significance threshold ($\alpha = 2.6 \times 10^{-6}$). 142

A.2 Microbiome GWAS results of *MTMR12*, based on the CLR-linear kernel. Panel **A**: Manhattan plot and linkage disequilibrium (LD; R^2) heatmap from the second-stage variant-level analysis of the HCHS/SOL data, using the PC-adjusted KRV. The red line represents variant-level significance ($\alpha = 1.08 \times 10^{-4}$) used in the main analysis. Panel **B**: PC2 vs. PC1 from kernel PCA on the CLR-linear kernel, colored by genotype of the top variant from *MTMR12*. The percent of variance captured by each kernel PC was provided in the axis labels. 143

A.3 Illustration of procedures to identify specific microbial taxa involved in the community-level microbiome GWAS associations. 144

A.4 Manhattan plots from alternative analysis of the HCHS/SOL data, via linear regression of the top PC of the community-level microbiome kernel matrix on the top PC of the gene-level genotype kernel matrix. Each panel corresponds to a distinct microbiome kernel. The top 5 PCs of genome-wide genetic variability were adjusted. The red lines represent the genome-wide significance threshold ($\alpha = 2.6 \times 10^{-6}$). 145

A.5 Manhattan plots from alternative analysis of the HCHS/SOL data, via SKAT test of the top PC of the community-level microbiome kernel matrix on gene-level genetic variation. Each panel corresponds to a distinct microbiome kernel. The top 5 PCs of genome-wide genetic variability were adjusted. The red lines represent the genome-wide significance threshold ($\alpha = 2.6 \times 10^{-6}$). . . 146

A.6 Empirical power of covariate-adjusted KRV and competing methods at nominal level $\alpha = 0.05$ for different microbiome kernels under large effect sizes. Panel **A**: A single SNP affects the abundance of common OTUs. Panel **B**: A single SNP affects the abundance of OTUs from a common phylogenetic cluster. Panel **C**: A single SNP affects the abundance of rare OTUs. In each scenario, linear kernel was used for genetic data. 147

A.7 PC2 vs. PC1 from kernel PCA on the Bray-Curtis microbiome kernel and the *IL23R-C1orf141* genotype kernel, colored by missing status of three covariates: age, gender and study site. Panel **A**: Kernel PCA was conducted on the Bray-Curtis microbiome kernel matrix. Panel **B**: Kernel PCA was conducted on the linear genotype kernel matrix, which was constructed based on common variants in the *IL23R-C1orf141* region. 148

B.1 P-value QQ-plots for $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{orig}}$ under Type I error simulation for three non-normal data scenarios. Simulation parameters are set as: $m = 500$, $d = 3$ and $\rho_c = 0.5$. The top row shows results based on the Gaussian kernel, and the bottom row shows results based on the linear kernel. 175

B.2 P-value QQ-plots for $\mathbf{HSIC_{cl}}$ under Type I error simulation with different cluster sizes. The Gaussian kernel is used. Simulation parameters are set as: $m = 500$, $\rho_c = 0.5$. 176

B.3 P-value QQ-plots for $\mathbf{HSIC_{cl}}$ under Type I error simulation with different cluster sizes. The linear kernel is used. Simulation parameters are set as: $m = 500$, $\rho_c = 0.5$. 177

B.4 Empirical power of $\mathbf{HSIC_{cl}}$ at nominal level $\alpha = 0.05$ under different cluster sizes. Simulation parameters are set as: $m = 500$, $\rho_c = 0.5$ and $\eta = 20\%$. **Power Scenario 1** is considered. 177

B.5 P-value QQ-plots for $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{perm}}$ under Type I error simulation. Simulation parameters are set as: $m = 100$, $d = 3$ and $\rho_c = 0.5$. The top row shows results based on the Gaussian kernel, and the bottom row shows results based on the linear kernel. 178

B.6 Venn diagrams for the number of metabolic pathways identified to be associated with the vaginal microbiome composition based on the MsFLASH data set ($\alpha = 5.3 \times 10^{-4}$). The results are separated by kernel. 182

B.7 Venn diagrams for the number of metabolic pathways identified to be associated with the vaginal microbiome composition based on the MsFLASH data set ($\alpha = 5.3 \times 10^{-4}$). The results are separated by method for HSIC test. . . 182

C.1 Example procedure for generating a genus-level network with species-level data available. 187

C.2 Patterns of heterogeneous relationships among species within different genera. Red and blue edges represent positive and negative conditional correlations, respectively. 189

C.3 ROC curves of different methods in recovering synthetic microbial networks of different topologies, averaged over 10 simulations. Conditional correlations of opposite directions are present among species within 10% of all genera. . . 191

C.4 ROC curves of different methods in recovering synthetic microbial networks of different topologies, averaged over 10 simulations. Conditional correlations of opposite directions are present among species within 30% of all genera. . . 192

C.5 Genus-level vaginal microbial networks based on a random subset of White women (Panel **A**; $n = 274$) vs. based on all Black women (Panel **B**; $n = 274$) of PIN study, with false discovery rate controlled at 0.2. The sizes of the nodes are proportional to their genus-level microbial abundance in the combined sample. 193

C.6 Heatmaps of network centrality measures for each genus in genus-level vaginal microbial networks based on a random subset of White women (Panel **A**; $n = 274$) vs. based on all Black women (Panel **B**; $n = 274$) of PIN study. . . 194

LIST OF TABLES

Table Number — Page

2.1 Significant genes identified from the first-stage (gene-level) analysis of the HCHS/SOL data, using the PC-adjusted KRV ($\alpha = 2.6 \times 10^{-6}$). 23

2.2 Empirical type I error rate of unadjusted and covariate-adjusted KRV at nominal level α under Type I Error Scenario 1. 30

3.1 Empirical type I error rate of $\mathbf{HSIC_{orig}}$ and $\mathbf{HSIC_{cl}}$ at nominal level α for normal data under simulation. 53

3.2 Number of metabolic pathways identified to be associated with the vaginal microbiome composition based on the MsFLASH data set ($\alpha = 5.3 \times 10^{-4}$). . 57

4.1 Average Hamming distance between the constructed network and the true graph for different methods under different topologies, when conditional correlations of opposite directions are present among species within 50% of all genera. 73

4.2 Selected conditional correlations, as measured by SEcov statistics, between common *Lactobacillus* species and pathogenic species. 81

5.1 Causal effects of gut microbial genera on systolic and diastolic blood pressure (SBP and DBP), based on Mendelian randomization analysis of the HCHS/SOL data. 105

A.1 P-values for the significant genes from Table 2.1 when additional covariates were adjusted in the first-stage KRV analysis of the HCHS/SOL data. . . . 141

A.2 Empirical type I error rate of unadjusted and covariate-adjusted KRV at nominal level α under Type I Error Scenario 2. 144

A.3 Specific microbial taxa involved in identified microbiome GWAS associations from the HCHS/SOL data. 149

B.1 Empirical power of $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{perm}}$ at nominal level $\alpha = 0.05$ under simulation ($m = 100$). 179

B.2 Average computation time (in seconds) of $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{perm}}$ (with 1000 permutations) for different number of clusters, with cluster size 3. 180

C.1 Average Hamming distance between the constructed network and the true graph for different methods under different topologies, when conditional correlations of opposite directions are present among species within 10% of all genera. 189
C.2 Average Hamming distance between the constructed network and the true graph for different methods under different topologies, when conditional correlations of opposite directions are present among species within 30% of all genera. 190

ACKNOWLEDGMENTS

Research is an exciting journey towards the unknown territory of science. It has been a challenging and rewarding experience to work through a series of statistical puzzles and be able to contribute useful quantitative tools to better understand the fundamental mechanisms of our body — specifically, in this case, the microorganisms that live as part of our body. I am glad to complete and present this dissertation, which would not be possible without the help and support of many people.

First, I want to thank my dissertation advisor, Michael Wu. Starting with a research project from Spring 2019, I have worked with Mike for more than four years. From detailed guidance on study background and statistical methodology in my first project to encouragement of independent thinking and novel ideas in my later projects, Mike's mentorship has helped me develop both technical and innovative skills as a researcher. Moreover, his enthusiasm and quips during our conversations always made my day better. In addition to guiding me through my research, Mike also offered me valuable opportunities and resources to participate in the greater academic community and navigate my future career path. It has been a great pleasure to work and chat with Mike, and I am grateful for his constant support throughout my PhD journey.

I would like to thank all the collaborators I have worked with on my projects, who generously offered help and guidance on various aspects of my research. Xiang Zhan and Ni Zhao provided help with method development in Project 1. Anna Plantinga helped with software implementation in Project 1 and provided advice on Project 2. Yunhua Xiang helped with method development and offered suggestions on Project 2 and 3. Ting Ye helped with method development and advised on data analysis in Project 4. I also thank all current

and previous members of Mike's group and Ni's group for the stimulating discussion during our group meetings, which serves as my research inspirations. Lastly, I would like to thank my committee members: Jing Ma, Wei Sun, Ting Ye and Sara Lindström for serving on my dissertation committee, asking insightful questions and providing valuable feedback and comments on my dissertation.

I am grateful to everyone who helped me grow as a researcher and biostatistician, and made my graduate experience fuller. I thank my undergraduate research supervisors, Aravinda Chakravarti and Zhirong Bao, for introducing me to the wonderful combination of biology and quantitative sciences. I thank my RA supervisor, Bruce Weir, for his mentorship on population genetics. I thank my cohort at UW Biostatistics for the camaraderie throughout our time in graduate school. I greatly enjoyed our time studying together, our discussion of statistics, and the fun we had outside school.

Finally, I want to thank my mother, Dezhen Chen, a brilliant scientist and mentor, for being my first inspiration to pursue research. I thank my father, Huiyou Liu, for always being passionate about my work and supportive of my decisions. Thank you to all the friends I made at various stages of my life for being there for me, even if we might live in different places and time zones now. I cherish all the laughter we shared.

DEDICATION

To those who strive for a space of their own in science and in life

Chapter 1

INTRODUCTION

The human microbiome is the collection of microorganisms (e.g., fungi, bacteria and viruses) that reside in various sites of the human body. Large-scale microbiome studies have been launched worldwide to understand the role of the microbiome in human health [1, 2, 3]. In these studies, high-throughput sequencing techniques, such as 16S ribosomal RNA (rRNA) sequencing and shotgun metagenomic sequencing, are used to generate detailed information on abundance and phylogeny of individual microbial taxa in the human microbiome. Through sample collection, microbial sequencing and bioinformatic processing of the sequencing data, we can typically obtain a taxon abundance table that records the observed count of each microbial taxon captured from the collected sample for each study participant. Such taxon abundance data allow us to investigate the relationships among microbial features as well as between the microbiome and host health.

A variety of association studies based on microbiome data has emerged in recent years. For example, studying taxon-taxon associations via constructing microbial association networks helps elucidate the global structure of the microbial community [4, 5]. Investigating the relationship between human genetic variation and the microbiome sheds light on the hereditary component of the microbiome composition [6, 7]. Associating the microbiome with various health outcomes has revealed an important role of the microbiome in metabolism, immune response [8, 9] and different diseases such as obesity [10], inflammatory bowel disease [11] and type 2 diabetes [12]. Furthermore, recent advances in Mendelian randomization enable us to examine the causal effect of microbial features on these health outcomes by using genetic variants as instrumental variables [13, 14].

While standard statistical tools exist for association analyses in biological and epidemi-

ological studies, challenges specific to microbiome data analysis call for novel statistical approaches. First, microbiome data, typically consisting of measurements for hundreds of microbial taxa, is inherently high-dimensional. Multivariate methods are a promising way to improve statistical power in microbial association analyses. Second, there are characteristics specific to microbiome data: phylogenetic relationships are inherently present among microbial taxa; microbial abundances derived from sequencing data often have sparsity and overdispersion. These data-specific characteristics can be addressed and incorporated into statistical methods. Finally, we can also accommodate certain practical needs that are common in association analyses, such as covariate adjustment and analysis of clustered data.

In this dissertation, we present four projects on novel statistical methods for association analyses of microbiome data. The proposed methods will be applied to study the genetics-microbiome association (Chapter 2), microbiome-metabolome association (Chapter 3), taxon-taxon association (Chapter 4) and causal effects of microbial taxa on health outcomes (Chapter 5). We now introduce the background on each project in the following sections.

1.1 Genetic Association Analysis for Microbiome Phenotypes

Studying the association between host genetic variation and the huamn microbiome helps elucidate the hereditary component of the microbiome and provides clues as to the biological mechanisms by which genetics may influence health outcomes [7]. Existing work often incorporates microbial features as phenotypes in genome-wide association studies (GWASs), where genetic variants along the genome are tested against the microbial features of interest. Typical analyses marginally test the association between abundances of individual microbial taxa and genotypes of individual genetic variants [15, 16, 17, 18]. Such analyses often suffer from a low statistical power, due to a large multiple-testing burden and failure to accommodate inherent structure in microbiome and genetic data, e.g., phylogenetic relationships among taxa and epistasis among genetic variants.

Since the microbiome functions as a community, we can consider the overall microbiome

composition as an alternative phenotype, which can be characterized using beta-diversity, the dissimilarity in overall microbial profiles between individuals. Beta-diversity analysis focuses on concerted changes in the microbial community rather than changes in individual taxa. While a few studies have considered beta-diversity as an outcome in microbiome GWAS [6, 19, 20], no standard approach exists as to the testing strategy.

In Chapter 2, we propose a novel approach for microbiome GWAS, where we assess the association between groups of variants at the gene level and the overall microbiome composition, characterized by beta-diversity, at the community level. By capturing innate structure within the data and reducing the multiple-testing burden, combining community-level analyses and multi-variant testing has the potential of improving statistical power of microbiome GWAS. Specifically, using the recently developed kernel RV (KRV) framework [21, 22], we evaluate the association between microbes and genetics by comparing similarity in microbial profiles to similarity in genetic profiles across all pairs of individuals via kernel functions. We further extend the original KRV framework to allow for flexible covariate adjustment. The proposed covariate-adjusted KRV test is evaluated in simulation studies and applied to the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) [23, 24] in a genome-wide association analysis for gut microbime beta-diversity. The content in Chapter 2 has been published in *Microbiome* [25].

1.2 Multivariate Independence Testing for Clustered Data

As high-dimensional omics data becomes increasingly available, we are often interested in studying the dependence between two multivariate biological features. In microbiome studies, for example, we may want to study the association between the overall microbiome composition, including hundreds of microbial taxa, and multiple host metabolites from a particular metabolic pathway [26, 27]. Such analyses can help us understand how the human microbiome contributes to the host metabolic environment. This type of multivariate analyses makes use of correlations between variables and aggregates weaker associations into more detectable signals, often resulting in a greater power than univariate analyses.

Meanwhile, correlated observations arise in many practical situations. For example, longitudinal data are common in epidemiological studies [28], where variables of interest are measured on the study subjects repeatedly over time. Such study designs introduce clustered dependence among the observations, where measurements from the same subject tend to be correlated with each other. We denote these types of data as cluster-correlated or clustered data. Standard statistical methods designed for independent observations often become invalid for clustered data and proper accommodation is needed. In this work, we aim to develop an approach for assessing the dependence between two multivariate variables, based on clustered data.

Current methods to deal with multivariate clustered data [29, 30] often extend upon existing tools for longitudinal data, such as generalized estimating equations (GEE) and random effects models. They generally apply to low-dimensional settings and are subject to parametric assumptions. Here we base our approach on the Hilbert-Schmidt Independence Criterion (HSIC) [31], a kernel-based measure for assessing the general dependence between two multivariate variables, where both variables can be high-dimensional. This measure makes no assumption on the distributions of the variables or the nature of dependence. By mapping the two variables into reproducing kernel Hilbert spaces (RKHS's), the HSIC can be viewed as a measure of maximized covariance between functions in the two RKHS's, allowing it to capture potential nonlinear relationships between the variables.

The original HSIC-based independence test [32] applies to independent and identically distributed (i.i.d.) observations. Several extensions [33, 34, 35] have been made to accommodate non-i.i.d. data, but none of the tests directly applies to clustered data at an observation level. In Chapter 3, we present a novel HSIC-based independence test for clustered data. Using the empirical HSIC [32] as our test statistic, we derive its asymptotic distribution under the null hypothesis of independence between the two variables but in the presence of clustered correlation among observations. We also establish the consistency of our test under the alternative hypothesis. We demonstrate the performance of our proposed test in both simulation studies and a real longitudinal microbiome-metabolite data set, where we assess

the association between the vaginal microbiome composition and groups of vaginal metabolites from different metabolic pathways. The content in Chapter 3 has been published in *Advances in Neural Information Processing Systems* 34 [105].

1.3 Multivariate Approach to Microbial Network Construction

Microbial association networks help elucidate the relationships among microbial taxa, shed light on the global structure of the microbial community, and provide clues to the mechanisms by which the microbiome affects host health [4, 5, 36]. In a microbial network, the nodes represent individual microbial taxa, and the edges connecting the nodes represent the association in abundance between a pair of taxa. Statistical association between a pair of taxa could indicate functional interactions between the taxa within the microbial community.

A key step in constructing a microbial network is to evaluate the association between each pair of microbial taxa. To achieve this, one popular way is to assess the conditional dependence between two taxa given all the other taxa in the community [37, 38]. This approach tends to capture direct interactions between taxa rather than indirect, spurious associations [39]. In our work, we also base our proposed approach for microbial network construction on a measure of conditional dependence.

While a microbial network can be constructed at different taxonomic levels, it is preferred to construct the network at a higher level (such as family or genus level) in order to reduce the dimension of microbial features and account for sparsity of microbial data [40, 41]. In these cases, conditional dependence is typically assessed based on the abundance data aggregated at the desired taxonomic level [38, 42]. However, such an aggregated approach could potentially cause power loss in edge detection when there are heterogeneous relationships present among the sub-taxa within the taxa of interest, as this approach implicitly assumes that the associations between the sub-taxa have the same directions.

In Chapter 4, we propose a multivariate approach to construct microbial association networks, which is based on a novel conditional independence test for multivariate variables, denoted as conditional RV (CRV). Suppose that we are interested in constructing a network

at the genus level while having species-level data available. Instead of assessing the conditional dependence between genus-level aggregated abundances, we propose to assess the conditional dependence between pairs of multivariate microbial features, where each feature represents the abundances of multiple species that belong to a particular genus. The proposed approach is able to preserve and aggregate species-level signals and can be especially helpful when associations of opposite directions are present among species within the pair of genera being studied. We evaluate the performance of CRV in simulation studies and apply CRV to construct vaginal microbial networks in pregnant women based on the Pregnancy, Infection, and Nutrition (PIN) Study [43].

1.4 Mendelian Randomization with a Microbial Exposure

Microbial association studies have revealed important associations between the human microbiome and various health outcomes such as type 2 diabetes [12] and inflammatory bowel disease [11]. However, existing work is largely based on observational studies, where the presence of unmeasured confounders prevents us from directly establishing causal relationships between microbial features and these outcomes. With the popularity of GWAS studies, where both the human microbiome and various health outcomes have been used as phenotypes of interest, Mendelian randomization (MR) emerges as a feasible framework to evaluate the causal effect of a microbial exposure on an outcome based on observational data, by using genetic variants as instruments [44]. Similar to the randomization step in a randomized controlled trial, the random segregation of genetic materials during gamete formation allows genetic variants to be free of confounding in the exposure-outcome relationship and serve as potential proxies (i.e., instrumental variables) for the microbial exposure.

Recently, an increasing number of MR studies has focused on the gut microbiome as an exposure and investigated the causal effect of gut microbial features on health outcomes such as metabolic traits and complex diseases [45, 13, 14]. These microbial features are typically the abundances of individual microbial taxa. As we mentioned above, microbial abundances obtained from microbial sequencing techniques are count data with unique characteristics,

often with overdispersion and zero-inflation. However, existing MR methods are usually based on a continuous exposure, without accommodation for the count nature of the microbial data or the potential nonlinear relationships between the microbial abundance and the genetic instruments. For example, two-stage least squares (2SLS) is a standard method in one-sample MR analysis [46, 47], where a linear relationship is assumed between the microbial abundance and the genetic instruments. While 2SLS can still provide valid inference when there is misspecification in the genetics-exposure model, methods that account for characteristics specific to microbial data have the potential of achieving a better efficiency.

In Chapter 5, we propose a novel approach to conduct MR analysis with a microbial exposure and a continuous outcome in the one-sample setting. We adapt an existing instrumental variable (IV) method from the econometrics literature, two-stage least squares with generated instruments (2SLS-GI) [46], to incorporate nonlinear models that account for characteristics of microbial count data and nonlinear relationships between the microbial abundances and genetic IVs. We demonstrate the power gain of 2SLS-GI in detecting causal effects compared to existing IV methods via simulation studies and apply 2SLS-GI to the HCHS/SOL study [23, 24] to identify causal effects of gut microbial taxa on systolic and diastolic blood pressure.

Chapter 2

KERNEL-BASED GENETIC ASSOCIATION ANALYSIS FOR MICROBIOME PHENOTYPES

2.1 Introduction

The human microbiome plays an important role in host health and is involved in fundamental body functions such as metabolism and immune response [8, 9]. While environmental factors have a large influence on microbiome composition [48], it is still of interest to study the effect of human genetic variation on the microbiome: such studies not only help us understand the hereditary component of the human microbiome, but also provide clues as to the biological mechanisms by which genetics may influence health outcomes. As a notable example, elevated abundance of *Bifidobacterium*, a genus of beneficial gut bacteria that utilizes lactose as an energy source, has been associated with a non-persistence genotype of the human lactase gene (LCT), which typically results in lactose intolerance [7, 15, 6]. Such an association implies a potential mediating role of the gut microbiome in the relationship between host genetics and metabolic outcomes, where the presence of Bifidobacteria may provide some level of lactose tolerance to lactase non-persistent individuals [7].

Many studies have sought to identify genetic variants that influence microbial composition, and most of them incorporate microbiome characteristics as phenotypes in genome-wide association studies (GWASs). Typical analyses marginally test the association between abundances of individual taxa and genotypes of individual genetic variants [15, 16, 17, 18]. Such analyses often suffer from a low statistical power, due to a large multiple-testing burden and failure to accommodate inherent structure in microbiome and genetic data, e.g., phylogenetic relationships among taxa and epistasis among genetic variants.

As the microbiome functions as a community, an alternative microbiome phenotype is

beta-diversity, the dissimilarity in overall microbiome profiles between individuals. Beta-diversity analysis represents a standard mode of analysis in microbiome profiling studies as it focuses on discovery of concerted shifts in the community rather than individual taxa. However, few studies have considered beta-diversity as a trait of interest in microbiome GWAS and there is no standard strategy. Some studies [6, 19] have performed principal coordinates analysis (PCoA) on the pairwise beta-diversity matrix and evaluated the association between the top principal coordinates (PCo's) and the genotype of each genetic variant. Such a strategy could suffer from power loss, as the top PCo's may not fully capture the variation within the microbiome data. Hua et al. [49] assumed a linear model between the pairwise beta-diversity and the pairwise genetic distance at each genetic variant and developed a score test called microbiomeGWAS. Rühlemann et al. [20] adopted a distance-based multivariate analysis of variance (MANOVA) approach called distance-based F test [50] and evaluated the difference in beta-diversity among the different genotype groups for each genetic variant. These approaches still test one variant at a time and are subject to a stringent genome-wide significance threshold. Studies using the above approaches have identified loci within genes involved in immunity [6, 20], vitamin metabolism [19] and complex diseases such as type 2 diabetes [51]. In our study, we aim to further improve statistical power with a novel approach and bring more discoveries from microbiome GWAS.

Here we propose to assess the association between groups of variants at the gene level and the overall microbiome composition, characterized by beta-diversity, at the community level. Community-level analyses and multi-variant testing have been shown to be powerful in microbiome [52, 53] and genetic studies [54], respectively, due to their ability to capture innate structure and correlation within the data, while reducing the multiple-testing burden. Using the recently developed kernel RV (KRV) framework [21, 22], we summarize individuals' microbiome (or genetic) characteristics by a pairwise similarity matrix called "kernel" matrix, where each entry in the matrix represents similarity in microbiome (or genetic) profiles between a pair of individuals. Microbiome similarity can be obtained by transforming known beta-diversity measures, while genetic similarity can also be characterized in various

ways, such as the average genotype matching over all genetic variants. The association between microbes and genetics is then assessed via comparing similarity in microbiome profiles to similarity in genetic profiles across all pairs of individuals. Intuitively, if the genetics is associated with the microbiome, we would expect the pairwise microbial profiles to be similar whenever the pairwise genetic profiles are similar. In particular, the test statistic is the normalized Frobenius inner product, a measure of correlation, between the two kernel matrices.

Although the KRV is a potentially powerful approach for microbiome GWAS, the KRV framework lacks a general strategy to control for covariates such as population structure, which is imperative for any genetic association analysis. Here we extend the original KRV framework to allow for flexible covariate adjustment.

We apply the covariate-adjusted KRV to the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) [23, 24] via a two-stage (first gene-level, then variant-level) genome-wide association analysis for gut microbiome. This is the first study to investigate the genetic effect on the overall gut microbiome composition, characterized by beta-diversity, in Hispanic/Latino populations. We have identified a gene (*IL23R*) reported in a previous microbiome genetic association study and discovered other novel genes related to immune functions. Furthermore, we have identified individual genetic variants and specific microbial taxa involved in these gene-microbiome associations. In addition, our simulation results show that the covariate-adjusted KRV maintains valid type I error rates in the presence of confounding and has a much greater power than other single-trait-based competing methods across a range of scenarios. Together, our proposed approach demonstrates good statistical properties and provides a powerful way to study the effect of human genetic variation on microbiome composition.

2.2 Methods

2.2.1 Overview of covariate-adjusted KRV

We aim to assess the covariate-adjusted association between genotypes of multiple genetic variants within a gene and abundances of microbial taxa at the community level, using the previously developed KRV framework. We now give an overview of the original KRV framework and extend it to allow for flexible covariate adjustment. The overall procedure for covariate-adjusted KRV in the context of microbiome GWAS is shown in Figure 2.1.

The KRV framework has been proposed by Zhan et al. [21, 22] to evaluate the general association between a group of genetic variants, G, and a group of traits, Y. Suppose we have genotype data of m genetic variants and phenotype data of q traits available for n unrelated individuals. For the ith subject, let $\boldsymbol{g}_i = (g_{i1}, \cdots, g_{im})^T$ be the set of genotypes, where $g_{il} \in \{0, 1, 2\}$ represents the number of minor alleles for the lth variant; let $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iq})^T$ be the set of traits. Example phenotypes in previous studies include expression values of multiple genes from a particular pathway [21] and levels of multiple amino acids [55]. In the context of microbiome GWAS, we treat the microbiome as the phenotype. Specifically, $\boldsymbol{g}_i$ represents the genotypes of m genetic variants within a particular gene, and $\boldsymbol{y}_i$ represents the abundances of q microbial taxa that form the microbiota.

Let $k(\boldsymbol{g}_i, \boldsymbol{g}_j)$ be a kernel function that measures the similarity in genetic profiles between individuals i and j. Let $\ell(\boldsymbol{y}_i, \boldsymbol{y}_j)$ be another kernel function that measures the similarity in phenotypic profiles between i and j. Specific choices of kernel functions in the context of microbiome GWAS are discussed in Section 2.2.2. We can then define a kernel matrix $\boldsymbol{K} \in \mathbb{R}^{n \times n}$, where the (i, j)-th entry of $\boldsymbol{K}$ is $k(\boldsymbol{g}_i, \boldsymbol{g}_j)$. Similarly, we define another kernel matrix $\boldsymbol{L} \in \mathbb{R}^{n \times n}$ such that $\boldsymbol{L}_{ij} := \ell(\boldsymbol{y}_i, \boldsymbol{y}_j)$. The matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ can be viewed as pairwise similarity matrices for genotypes and phenotypes, respectively. We further center the two kernel matrices: let $\tilde{\boldsymbol{K}} := \boldsymbol{HKH}$ and $\tilde{\boldsymbol{L}} := \boldsymbol{HLH}$, where $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n$ is a column-centering matrix. Then the KRV coefficient that evaluates the relationship between

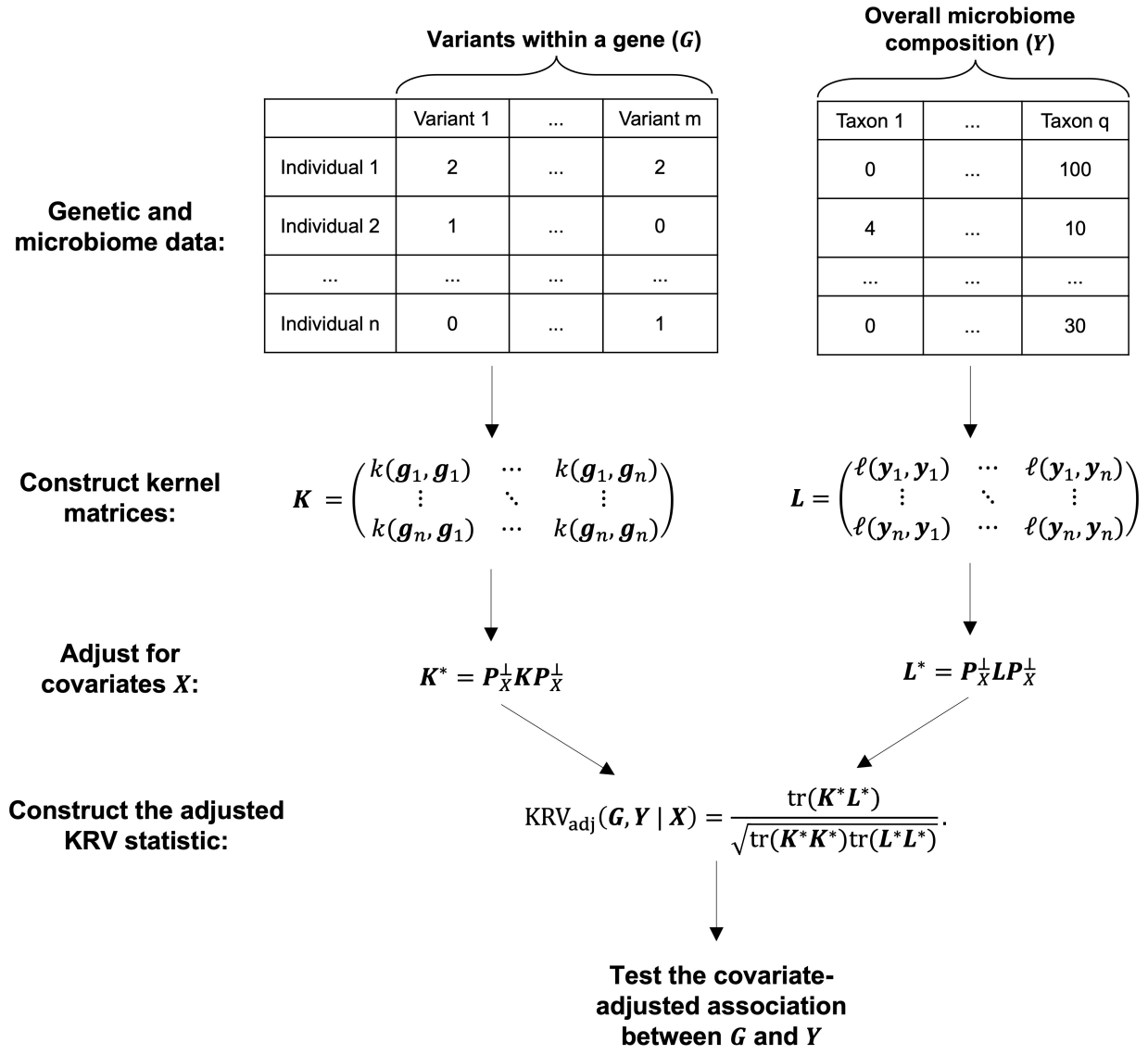

Figure 2.1: Illustration of covariate-adjusted KRV for microbiome genome-wide association studies.

the genetic variants and the traits is defined as

$$\mathrm{KRV}(G, Y) := \frac{\mathrm{tr}(\tilde{\boldsymbol{K}} \tilde{\boldsymbol{L}})}{\sqrt{\mathrm{tr}(\tilde{\boldsymbol{K}} \tilde{\boldsymbol{K}})\, \mathrm{tr}(\tilde{\boldsymbol{L}} \tilde{\boldsymbol{L}})}}. \tag{2.1}$$

Intuitively, the KRV coefficient compares genotypic similarity to phenotypic similarity across all pairs of individuals. A large KRV coefficient indicates that the pairwise similarity pattern in genetic profiles well resembles the pairwise similarity pattern in phenotypic profiles, which implies that the genetic variants are associated with the traits in a certain way. To perform hypothesis testing, the permutation distribution of the KRV statistic under the null hypothesis of no association between genetics and phenotypes can be approximated by a Pearson Type III distribution [21], allowing us to obtain a p-value and assess the significance of the association at a given significance level.

The above framework does not take into account any covariates that might be involved in a typical genetic association study. Now suppose that, for each individual i, we have a set of covariates $\boldsymbol{x}_i = (1, x_{i1}, \cdots, x_{ip})^T \in \mathbb{R}^{p+1}$; let $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$ be the sample covariates matrix such that the i-th row of $\boldsymbol{X}$ is $\boldsymbol{x}_i^T$. Assume that $\boldsymbol{X}$ has full rank. We intend to assess the association between the genetic variants and the phentoypes, after adjusting for the effects of covariates $\boldsymbol{X}$. Previous studies, including the original KRV framework, have suggested using a residual-based approach [54, 56, 21], where we first regress out the covariates from each raw phenotype and then construct the phenotype kernel matrix using the resulting residuals. Such an approach is not universally feasible for all microbiome kernels, as certain popular microbiome kernels (e.g., the Bray-Curtis kernel and the unweighted UniFrac kernel) require the input to be discrete taxa count data or taxa presence/absence data, which is not satisfied by the covariate-adjusted residuals. Furthermore, adjustment based on linear regression may not account for the potentially nonlinear relationships between the genetics/microbiome and the covariates.

To adjust for covariates in a general way, we propose a novel adjustment approach that applies to all possible kernel types, regardless of the requirement for input data. Our approach is based on kernel principal component analysis (kernel PCA) [57], a general and nonlinear extension of regular PCA, of the kernel matrices. Specifically, we first perform a kernel PCA on the constructed phenotype kernel matrix and treat the resulting kernel PCs as surrogate phenotypes, which could capture both linear and nonlinear features of the original

phenotype data depending on the kernel function used. We then regress out the covariates from all kernel PCs and reconstruct the phenotype kernel matrix with the adjusted PCs. By adjusting the covariates on all kernel PCs, we are able to fully account for the variation within the phenotype data. The same procedure is performed on the genotype kernel matrix. After algebraic manipulation (see Section A.1), the adjusted KRV coefficient is of the form:

$$\mathrm{KRV}_{adj}(G, Y|X) := \frac{\mathrm{tr}(\boldsymbol{K}^*\boldsymbol{L}^*)}{\sqrt{\mathrm{tr}(\boldsymbol{K}^*\boldsymbol{K}^*)\,\mathrm{tr}(\boldsymbol{L}^*\boldsymbol{L}^*)}},$$

where $\boldsymbol{K}^* := \boldsymbol{P}_X^{\perp}\boldsymbol{K}\boldsymbol{P}_X^{\perp}$, $\boldsymbol{L}^* := \boldsymbol{P}_X^{\perp}\boldsymbol{L}\boldsymbol{P}_X^{\perp}$, $\boldsymbol{P}_X^{\perp} := \boldsymbol{I} - \boldsymbol{P}_X$ and $\boldsymbol{P}_X$ is the projection matrix onto the column space of $\boldsymbol{X}$. We adjust for covariates on both the phenotype kernel and the genotype kernel, due to the symmetry of the KRV coefficient. Our proposed approach for covariate adjustment is able to capture both linear and nonlinear relationships between the genetics/microbiome and the covariates, and thus can be viewed as a general extension of the previous residual-based approach. When a linear kernel is used, our strategy is exactly equivalent to the residual-based approach (see Section A.1.1).

The usual hypothesis testing procedure in the KRV framework can be applied to the adjusted KRV statistic to obtain a p-value. In this case, the null hypothesis is that there is no association between the genetics and the phenotypes after adjusting for the effects of the covariates.

2.2.2 Choice of kernels

In the KRV framework, kernel functions are used to summarize pairwise similarities in genotype and phenotype profiles among the subjects. In order to improve the statistical power in hypothesis testing, we would like to choose kernels that better reflect the actual structure within the genetic and phenotype data as well as the patterns of association [52, 58]. For the KRV statistic in (2.1) to be well-defined theoretically, the kernel matrices need to be positive semi-definite. We now review some of the common kernels used for genetic and microbiome data.

For genotype data, popular kernel functions include the linear kernel $k(\boldsymbol{g}_i, \boldsymbol{g}_j) = \boldsymbol{g}_i^T \boldsymbol{g}_j$ and the identity-by-state (IBS) kernel $k(\boldsymbol{g}_i, \boldsymbol{g}_j) = \frac{1}{2m} \sum_{l=1}^{m} (2 - |g_{il} - g_{jl}|)$. The linear kernel assumes that the genetic variants are associated with the traits in a linear fashion. The IBS kernel defines pairwise similarity as the pairwise genotype matching averaged over all genetic variants, and is useful when there are epistatic effects among the variants [54]. Depending on analysis interests (e.g. rare-variant analysis), it is also possible to incorporate a weight for each variant in the linear and IBS kernels [54].

For microbiome data at the community level, the kernel matrix can be obtained by transforming known ecological or phylogenetic dissimilarity measures (i.e., beta-diversity measures). For example, Bray-Curtis dissimilarity quantifies the dissimilarity between two microbial communities based on the difference in counts at each taxon between the two communities. The UniFrac distances are dissimilarity measures based on the phylogenetic structure of the taxa [59, 60, 61]: the unweighted UniFrac distance is calculated as the fraction of branch lengths within the phylogenetic tree that are not shared between the two communities; the weighted UniFrac distance further incorporates taxa abundance information on the basis of the unweighted distance; the generalized UniFrac distance is a compromise between weighted and unweighted UniFrac distances.

While the Bray-Curtis dissimilarity and UniFrac distances take scaled or rarefied microbial counts or presence/absence information as input, microbial dissimilarity can also be calculated from other types of transformed abundance data. For example, the centered log-ratio (CLR) transformation [62, 63] and phylogenetic isometric log-ratio (PhILR) transformation [64] have been proposed to address the compositional nature of microbiome data, where PhILR further incorporates phylogenetic information into the transformed data. As these log-ratio-based transformations encourage normality, Euclidean distances can then be calculated based on the CLR-transformed or PhILR-transformed data as measures of dissimilarity.

Given a pairwise dissimilarity matrix $\boldsymbol{D}$, the corresponding kernel matrix can be con-

structed as:

$$\boldsymbol{L} = -\frac{1}{2}\Big(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\Big)\boldsymbol{D}^2\Big(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\Big),$$

where $\boldsymbol{D}^2$ is the element-wise square of $\boldsymbol{D}$. To ensure that the kernel matrix $\boldsymbol{L}$ is positive semi-definite, we further apply a correction procedure as implemented in the MiRKAT R package [52], where we perform an eigendecomposition of $\boldsymbol{L}$, convert any negative eigenvalues to zero and then reconstruct the kernel matrix.

We note that taking Euclidean distances followed by kernel matrix transformation is equivalent to constructing a linear kernel matrix based on the same data (see Section A.1.2). Therefore, the kernels derived from Euclidean distances of CLR- and PhILR-transformed data can be viewed as linear kernels directly applied to these transformed data. We denote the resulting kernel matrices as CLR-linear and PhILR-linear kernels, respectively.

2.2.3 Description of the HCHS/SOL study

Hispanic Community Health Study/Study of Latinos (HCHS/SOL) is a community-based prospective cohort study aimed to identify risk factors for health outcomes in Hispanic/Latino populations in the United States. The study recruited 16,415 Hispanic/Latino adults aged 18 - 74 years, representing diverse ethnic background, at four U.S. field centers (Bronx, NY, Chicago, IL, Miami, FL, and San Diego, CA), using a two-stage probability sampling design [23].

12,803 participants consented to genetic studies. Genotyping was performed on an Illumina custom array, SOL HCHS Custom 15041502 B3, which consisted of the Illumina Omni 2.5M array (HumanOmni2.5-8v1-1) and ∼150,000 custom SNPs [65]. Quality control, genotype imputation and estimation of pairwise kinship coefficients and PCs of genome-wide genetic variability were described in detail by Conomos et al. [65]. In addition to the quality control procedures described in [65], prior to the microbiome GWAS analysis, we also filtered imputed genetic variants based on an "effective minor allele count": $N_{\text{eff}} = 2\hat{p}(1-\hat{p})Nv$, where $\hat{p}$ is the estimated minor allele frequency, N is the sample size and v is the ratio of

observed variance of imputed dosages to the expected binomial variance [66]. We retained variants with sufficient minor allele counts and excluded any variants with $N_{\text{eff}} < 30$.

As an ancillary study, the HCHS/SOL Gut Origins of Latino Diabetes (GOLD) study was further conducted to investigate the role of gut microbiome composition in diabetes and other health outcomes in Hispanic/Latino individuals [24]. Gut microbiome profiles were available in 1674 participants, a subset of the HCHS/SOL participants. Based on the collected stool samples, DNA extraction and 16S rRNA gene sequencing were performed according to the Earth Microbiome Project (EMP) standard protocols [67]. Subsequent bioinformatic processing of the microbiome sequencing data was described in detail by Kaplan et al. [24].

The HCHS/SOL study was approved by the Institutional Review Boards of all participating institutions, and written informed consent was obtained from all participants.

2.2.4 Microbiome GWAS analysis of HCHS/SOL data

To identify genetic variants associated with the overall gut microbiome composition in Hispanic/Latino individuals, we applied the covariate-adjusted KRV test to the HCHS/SOL study in a genome-wide association analysis for gut microbiome beta-diversity.

We considered genetic variants (including both single-nucleotide polymorphisms, or SNPs, and insertion/deletion variants, or indels) within ±10 kb of gene regions along Chromosomes 1-22 and grouped the variants into gene-level variant-sets correspondingly. The microbiome operational taxonomic units (OTUs) were collapsed at the genus level. We used a linear kernel for the genetic data and six different kernels for the microbiome data, including Bray-Curtis, unweighted UniFrac, weighted UniFrac, generalized UniFrac, CLR-linear and PhILR-linear, as described in Section 2.2.2. Rarefied microbial abundance data were used to construct Bray-Curtis and UniFrac kernels, while absolute abundance data were used to construct CLR-linear and PhILR-linear kernels, where a unit pseudo-count was added to address zero entries before CLR and PhILR transformations. The weightings used in PhILR transformation were the same as those proposed in [64].

For each gene, we assessed the association between common variants (with minor allele

frequency, or MAF, ≥ 0.05) within the gene and the community-level microbiome profile, using both adjusted and unadjusted KRV tests. In the adjusted KRV, we mainly controlled for the top 5 PCs of genome-wide genetic variability (denoted as the PC-adjusted KRV), as they were shown to well capture the population structure of the sample based on a previous genetic study of HCHS/SOL data [65]. Individuals from different populations and ethnic groups often have systematic differences in their genetic and microbiome profiles [68, 69], so population structure is an important confounder in our analysis. We also performed additional analyses that adjusted for other non-confounding covariates including age, gender and study sites.

To avoid confusion, we emphasize the distinction between (1) kernel PCs derived from the kernel matrices, as mentioned in Section 2.2.1 and (2) genome-wide genetic PCs. In the context of our gene-level microbiome GWAS, the kernel PCs of the genotype kernel matrix capture information of a particular gene that we are interested in testing against the microbiome. On the other hand, the genome-wide genetic PCs capture genetic information along the entire genome and are used as covariates to measure population structure. In the PC-adjusted KRV analysis, the top 5 genome-wide genetic PCs were regressed out from all kernel PCs of the gene-level genotype kernel matrix and all kernel PCs of the community-level microbiome kernel matrix.

Our investigation of the genetic effect on the microbiome involved two stages. In the first stage, we tested the association between the variants in each gene and the microbiome profile at the community level. In the second stage, for any genes called significant in the first stage, we marginally assessed the association between each of the individual variants within those genes and the community-level microbiome profile to look for significant variants, again using the covariate-adjusted KRV. Bonferroni correction was applied in both stages. Since this was a nested hypothesis testing approach, the second-stage test only required correction for the number of variants in the genes that were called significant in the first stage. All analyses were performed on unrelated individuals (pairwise kinship coefficient ≤ 0.05) where genetic data, microbiome data and covariates data were available.

As a comparison to our proposed covariate-adjusted KRV approach, we applied additional microbiome GWAS approaches to the same sample. First, we consider two methods that still analyze the association between gene-level genetic variation and community-level microbiome composition but use univariate approaches. One method was linear regression, where we performed kernel PCA on both the gene-level genotype kernel matrix and the community-level microbiome kernel matrix and regressed the top kernel PC of the microbiome kernel on the top kernel PC of the genotype kernel, while adjusting for covariates. The other method was SNP-set kernel association test (SKAT) [54], a kernel machine regression framework for assessing the general association between a univariate trait and multiple genetic variants. Here we performed kernel PCA on the community-level microbiome kernel matrix and used the SKAT test to regress the top kernel PC of the microbiome kernel on the genetic variants within each gene, while adjusting for covariates; a linear kernel was used for genetic data in the SKAT test. In addition to gene-based community-level competing methods, we also conducted a traditional variant-based taxon-level microbiome GWAS, where we tested the association between individual genetic variants along the genome and individual microbial genera present in $\geq 10\%$ of all participants. A detailed analysis procedure for the taxon-level analysis is described in Section A.2. In all the competing methods, the top 5 PCs of genome-wide genetic variability were adjusted as covariates.

2.2.5 Simulation studies

We conducted simulation studies to further evaluate the type I error rate and power of the covariate-adjusted KRV test. We simulated genotype data and microbial OTU count data under realistic settings, and introduced population stratification as a confounder that affected both genetic and microbiome data.

The general simulation setting is as following. We considered a sample size of 1000. SNP genotype data over a 1 Mb chromosome were simulated for 500 individuals of African ancestry and 500 individuals of European ancestry. Specifically, we first generated 10,000 haplotypes of African ancestry and another 10,000 haplotypes of European ancestry over a

1 Mb chromosome according to coalescent theory using the *cosi2* program [70]. To form a sample, we then generated the genotype of each African individual in the sample by randomly selecting and pairing 2 haplotypes from the 10,000 founding African haplotypes. A similar procedure was used to generate the genotypes of European individuals.

We used a Dirichlet-multinomial distribution to generate microbial OTU counts for each individual in the sample, as this distribution well accommodates the over-dispersion of microbiome count data [52, 71]. To ensure a realistic simulation of OTU counts, we estimated the parameters of the Dirichlet-multinomial distribution from a real upper-respiratory-tract microbiome data set [72], which consisted of 856 OTUs. This data set is publicly available as part of the GUniFrac R package. We assumed 1000 total OTU counts per individual. Population structure was introduced into the OTU count data in two ways, as described below.

Both unadjusted and adjusted KRV tests were performed to test the association between the overall microbiome composition (composed of 856 OTUs) and common SNPs (with MAF ≥ 0.05) within an 8 kb subregion of the 1 Mb chromosome. This 8 kb subregion can be considered as a simulated gene region. In the adjusted KRV test, the top PC of genetic variability (obtained from PCA on SNP data over the entire 1 Mb region) was used as the covariate, a surrogate for population structure. We considered a linear kernel for genetic data and six different kernels for microbiome data: Bray-Curtis, unweighted UniFrac, weighted UniFrac, generalized UniFrac, CLR-linear and PhILR-linear.

To evaluate type I error rates in the presence of confounding, we introduced population structure into the OTU count data in two scenarios (denoted as Type I Error Scenario 1 and 2). In Type I Error Scenario 1, we increased the abundance of the 10 most common OTUs by 10% in African individuals and then rarefied the abundance back to 1000 total counts per individual. In Type I Error Scenario 2, we increased the abundance of 10 rare OTUs (chosen randomly from the top 40 rarest OTUs) in African individuals by adding a unit count before rarefying the abundance back to 1000 total counts per individual. These two scenarios were not meant to reflect the microbiome difference between African and European individuals in

reality, but they served as hypothetical situations to introduce confounding effect into the genetics-microbiome relationship. Here we used the estimated mean proportion parameters of the Dirichlet-multinomial distribution as a measure of OTU prevalence. 10,000 simulations were performed for each type I error scenario.

To evaluate the power of the covariate-adjusted KRV, we based our simulation setting on Type I Error Scenario 1 and further introduced genetic effect on the microbiome in three different power scenarios, where a single SNP affected the abundance of multiple microbial OTUs (i.e., a pleiotropy effect). Let g_i be the genotype (0, 1 or 2) of individual i at a chosen common SNP. In Power Scenario 1, for each individual i, we increased the counts of the 11th - 20th most common OTUs by a factor of f_i, where $f_i = 1 + c_1 g_i$. In Power Scenario 2, utilizing the available phylogenetic tree for the 856 OTUs [72], we increased the counts of OTUs from a relatively abundant cluster (representing 10.3% abundance of the total OTU counts) by a factor of f_i for each individual i, where $f_i = 1 + c_2 g_i$. In Power Scenario 3, for each individual i, we increased the counts of 5 rare OTUs (chosen randomly from the top 40 rarest OTUs) by an addition of a_i, where $a_i = c_3 g_i$. We considered two sets of effect sizes: (a) small effect sizes: $c_1 = c_2 = 0.3, c_3 = 0.5$ and (b) large effect sizes: $c_1 = 0.8, c_2 = 0.7, c_3 = 1$. After introducing these genetic effects on the microbiome, we again rarefied the OTU counts to 1000 total counts per individual. For each power scenario, 1000 simulations were performed.

In the power simulation, we also considered two competing methods that analyze the association between a group of variants and the overall microbiome composition but rely on univariate microbiome phenotypes, as described in Section 2.2.4. The first method was linear regression, where we regressed the top kernel PC of the community-level microbiome kernel matrix on the top kernel PC of the gene-level genotype kernel matrix, while adjusting for covariates. The second method was SKAT, where we applied the SKAT test to regress the top kernel PC of the microbiome kernel on the genetic variants within the pre-specified gene region, while adjusting for covariates; we used a linear kernel for genetic data in the SKAT test.

2.2.6 Computation time

We estimated the computation time of the covariate-adjusted KRV test for different sample sizes. For each sample size, we simulated 10 data sets and reported the average computation time. Given constructed genotype and microbiome kernel matrices and 10 covariates, the average computation times are 0.09, 1.23, 12.58 and 97.57 seconds on a laptop (2.7 GHz CPU and 16 GB memory) for sample sizes of 200, 500, 1000 and 2000, respectively. The gene-level analysis of the HCHS/SOL data set (with one genotype kernel, 6 microbiome kernels and 19223 variant-sets) took approximately 8 hours on a high-performance computing cluster (each node with 24 cores, 3.00 GHz CPU and 384 GB memory), with computing jobs divided by chromosome.

2.3 Results

2.3.1 Application of covariate-adjusted KRV to HCHS/SOL

We performed our microbiome GWAS analyses on 1219 unrelated participants from HCHS/SOL where all relevant data were available. Among these individuals, 47.0% identified their background as Mexican, 14.8% as Cuban, 12.7% as Puerto Rican, 10.3% as Central American, 7.7% as South American and 7.5% as Dominican. Microbiome count data were obtained on 408 genera, rarefied to 10,000 total counts per individual to construct Bray-Curtis and UniFrac kernels. A total of 19223 gene-level variant-sets that contained at least one common variant were available. Figure 2.2 shows the p-value QQ-plots of the first-stage gene-level analysis results. For all microbiome kernels, the unadjusted KRV produces highly anti-conservative p-values (with large genomic inflation factors), while the PC-adjusted KRV has well-controlled type I error rates (with genomic inflation factors ≤ 1.05), confirming that population structure is the major confounder in our study. The gene-level Manhattan plots based on the PC-adjusted KRV are shown in Figure A.1.

Table 2.1 shows the genes identified at a genome-wide significance in the PC-adjusted first-stage analysis ($\alpha = 0.05/19223 = 2.6 \times 10^{-6}$). We have found two genes, *IL23R* and

C1orf141, using the Bray-Curtis kernel and two genes, *MTMR12* and *ZFR*, using the unweighted UniFrac kernel. *MTMR12* is also identified by the CLR-linear kernel. When the analysis is performed on a reduced set of individuals (n=1096) where additional covariates (age, gender and study sites) are available and adjusted, *IL23R* and *C1orf141* are no longer genome-widely significant (Table S1). Similar non-significant results are observed for *IL23R* and *C1orf141* when only genome-wide genetic PCs are adjusted in the same subsample. To investigate the reason for this power loss, we perform PC-adjusted analyses on random subsamples of the same size from the original 1219 individuals. Around half of the times, at least two out of the four genes no longer have genome-wide significance, indicating that the non-significant results in the reduced sample are likely due to sample size loss, rather than systematic differences between the reduced sample and the original sample. Nevertheless, the results from the two adjusted analyses are similar in both their observed KRV statistics (*IL23R*: 0.017 in the original sample vs. 0.016 in the reduced sample; *C1orf141*: 0.018 in the original sample vs. 0.016 in the reduced sample) and the order of magnitude of their p-values (10^{-6} in the original sample vs. 10^{-5} in the reduced sample). Additional analyses to assess the robustness of these two signals are reported in Section A.3.

Table 2.1: Significant genes identified from the first-stage (gene-level) analysis of the HCHS/SOL data, using the PC-adjusted KRV ($\alpha = 2.6 \times 10^{-6}$).

Microbiome kernel	Significant genes	Number of common variants	P-value
Bray-Curtis	*C1orf141*	484	1.1×10^{-6}
	IL23R	284	2.4×10^{-6}
Unweighted UniFrac	*MTMR12*	174	6.5×10^{-8}
	ZFR	288	2.5×10^{-9}
CLR-linear	*MTMR12*	174	1.7×10^{-6}

The top 5 PCs of genome-wide genetic variability were adjusted.

Among these genes, *IL23R* is of considerable interest: it encodes one part of the receptor for interleukin-23 (IL-23), a pro-inflammatory cytokine closely involved in autoimmunity

[73]. The *IL23R* gene has been associated with inflammatory bowel diseases (IBD) including Crohn's disease and ulcerative colitis [74, 75]. In a previous genetic association study of microbiome composition [76], the protective variant of the *IL23R* gene (rs11209026) was associated with a higher microbiome diversity and richness and a higher abundance of beneficial gut bacteria in the ileum of healthy individuals, suggesting the influence of host genetics on the microbiome prior to onset of IBD. In addition, a mouse-based experimental study [77] showed that mice deficient in intestinal *IL23R* expression had altered gut microbiota and were susceptible to colonic inflammation, where increased disturbance of gut microbiota exacerbated the disease activity. Coupled with these results, our finding further supports that the gut microbiome may mediate the host genetic effect on the development of inflammatory diseases like IBD. In its normal function, the *IL23R* gene likely helps shape the overall gut microbiota towards a healthy composition, which may in turn support normal immune activities and prevent gut inflammation.

The other genes are also interesting to further explore. The *C1orf141* gene, with uncharacterized protein function, has overlapping regions with *IL23R*. Variants in the *IL23R-C1orf141* region have been associated with susceptibility to Vogt-Koyanagi-Harada disease, a multi-system autoimmune disorder that affects pigmented tissues, in Chinese and Japanese populations [78, 79]. The *ZFR* gene encodes the highly conserved zinc finger RNA-binding protein, which is shown to prevent excessive type I interferon activation by regulating alternative pre-mRNA splicing [80]. Prevention of excessive type I interferon activation is important for the regulation of immune responses. The *MTMR12* gene encodes an adapter protein for myotubularin-related phosphatases and is likely involved in skeletal muscle functions [81]. Overall, most of the significant genes have a role in immunity, indicating an interaction between the host genetics and the gut microbiome in facilitating immune responses or developing autoimmune disorders.

As *MTMR12* is more significant with the unweighted UniFrac kernel than with the CLR-linear kernel, we focus on unweighted UniFrac for our subsequent analysis of *MTMR12*. Figure 2.3 shows the Manhattan plots and linkage disequilibrium (LD) heatmaps from the

second-stage variant-level analysis of the HCHS/SOL data, using the PC-adjusted KRV. The *IL23R* and *C1orf141* genes were combined into a single *IL23R-C1orf141* region due to overlapping variants. Based on the analysis using the Bray-Curtis kernel, there are 72 significant variants (out of 557 common variants) in the *IL23R-C1orf141* region ($\alpha = 0.05/557 = 8.98 \times 10^{-5}$). Based on the analysis using the unweighted UniFrac kernel, there are 114 significant variants (out of 288 common variants) in *ZFR* and 125 significant variants (out of 174 common variants) in *MTMR12* ($\alpha = 0.05/(288 + 174) = 1.08 \times 10^{-4}$). In addition, the Manhattan plot for *MTMR12* based on the CLR-linear kernel shows similar association patterns to the result based on unweighted UniFrac (Figure A.2). From the LD heatmaps, in each gene, the significant variants share a high level of linkage disequilibrium with one another. Future fine mapping of causal variants that affect the microbiome composition will be needed.

To confirm the validity of the covariate-adjusted KRV approach, we conduct kernel PCA on the Bray-Curtis and unweighted UniFrac kernel matrices, and check whether individuals' microbiome profiles, captured by the top two kernel PCs, differ by genotypes of the top (most significant) variant from each identified gene. This is similar to a PCoA analysis. Figure 2.4 shows that, for each top variant, the 95% confidence ellipses for different genotypes are well separated from one another, corroborating the findings by the adjusted KRV. Similar results are found for the CLR-linear kernel with respect to the top variant from *MTMR12* (Figure A.2).

2.3.2 Specific taxa involved in microbiome GWAS associations

To further understand how the discovered genes drive differences in gut microbiome composition, we conduct an exploratory analysis to identify specific microbial taxa involved in the microbiome GWAS associations. Our strategy is to perform dimension reduction on both genetic and microbiome data and use correlation analyses to complement and help interpret our community-level analysis results.

The general analysis procedure is summarized in Figure A.3. As each gene-microbiome

association signal appears to be driven by a single locus (as shown in the LD heatmaps from Figure 2.3), we focus on the top variant from each identified gene for our analysis. On the other hand, we also use the leading 10 kernel PCs from each microbiome kernel to capture the major variation from the overall microbiome composition. For each gene-microbiome association, the specific variant and microbiome kernel used in the analysis are consistent with the association results in Table 2.1. In Step 1, among the top 10 microbiome kernel PCs, we identify kernel PCs that are significantly correlated with the top variant after adjusting for population structure (with false discovery rate (FDR) corrected p-value < 0.05 from linear regression): these kernel PCs represent the microbial community profiles that mainly drive the gene-microbiome associations. In Step 2, we inspect genus-level microbial abundance data and identify taxa that contribute the most to the significant kernel PCs from Step 1 (with absolute correlation between taxon abundance and kernel PC ≥ 0.5): these taxa dominate the microbial profiles captured by the kernel PCs and in turn drive the gene-microbiome associations.

The microbial taxa identified for each gene-microbiome association signal are listed in Table A.3. Due to roles in immunity, we focus on findings related to *IL23R* and *ZFR* for a detailed discussion. We first discuss the taxa involved in the association between *IL23R* and the Bray-Curtis kernel. Allele A (vs. Allele G) of the top variant, rs10789226, from *IL23R* is positively associated with the abundance of *Bacteroides* and *Blautia*, while being negatively associated with the abundance of *Prevotella*. *Bacteroides* and *Prevotella* are the most abundant genera in this study (representing 23.7% and 25.0% abundances of all microbial taxa) and dominate the first PC of the Bray-Curtis kernel. These two genera have been studied extensively as metrics for dietary patterns [82, 83]. Interestingly, a higher *Prevotella*-to-*Bacteroides* ratio is associated with greater obesity in Hispanic/Latino populations based on a previous study using HCHS/SOL data [24]. In terms of relation to immunity disorders, a meta-analysis [84] suggests that patients with IBD are associated with a lower abundance of *Bacteroides* compared to healthy individuals, although mixed roles of *Bacteroides* have been reported in other studies [85]. On the other hand, while

Prevotella species are classically considered as commensal bacteria, increased abundance of certain *Prevotella* strains has been associated with mucosal inflammation and linked to chronic inflammatory diseases [86]. Based on these findings, it appears that Allele A of rs10789226 might be associated with an overall healthier gut microbiome composition in Hispanic/Latino populations.

We next look at the taxa involved in the association between *ZFR* and the unweighted UniFrac kernel. Allele T (vs. Allele A) of the top variant, rs2113093, from *ZFR* is positively associated with the abundance of two unidentified genera from *Clostridiales* and *Ruminococcaceae*. As *Ruminococcaceae* is an order that belongs to the *Clostridiales* family, this result is consistent with the strength of the unweighted UniFrac kernel in utilizing phylogenetic information. *Ruminococcaceae* helps maintain the gut health by producing short-chain fatty acids (SCFAs) [87], and a decreased abundance of *Ruminococcaceae* has been associated with IBD disorders [88] and inflammation in hepatic encephalopathy [89]. On the other hand, several commensal *Clostridiales* strains have been shown to mediate effective immune response against colorectal cancer in mouse models [90]. These findings support the potential roles of *Clostridiales* and *Ruminococcaceae* bacteria in mediating the effect of *ZFR* in regulating innate immune response, and Allele T of rs2113093 is likely associated with a more favorable gut microbiome composition.

Overall, the above findings offer us a better understanding of the identified community-level associations. Nevertheless, due to heterogeneity in functions of individual bacterial species and strains, a higher study resolution will be required to further elucidate the mechanisms underlying the association between the identified genes and the gut microbiome.

2.3.3 Comparison to competing methods and previous studies

As a comparison to our proposed covariate-adjusted KRV approach, we applied additional competing methods of microbiome GWAS to the same set of HCHS/SOL data ($n = 1219$). We first performed two gene-based community-level analyses that rely on univariate microbiome phenotypes (i.e., only using the top kernel PC of the microbiome kernel matrix),

denoted as linear regression and SKAT. Neither of the methods has identified any genome-widely significant signals (Manhattan plots in Figure A.4 and Figure A.5). Therefore, compared to univariate methods that identify the same type of genetic features (i.e., genes associated with the overall microbiome composition), our proposed KRV framework has a superior power in detecting associations.

We also performed a traditional variant-based taxon-level analysis to identify individual genetic variants associated with individual microbial genera. 89 relatively common genera (present in $\geq 10\%$ of all individuals) were tested in the analysis.

At a study-wide significance level ($\alpha = 5 \times 10^{-8}/89 = 5.6 \times 10^{-10}$), we have identified two associations that involve two genetic loci. The first association signal is between a block of $\sim$1 Mb region located at Chromosome 2 q21.3-q22.1, including 58 significant variants, and the abundance of *Bifidobacterium*. This locus involves the *LCT* gene and 8 other genes, exhibiting high-level LD among the significant variants. The top variant from this locus is rs4988235 (p-value $= 4.2 \times 10^{-17}$), a functional variant associated with lactase persistence [91]. This signal was also reported by Kurilshikov et al. [15], who analyzed a sample of 18,340 individuals which consisted of 24 multi-ancestry cohorts including the HCHS/SOL GOLD cohort. In our gene-level analysis using the PC-adjusted KRV, the *LCT* gene is nominally significant based on the unweighted UniFrac kernel (p-value = 0.013), the CLR-linear kernel (p-value = 0.027) and the PhILR-linear kernel (p-value = 0.015), but not significant at the genome-wide level.

The second association signal is between a locus at Chromosome 18 q11.2, including 2 significant variants, and the presence/absence of *Christensenella* (top variant: rs1607482; p-value $= 2.2 \times 10^{-10}$). This locus is intergenic, located between two RNA genes, *LINC01908* and *LOC105372038*. As our proposed analysis approach focused on gene regions only, these variants were not covered in our community-level analysis.

We next investigate the replication of signals found by previous gut microbiome GWAS studies in our analysis. We have examined the significance of 63 previously reported genes that harbor variants associated with gut microbiome beta-diversity [19, 20, 51, 92, 93]. 59

out of 63 genes include at least one common variant in the HCHS/SOL data. Five genes are replicated with nominal significance (p-value < 0.05) based on various microbiome kernels: *BANK1* (unweighted UniFrac, weighted UniFrac), *MAST3* (weighted UniFrac, generalized UniFrac), *POMC* (CLR-linear), *C1orf21* (CLR-linear) and *AHSA2* (PhILR-linear). Among these genes, *POMC* produces peptides involved in anti-inflammatory actions [94], *BANK1* is associated with systemic lupus erythematosus [95], and *MAST3* and *AHSA2* are associated with IBD [96, 97], corroborating the role of immunity-related genes in shaping gut microbiota. However, none of the genes are significant at the genome-wide level.

2.3.4 Simulation results

We have conducted simulation studies to further evaluate the performance of our proposed covariate-adjusted KRV test in terms of type I error rate and power. Table 2.2 shows the empirical type I error rates of both unadjusted and adjusted KRV tests at different significance levels under Type I Error Scenario 1. The unadjusted KRV has inflated type I error rates for all microbiome kernels except unweighted UniFrac. In contrast, the adjusted KRV maintains valid type I error rates for all microbiome kernels. Note that for Type I Error Scenario 1, population structure affected the abundance of common OTUs, which was unlikely to change these OTUs' presence. Since the unweighted UniFrac kernel only captures presence/absence, but not abundance information of a taxon, the population stratification of microbiome profiles is not reflected in the unweighted UniFrac kernel. This absence of confounding effect leads to a valid type I error rate for the unweighted UniFrac kernel even when the unadjusted KRV is used.

Under Type I Error Scenario 2 (Table A.2), where population structure affected the abundance of rare OTUs, the unadjusted KRV has highly inflated type I error rates for all microbiome kernels. Again, the adjusted KRV is able to maintain valid type I error rates for all microbiome kernels.

Figure 2.5 shows the empirical power of the covariate-adjusted KRV test and competing methods under small effect sizes, at the nominal level $\alpha = 0.05$. In general, for each power

Table 2.2: Empirical type I error rate of unadjusted and covariate-adjusted KRV at nominal level α under Type I Error Scenario 1.

Method	Microbiome kernel	α		
		0.05	0.01	0.001
Unadjusted KRV	Bray-Curtis	0.2403	0.0936	0.0255
	Unweighted UniFrac	0.0484	0.0094	0.0011
	Weighted UniFrac	0.1371	0.0371	0.0057
	Generalized UniFrac	0.1412	0.0416	0.0063
	CLR-linear	0.0811	0.0178	0.0016
	PhILR-linear	0.1389	0.0434	0.0076
Adjusted KRV	Bray-Curtis	0.0473	0.0114	0.0012
	Unweighted UniFrac	0.0523	0.0115	0.0009
	Weighted UniFrac	0.0507	0.0095	0.0012
	Generalized UniFrac	0.0499	0.0097	0.0011
	CLR-linear	0.0450	0.0091	0.0011
	PhILR-linear	0.0482	0.0093	0.0015

Linear kernel was used for genetic data.

scenario, the adjusted KRV has a much higher power than linear regression and SKAT, regardless of the microbiome kernel being used (with the exception of unweighted UniFrac in Power Scenario 1 and 2). Next we focus on the adjusted KRV and compare across microbiome kernels: in Power Scenario 1, the Bray-Curtis kernel has the highest power; in Power Scenario 2, the weighted UniFrac kernel has the highest power; in Power Scenario 3, the unweighted UniFrac kernel has the highest power. These results are consistent with the ways these microbiome similarity measures are constructed and can serve as clues as to which microbial features are affected when we use these kernels to detect associations in practice. The Bray-Curtis kernel is efficient in detecting abundance changes in common OTUs. The weighted UniFrac kernel has more power to detect abundance changes in common phylogenetic clusters, and the unweighted UniFrac kernel is more efficient in detecting changes in rare lineages. Again, due to the nature of unweighted UniFrac, all three methods based on this kernel have little power in Power Scenario 1 and 2, where the SNP effect on common

OTUs or common phylogenetic clusters is unlikely to change their presence.

Under large effect sizes (Figure A.6), while the covariate-adjusted KRV displays a clear improvement in power, the overall patterns are similar to those under small effect sizes and again highlight the power gain of our proposed approach over univariate phenotype-based competing methods.

2.4 Discussion

Given the importance of the microbiome in human health, there is an emerging interest in studying the relationship between host genetic variation and human microbiome. Our methodological contribution in this work is twofold. First, we have proposed a novel microbiome GWAS approach to evaluate the association between gene-level genetic variation and community-level microbiome composition. Second, we have proposed a novel multivariate statistic, the covariate-adjusted KRV, to implement this approach with flexible covariate adjustment. By reducing the multiple-testing burden and aggregating small effect sizes between the genetics and the microbiome, our proposed approach improves statistical power and thus requires fewer samples to detect associations compared to the traditional marginal testing approach.

Simulation studies show that the covariate-adjusted KRV maintains valid type I error rates in the presence of confounders and has a much higher power compared to other microbiome GWAS methods that rely on univariate microbiome phenotypes. In a genome-wide analysis of the HCHS/SOL data, we have identified four genes associated with gut microbiome beta-diversity. We have also identified individual variants within these genes and specific microbial taxa involved in the associations, which will be useful for future investigation of the mechanisms underlying the genetics-microbiome relationships.

Most of the identified genes based on the HCHS/SOL data have been previously implicated in immune functions or immunity-related disorders. This is consistent with the works by Blekhman et al. [6] and Rühlemann et al. [20], where loci in immunity-related genes and pathways have been shown to correlate with gut microbiome composition. The *IL23R* gene is

especially interesting for future study, due to its recognition in previous microbiome genetic association studies [76] and its role in IBD, a chronic inflammatory disease that involves both genetic and microbial factors. Many genetic markers associated with IBD are involved in the interactions between the immune system and the microbiome [98, 99]. Furthermore, IBD is characterized by shift in the gut microbiome composition [100, 101], and specific microbes have also been shown to predict response to therapy [102] and postoperative disease recurrence [103] in patients with IBD. Therefore, our finding supports previous work and could contribute to future investigation of the disease etiology. Finally, as HCHS/SOL is one of the most comprehensive studies of Hispanic/Latino populations in the U.S., the results from our analysis will help inform important genetic risk factors for gut-microbiome-related health outcomes in Hispanic/Latino individuals.

Although the covariate-adjusted KRV has valid type I error rates regardless of the kernels used, selecting appropriate kernels that reflect the actual patterns of association is important for maintaining a good statistical power. Different kernels measure different aspects of the structure within the data and assume different association patterns. For example, as we see from previous studies [52] and our simulations results, the Bray-Curtis kernel is more powerful in detecting associations where genetic variation affects common microbial taxa, whereas the unweighted UniFrac kernel is more powerful when genetics affects rarer phylogenetic clusters. In the analysis of the HCHS/SOL data, using different microbiome kernels, we discovered distinct significant genes. This is likely because these genes affect different aspects of the microbiome composition. For example, variants in the *IL23R-C1orf141* region, identified using Bray-Curtis, mainly associate with abundances of *Bacteroides* and *Prevotella* (Table A.3), which are the most abundant genera in this data set. Variants in *ZFR* and *MTMR12*, identified using unweighted UniFrac, associate with genera from less abundant microbial lineages such as *Clostridiales* and *Ruminococcaceae* (Table A.3). Often, we do not have prior knowledge on the ways genetics is associated with the microbiome. A possible extension would be to use an omnibus test that accommodates multiple possible kernels. For example, as proposed by Zhan et al. [22], we could construct an omnibus kernel matrix via a weighted

sum of multiple candidate kernel matrices. Another approach would be to combine p-values obtained using different candidate kernels into a single p-value, such as the Cauchy p-value combination method [104].

While we mainly adjusted for population structure, a major confounder in the genetics-microbiome relationship, in our analysis of the HCHS/SOL data, adjusting for additional covariates (age, gender and study sites) in a reduced sample revealed similar results. However, the signal from the *IL23R-C1orf141* region based on the Bray-Curtis kernel no longer has genome-wide significance in the latter analysis, which is a limitation of our study. Further analyses (Section A.3) suggest that this loss of power is likely due to sample size loss, rather than additional confounding or systematic differences from sub-sampling. Previous studies have reported that Bray-Curtis dissimilarity is less stable to sub-setting and aggregation of data than other types of dissimilarity/distance measures [63], which might also contribute to this reduced significance.

We have compared our gene-based community-level analysis to a traditional variant-based taxon-level microbiome GWAS conducted on the same data. While we identified an association between the *LCT* locus and *Bifidobacterium* abundance at a study-wide significance in the taxon-level analysis, the *LCT* gene was not genome-widely significant in the community-level analysis. *Bifidobacterium* was a relatively common genus (representing 1.04% abundance of all microbial genera) in the HCHS/SOL data. However, when we analyzed the microbiome as a whole and used microbiome kernels that are efficient in detecting abundance changes in common taxa, such as Bray-Curtis and weighted UniFrac, abundance differences in *Bifidobacterium* were likely overshadowed by those in the most abundant genera such as *Bacteroides* and *Prevotella*. This discrepancy in results reflects the inherent difference between taxon-level and community-level analyses. On the other hand, none of the genes identified in our community-level analysis was replicated in the taxon-level analysis, highlighting the value of our proposed approach in discovering gene-microbiome associations that involve concerted shifts in the microbial community. Nevertheless, our proposed KRV framework is not meant to replace the existing taxon-level microbiome GWAS approaches,

as the two modes of analysis focus on distinct types of genetic features. If one is interested in identifying both loci associated with individual taxa and loci associated with the overall microbiome composition, our proposed framework can be applied in conjunction with existing taxon-level GWAS approaches to provide comprehensive results.

We have also investigated the replication of signals from previous gut microbiome GWAS studies. Five previously reported beta-diversity-associated genes [19] have been replicated in our analyses at a nominal significance, but none of the previous signals [19, 20, 51, 92, 93] reaches genome-wide significance. There are several possible reasons. First, compared to environmental effect, most host genetic influences on gut microbiome composition have relatively small effect sizes [48]. The sample sizes of current microbiome GWAS studies, including our study, are still too small to achieve enough statistical power. Second, there is considerable variation across studies in the collection and processing of microbiome data, leading to difficulties in reproducibility. Lastly, certain genetics-microbiome associations might be specific to ancestry or populations. In addition, since we focused on genetic loci within or close to gene regions, we were unable to evaluate the significance of previously identified loci that fell in intergenic regions.

While we have focused on the application of our proposed approach to microbiome GWAS in this work, the covariate-adjusted KRV can also be applied to investigate the relationships among other types of multivariate omics data. For example, we can investigate microbiome-metabolome relationships by examining the association between microbiome composition and groups of host metabolites that belong to distinct metabolic pathways. Such an analysis was described in one of our previous works [105], where we used a similar multivariate testing strategy to identify metabolic pathways associated with the vaginal microbiome. The advantages of reduced multiple testing burden and better captured data structure in our proposed approach can be readily carried over to other types of omics data.

In conclusion, we have proposed a promising approach, the covariate-adjusted KRV framework, to study the covariate-adjusted association between host genetic variation and community-level microbiome composition, which demonstrates good performances in both

simulations and real data analysis. The genes and loci identified using our approach will help elucidate the complex interactions among host genetics, gut microbiome and host immune systems. With the increasing collection of various omics data and high-dimensional traits, we expect the covariate-adjusted KRV to bring more discoveries by taking advantage of the innate structure within the omics and phenotypic data.

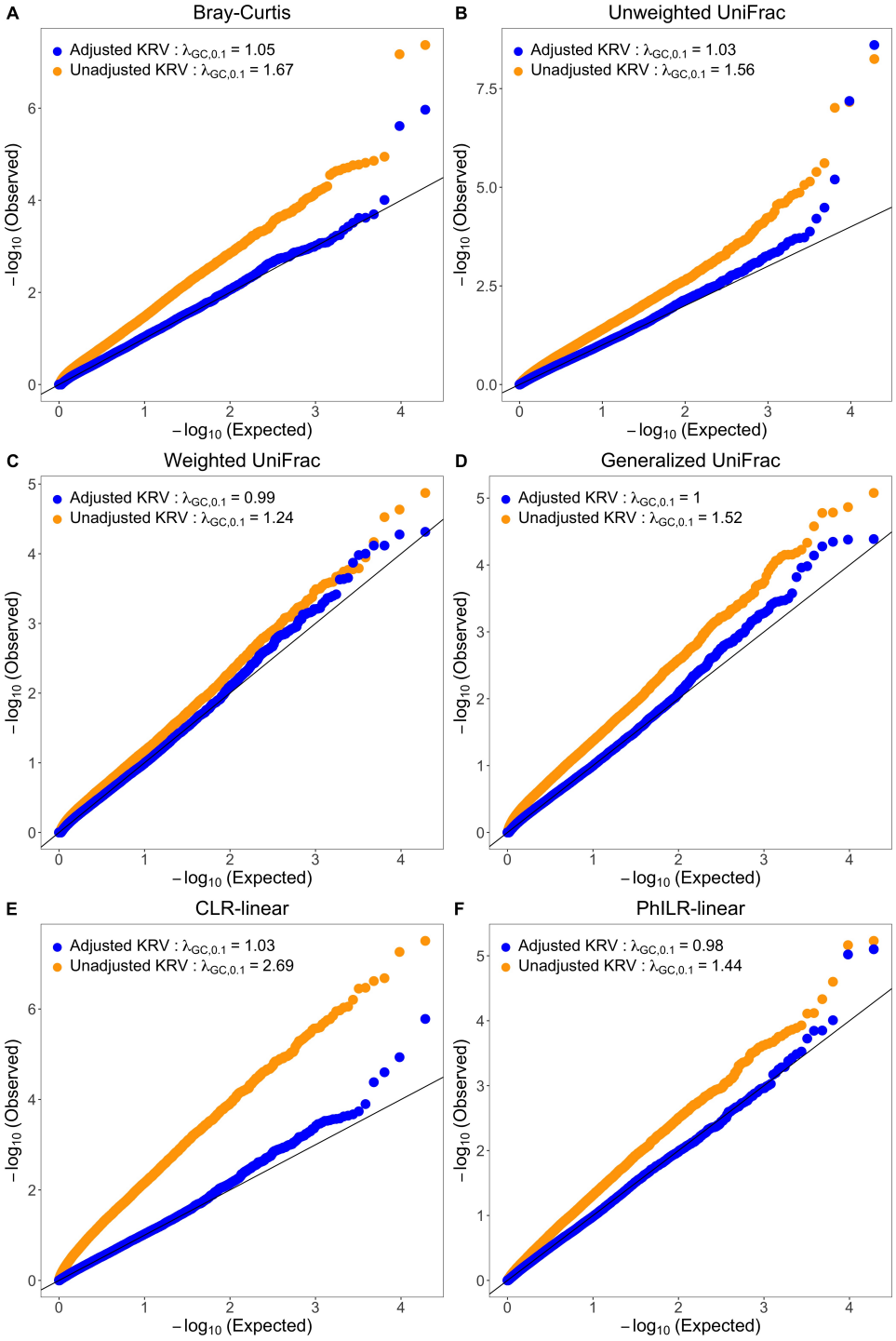

Figure 2.2: P-value QQ-plots from the first-stage gene-level analysis of the HCHS/SOL data. Each panel corresponds to a QQ-plot based on a distinct microbiome kernel. In the adjusted KRV, the top 5 PCs of genome-wide genetic variability were adjusted. $\lambda_{GC,0.1}$ represents the genomic inflation factor evaluated at the upper 10th percentile.

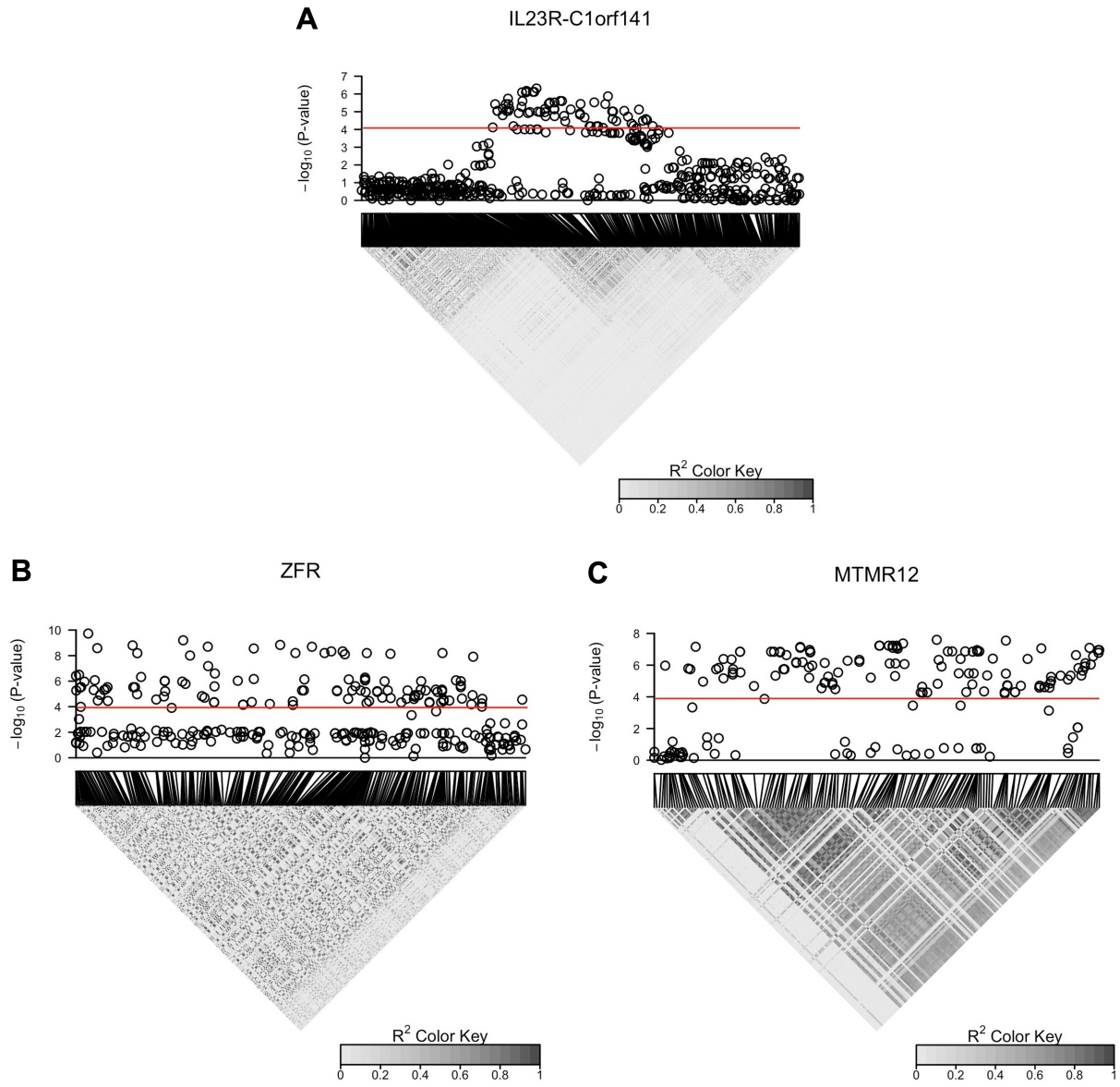

Figure 2.3: Manhattan plots and linkage disequilibrium (LD; R^2) heatmaps from the second-stage variant-level analysis of the HCHS/SOL data, using the PC-adjusted KRV. Each panel corresponds to a distinct gene or gene region. The Bray-Curtis kernel was used for analysis of variants in the *IL23R-C1orf141* region; the unweighted UniFrac kernel was used for analysis of variants in *ZFR* and *MTMR12*. The top 5 PCs of genome-wide genetic variability were adjusted. The red lines represent variant-level significance after Bonferroni correction ($\alpha = 8.98 \times 10^{-5}$ for variants in the *IL23R-C1orf141* region, and 1.08×10^{-4} for variants in *ZFR* and *MTMR12*). A large R^2 value indicates high LD.

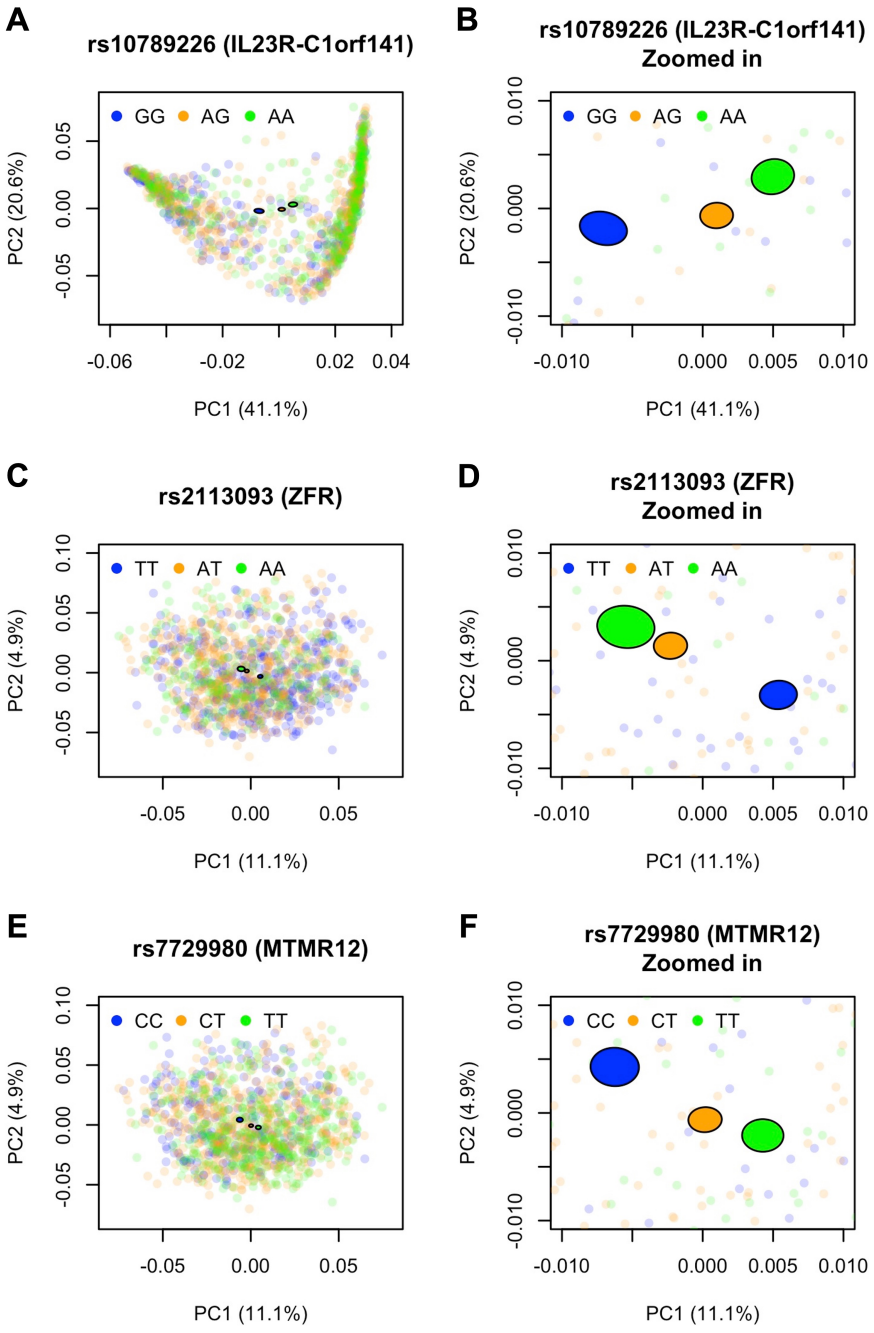

Figure 2.4: PC2 vs. PC1 from kernel PCA on the microbiome kernel, colored by genotype of top variants from the significant genes in the HCHS/SOL study. For each variant, a 95% confidence ellipse (shown as a filled ellipse with black borders) was constructed for individuals from each genotype. The Bray-Curtis kernel was used for the top variant in the *IL23R-C1orf141* region; the unweighted UniFrac kernel was used for the top variants in *ZFR* and *MTMR12*. The percent of variance captured by each kernel PC was provided in the axis labels. Panels **B**, **D**, **F** show enlarged versions of the confidence ellipses from panels **A**, **C**, **E**.

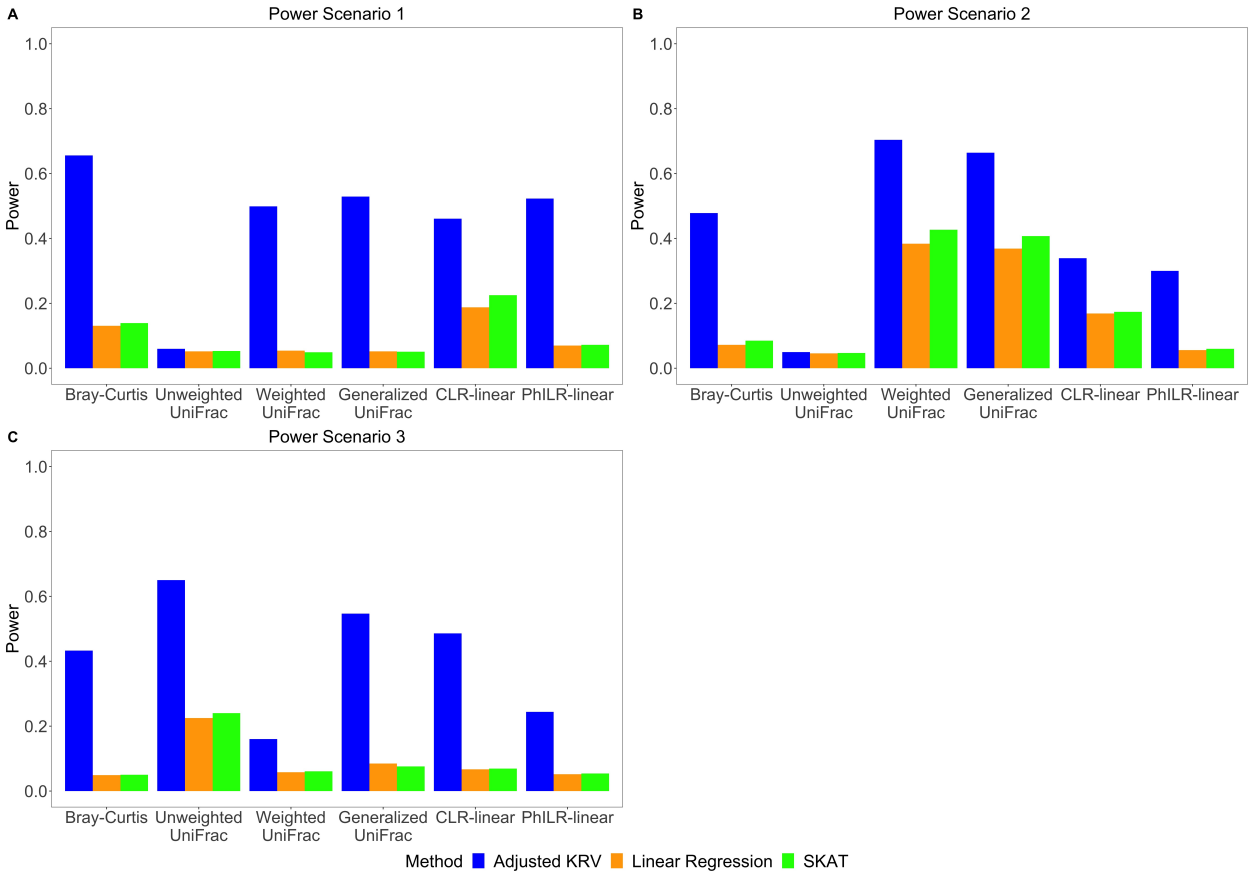

Figure 2.5: Empirical power of covariate-adjusted KRV and competing methods at nominal level $\alpha = 0.05$ for different microbiome kernels under small effect sizes. Panel **A**: A single SNP affects the abundance of common OTUs. Panel **B**: A single SNP affects the abundance of OTUs from a common phylogenetic cluster. Panel **C**: A single SNP affects the abundance of rare OTUs. In each scenario, linear kernel was used for genetic data.

Chapter 3

A KERNEL-BASED TEST OF INDEPENDENCE FOR CLUSTER-CORRELATED DATA

3.1 Introduction

We are often interested in studying the dependence between two multivariate variables. For example, in genetic studies, we may want to assess the association between multiple genetic variants within a gene and a group of traits that likely share a common genetic mechanism [56, 21]. In microbiome studies, we may wish to investigate the association between the overall composition of human microbiota, including hundreds of microbial taxa, and multiple host metabolites from a particular metabolic pathway [26, 27]. Such multivariate analyses aggregate information across variables and are often more powerful than univariate analyses. Meanwhile, correlated observations arise in many practical situations. Family-based designs are common in genetic studies [106], where multiple family members are recruited together into the study. Longitudinal data are common in epidemiological studies [28], where variables of interest are measured on the subjects repeatedly over time. Such study designs introduce clustered dependence among the observations and require proper accommodation. In this work, we aim to develop an approach for assessing the dependence between two multivariate variables, based on cluster-correlated data.

A variety of parametric and semi-parametric methods has been proposed to study the association between one or multiple exposure variables and a multivariate longitudinal (or other cluster-correlated) outcome. These methods often extend upon existing tools for univariate longitudinal data. For example, many studies stack the multivariate outcome into a single response vector and then apply the usual methods that account for clustered data, such as generalized estimating equations (GEE) [107, 29] and random effects models [108, 30].

Such approaches generally apply to a low-dimensional setting [109] and are subject to limitations typical of parametric and semi-parametric methods. Random-effect-based methods require assumptions on the distribution of the multivariate outcome. GEE-based methods rely on good estimation of the correlation structure within clusters as well as across different outcome variables to achieve a high efficiency. Finally, both approaches assume parametric (often linear) relationships between the exposure and the outcome. Therefore, they can only evaluate a limited number of dependence patterns, and are not sufficient as independence tests.

Here we base our approach on the Hilbert-Schmidt Independence Criterion (HSIC), a non-parametric kernel-based measure for assessing the generalized dependence between two multivariate, and potentially high-dimensional, variables [31]. By mapping the two variables into reproducing kernel Hilbert spaces (RKHS's), the population HSIC can be viewed as a measure of maximized covariance between functions in the two RKHS's. When the RKHS's being used are characteristic [110], the population HSIC is zero if and only if the two variables are independent. This measure makes no assumption on the distributions of the variables or the nature of the dependence.

The original HSIC-based independence test [32] applies to independent and identically distributed (i.i.d.) observations. Several extensions have been made to accommodate non-i.i.d. data, but none of the tests, to our knowledge, directly applies to clustered data at an observation level. Zhang et al. (2008) [33] extended the HSIC to certain sequence data, such as XOR sequence and Gaussian process, by specifying the correlation structure of the data as a graphical model and deriving the test statistic based on the maximal cliques. Chwialkowski et al. [34] and Wang et al. [111] developed HSIC-based tests to evaluate the dependence between two time series or random processes in general. Flaxman et al. [35] considered spatial and temporal data; they proposed to first use Gaussian process regression to remove dependence on space and time from the raw variables, and then perform the HSIC test on the resulting de-correlated residuals. However, their approach generally applies to independence testing between two univariate variables. A study with an aim closest to ours

is by Rudra et al. [112]: They analyzed the association between multiple genetic variants and a multivariate longitudinal outcome. Rudra et al. concatenated the outcome measurements from different time points at the subject level, and then applied the HSIC test to the subject-level data. Although this is a straightforward approach to deal with clustered correlation, there could be a loss of statistical power by analyzing data at the subject/cluster level rather than observation level.

In this work, we present the first HSIC-based independence test for cluster-correlated data. Using the empirical HSIC [32] as our test statistic, we derive its asymptotic distribution under the null hypothesis of independence between the two variables but in the presence of clustered correlation among observations. We also examine the behavior of the test statistic under the alternative hypothesis and establish the consistency of our test. Furthermore, we provide a way to approximate the asymptotic null distribution of the test statistic and allow for statistical testing in practice. In simulation studies, our proposed test controls type I error rates well and has a much higher statistical power than competing methods across a range of scenarios. In an application to a longitudinal microbiome-metabolite data set, compared to other approaches, our proposed test identifies a larger number of metabolic pathways significantly associated with the overall microbiome composition, highlighting the value of our test in scientific studies.

The remaining sections are organized as following. In Section 3.2, we provide our background assumption on clustered data and an overview of the HSIC statistic. In Section 3.3, we study the asymptotic behavior of the HSIC statistic under null and alternative hypotheses, and construct a statistical test of independence for cluster-correlated data. In Section 3.4 and 3.5, we demonstrate the performance of our proposed test on both simulated and real data. In Section 3.6, we summarize our work, discuss the limitations of our proposed test and provide a conclusion.

3.2 Background

In this section, we introduce our assumption on cluster-correlated data and give an overview of the HSIC statistic.

3.2.1 General setting

Let P_{XY} be a probability measure defined on a sample space $\mathcal{X} \times \mathcal{Y}$, where both $\mathcal{X}$ and $\mathcal{Y}$ can be multi-dimensional. Let P_X and P_Y be the marginal distributions on $\mathcal{X}$ and $\mathcal{Y}$, respectively. The variables X and Y are statistically independent if $P_{XY} = P_X P_Y$ (equivalently, we can write $X \perp\!\!\!\perp Y$).

We consider a sample of clustered data $\{(X_j, Y_j)\}_{j=1}^n$ drawn from P_{XY}, where the pattern of clustered correlation is balanced and complete:

Assumption 3.2.1. *The observations $(X_1, Y_1), \cdots, (X_n, Y_n)$ are identically distributed according to P_{XY}, and can be divided into m clusters of fixed size d (i.e., $n = md$). In particular, the m clusters*

$$\left\{ \left[(X_{di-d+1}, Y_{di-d+1}), \cdots, (X_{di}, Y_{di}) \right] \right\}_{i=1}^m$$

are independent from one another while having identical within-cluster correlation structure.

The specific correlation structure among the observations in each cluster can be arbitrary. We are interested in studying the dependence between X and Y based on the sample $\{(X_j, Y_j)\}_{j=1}^n$.

3.2.2 Hilbert-Schmidt Independence Criterion

We briefly review the Hilbert-Schmidt Independence Criterion (HSIC) proposed by Gretton et al. (2005a) [31]. The HSIC measures the generalized dependence between two variables X and Y, by embedding X and Y into reproducing kernel Hilbert spaces (RKHS's) and maximizing the covariance between functions of X and Y in the RKHS's.

Let $\mathcal{H}_X$ be an RKHS on $\mathcal{X}$ with associated kernel function (i.e., inner product in the RKHS) $k_X : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and let $\mathcal{H}_Y$ be an RKHS on $\mathcal{Y}$ with associated kernel function $k_Y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Following Gretton et al. (2007) [32], the cross-covariance operator $C_{XY} : \mathcal{H}_Y \to \mathcal{H}_X$ can be defined such that, for any $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$,

$$\langle f, C_{XY} g \rangle_{\mathcal{H}_X} = \mathbb{E}_{XY}\left[\Big(f(X) - \mathbb{E}_X[f(X)]\Big)\Big(g(Y) - \mathbb{E}_Y[g(Y)]\Big)\right] = \mathbb{C}\mathrm{ov}\Big(f(X), g(Y)\Big).$$

As shown by Gretton et al. (2005b) [113], the operator norm (i.e., the largest singular value) of C_{XY}, defined by $\|C_{XY}\| := \sup_{f \in \mathcal{H}_X, g \in \mathcal{H}_Y, \|f\|_\infty \leq 1, \|g\|_\infty \leq 1} \mathbb{C}\mathrm{ov}\big(f(X), g(Y)\big)$, is zero if and only if $X \perp\!\!\!\perp Y$, given that the kernels k_X and k_Y are *universal* (see Definition 5 of [113]). In this sense, $\|C_{XY}\|$ is a measure of independence between X and Y.

The largest singular value of C_{XY} becomes zero when the sum of all squared singular values, denoted as the squared Hilbert-Schmidt norm [31], is zero. Therefore, the squared Hilbert-Schmidt norm of C_{XY}, $\|C_{XY}\|_{HS}^2$, is also an independence criterion. This measure is defined as the population HSIC, which can be expressed conveniently in terms of kernel functions:

$$\begin{aligned}\mathrm{HSIC}(P_{XY}) := \|C_{XY}\|_{HS}^2 &= \mathbb{E}_{XX'YY'}[k_X(X, X')k_Y(Y, Y')] \\ &+ \mathbb{E}_{XX'}[k_X(X, X')]\,\mathbb{E}_{YY'}[k_Y(Y, Y')] - 2\,\mathbb{E}_{XY}\Big[\mathbb{E}_{X'}[k_X(X, X')]\,\mathbb{E}_{Y'}[k_Y(Y, Y')]\Big],\end{aligned}$$

where X' is an independent copy of X. It is obvious that, if X is independent from Y, then we have $\mathrm{HSIC}(P_{XY}) = 0$. Furthermore, for certain *characteristic* k_X and k_Y [110], $\mathrm{HSIC}(P_{XY}) = 0$ if and only if $X \perp\!\!\!\perp Y$. Example characteristic kernels include Gaussian kernels and Laplacian kernels [114].

To estimate the population HSIC from a sample $\{(X_j, Y_j)\}_{j=1}^n$, the empirical HSIC can

be used:

$$\text{HSIC}(P_n) := \frac{1}{n^2}\sum_{i,j}^{n} k_X(X_i, X_j)k_Y(Y_i, Y_j) + \frac{1}{n^4}\sum_{i,j,q,r}^{n} k_X(X_i, X_j)k_Y(Y_q, Y_r) - \frac{2}{n^3}\sum_{i,j,q}^{n} k_X(X_i, X_j)k_Y(Y_i, Y_q).$$

Define the kernel matrices $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ such that the (i, j)-th element of $\boldsymbol{K}_X$ is $k_X(X_i, X_j)$ and the (i, j)-th element of $\boldsymbol{K}_Y$ is $k_Y(Y_i, Y_j)$. Then the empirical HSIC can also be written in terms of $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$:

$$\text{HSIC}(P_n) = \frac{1}{n^2}\,\text{tr}(\boldsymbol{H}\boldsymbol{K}_X\boldsymbol{H}\boldsymbol{K}_Y),$$

where $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is a centering matrix.

Both Gretton et al. (2007) [32] and Zhang et al. (2012) [115] have derived the asymptotic distribution of $\text{HSIC}(P_n)$ under the null hypothesis of independence between X and Y as a weighted sum of chi-square variables, when the observations are i.i.d. In Section 3.3, we examine the asymptotic behavior of $\text{HSIC}(P_n)$ based on cluster-correlated observations. It turns out that the null distribution in this case is still a weighted sum of chi-square variables, where the weights are now modified.

3.3 HSIC for Cluster-correlated Data

Based on the clustered data setting in Section 3.2.1, we aim to test the null hypothesis $H_0 : X \perp\!\!\!\perp Y$ using the empirical HSIC statistic. We first define some useful parameters and statistics.

Assume that the kernel matrices $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ defined in Section 3.2.2 are positive semi-definite. We focus our attention on the centered kernel matrices: $\widetilde{\boldsymbol{K}}_X := \boldsymbol{H}\boldsymbol{K}_X\boldsymbol{H}$ and $\widetilde{\boldsymbol{K}}_Y := \boldsymbol{H}\boldsymbol{K}_Y\boldsymbol{H}$. Let $\tilde{k}_X$ and $\tilde{k}_Y$ be the centered kernel functions [1] derived from k_X and

[1] For the kernel function k_X, the corresponding centered kernel function $\tilde{k}_X$ is: $\tilde{k}_X(x, x') = k_X(x, x') - \mathbb{E}_X[k_X(X, x')] - \mathbb{E}_{X'}[k_X(x, X')] + \mathbb{E}_{XX'}[k_X(X, X')]$.

k_Y, with associated RKHS's $\widetilde{\mathcal{H}}_X$ and $\widetilde{\mathcal{H}}_Y$, respectively. Note that the empirical HSIC can be written as $\mathrm{HSIC}(P_n) = \frac{1}{n^2}\operatorname{tr}(\widetilde{\boldsymbol{K}}_X\widetilde{\boldsymbol{K}}_Y)$.

Let $\gamma_{X,r}$ be the r-th largest eigenvalue and $\boldsymbol{u}_{X,r} = (u_{X,r}(X_1), \cdots, u_{X,r}(X_n))^T$ be the r-th eigenvector of $\widetilde{\boldsymbol{K}}_X$. Similarly, we define the eigenvalues $\gamma_{Y,r}$'s and eigenvectors $\boldsymbol{u}_{Y,r}$'s for $\widetilde{\boldsymbol{K}}_Y$. On the other hand, let $\lambda_{X,r}$ be the r-th largest eigenvalue of the kernel $\tilde{k}_X$ with respect to P_X, with associated eigenfunction $\phi_{X,r}(\cdot)$, such that $\int \tilde{k}_X(x,x')\phi_{X,r}(x')dP_X(x') = \lambda_{X,r}\phi_{X,r}(x)$. Similarly, we define the eigenvalues $\lambda_{Y,r}$'s and eigenfunctions $\phi_{Y,r}$'s for the kernel $\tilde{k}_Y$ with respect to P_Y.

For any fixed $R \in \mathbb{N}$, let $\tilde{k}_{X,R}(x,x') := \sum_{r=1}^{R}\lambda_{X,r}\phi_{X,r}(x)\phi_{X,r}(x')$. For each r where $\gamma_{X,r} > 0$, let $g_{X,r}(x) := \frac{\sqrt{n}}{\gamma_{X,r}}\sum_{j=1}^{n}\tilde{k}_X(x,X_j)u_{X,r}(X_j)$. Define $\tilde{k}_{Y,R}$ and $g_{Y,r}$ similarly. The upcoming theorems will rely on the following assumption:

Assumption 3.3.1. *Suppose that $\mathbb{E}[\tilde{k}_X^2(X,X')] < \infty$ and $\mathbb{E}[\tilde{k}_Y^2(Y,Y')] < \infty$. Assume that, for each $R \in \mathbb{N}$, the classes $\mathcal{C}_X := \{x \mapsto (\tilde{k}_X - \tilde{k}_{X,R})^2(x,x') : x' \in \mathcal{X}\}$ and $\mathcal{C}_Y := \{y \mapsto (\tilde{k}_Y - \tilde{k}_{Y,R})^2(y,y') : y' \in \mathcal{Y}\}$ are P_X-Donsker and P_Y-Donsker [116], respectively. Further assume that, for each r, the functions $x \mapsto g_{X,r}(x)$ and $y \mapsto g_{Y,r}(y)$ converge uniformly in probability as $m \to \infty$, with their limit functions in $L_2(P_X)$ and $L_2(P_Y)$, respectively.*

In general, Assumption 3.3.1 ensures that the data-dependent eigenvalues and (elements of) eigenvectors of the kernel matrices $\widetilde{\boldsymbol{K}}_X$ and $\widetilde{\boldsymbol{K}}_Y$ converge in probability to eigenvalues and eigenfunctions of the kernels $\tilde{k}_X$ and $\tilde{k}_Y$. We show in Section B.1 that the Donsker class condition in Assumption 3.3.1 holds for Gaussian kernels. Now we can establish the asymptotic distribution of $\mathrm{HSIC}(P_n)$ under $H_0 : X \perp\!\!\!\perp Y$ based on clustered data.

Theorem 3.3.2. *Suppose that, for two multivariate random variables X and Y, we have centered kernels $\tilde{k}_X$ and $\tilde{k}_Y$ with discrete eigenvalues. Suppose that Assumption 3.2.1 and Assumption 3.3.1 hold. Under the null hypothesis $H_0 : X \perp\!\!\!\perp Y$, we have*

$$n\,\mathrm{HSIC}(P_n) = \frac{1}{n}\operatorname{tr}(\widetilde{\boldsymbol{K}}_X\widetilde{\boldsymbol{K}}_Y) \xrightarrow{d} \sum_{t=1}^{\infty}\ell_t z_t^2 \text{ as } m \to \infty, \tag{3.1}$$

where z_t's are i.i.d. standard normal variables, and ℓ_t's are the solutions to the eigenvalue problem

$$\ell_t \psi_{t,rs} = \frac{1}{d} \sum_{p,q=1}^{\infty} \mathbb{E}\left[\left(\sum_{i=1}^{d} \sqrt{\lambda_{X,r}\lambda_{Y,s}} \phi_{X,r}(X_i)\phi_{Y,s}(Y_i) \right) \left(\sum_{i=1}^{d} \sqrt{\lambda_{X,p}\lambda_{Y,q}} \phi_{X,p}(X_i)\phi_{Y,q}(Y_i) \right) \right] \psi_{t,pq}$$

for some double sequence $\{\psi_{t,rs}\}_{r,s=1}^{\infty} \in \mathbb{R}$.

The proof of Theorem 3.3.2 is provided in Section B.2. To prove the theorem, we first show the convergence of eigenvalues and eigenvectors of $\widetilde{\boldsymbol{K}}_X$ and $\widetilde{\boldsymbol{K}}_Y$ in the presence of clustered data. We then adopt a strategy similar to that of Zhang et al. (2012) [115]: The test statistic $n\,\mathrm{HSIC}(P_n)$ can be expressed as a sum of squared terms, $\sum_{r,s=1}^{n} Q_{rs}^2$, where the terms Q_{rs}'s depend on eigenvalues and eigenvectors of the kernel matrices. We could show that Q_{rs}'s are asymptotically jointly normal with mean zero under H_0, and the asymptotic variances and covariances of these terms depend on eigenvalues and eigenfunctions of the kernels $\tilde{k}_X$ and $\tilde{k}_Y$.

As a result, the asymptotic distribution of $\mathrm{HSIC}(P_n)$ under H_0 is a weighted sum of chi-square variables. In particular, we require the number of clusters, m, to be sufficiently large. Knowing the null distribution of the test statistic enables us to construct a statistical test at a given significance level. To examine the power of the proposed test, we further explore the behavior of the test statistic when the null hypothesis is violated. The next theorem states the asymptotic behavior of $\mathrm{HSIC}(P_n)$ under the alternative hypothesis $H_1 : X \not\perp\!\!\!\perp Y$.

Theorem 3.3.3. *Suppose that, for two multivariate random variables X and Y, we have centered kernels $\tilde{k}_X$ and $\tilde{k}_Y$ with discrete eigenvalues. Suppose that Assumption 3.2.1 and Assumption 3.3.1 hold. If*

$$\text{there exists some } r, s \in \mathbb{N} \text{ such that } \mathbb{E}[\phi_{X,r}(X)\phi_{Y,s}(Y)] \neq 0, \tag{3.2}$$

then

$$n\,\mathrm{HSIC}(P_n) = \frac{1}{n}\,\mathrm{tr}(\widetilde{\boldsymbol{K}}_X\widetilde{\boldsymbol{K}}_Y) \xrightarrow{p} \infty \text{ as } m \to \infty.$$

When $\tilde{k}_X$ and $\tilde{k}_Y$ are characteristic kernels, Condition (3.2) is equivalent to $H_1 : X \not\perp\!\!\!\perp Y$.

Here Condition (3.2) is a sufficient condition for $X \not\perp\!\!\!\perp Y$: If $X \perp\!\!\!\perp Y$, then $\mathbb{E}[\phi_{X,r}(X)\phi_{Y,s}(Y)] = \mathbb{E}[\phi_{X,r}(X)]\,\mathbb{E}[\phi_{Y,s}(Y)] = 0$ for all $r, s \in \mathbb{N}$; as a contrapositive, (3.2) implies $X \not\perp\!\!\!\perp Y$. When characteristic kernels are used, based on the definition and property of the population HSIC, we can show that $X \not\perp\!\!\!\perp Y$ also implies (3.2).

The proof of Theorem 3.3.3 is provided in Section B.3. To prove the theorem, we show that, under Condition (3.2), there exists a statistic smaller than $d\,\mathrm{HSIC}(P_n)$ that converges in probability to a positive constant, which results in $n\,\mathrm{HSIC}(P_n) = md\,\mathrm{HSIC}(P_n)$ going to infinity, as the number of clusters (m) goes to infinity. When the test statistic goes to infinity, the rejection rate of the test would approach one. Hence, based on Theorem 3.3.3, we have established the consistency of the proposed test.

In practice, the weights ℓ_t's in (3.1) of Theorem 3.3.2 are unknown and we need to estimate them with empirical counterparts. In a similar spirit to Theorem 4 of Zhang et al. (2012) [115], the following proposition provides an approximation for the asymptotic null distribution of $\mathrm{HSIC}(P_n)$ and allows for independence testing in clustered data.

Proposition 3.3.4. *Assume that the conditions in Theorem 3.3.2 hold. To test the null hypothesis $H_0 : X \perp\!\!\!\perp Y$ at a significance level α, we can compare the statistic $n\,\mathrm{HSIC}(P_n) = \frac{1}{n}\,\mathrm{tr}(\widetilde{\boldsymbol{K}}_X\widetilde{\boldsymbol{K}}_Y)$ against the $(1-\alpha)$-quantile of the distribution of*

$$\widetilde{T} = \frac{1}{m}\sum_{t=1}^{n^2} \tilde{\ell}_t z_t^2,$$

where z_t's are i.i.d. standard normal variables and $\tilde{\ell}_t$'s are eigenvalues of $\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^T$, with $\widetilde{\boldsymbol{V}} = [\tilde{\boldsymbol{v}}_1, \cdots, \tilde{\boldsymbol{v}}_m]$. Each vector $\tilde{\boldsymbol{v}}_i$ is obtained by vectorizing (i.e., stacking the columns of) the

$n \times n$ matrix $\widetilde{\boldsymbol{M}}_i$, whose (r,s)-th entry is

$$\widetilde{M}_{i,rs} = \frac{1}{\sqrt{d}} \sum_{j=di-d+1}^{di} \sqrt{\gamma_{X,r}\gamma_{Y,s}} u_{X,r}(X_j) u_{Y,s}(Y_j).$$

The proof of Proposition 3.3.4 is provided in Section B.4, where we show that $\widetilde{T}$ has the same asymptotic distribution as $n\,\mathrm{HSIC}(P_n)$ under H_0. Note that the eigenvalues of $\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^T$, an $n^2 \times n^2$ matrix, are the same as the eigenvalues of $\widetilde{\boldsymbol{V}}^T\widetilde{\boldsymbol{V}}$, an $m \times m$ matrix. In practice, we can calculate the eigenvalues of $\widetilde{\boldsymbol{V}}^T\widetilde{\boldsymbol{V}}$ instead to avoid excessive computational burden.

The distribution of $\widetilde{T}$, which is a mixture of chi-square variables, can be efficiently approximated by Davies' exact method [117]. This method is shown to work well in previous studies [56, 54] that have statistical tests based on a mixture of chi-square distributions.

3.4 Simulation Studies

3.4.1 Simulation Methods

We now conduct simulation studies to demonstrate the performance of our proposed test. We consider a longitudinal data setting, where a set of exposure variables $X \in \mathbb{R}^p$ and a set of outcome variables $Y \in \mathbb{R}^q$ are measured on m subjects at 3 time points. In other words, the observations $\{(X_j, Y_j)\}_{j=1}^n$ are grouped into m clusters with cluster size $d = 3$. To introduce correlation across different time points as well as across different variables, we use a Kronecker product-based covariance structure [107], which has often been used to model multivariate longitudinal data [109].

The general simulation setting is as following. For each cluster, let x_{ij} denote the i-th variable in X measured at the j-th time point, for $i = 1, \cdots, p$ and $j = 1, 2, 3$. Let y_{ij} be defined similarly. Within each cluster, we let $(x_{11}, x_{12}, x_{13}, \cdots, x_{p1}, x_{p2}, x_{p3})^T \sim \mathcal{N}(5 \times$

$\mathbf{1}_{3p}, \boldsymbol{\Sigma}_X)$, where $\boldsymbol{\Sigma}_X = \boldsymbol{R}_X \otimes \boldsymbol{R}_{cl}$, with

$$\boldsymbol{R}_X = \begin{pmatrix} 1 & \rho_X & \cdots & \rho_X \\ \rho_X & 1 & \cdots & \rho_X \\ \vdots & \vdots & \ddots & \vdots \\ \rho_X & \rho_X & \cdots & 1 \end{pmatrix}_{p\times p}, \quad \boldsymbol{R}_{cl} = \begin{pmatrix} 1 & \rho_c & \rho_c^2 \\ \rho_c & 1 & \rho_c \\ \rho_c^2 & \rho_c & 1 \end{pmatrix}.$$

Here $\otimes$ is the Kronecker product. Marginally, we have imposed an exchangeable correlation structure $\boldsymbol{R}_X$ across the p variables in X and an AR(1) correlation structure $\boldsymbol{R}_{cl}$ across the three time points. The correlations between distinct variables at different time points are products of the marginal correlations: e.g., $\mathbb{C}\text{or}(x_{11}, x_{22}) = \rho_X \rho_c$.

We simulate a situation where a single exposure (say, the r-th variable in X) affects multiple outcomes, with different effect sizes on different outcomes. Within each cluster, we use the model:

$$\begin{aligned} &(y_{11}, y_{12}, y_{13}, \cdots, y_{q1}, y_{q2}, y_{q3})^T \\ &= (\beta_1 f(x_{r1}), \beta_1 f(x_{r2}), \beta_1 f(x_{r3}), \cdots, \beta_q f(x_{r1}), \beta_q f(x_{r2}), \beta_q f(x_{r3}))^T + \boldsymbol{\epsilon}, \end{aligned} \tag{3.3}$$

where β_s, with $s = 1, \cdots, q$, is the effect size of the chosen exposure on the s-th outcome, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_Y)$, with $\boldsymbol{\Sigma}_Y = \boldsymbol{R}_Y \otimes \boldsymbol{R}_{cl}$. $\boldsymbol{R}_Y$ is the correlation matrix for the q variables in Y. In simulations, we set $p = q = 20$, $\rho_X = 0.5$ and consider various levels of within-cluster correlation: $\rho_c = 0.3, 0.5$ or 0.7. We also let $\boldsymbol{R}_Y = \boldsymbol{R}_X$.

Type I error simulation To evaluate the type I error rate (rejection rate under H_0), we let $\beta_1 = \cdots = \beta_q = 0$, so that the null hypothesis $H_0 : X \perp\!\!\!\perp Y$ is true. We perform both the proposed HSIC test with proper accommodation for clustered correlation ($\mathbf{HSIC_{cl}}$), and the original HSIC test without any adjustment ($\mathbf{HSIC_{orig}}$) as in [115] and [56]. These two methods are applied to the data $\{(X_j, Y_j)\}_{j=1}^n$ at the observation level. We consider $m =$ 500, 1000 or 1500 clusters and calculate the empirical type I error rates in each setting based

on 1000 simulated data sets.

From Model (3.3), both X and Y have multivariate normal distributions under H_0. While we focus on normal data here, additional Type I error simulations based on non-normal data are considered in Section B.6.1.

Power simulation To evaluate the power (rejection rate under H_1), we randomly select one exposure variable from X to be the causal exposure, and make the first η proportion ($\eta = 10\%, 20\%, 30\%, 40\%$) of outcomes in Y depend on that exposure (with nonzero β_s's). We let the function $f(x)$ take two forms: $f(x) = x$ (**Power Scenario 1**) and $f(x) = \log((x-4)^2)$ (**Power Scenario 2**). For $s = 1, \cdots, \eta q$, the effect sizes β_s's are generated from a Uniform$(0, \sqrt{25/m})$ distribution under **Power Scenario 1**, and from Uniform$(0, \sqrt{10/m})$ under **Power Scenario 2**.

In the power simulation, we perform $\mathbf{HSIC_{cl}}$ and two other HSIC-based competing methods. The two competing methods analyze data at the cluster/subject level. In the first method ($\mathbf{HSIC_{mean}}$), for each cluster, we take an average of observations at different time points: We consider the new variables $X^* := (\frac{1}{3}\sum_{j=1}^3 x_{1j}, \cdots, \frac{1}{3}\sum_{j=1}^3 x_{pj})^T$ and $Y^* := (\frac{1}{3}\sum_{j=1}^3 y_{1j}, \cdots, \frac{1}{3}\sum_{j=1}^3 y_{qj})^T$ and then perform the original HSIC test based on $\{(X_i^*, Y_i^*)\}_{i=1}^m$. In the second method ($\mathbf{HSIC_{cat}}$), we follow the strategy of Rudra et al. [112] and concatenate the observations at different time points for each cluster: We consider the new variables $X^{**} := (x_{11}, x_{12}, x_{13}, \cdots, x_{p1}, x_{p2}, x_{p3})^T$ and $Y^{**} := (y_{11}, y_{12}, y_{13}, \cdots, y_{q1}, y_{q2}, y_{q3})^T$ and then perform the original HSIC test based on $\{(X_i^{**}, Y_i^{**})\}_{i=1}^m$.

We consider $m = 500$ clusters and calculate the empirical power in each setting based on 1000 simulated data sets.

Kernel choices For both X and Y, we consider two different kernels: the Gaussian kernel $k_X(z_1, z_2) = k_Y(z_1, z_2) = \exp(-\|z_1 - z_2\|_2^2/\tau)$ and the linear kernel $k_X(z_1, z_2) = k_Y(z_1, z_2) = z_1^T z_2$. For the Gaussian kernel, the shape parameter τ is chosen as the median of the Euclidean distance between each sample pair. While the Gaussian kernel is a characteristic kernel

[114], the linear kernel is not characteristic and is designed to detect linear or close-to-linear relationships between X and Y. Nevertheless, linear kernels have been shown to be reasonably powerful in previous association studies [54, 21] and can be computationally efficient (see Section B.7.2).

Additional simulation studies are provided in Section B.6. Additional implementation details including computation time and code availability are provided in Section B.7.

3.4.2 Simulation Results

Table 3.1 shows the empirical type I error rates of $\mathbf{HSIC_{orig}}$ and $\mathbf{HSIC_{cl}}$ for normal data. The type I error rate of $\mathbf{HSIC_{orig}}$ is inflated in each setting, where the inflation becomes greater as the within-cluster correlation (ρ_c) increases. In contrast, $\mathbf{HSIC_{cl}}$ has a well-controlled type I error rate across all levels of within-cluster correlation. Using the linear kernel, $\mathbf{HSIC_{cl}}$ has type I error rates close to the nominal α in all situations. Using the Gaussian kernel, $\mathbf{HSIC_{cl}}$ is conservative when the number of clusters is moderate ($m = 500$), but its type I error rate gets close to the nominal α at a larger sample size ($m = 1500$). For non-normal data (Figure B.1), the pattern is similar: $\mathbf{HSIC_{cl}}$ is able to control the type I error rate, either with the Gaussian kernel or with the linear kernel.

$\mathbf{HSIC_{cl}}$ based on the Gaussian kernel is more conservative, likely because the Gaussian kernel is associated with a larger number of non-zero eigenvalues in finite samples than the linear kernel in our simulation setting. The null distribution for Gaussian-kernel-based $\mathbf{HSIC_{cl}}$ thus involves more terms in the weighted sum of chi-square variables (as the weights depend on eigenvalues), which might aggregate more finite-sample errors and make the test statistic converge slower to the asymptotic distribution.

Figure 3.1 shows the empirical power of $\mathbf{HSIC_{cl}}$ and the two competing methods under **Power Scenario 1**. In all situations, $\mathbf{HSIC_{cl}}$ has a higher power than both $\mathbf{HSIC_{mean}}$ and $\mathbf{HSIC_{cat}}$, regardless of the level of within-cluster correlation or the kernel being used. In addition, the power of $\mathbf{HSIC_{cl}}$ improves quickly as a higher proportion of variables in Y is associated with X. Since X and Y are linearly associated in Scenario 1, it is not surprising

Table 3.1: Empirical type I error rate of $\mathbf{HSIC_{orig}}$ and $\mathbf{HSIC_{cl}}$ at nominal level α for normal data under simulation.

α	m	ρ_c	Gaussian kernel		Linear kernel	
			$\mathbf{HSIC_{orig}}$	$\mathbf{HSIC_{cl}}$	$\mathbf{HSIC_{orig}}$	$\mathbf{HSIC_{cl}}$
0.05	500	0.3	0.119	0.024	0.068	0.047
		0.5	0.603	0.031	0.141	0.044
		0.7	1.000	0.030	0.330	0.044
	1000	0.3	0.115	0.029	0.070	0.043
		0.5	0.591	0.034	0.166	0.054
		0.7	1.000	0.034	0.348	0.043
	1500	0.3	0.113	0.047	0.082	0.053
		0.5	0.608	0.043	0.145	0.053
		0.7	1.000	0.044	0.352	0.052
0.01	500	0.3	0.018	0.005	0.021	0.013
		0.5	0.190	0.005	0.035	0.011
		0.7	0.998	0.009	0.114	0.008
	1000	0.3	0.022	0.006	0.015	0.010
		0.5	0.180	0.010	0.050	0.008
		0.7	0.999	0.010	0.111	0.010
	1500	0.3	0.027	0.007	0.019	0.010
		0.5	0.209	0.008	0.047	0.010
		0.7	0.999	0.008	0.117	0.009

that the linear kernel is powerful in detecting this dependence. The Gaussian kernel has a comparable performance as the linear kernel.

Figure 3.2 shows the empirical power under **Power Scenario 2**, where X and Y have a non-linear relationship. Similar to **Power Scenario 1**, for both the Gaussian kernel and the linear kernel, $\mathbf{HSIC_{cl}}$ achieves a higher power than the competing methods under all levels of within-cluster correlation. When compared between kernels, $\mathbf{HSIC_{cl}}$ based on the Gaussian kernel is more powerful than $\mathbf{HSIC_{cl}}$ based on the linear kernel in each setting, showing the advantage of the Gaussian kernel as a characteristic kernel to detect general dependence patterns.

Overall, both **Power Scenario 1** and **2** show the considerable power gain of $\mathbf{HSIC_{cl}}$ over

analyzing data at the cluster level. We also note that, the power gain of $\mathbf{HSIC_{cl}}$ decreases as the within-cluster correlation increases. This is expected since there will be less pronounced information loss in averaging or concatenating the data at the cluster level if observations within a cluster are highly correlated.

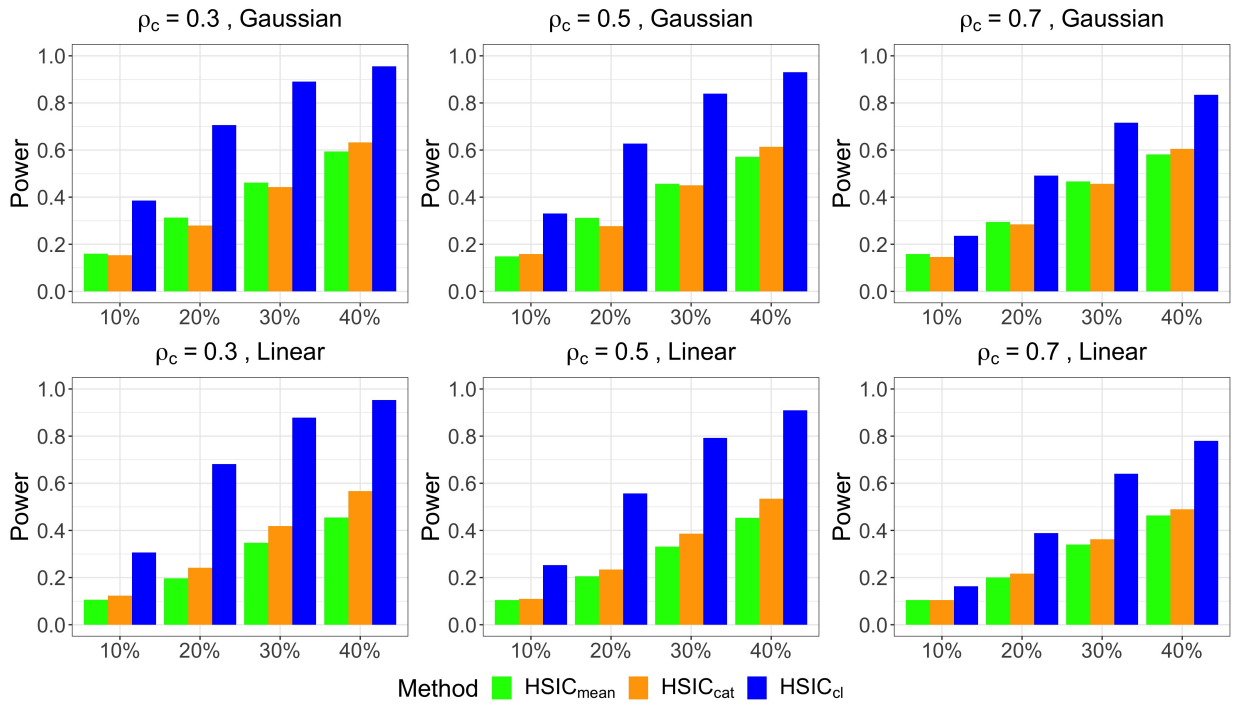

Figure 3.1: Empirical power of $\mathbf{HSIC_{cl}}$ and competing methods at nominal level $\alpha = 0.05$ under **Power Scenario 1**. The x-axis represents the proportion of variables in Y that are associated with X. The top row shows results based on the Gaussian kernel, and the bottom row shows results based on the linear kernel.

While the above results are based on a fixed cluster size, we have also investigated the effect of cluster size on the performance of $\mathbf{HSIC_{cl}}$ (Section B.6.2). When the number of clusters (m) and the level of within-cluster correlation (ρ_c) are fixed, type I error control is similar for different cluster sizes (Figure B.2-B.3), suggesting that the convergence speed of the test statistic under H_0 is likely not affected by cluster size. However, a larger cluster size tends to result in a higher statistical power (Figure B.4), possibly due to an increase in the

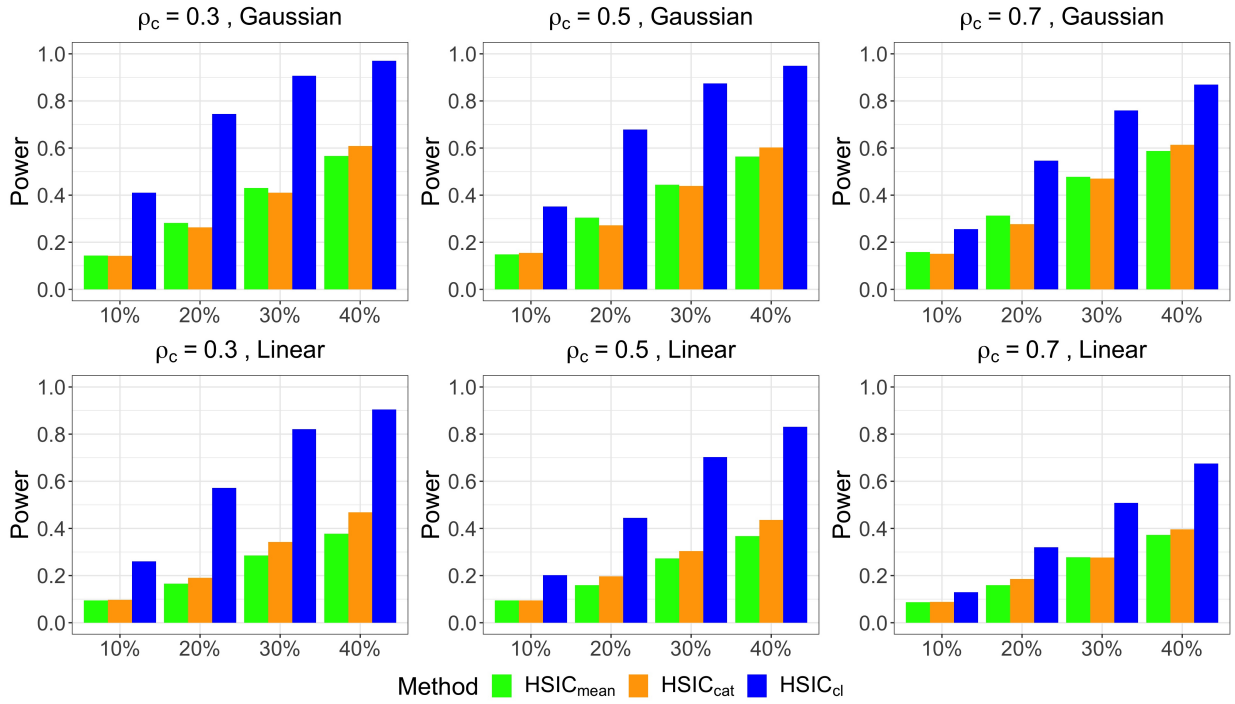

Figure 3.2: Empirical power of $\mathbf{HSIC_{cl}}$ and competing methods at nominal level $\alpha = 0.05$ under **Power Scenario 2**. The x-axis represents the proportion of variables in Y that are associated with X. The top row shows results based on the Gaussian kernel, and the bottom row shows results based on the linear kernel.

overall sample size, which allows for additional information gain.

3.5 Application to Real Data

The vaginal microbiota plays an important role in maintaining vaginal homeostasis. Common vaginal conditions, such as bacterial vaginosis, are characterized by shifts in vaginal microbiome composition and changes in vaginal metabolites [27, 118]. Studying the association between the microbiome and the metabolites helps us better understand how the vaginal microbiota contributes to the host metabolic environment, and identifies potential metabolic biomarkers for vaginal conditions [27]. Here we apply $\mathbf{HSIC_{cl}}$ and competing methods to test the dependence between the overall vaginal microbiome composition and different metabolic pathways, using data from the Menopause Strategies: Finding Lasting

Answers for Symptoms and Health (MsFLASH) Vaginal Health Trial [119].

The MsFLASH trial was a 12-week randomized clinical trial to evaluate the treatment effect of vaginal estradiol vs. placebo on vaginal discomfort in postmenopausal women [119] (see Section B.8.1 for more details). As part of an effort to investigate the mechanism of postmenopausal vaginal discomfort, vaginal microbiota and vaginal fluid metabolites were characterized longitudinally and available in 141 participants at baseline, 4 and 12 weeks [120]. The vaginal microbiome profiles included abundance data of 381 taxa. The metabolome profiles included abundance data of 171 metabolites, which were grouped into 95 metabolic pathways. We apply $\mathbf{HSIC_{cl}}$, $\mathbf{HSIC_{mean}}$ and $\mathbf{HSIC_{cat}}$ to assess the dependence between the overall vaginal microbiome composition and the metabolites in each pathway, across all 95 pathways. In other words, for each test, we have $m = 141$, $d = 3$, $X \in \mathbb{R}^{381}$ and $Y \in \mathbb{R}^q$, where q is the number of metabolites in a pathway, ranging from 1 to 21 in this data set; 95 tests are performed in total.

Table 3.2 shows the number of metabolic pathways identified to be associated with the vaginal microbiome composition, at a Bonferroni-corrected significance level $\alpha = 0.05/95 = 5.3 \times 10^{-4}$. Due to the close relationship between vaginal microbiota and vaginal metabolites, all methods have identified a considerable number of significant metabolic pathways. Still, $\mathbf{HSIC_{cl}}$ identifies a larger number of pathways than $\mathbf{HSIC_{mean}}$ and $\mathbf{HSIC_{cat}}$, either with the Gaussian kernel or with the linear kernel. In particular, based on the Gaussian kernel, $\mathbf{HSIC_{cl}}$ successfully identifies all the significant pathways discovered by $\mathbf{HSIC_{mean}}$ and $\mathbf{HSIC_{cat}}$, and discovers 4 (7) additional pathways compared to $\mathbf{HSIC_{cat}}$ ($\mathbf{HSIC_{mean}}$) (Figure B.6). 67 out of 68 pathways discovered by $\mathbf{HSIC_{cl}}$ using the linear kernel are also identified by $\mathbf{HSIC_{cl}}$ using the Gaussian kernel (Figure B.7). For this data set, the Gaussian kernel appears to be more powerful in detecting dependence than the linear kernel, indicating a possibly non-linear relationship between certain metabolites and microbial taxa abundances.

We focus on some of the top pathways (with high statistical significance) identified using $\mathbf{HSIC_{cl}}$ and highlight their biological relevance. The top pathways include multiple

Table 3.2: Number of metabolic pathways identified to be associated with the vaginal microbiome composition based on the MsFLASH data set ($\alpha = 5.3 \times 10^{-4}$).

Kernel	$\mathbf{HSIC_{mean}}$	$\mathbf{HSIC_{cat}}$	$\mathbf{HSIC_{cl}}$
Gaussian	68	71	75
Linear	64	64	68

metabolic pathways for amino acids. The human vaginal microbiota is dominated by bacteria in the *Lactobacillus* genus [121], which are known to produce branched-chain amino acids including valine, leucine and isoleucine [122]. All these amino acids are present in our top pathways. In particular, one pathway related to leucine metabolism is only identified by $\mathbf{HSIC_{cl}}$ but not by $\mathbf{HSIC_{mean}}$ or $\mathbf{HSIC_{cat}}$. Therefore, our finding is consistent with previous studies on bacterial metabolism, confirming the power improvement in using $\mathbf{HSIC_{cl}}$ for scientific discovery.

3.6 Discussion

We have introduced a novel kernel-based approach, $\mathbf{HSIC_{cl}}$, to evaluate the generalized dependence between two multivariate variables based on cluster-correlated data. Using the previously developed HSIC statistic as our test statistic, we have derived its asymptotic null distribution in the presence of clustered correlation and constructed a statistical test of independence accordingly. We have also established the consistency of the proposed test. Both simulation studies and application to real longitudinal data demonstrate the power gain in using our proposed test, compared to methods based on measurements averaged or concatenated at the cluster level.

One limitation of our framework is that the proposed test only applies to balanced and complete clustered data, which might not be always available in practice. In longitudinal studies, for example, subjects might be followed at different time points from one another (resulting in unbalanced data), or become lost to follow-up (resulting in incomplete data).

For incomplete data, one solution is to impute the missing data before applying $\mathbf{HSIC_{cl}}$. Further extension of the test for unbalanced or incomplete clustered data will be interesting for future study.

Another limitation is that our proposed test relies on asymptotic results, and the null distribution might not be accurately approximated when the number of clusters is small, which is likely true of many family-based or longitudinal studies. Permutation-based approaches could be a surrogate for $\mathbf{HSIC_{cl}}$ at small sample sizes (see Section B.6.3), although their computational burden is large compared to $\mathbf{HSIC_{cl}}$ (see Table B.2). Computationally efficient adaptations of $\mathbf{HSIC_{cl}}$ to small sample sizes, such as those proposed by Lee et al. [123] and Zhan et al. [22], would be another useful extension.

With the continuing emergence of high-dimensional data and the prevalence of cluster-correlated data in different scientific fields, our proposed test is a promising approach to discover novel associations and bring new scientific insights in these settings. Meanwhile, we need to be cautious about potential risks to society that might result from misuse or misinterpretation of our proposed test. For example, confounding is an important factor to consider in genetic and epidemiological studies. A confounder affects both the exposure and the outcome, and could lead to spurious associations between the two variables even if the variables themselves do not have causal relationships. Therefore, as one applies our proposed test to evaluate the association between two variables, it is important to consider the presence of potential confounders and be careful in interpreting the test results. Mistaking certain observed correlation for causation could lead to misinformation in the scientific community and would be especially concerning when the studies being conducted directly influence people's life.

Chapter 4

MULTIVARIATE CONDITIONAL INDEPENDENCE TESTING FOR VAGINAL MICROBIAL NETWORK CONSTRUCTION

4.1 Introduction

The human microbiome, composed of microorganisms that reside on and within different parts of the human body, plays important roles in host health. Specifically, the vaginal microbiome is involved in health outcomes of the female reproductive tract. A vaginal microbiome dominated by species in the *Lactobacillus* genus is considered to help maintain vaginal health by protecting against pathogenic bacteria through lactic acid production [124, 125]. On the other hand, increased diversity of non-*Lactobacillus* species has been associated with urogenital diseases including bacterial vaginosis (BV), sexually transmitted infections (STIs) and HIV infection, as well as adverse pregnancy outcomes such as preterm birth [124, 126, 127]. While changes in hormone levels and pregnancy status could cause fluctuations in the vaginal microbiome [128, 129], the vaginal microbiome also differs among individuals by race, ethnicity and various environmental and socioeconomic factors [130, 131, 129].

Microbial sequencing techniques, such as 16S ribosomal RNA (rRNA) sequencing, enable us to characterize microbial communities in their natural environment, by providing detailed information on abundance and phylogeny of individual microbial taxa within the community. Due to complex interactions among vaginal microbial taxa, such as the competitive exclusion between *Lactobacillus* species and other pathogenic bacteria, microbial association networks are a promising tool to understand the vaginal microbiome [132]. As a standard analysis in microbiome profiling studies, microbial networks help elucidate the relationships among microbial taxa, shed light on the global structure of the microbial community, and provide clues to the mechanisms by which the microbiome affects host health [133, 5, 36]. A microbial

network is an undirected graph, where the nodes represent individual microbial taxa and the edges connecting the nodes represent the association in abundance between a pair of taxa. Statistical association between a pair of taxa could indicate functional interactions between the taxa within the microbial community.

One popular way to evaluate the association between two microbial taxa is assessing the conditional dependence between the two taxa given all the other taxa in the community, as this approach tends to capture direct interactions between taxa rather than indirect, spurious associations [37, 39]. Existing work [37, 134, 38, 135] often utilizes methods in graphical modeling to construct microbial networks based on conditional dependence, such as neighborhood selection for estimating local conditional dependence structure for each node [136] and penalized maximum likelihood for estimating a sparse inverse covariance matrix globally [137]. Recently, phylogenetic relationships among the microbial taxa have also been incorporated to improve the accuracy of network estimation [42].

Based on 16S sequencing data, vaginal microbial taxa can be classified as accurate as to the species level [138]. While different taxonomic levels are available, it is preferred to construct a network at a higher level (such as family or genus level) in order to reduce the dimension of microbial features and account for sparsity of microbial data [40, 41]. In these cases, conditional dependence is typically assessed based on abundance data aggregated at the desired taxonomic level [38, 42]. However, such an aggregated approach will result in power loss when there are heterogeneous relationships present among the sub-taxa within the taxa of interest, since this approach implicitly assumes that the associations between these sub-taxa have the same directions. Different species and strains within the same bacterial genus can differ in their functions and activities. For example, among common *Lactobacillus* species, while *L. crispatus* and *L. gasseri* mostly serve beneficial roles in vaginal health [125], *L. iners* is postulated to contain both beneficial and pathogenic features [139]. As a consequence, to construct a genus-level network, simply aggregating abundance data at the genus level could obscure species-level signals and omit important associations.

In this work, we propose a multivariate approach to construct microbial association net-

works based on conditional dependence. Without loss of generality, suppose that we are interested in constructing a network at the genus level while having species-level data available. Instead of assessing the conditional dependence between genus-level aggregated abundances, we assess the conditional dependence between pairs of multivariate variables, where each multivariate variable represents the abundances of multiple species that belong to a particular genus. This multivariate approach improves power by utilizing abundance information at a finer level and accumulating multiple weak species-level associations into a more detectable signal at the genus level. The proposed approach is especially advantageous when there are (conditional) correlations of opposite directions among species within the pair of genera being studied, as shown in Figure 4.1.

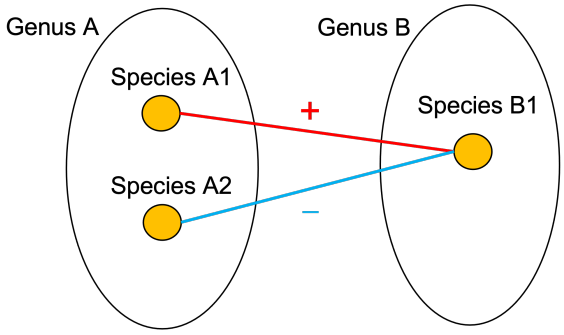

Figure 4.1: Example relationship between two microbial taxa, Genus A and Genus B, where correlations of opposite directions are present among the sub-taxa: positive association between Species A1 and Species B1; negative association between Species A2 and Species B1.

In contrast to previous methods that focus on estimation or prediction [37, 134, 38], here we assess the pairwise conditional dependence via hypothesis testing and provide a measure of statistical significance. We develop a novel multivariate statistic, denoted as Conditional RV (CRV), to conduct multivariate conditional independence testing. CRV is constructed both as an extension of the well-known unconditional RV coefficient [140] and as an extension of the recently proposed univariate conditional independence test, SEcov [141]. As a nonparametric and flexible measure of conditional dependence, CRV requires no assumption on the distribution of the two multivariate variables being tested and is

able to incorporate common machine learning approaches for estimating conditional means. We derive the asymptotic distribution of the CRV statistic under the null hypothesis of conditional independence between two variables and construct a statistical test of conditional independence accordingly. The performance of CRV is evaluated in simulation studies, and we apply CRV to investigate the vaginal microbiome in Black and White pregnant women based on the Pregnancy, Infection, and Nutrition (PIN) Study [43]. We construct a genus-level microbial network based on the entire sample to identify important vaginal microbial interactions. The network is also constructed within Black and White women separately to identify any racial differences in network structure.

The rest of the article is organized as follows. Section 4.2 describes the microbiome data from the PIN study and specifies the questions of interest. Section 4.3 introduces the CRV statistic and the conditional independence test based on CRV. Section 4.4 presents simulation studies to evaluate the performance of CRV in both conditional independence testing and microbial network construction. Section 4.5 applies CRV to construct vaginal microbial networks based on the PIN study. Finally, we conclude with a discussion in Section 4.6.

4.2 Data and Problem Description

Our study of the vaginal microbiome involves two goals. First, the vaginal microbiome is not only associated with female urogenital conditions, but also has an impact on pregnancy outcomes and neonatal health [127, 142]. To further characterize the vaginal microbiome during pregnancy, we are interested in understanding the global structure of the vaginal microbial community and identifying important vaginal microbial interactions in healthy pregnant women via microbial networks. Second, the vaginal microbiome displays racial differences as well [130, 129]. Black women are more likely to harbor a diverse vaginal microbiome than White women, and are also more likely to experience adverse birth outcomes [130]. Previous studies have postulated that this racial disparity in pregnancy outcomes could be attributed to the racial difference in vaginal microbiome composition [130, 143].

As a secondary goal, we are interested in comparing microbial networks between Black and White women to identify any racial difference in network structure.

To address the above goals, we utilize the vaginal microbiome data from the PIN study, a prospective cohort study of pregnant women conducted in central North Carolina in the United States [143]. The original study recruited 3,163 pregnant women with singleton pregnancies from prenatal clinics from August 1995 to February 2001. For our current analysis, we consider 652 women from a recent subset study of the PIN data [143], including 375 White women and 277 Black women who experienced a term birth and have vaginal microbiome data available. A detailed description of this sample is provided in Section C.4.1. For these women, vaginal swabs were collected from the participants between 24 and 29 weeks of gestation, and 16S rRNA sequencing was performed on the swab samples to characterize the vaginal microbial profiles. High-quality sequences were mapped at the species level using the STIRRUPS platform [138]. A detailed description of swab collection, DNA extraction and sequencing, and bioinformatic processing of 16S data can be found in [143].

After standard data filtering procedures (see Section 4.5 for details), we obtain species-level abundance data on 186 species from 24 genera in 647 subjects, where the number of species in each genus ranges from 1 to 38. Our analysis of the vaginal microbiome data involves three parts. First, we construct a genus-level microbial network based on the entire sample of 647 women, using the species-level abundance data. Second, we construct networks within Black and White women separately and compare the two networks qualitatively to identify racial differences in network structure. Finally, we investigate conditional relationships between individual species based on the constructed genus-level network from the combined sample to identify potential heterogeneous species-species interactions, which highlight the advantage of our proposed CRV approach.

4.3 Methods

We introduce a novel multivariate measure, Conditional RV (CRV), to evaluate the conditional dependence between two multivariate variables given a third, potentially high-dimensional variable. CRV can be viewed as a conditional extension of the multivariate RV coefficient [140, 144]. We first give an overview of CRV and then construct a conditional independence test based on CRV. Finally, we discuss special considerations when applying CRV to microbial network construction.

4.3.1 Overview of CRV

Consider three multivariate random variables $\boldsymbol{Y} = (Y_1, \cdots, Y_p)^T \in \mathbb{R}^p$, $\boldsymbol{Z} = (Z_1, \cdots, Z_q)^T \in \mathbb{R}^q$, $\boldsymbol{X} \in \mathbb{R}^m$ with a joint distribution P_{YZX}. Suppose that we observe a sample $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P_{YZX}$. We are interested in assessing the conditional dependence between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ given $\boldsymbol{X}$. In general, $\boldsymbol{X}$ is high-dimensional in a graphical modeling setting. In the context of a genus-level microbial network, when we assess the association between two genera, $\boldsymbol{Y}$ represents the abundances of all species that belong to the first genus, $\boldsymbol{Z}$ represents the abundances of species that belong to the second genus, and $\boldsymbol{X}$ represents the abundances of all the other species in the microbial community that do not belong to the two genera under study.

First, we define a population CRV coefficient as a population measure of conditional dependence between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ given $\boldsymbol{X}$:

$$\mathrm{CRV}_{P_{YZX}}(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X}) = \frac{\mathrm{tr}(\Sigma_{YZ|X}\Sigma_{ZY|X})}{\sqrt{\mathrm{tr}(\Sigma_{YY|X}^2)\,\mathrm{tr}(\Sigma_{ZZ|X}^2)}}, \tag{4.1}$$

where $\Sigma_{YZ|X} = \mathbb{E}[(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}])(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}|\boldsymbol{X}])^T]$ and $\Sigma_{YY|X} = \mathbb{E}[(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}])(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}])^T]$. Here we can rewrite the numerator of $\mathrm{CRV}_{P_{YZX}}(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$ as $\mathrm{tr}(\Sigma_{YZ|X}\Sigma_{ZY|X}) = \sum_{j=1}^p \sum_{k=1}^q \mathbb{E}^2\left[\mathbb{C}\mathrm{ov}(Y_j, Z_k|\boldsymbol{X})\right]$. Hence the population CRV characterizes the conditional dependence by summing over the expected conditional covariance between individual elements in $\boldsymbol{Y}$ and $\boldsymbol{Z}$ given $\boldsymbol{X}$. Element-level associations can thus be accumulated into

a larger, more detectable signal. By averaging the conditional covariance over values of $\boldsymbol{X}$, CRV provides a global, rather than local, assessment of conditional dependence. It is easy to see that, if $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}$, then the numerator of $\mathrm{CRV}_{P_{YZX}}(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$ becomes zero, causing $\mathrm{CRV}_{P_{YZX}}(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$ to be zero. Similar to the RV coefficient, the denominator of $\mathrm{CRV}_{P_{YZX}}(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$ acts as a scaling factor to provide a standardized measure of association such that CRV takes values in $[0, 1]$.

Next, we define a vector-valued conditional mean function $\boldsymbol{\mu}_Y(\boldsymbol{x}) = (\mu_{Y_1}(\boldsymbol{x}), \cdots, \mu_{Y_p}(\boldsymbol{x}))^T := \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}]$ and define $\boldsymbol{\mu}_Z(\boldsymbol{x})$ similarly. Then, based on the sample $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i)\}_{i=1}^n$, the population CRV can be estimated by the empirical CRV statistic defined as:

$$\mathrm{CRV}_{P_n}(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X}) = \frac{\mathrm{tr}(S_{YZ|X} S_{ZY|X})}{\sqrt{\mathrm{tr}(S_{YY|X}^2)\,\mathrm{tr}(S_{ZZ|X}^2)}}, \tag{4.2}$$

where $S_{YY|X} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_Y(\boldsymbol{X}_i))(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_Y(\boldsymbol{X}_i))^T$ and $S_{YZ|X} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_Y(\boldsymbol{X}_i))(\boldsymbol{Z}_i - \hat{\boldsymbol{\mu}}_Z(\boldsymbol{X}_i))^T$. Here $\hat{\boldsymbol{\mu}}_Y$ and $\hat{\boldsymbol{\mu}}_Z$ are predictive models for the conditional means $\boldsymbol{\mu}_Y$ and $\boldsymbol{\mu}_Z$. To accommodate the high-dimensional nature of $\boldsymbol{X}$, as well as potential nonlinear relationships between $\boldsymbol{Y}$ (or $\boldsymbol{Z}$) and $\boldsymbol{X}$, we use common machine learning approaches for predictive modeling (e.g., random forest and lasso [145]) to derive these conditional mean estimators. In the next section, we will see that, when the conditional means are estimated sufficiently well, we could establish the asymptotic results for $\mathrm{CRV}_{P_n}(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$ and construct a test of conditional independence accordingly.

4.3.2 Conditional Independence Testing Based on CRV

We will use $\mathrm{CRV}_{P_n}(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$ as a empirical measure of conditional dependence and study its asymptotic behavior. Our results are built upon the previous work of Xiang et al. [141], which provides asymptotic results for a univariate measure of conditional dependence, SEcov, that is close in form to the CRV statistic. In particular, SEcov calculates a scaled expected conditional covariance between two univariate variables given a third multivariate variable.

When both $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are 1-dimensional, the CRV statistic is exactly equal to the squared SEcov statistic (see Section C.1 for details).

First, our asymptotic results will rely on the following assumption adapted from [141]:

Assumption 4.3.1. *Let $\hat{\mu}_{Y_j}$ and $\hat{\mu}_{Z_k}$ be estimators of the conditional means μ_{Y_j} and μ_{Z_k} for $j = 1, \ldots, p$ and $k = 1, \cdots, q$, based on the sample $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i)\}_{i=1}^n$. We assume that these estimators each fall in a P-Donsker class [116]. Furthermore, for each j and k, we assume that*

$$\int [\hat{\mu}_{Y_j}(\boldsymbol{x}) - \mu_{Y_j}(\boldsymbol{x})]^2 dP_{YZX}(\boldsymbol{x}) = o_p(n^{-1/2}),$$
$$\int [\hat{\mu}_{Z_k}(\boldsymbol{x}) - \mu_{Z_k}(\boldsymbol{x})]^2 dP_{YZX}(\boldsymbol{x}) = o_p(n^{-1/2}).$$

Assumption 4.3.1 mainly requires that the conditional mean estimators converge sufficiently fast to the true conditional means, which is satisfied by estimators derived from correctly specified low-dimensional parametric models such as linear and logistic regression and also high-dimensional methods such as lasso and neural network [141].

Next, since the CRV statistic is zero if and only if its numerator is zero, we focus on the numerator for the purpose of hypothesis testing. Let Ψ be the numerator of $\mathrm{CRV}_{P_{YZX}}(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$ and $\hat{\Psi}$ be the numerator of $\mathrm{CRV}_{P_n}(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$. We can establish the asymptotic distribution of $\hat{\Psi}$ under $H_0 : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}$ as:

Theorem 4.3.2. *Suppose Assumption 4.3.1 holds. Under $H_0 : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}$ such that $\Psi = 0$, we have*

$$n\hat{\Psi} \xrightarrow{d} \sum_{t=1}^{pq} \ell_t \tau_t,$$

where τ_t's are i.i.d. χ_1^2 variables and ℓ_t's are eigenvalues of the matrix $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]$, with $\boldsymbol{w} = \Big([Y_1 - \mu_{Y_1}(\boldsymbol{X})][Z_1 - \mu_{Z_1}(\boldsymbol{X})], [Y_1 - \mu_{Y_1}(\boldsymbol{X})][Z_2 - \mu_{Z_2}(\boldsymbol{X})], \cdots, [Y_p - \mu_{Y_p}(\boldsymbol{X})][Z_q - \mu_{Z_q}(\boldsymbol{X})]\Big)^T$ being the $(pq \times 1)$ vector of products between $Y_j - \mu_{Y_j}(\boldsymbol{X})$ and $Z_k - \mu_{Z_k}(\boldsymbol{X})$ for all j and k.

The proof of Theorem 4.3.2 is provided in Section C.2. Under H_0, the test statistic $n\hat{\Psi}$ follows a weighted sum of chi-square distributions asymptotically. Based on Theorem

4.3.2, we can construct a conditional independence test accordingly. In practice, the weights ℓ_t's can be estimated using their empirical counterparts. Specifically, we reject H_0 at a significance level α if $n\hat{\Psi}$ is larger than the $(1-\alpha)$-quantile of $\sum_{t=1}^{pq} \hat{\ell}_t \tau_t$, where $\hat{\ell}_t$'s are eigenvalues of the matrix $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}_i \boldsymbol{v}_i^T$, with $\boldsymbol{v}_i = \Big([Y_{1i} - \hat{\mu}_{Y_1}(\boldsymbol{X}_i)][Z_{1i} - \hat{\mu}_{Z_1}(\boldsymbol{X}_i)], [Y_{1i} - \hat{\mu}_{Y_1}(\boldsymbol{X}_i)][Z_{2i} - \hat{\mu}_{Z_2}(\boldsymbol{X}_i)], \cdots, [Y_{pi} - \hat{\mu}_{Y_p}(\boldsymbol{X}_i)][Z_{qi} - \hat{\mu}_{Z_q}(\boldsymbol{X}_i)]\Big)^T$. Following previous studies [54, 56] with similar forms of test statistic, we can efficiently approximate the mixture of chi-square variables, $\sum_{t=1}^{pq} \hat{\ell}_t \tau_t$, using Davies' exact method [117].

4.3.3 CRV for Microbial Network Construction

We next discuss several practical considerations when applying CRV to microbial network construction. First, to ensure valid inference, we need to perform appropriate normalization and transformation of microbial abundance data derived from high-throughput sequencing techniques before conditional independence testing [146, 147]. The centered log-ratio (CLR) transformation is a common transformation approach to address both differential read depths across individuals and compositionality of microbial data [62, 63, 37]. Let $\boldsymbol{r} = (r_1, \cdots, r_s)^T$ be the raw abundance vector for s microbial taxa observed in one individual. The CLR-transformed abundance vector is $\text{CLR}(\boldsymbol{r}) = (\log(r_1/g(\boldsymbol{r})), \cdots, \log(r_s/g(\boldsymbol{r})))^T$, where $g(\boldsymbol{r}) = [\Pi_{j=1}^{s} r_j]^{1/s}$ is the geometric mean of $\boldsymbol{r}$. In our analysis, we CLR-transform the raw microbial abundance data (with unit pseudo counts added to all entries) for each individual before applying the CRV test.

Second, since we apply CRV to test the conditional dependence between all pairs of microbial taxa, it is important to set a reasonable significance level in order to account for multiple testing. As a conservative approach, Bonferroni correction could be applied to control for the family-wise error rate. Alternatively, methods that control for the false discovery rate (FDR) could be used. As a general procedure to construct microbial networks in practice, we perform the CRV test on all pairs of taxa to obtain pair-specific p-values and only add an edge between a pair of taxa if their p-value is smaller than the pre-specified

significance level.

4.4 Simulation Studies

We conduct simulation studies to evaluate the performance of CRV and compare it against existing methods. We first evaluate CRV in conditional independence testing based on multivariate data and then examine the performance of CRV in recovering synthetic microbial networks.

4.4.1 Conditional Independence Testing

In this simulation, we examine the type I error rate and power of CRV in conditional independence testing. We consider multivariate variables $\boldsymbol{Y} = (Y_1, Y_2)^T \in \mathbb{R}^2$, $\boldsymbol{Z} = (Z_1, Z_2)^T \in \mathbb{R}^2$ and $\boldsymbol{X} \in \mathbb{R}^m$ and wish to assess whether $\boldsymbol{Y}$ is independent from $\boldsymbol{Z}$ given $\boldsymbol{X}$. To generate the observed data, we first simulate the conditional distribution $(\boldsymbol{Y}, \boldsymbol{Z})|\boldsymbol{X}$ according to a multivariate normal distribution with zero mean. We then introduce effects from $\boldsymbol{X}$ by adding linear or nonlinear functions of $\boldsymbol{X}$ to elements of $\boldsymbol{Y}$ and $\boldsymbol{Z}$, yielding the joint distribution $(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X})$.

To examine the type I error, we let $\mathbb{C}\text{or}(Y_j, Z_k|\boldsymbol{X}) = 0$ for $j, k \in \{1, 2\}$ such that $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z}|\boldsymbol{X}$ holds true. To examine the power, we consider two scenarios: (a) $\mathbb{C}\text{or}(Y_1, Z_1|\boldsymbol{X}) = \mathbb{C}\text{or}(Y_1, Z_2|\boldsymbol{X}) = \rho \in (0, 1)$, i.e., the conditional correlations between elements in $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are in the same direction; (b) $\mathbb{C}\text{or}(Y_1, Z_1|\boldsymbol{X}) = \rho$ and $\mathbb{C}\text{or}(Y_1, Z_2|\boldsymbol{X}) = -\rho$ with $\rho \in (0, 1)$, i.e., the conditional correlations are in the opposite direction. In addition, we set $\mathbb{C}\text{or}(Y_1, Y_2|\boldsymbol{X}) = \mathbb{C}\text{or}(Z_1, Z_2|\boldsymbol{X}) = 0.5$ in general. In the type I error simulation, we consider sample sizes of 100, 200 and 300. In the power simulation, we fix the sample size as 200.

We also consider two situations for the conditioning set $\boldsymbol{X}$. In Situation 1, we consider a low-dimensional $\boldsymbol{X}$. We let $X \in \mathbb{R}$ with $X \sim N(0, 1)$ and set $\mathbb{E}[Y_j|X] = \beta_{Y_j} \sin(X)$, $\mathbb{E}[Z_k|X] = \beta_{Z_k} \sin(X)$ for $j, k \in \{1, 2\}$, where $\beta_{Y_j}, \beta_{Z_k} \sim Unif(-1, 1)$. Here we use LOESS regression to estimate the conditional means in CRV. In Situation 2, we consider a high-dimensional $\boldsymbol{X}$. We let $\boldsymbol{X} \in \mathbb{R}^{100}$ with $\boldsymbol{X} \sim MVN(\boldsymbol{0}, \boldsymbol{I}_{100})$ and set $\mathbb{E}[Y_j|\boldsymbol{X}] = \beta_{Y_{j1}} X_1 + \cdots +$

$\beta_{Y_{j10}} X_{10}$, $\mathbb{E}[Z_k|\boldsymbol{X}] = \beta_{Z_{k1}} X_1 + \cdots + \beta_{Z_{k10}} X_{10}$ for $j, k \in \{1, 2\}$, where $\beta_{Y_{js}}, \beta_{Z_{ks}} \sim Unif(-1, 1)$. In this case, we use lasso to estimate the conditional means in CRV.

In addition to CRV, we consider several competing methods. In type I error simulation, the competing methods include the multivariate unconditional RV test [148] and the univariate conditional independence test, SEcov (with the same conditional mean estimators as in CRV). In power simulation, we compare CRV to SEcov only. While CRV and RV are applied to $\{(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{X}_i)\}_{i=1}^n$, SEcov is applied to the aggregated data $\{(Y_{1i} + Y_{2i}, Z_{1i} + Z_{2i}, \boldsymbol{X}_i)\}_{i=1}^n$.

Figure 4.2 shows the simulation results based on 1000 simulated data sets. As expected, RV has inflated type I errors since it does not adjust for the effects of $\boldsymbol{X}$, which correlates with both $\boldsymbol{Y}$ and $\boldsymbol{Z}$. In contrast, both CRV and SEcov are able to maintain valid type I error rates, especially in larger sample sizes. In power simulation, CRV achieves a greater power than SEcov in each situation, and the power gain is especially evident when the conditional correlations between elements in $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are of opposite directions. With SEcov, these opposite signals were likely canceled out through the aggregation of data, leading to a low power.

4.4.2 Synthetic Microbial Network

Next, we evaluate the performance of CRV in recovering synthetic microbial networks. We follow the strategy of Kurtz et al. [37] to generate synthetic microbial abundance data that reflects real-world microbial characteristics but with custom network topologies. A detailed simulation procedure in provided in Section C.3. Briefly, the desired network topology is first stored in a precision matrix, where its non-zero pattern corresponds to the adjacency matrix for the network under normality assumptions. Multivariate normal data are generated according to the specified precision matrix and then quantile-transformed to zero-inflated negative binomial (ZINB) distributions marginally, which serve as the simulated microbial counts. Here the parameters of the ZINB distributions are fitted from real microbiome data of the American Gut Project [149].

On the basis of the above simulation strategy, we further introduce a multivariate aspect

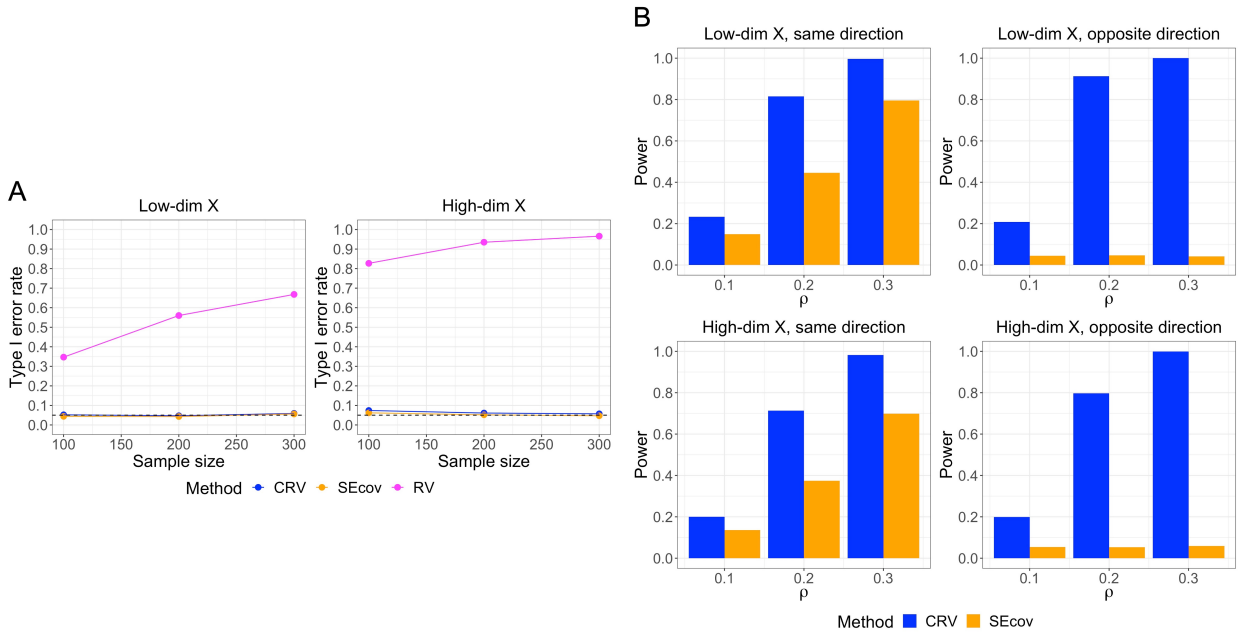

Figure 4.2: Empirical type I error rates (Panel **A**) and power (Panel **B**) of CRV at a significance level of 0.05 in conditional independence testing.

to the synthetic network in order to better evaluate our proposed CRV approach. Specifically, we consider a genus-level network where each genus is composed of multiple species, with species-level data available (see Section C.3 for details). As a general setting, we consider 40 genera (i.e., nodes) within the network, where each genus has 1-8 species. The number of true edges is set as 40, with three different possible topologies: band-like, cluster and scale-free (see Figure 3b from [37] for examples of these topologies). We further introduce conditional correlations of opposite directions among the species within 10%, 30%, or 50% of all genera (Figure C.2). The three topologies considered here are representative models of realistic ecological scenarios [37] and display different maximum network degrees (band $<$ cluster $<$ scale-free), which affects the difficulty of network recovery [150]. We set the sample size as 500 in general.

In our simulation, we apply CRV and four univariate competing methods for microbial network construction: SEcov, SPIEC-EASI with neighborhood selection based on the Mein-

shausen and Bühlmann approach (MB) [136], SPIEC-EASI with graphical lasso (glasso) [137, 37] and SPRING [38]. Both SPIEC-EASI and SPRING employ graphical modeling methods with regularization (either neighborhood selection or graphical lasso) to estimate a sparse network, where SPRING further accounts for zero-inflation in microbiome data. While CRV is applied to species-level count data, SEcov, SPIEC-EASI and SPRING are applied to count data aggregated at the genus level. For each method, the count data is CLR-transformed beforehand. In CRV and SEcov, we use random forest to estimate conditional means, as it is able to accommodate a moderately large conditioning set and potential nonlinear relationships between the microbial features.

Figure 4.3 shows example networks constructed by different methods in a single simulation under a cluster network topology, where heterogeneous conditional correlations are present in 50% of all genera. The false discovery rate (FDR) is controlled at 0.2 for a fair comparison across methods: we adjust for the FDR in CRV using the Benjamini–Hochberg method [151] and manually controlled for the FDR as close as possible in SEcov, SPIEC-EASI and SPRING according to the true graph, where the optimal tuning parameter is selected in SPIEC-EASI and SPRING (note that in practice, where the true graph is unknown, we cannot control for FDR in these two methods). Based on Figure 4.3, we see that CRV is able to identify a higher number of correct edges and produce a network closest to the true graph, while the networks produced by the other methods are less similar to the true graph.

To quantitatively measure how well the true graph is recovered, we calculate the Hamming distance for different methods, defined as the number of edges that disagree between the constructed network and the true graph. Table 4.1 shows the average Hamming distance over 10 simulations for different methods under different network topologies, when heterogeneous conditional correlations are introduced to 50% of all genera. The FDR is controlled at 0.2 as described before. For each topology, CRV has the lowest Hamming distance on average among all methods, confirming that the networks produced by CRV indeed have the most resemblance to the true graph.

We further calculate the true positive rates (TPRs) and false positive rates (FPRs) in

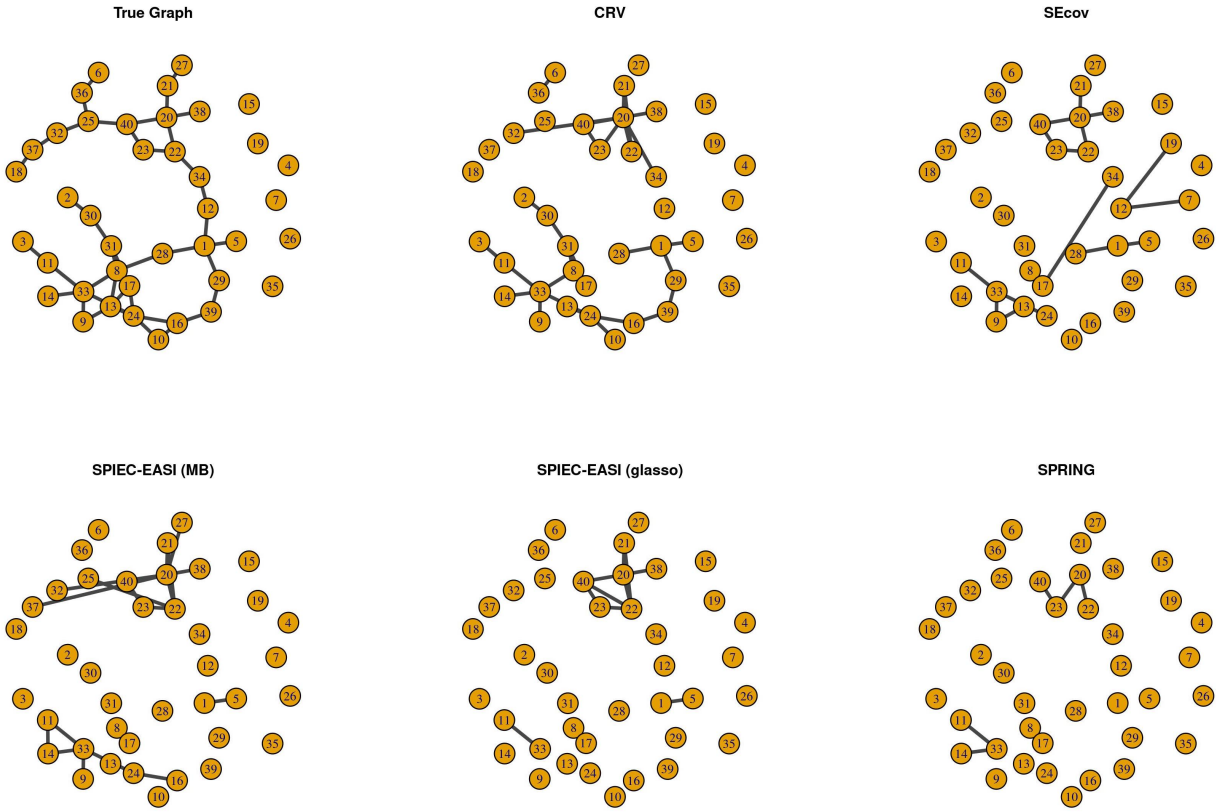

Figure 4.3: Performance of different methods for recovering a cluster-type synthetic microbial network, with false discovery rate controlled at 0.2. Conditional correlations of opposite directions are present among species within 50% of all genera.

predicting the edges. Figure 4.4 shows the average receiver operating characteristic (ROC) curves of different methods over 10 simulations in recovering microbial networks of different topologies, where heterogeneous conditional correlations are present in 50% of all genera. For CRV and SEcov, the TPRs and FPRs are calculated at various significance levels; for SPIEC-EASI and SPRING, the TPRs and FPRs are calculated at various choices of the tuning parameter. Based on Figure 4.4, for all network topologies, CRV achieves a better performance than other methods in recovering the true network in terms of TPR and FPR. When compared between topologies, CRV has the best performance in recovering band-like networks, followed by cluster networks and performs worst in scale-free networks. This is

Table 4.1: Average Hamming distance between the constructed network and the true graph for different methods under different topologies, when conditional correlations of opposite directions are present among species within 50% of all genera.

Method	Band	Cluster	Scale-free
CRV	19.8	26.0	36.2
SEcov	37.1	35.5	38.9
SPIEC-EASI (MB)	37.9	36.5	52.0
SPIEC-EASI (glasso)	38.5	36.8	51.1
SPRING	37.9	36.4	58.4

The false discovery rate is controlled at 0.2.

expected since scale-free networks tend to have the highest maximum network degree, making recovery of the true network more difficult.

Similar patterns in the Hamming distance and ROC curves are observed when heterogeneous conditional correlations are present in 10% or 30% of all genera (see Table C.1 and C.2; Figure C.3 and C.4), where CRV again performs better than the competing univariate methods. We note that, as the proportion of heterogeneous relationships increases, the performance gain of CRV becomes more evident. Overall, our simulation studies show an advantage of using CRV in recovering microbial networks at a taxonomic level where sub-taxa data are available and heterogeneous relationships are present among the sub-taxa.

4.5 Data Application

We apply our proposed CRV framework to construct microbial networks based on vaginal microbiome data from the PIN study. As described in Section 4.2, our sample includes 652 women (375 White women and 277 Black women) who experienced a term birth. The vaginal microbiome data were collected during the second trimester (between 24 and 29 weeks) of pregnancy. We have initial species-level abundance data on 465 species that belong to 137 genera.

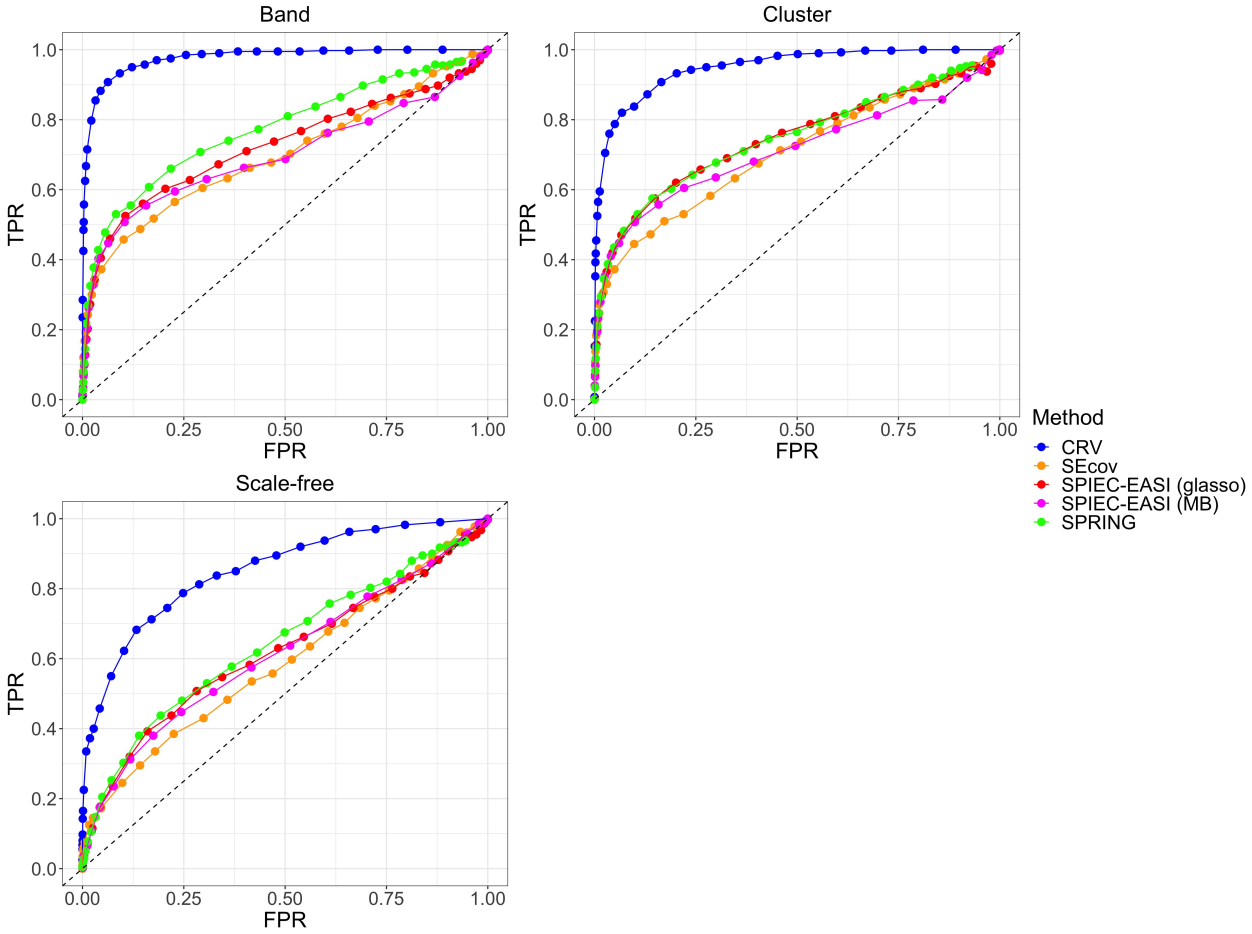

Figure 4.4: ROC curves of different methods in recovering synthetic microbial networks of different topologies, averaged over 10 simulations. Conditional correlations of opposite directions are present among species within 50% of all genera.

Consistent with previous microbial network studies [38, 42], before constructing the networks, we perform standard data filtering to focus on subjects with high sequencing depths (i.e., total microbial abundances) and genera that are sufficiently common among the subjects. Our filtering involves two steps: (1) we exclude subjects whose sequencing depths are $\leq$ 10,000; (2) we exclude all genera that are present in $<$ 25% of subjects, based on the genus-level abundance information. As a consequence, for our analysis, we obtain abundance data on 186 species from 24 genera in 647 subjects (373 White women and 274 Black

women).

As detailed in the following subsections, we first construct a genus-level microbial network based on the combined sample of 647 subjects; we then construct networks for Black and White women separately and compare the networks between the two race groups; lastly, we investigate conditional dependence between individual species based on the network constructed from the combined sample. We use an FDR threshold of 0.2 in the CRV test throughout our network construction based on the PIN data.

To characterize the networks quantitatively and understand the roles of individual genera within the microbial community, we calculate several network centrality measures, including degree, betweenness and eigenvector centrality, for each genus in the constructed networks. Briefly, degree centrality for a node is the number of nodes connected to it [152]. Betweenness centrality for a node is the fraction of times that node lies on the shortest path between all other nodes [152]; a node is central based on betweenness centrality if it is able to connect sub-networks [153]. Eigenvector centrality for a node is its eigenvector entry that corresponds to the largest eigenvalue derived from eigendecomposition of the adjacency matrix [154, 155]; a node is central based on eigenvector centrality if it is connected to other nodes that are central themselves. Nodes with high centrality values can be considered as hubs in the network, which correspond to genera that play important roles in the microbial community [156, 157]. All centrality measures are normalized to be within the range of [0, 1].

4.5.1 Overall Vaginal Microbial Network

To understand the global structure and genus-genus interactions of the vaginal microbial community during pregnancy, we construct a genus-level vaginal microbial network based on the combined sample of 647 subjects, as shown in Figure 4.5. By inspecting the network structure and centrality measures of each node, we see that *Lactobacillus* is the hub genus within the network, with the highest degree, betweenness and eigenvector centrality among all genera. This is consistent with previous studies on vaginal microbiota: a healthy vaginal microbiome tends to be dominated by *Lactobacillus* bacteria [124]; furthermore, dur-

ing normal pregnancy, the vaginal microbiome becomes even more *Lactobacillus*-dominated compared to non-pregnant women [129]. The central status of *Lactobacillus* in our current network agrees with its crucial role in maintaining vaginal homeostasis.

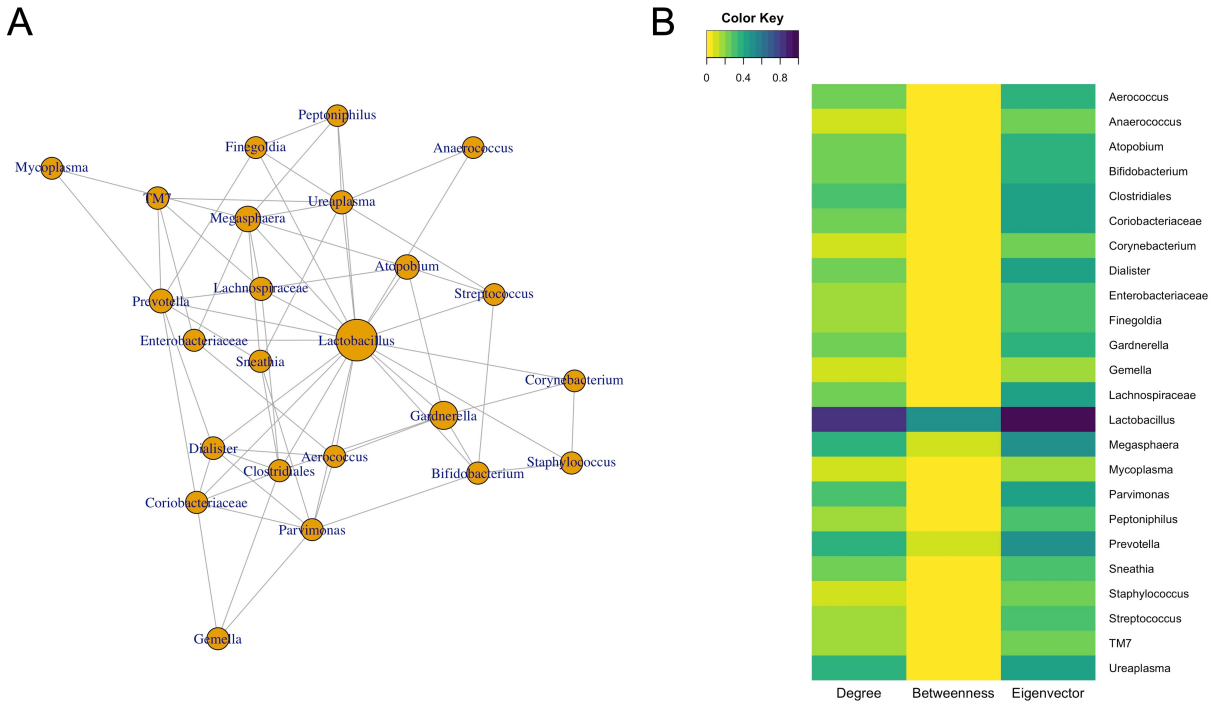

Figure 4.5: Genus-level vaginal microbial network based on the combined sample of 647 subjects of PIN study, with false discovery rate controlled at 0.2. Panel **A**: Constructed genus-level network; the sizes of the nodes are proportional to their genus-level microbial abundance among all subjects. Panel **B**: Heatmap of three network centrality measures for each genus in the constructed network.

Lactobacillus is connected to 19 other genera. Many of these genera, such as *Gardnerella*, *Atopobium*, *Prevotella*, *Megasphaera* and *Clostridiales*, harbor pathogenic species that are involved in bacterial vaginosis (BV), the most common vaginal infection among women of reproductive age [158, 159]. These connections are consistent with known functions of *Lactobacillus*. Through lactic acid production, lactobacilli contribute to an acidic vaginal environment that prevents the growth of pathogens [160, 125]. Lactobacilli can also

protect against pathogens by effectively competing for resources with them in the microenvironment (i.e., competitive exclusion) [161]. On the other hand, BV is characterized by excessive growth of the aforementioned pathogenic bacteria and a reduction of lactobacilli [162] in the vaginal microbiome. Therefore, our constructed network successfully identifies the important interactions between *Lactobacillus* and the other pathogenic bacterial genera.

Overall, based on the combined sample, our network displays a dominant role of *Lactobacillus* in the vaginal microbial community during the second trimester of a normal pregnancy.

4.5.2 Comparison between Black and White Women

Due to previously reported racial differences in the vaginal microbiome, we next construct separate networks in Black and White women and compare the networks between the two groups qualitatively. Figure 4.6 shows the genus-level vaginal microbial networks based on 373 White women and 274 Black women, respectively. Figure 4.7 shows the corresponding centrality measures of each node in the networks. Both race-specific networks are considerably sparser than the network based on the combined sample (Figure 4.5), which is likely due to power loss of the CRV test from reduced sample sizes of the subgroups. In White women, *Lactobacillus* still has the highest degree, betweenness and eigenvector centrality among all genera (Figure 4.7: Panel **A**), assuming a dominant role in the vaginal microbial community. In Black women, however, *Lactobacillus* no longer has the central status; instead, *Megasphaera* and *Lachnospiraceae* appear to have more important roles based on the centrality measures (Figure 4.7: Panel **B**). Black women also have a greater number of genus-genus interactions than White women (network density: 0.069 in Black women vs. 0.054 in White women), despite the smaller sample size.

The differences in microbial networks between Black and White women are consistent with previously identified racial differences in vaginal microbiome composition. While White women tend to have a *Lactobacillus*-dominated vaginal microbiome, Black women are more likely to harbor a diverse microbial profile [130, 129]. Correspondingly, our constructed

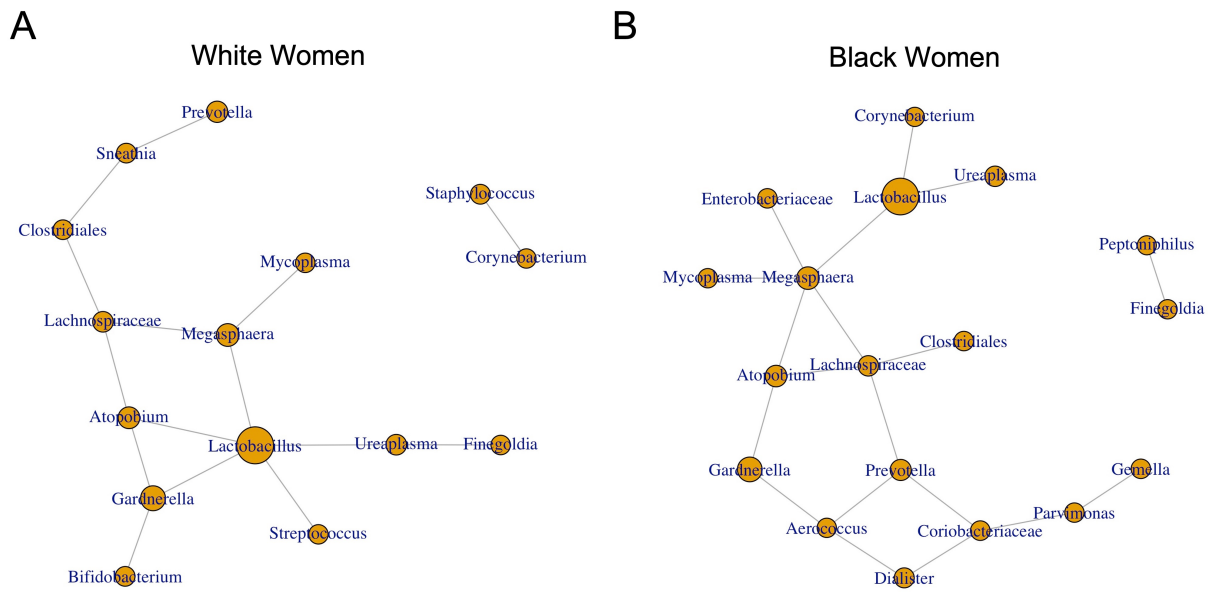

Figure 4.6: Genus-level vaginal microbial networks in White women (Panel **A**; $n = 373$) vs. Black women (Panel **B**; $n = 274$) of PIN study, with false discovery rate controlled at 0.2. The sizes of the nodes are proportional to their genus-level microbial abundance in the combined sample.

network in Black women displays a greater network density and a lack of centrality by *Lactobacillus* compared to White women. In particular, the two hub genera in Black women, *Megasphaera* and *Lachnospiraceae*, have been associated with BV and other adverse reproductive health conditions: *Megasphaera* phylotypes 1 and 2 are associated with BV, preterm birth and trichomoniasis [158, 163]; the *Lachnospiraceae* family harbors bacterial vaginosis-associated bacterium-1 (BVAB1), which is associated with BV and HPV infection [158, 128]. Therefore, compared to *Lactobacillus*'s dominance in the vaginal microbiome of White women, pathogenic genera appear to participate in more interactions within the vaginal microbial community of Black women.

Interestingly, one previous study [129] observes that the vaginal microbiome shifts significantly during early pregnancy towards *Lactobacillus*-dominated profiles in Black women. Based on our network analysis, however, it appears that there is still considerable diver-

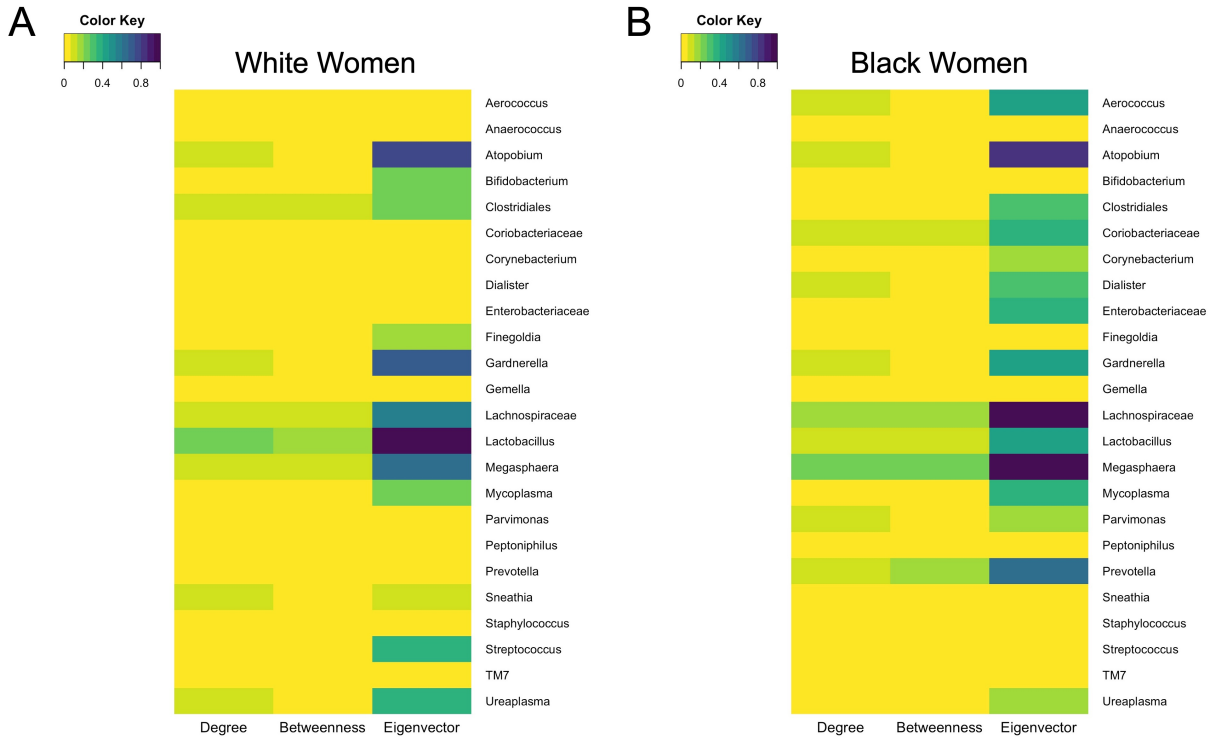

Figure 4.7: Heatmaps of network centrality measures for each genus in genus-level vaginal microbial networks based on White women (Panel **A**; $n = 373$) vs. Black women (Panel **B**; $n = 274$) of PIN study.

sity and active interactions among pathogenic taxa in Black women even during the second trimester of pregnancy.

As the power of the CRV test increases with sample size, constructing a network based on a larger sample size will typically result in a greater number of edges. Due to unequal sample sizes between Black and White women, a direct comparison between the two race-specific networks could be subject to bias caused by the power difference in edge detection. To assess the robustness of our results, we further construct a network based on a random subset of White women, with the same sample size as Black women ($n = 274$), and compare it against the network based on the entire set of Black women. The resulting networks and heatmaps

for centrality measures of each node are reported in Figure C.5 and C.6. The network based on a reduced subsample of White women is even sparser than before (network density = 0.029), and *Lactobacillus* still maintains its dominant role among all genera. Therefore, our previous results on racial differences in network structure still hold.

4.5.3 Conditional Dependence between Individual Species

Individual species within the same genus could have distinct functions and thus have heterogeneous interactions with species in other genera. We further investigate the conditional dependence between individual species in different genera, based on the network constructed from the combined sample ($n = 647$). Due to the important role of *Lactobacillus* in vaginal health and previously reported distinct characteristics of individual *Lactobacillus* species [125], here we focus on species in the *Lactobacillus* genus and assess their conditional correlations with individual species in other genera that are connected to *Lactobacillus* in our constructed network (Figure 4.5). We use the SEcov test [141], the univariate version of CRV, to test the conditional dependence based on a significance level of 0.05. For each pair of species-species association, we condition on all the other species within the microbial community that do not belong to the same genera as the species under analysis. As the SEcov statistic provides the direction of conditional association between two univariate variables, we are able to identify potential heterogeneous associations between individual species.

We report selected conditional associations between five common *Lactobacillus* species [125] and species in four pathogenic genera: *Aerococcus*, *Gardnerella*, *Megasphaera* and *Ureaplasma*, where associations of opposite directions are observed among the *Lactobacillus* species. The results are shown in Table 4.2, where the SEcov statistics are reported as measures of conditional correlations.

For *Aerococcus christensenii*, *Megasphaera OTU70 type1* and *Ureaplasma cluster23*, while *L. iners* exhibits a positive association, all the other common *Lactobacillus* species show negative associations. This pattern is consistent with previously reported differences between *L. iners* and other *Lactobacillus* species. While vaginal microbiome profiles dominated by

Table 4.2: Selected conditional correlations, as measured by SEcov statistics, between common *Lactobacillus* species and pathogenic species.

	Aerococcus christensenii	*Gardnerella vaginalis*	*Megasphaera OTU70 type1*	*Ureaplasma cluster23*
L. gasseri	-0.09	0.14	-0.14	–
L. iners	0.13	-0.17	0.16	0.15
L. jensenii	-0.08	–	–	-0.09
L. crispatus	–	-0.14	–	–
L. vaginalis	–	-0.09	-0.18	–

The reported conditional correlations are based on a significance level of 0.05.
"–" indicates lack of significance.

L. crispatus, *L. gasseri* and *L. jensenii* are associated with healthy vaginal states and a stable vaginal community [125], a *L. iners*-dominated vaginal microbiome tends to be less stable and more often associated with BV and vaginal dysbiosis [139, 164]. Genetic and biochemical analyses also suggest that *L. iners* contains both probiotic and pathogenic features [139]. Our results thus further support the possibility that *L. iners* maintains distinct positive interactions or co-existence with certain pathogenic bacteria, in contrast to the other protective *Lactobacillus* species that mainly inhibit the growth of pathogens.

Interestingly, for *Gardnerella vaginalis*, while *L. gasseri* shows a positive association, all the other *Lactobacillus* species, including *L. iners*, display negative associations. *G. vaginalis* is a predominant species associated with BV. Previous studies indicate that *L. iners* might enhance the adhesion of specific *G. vaginalis* strains to cervical epithelial cells [165]. Here our analysis suggests that *L. gasseri* might also possess some features that allows it to cohabitate with *G. vaginalis*, which will be interesting for further investigation.

Based on the species-level conditional dependence analysis, we confirm the presence of heterogeneous interactions among individual *Lactobacillus* species due to their distinct functions. This again showcases the advantage of using a multivariate approach to construct a

higher-level network via the CRV test.

4.6 Discussion

Microbial association networks are useful tools to understand the global structure of human microbial communities and detect important taxon-taxon interactions within the community. In this work, we have proposed a novel approach to construct microbial networks at a taxonomic level with sub-taxa available, via the multivariate conditional independence test based on CRV. The proposed CRV test achieves a superior performance in recovering simulated microbial networks compared to existing unvariate methods for microbial network construction, especially when heterogeneous associations are present among the sub-taxa under the taxonomic level of interest. By applying the CRV test to vaginal microbiome data of the PIN study, we have constructed microbial networks consistent with previous knowledge on vaginal microbiota and detected racial differences in network structure between Black and White women.

Based on the combined sample of the PIN study, *Lactobacillus* shows a central role in the genus-level microbial network with the highest number of interactions among all genera, consistent with the importance of *Lactobacillus* in maintaining vaginal health. When further compared between Black and White women, we have discovered racial differences in network structure: while *Lactobacillus* maintains its central status in White women, pathogenic genera such as *Megasphaera* and *Lachnospiraceae* are more dominant and associated with more interactions in Black women. Such network differences are consistent with previously reported racial differences in vaginal microbiome composition, and we have confirmed that these racial differences still exist during the second trimester of a normal pregnancy. It is important to further understand the factors that contribute to these racial differences in vaginal microbiota, which will help us develop effective strategies to improve reproductive health outcomes.

Our multivariate CRV test has the advantage of utilizing taxon abundance at a greater resolution without the need for data aggregation. By accumulating weak, species-level signals

into a stronger, genus-level association, we can improve the power in detecting edges while accounting for potential heterogenous relationships at the species level. In the PIN study, we have shown that such heterogenous associations indeed exist among the species due to their distinct functions and activities. In addition, since CRV is developed as a hypothesis test, we can construct networks at a given significance level, without the need to choose tuning parameters as in previous microbial construction approaches that rely on regularized prediction methods such as neighborhood selection and graphical lasso.

One limitation of our CRV approach is that the number of edges detected appears to be sensitive to power loss due to smaller sample sizes, as seen in the race-specific analysis of the PIN study. As we perform hypothesis testing between each pair of taxa, there could be a large multiple testing burden, and the resulting significance level after multiple-testing adjustment could be very stringent to achieve for a smaller sample size. Therefore, a moderate to large sample size is typically recommended.

Overall, we have developed a promising novel framework for microbial network construction based on multivariate conditional independence testing. The vaginal microbial networks constructed using CRV contribute to our understanding of the vaginal microbial community during pregnancy and will be useful for further investigation of the impact of vaginal microbiome on reproductive health. Our proposed CRV approach will also be a useful tool for network analysis of other types of microbiome data, such as human microbiome from other body sites and environmental microbiome.

Chapter 5

MENDELIAN RANDOMIZATION WITH A MICROBIAL EXPOSURE

5.1 Introduction

The microbiome is an integral part of the human body. In recent years, there has been a great effort to investigate the role of the human microbiome in various health outcomes, often via observational studies where the association between the microbial features and the outcome is assessed. While such microbial association studies have revealed important connections between the microbiome and various phenotypes such as BMI [10], blood pressure [166] and conditions like inflammatory bowel disease [11] and type 2 diabetes [12], the presence of unmeasured confounders in observational studies prevents us from directly establishing causal relationships between specific microbial features and these outcomes. With advances in next-generation sequencing technology and popularity of genome-wide association studies (GWASs), Mendelian randomization (MR) emerges as a feasible framework to estimate and test the causal effect of an microbial exposure on an outcome based on observational data, by using genetic variants as instruments [44].

MR is a specific type of instrumental variable (IV) analysis, which originated in econometrics [167] and was later popularized in genetic epidemiology [168]. The idea of MR is analogous to that of a randomized controlled trial (RCT). During gamete formation from meiosis, offsprings randomly inherit one copy of genetic materials from each of their parents, based on Mendel's laws; this random segregation of genetic materials is similar to the randomization step in RCTs. If certain genetic variants are associated with the exposure of interest, we can then use these variants as a proxy (i.e., an instrumental variable, or IV) to investigate the causal effect of that exposure on the outcome. Since the randomly inherited

genetic variants are not associated with (most) confounders in the exposure-outcome relationship, differences in the outcome among individuals who carry distinct variant alleles can thus be attributed to the difference in the exposure.

Recently, an increasing number of MR studies has focused on the gut microbiome as an exposure and investigated the causal effect of gut microbial features on health outcomes such as metabolic traits and complex diseases [45, 13, 14]. These microbial features are typically the abundances of individual microbial taxa. Microbial abundances obtained from microbial sequencing techniques such as 16S rRNA sequencing and metagenomic sequencing are count data with unique characteristics, often with overdispersion and zero-inflation [169]. However, existing MR methods are usually based on a continuous exposure, without accommodation for the count nature of the microbial data or the potential nonlinear relationships between the microbial abundance and the genetic instruments as well as covariates.

The majority of existing MR studies with a microbial exposure is conducted in a two-sample setting [45, 13, 170, 171], where the genetics-microbe association and the genetics-outcome association are estimated from two separate data sources [47]. Often times, only GWAS summary statistics, rather than individual-level data, are available. Typically, the causal estimates are obtained by combining ratio-of-coefficients estimates [47] derived from individual genetic IVs, using two-sample MR methods such as the inverse variance weighted (IVW) estimator [172] and MR-Egger [173]. Since only summary statistics are available and the summary statistics from microbiome GWAS are often derived from linear models [92, 15], it is difficult to further incorporate nonlinear models for the genetics-microbe relationship in this setting.

One previous MR study considered a microbial exposure in the one-sample setting [14], where individual-level data on genetic variants, gut microbial features and the outcomes were available. This study used the two-stage least squares (2SLS) method, a standard approach in IV analysis [46], for the MR analysis. In 2SLS, through linear regression, the exposure is regressed on the IVs to obtain fitted values of the exposure in the first stage, and the outcome is then regressed on the fitted exposure values to provide a causal estimate in the

second stage. When applied to a microbial exposure such as taxon abundance, however, this approach does not account for the count nature of microbial data. 2SLS also assumes a linear relationship between the microbial abundance and the genetic IVs. While 2SLS can still provide valid inference when there is misspecification in its first-stage model, methods that account for unique characteristics of microbial data might achieve a better efficiency.

In this work, we propose a novel statistical approach to conduct MR analysis with a microbial exposure and a continuous outcome in the one-sample setting. We adapt an existing IV method, two-stage least squares with generated instruments (2SLS-GI) [46], to incorporate nonlinear models that account for characteristics of microbial count data including overdispersion and zero-inflation as well as nonlinear relationships between the microbial abundances and genetic IVs and other covariates. The 2SLS-GI method can be considered as an extension of 2SLS. It has been applied in the econometrics literature [174], but, to our knowledge, no application is seen in the MR literature so far. In the case of a microbial count exposure, we demonstrate the power gain of 2SLS-GI in detecting causal effects compared to other IV methods via simulation studies. Furthermore, we apply 2SLS-GI to the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) [23, 24] and investigate the causal effect of different gut microbial taxa on blood pressure, where 2SLS-GI identifies a greater number of significant causal relationships than competing methods.

The rest of the article is organized as follows. Section 5.2 provides an introduction of 2SLS-GI with accommodation for a microbial count exposure. In Section 5.3, we conduct simulation studies to evaluate the performance of 2SLS-GI in MR analysis with a microbial exposure, under different data generation scenarios for microbial abundances. In Section 5.4, we apply 2SLS-GI to evaluate the causal effect of different gut microbial genera on systolic and diastolic blood pressure based on the HCHS/SOL study. Finally, we conclude with a discussion in Section 5.5.

5.2 Methods

We provide an introduction of two-stage least squares with generated instruments (2SLS-GI), an IV method that can be employed to account for specific characteristics of a microbial count exposure. First, we introduce the general assumptions on genetic IVs in MR analysis and give an overview of 2SLS, the standard approach in one-sample MR analysis. Next, we explain the extension of 2SLS to 2SLS-GI and illustrate how 2SLS-GI is adapted to accommodate a microbial exposure.

5.2.1 Assumptions on Genetic Instruments

Let $X_E \in \mathbb{R}$ denote the exposure of interest, $Y \in \mathbb{R}$ denote a continuous outcome, $G \in \mathbb{R}^m$ denote the genetic instruments (IVs) and U denote the unmeasured confounders in the exposure-outcome relationship. In the context of MR analysis with a microbial exposure, X_E typically corresponds to the abundance of a particular microbial taxon, e.g., a bacterial genus. Often, we are able to identify multiple genetic IVs that are associated with the exposure such that $m \geq 1$.

The core assumptions on genetic IVs are illustrated in Figure 5.1. The first assumption requires that G is not associated with U, which ensures validity of the genetic IVs. The second assumption requires that G is associated with X in some way, which ensures relevance of the genetic IVs. Finally, the third assumption requires that G is not associated with Y conditional on X, which ensures that the genetic IVs can only affect the outcome through the exposure of interest, but not through other pathways (i.e., horizontal pleiotropy).

Typically, the validity of these IV assumptions depends on prior biological knowledge. In particular, the first and third assumptions are hard to verify in practice. In this work, however, our focus is not on dealing with invalid genetic IVs, and we will assume that these assumptions hold true throughout our description of the proposed approach.

The specific definition of association in the above assumptions can vary depending on the context. Most strictly, no association between the genetic IVs and the confounders can

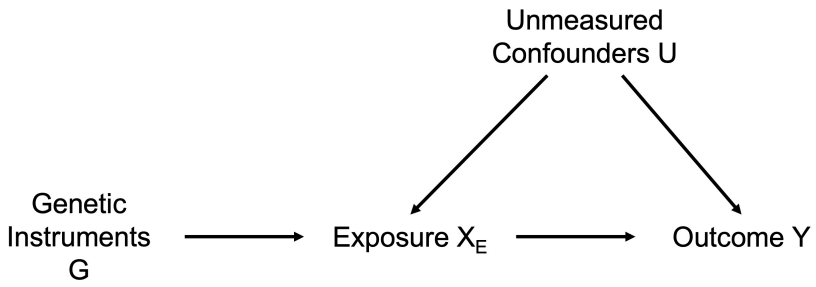

Figure 5.1: Graphical representation of assumptions on genetic instruments in Mendelian randomization.

be defined as statistical independence between G and U. However, relaxed criteria such as uncorrelatedness and conditional mean independence will satisfy in many IV methods, as described in the following sections.

5.2.2 Overview of 2SLS

Next, we give a brief overview of the standard 2SLS method. We now introduce some new variables and notations. Let $C \in \mathbb{R}^k$ represent a set of measured covariates that we want to adjust in the exposure-outcome relationship. In the context of MR analysis, adjusting for population structure as a covariate is necessary to ensure validity of the genetic IVs. Since different populations tend to vary in both genetics and various health outcomes, the IV assumptions from Section 5.2.1 can only hold when conditioning on population structure [47]. In addition, adjusting for other known confounders in the exposure-outcome relationship can also improve efficiency in causal estimation.

Let $X := (1, X_E, C^T)^T \in \mathbb{R}^{k+2}$ and $Z := (1, G^T, C^T)^T \in \mathbb{R}^{k+m+1}$. Here Z can be considered as a more general definition of instruments, where the genetic variants G serve as IVs for the exposure X_E and the measured covariates C and the constant 1 serve as their own IVs. We consider the following linear model for the exposure-outcome relationship:

$$Y = X^T\boldsymbol{\beta} + u = \beta_0 + X_E\beta_1 + C^T\boldsymbol{\beta}_2 + u, \tag{5.1}$$

where $u \in \mathbb{R}$ is the error term, $\mathbb{E}[u] = 0$, $\mathbb{C}\text{or}(X_E, u) \neq 0$ and $\mathbb{C}\text{or}(C, u) = \mathbf{0}$. Here β_1 corresponds to the causal effect of X_E on Y and is our parameter of interest. Since X_E is associated with the error term, implying unmeasured confounding between X_E and Y, the usual ordinary least squares (OLS) estimator will produce a biased causal estimate.

Consider the observed data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ in n individuals. Let $\boldsymbol{X} \in \mathbb{R}^{n\times(k+2)}$, $\boldsymbol{Y} \in \mathbb{R}^n$ and $\boldsymbol{Z} \in \mathbb{R}^{n\times(k+m+1)}$ denote the observed data matrix for X, Y and Z. The 2SLS estimator for $\boldsymbol{\beta}$ based on the observed data is defined as

$$\hat{\boldsymbol{\beta}}_{\text{2SLS}} = (\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{Y}, \tag{5.2}$$

where $\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}$. The causal parameter of interest, β_1, is estimated by $\hat{\beta}_{1,\text{2SLS}}$, the second component of $\hat{\boldsymbol{\beta}}_{\text{2SLS}}$. The procedure for obtaining the 2SLS estimator can alternatively be described as the following two steps:

1. Perform linear regression of X_E on genetic instruments G and covariates C to obtain the fitted exposure $\hat{X}_E$.

2. Perform linear regression of Y on $\hat{X}_E$ and C to obtain the causal estimate $\hat{\beta}_{1,\text{2SLS}}$.

The properties of the 2SLS estimator have been well studied. Suppose the following three assumptions hold true: (1) $\mathbb{E}[Zu] = \mathbf{0}$ (validity of IVs); (2) $\mathbb{E}[ZX^T]$ has full rank (relevance of IVs); and (3) $\mathbb{E}[u^2|Z] = \sigma^2$ (homoscedasticity). Then we can obtain asymptotic normality of the 2SLS estimator $\hat{\boldsymbol{\beta}}_{\text{2SLS}}$ (see Theorem 5.2 of [46]), which can be used to perform hypothesis testing on the causal parameter β_1. Under the same set of assumptions, the 2SLS estimator is efficient in the class of all IV estimators using instruments that are linear in Z (Theorem 5.3 of [46]).

5.2.3 Extension to 2SLS-GI

The 2SLS estimator has the best efficiency among IV estimators with instruments that are linear in Z. However, to account for potential nonlinear relationships between the microbial

exposure and the genetic IVs and covariates, we can further consider new instruments that are nonlinear functions of Z to improve efficiency. To this end, we interpret 2SLS as a special case of the generalized method of moments (GMM) estimator [175]. GMM is a semiparametric IV method, where the no-association assumption between the IVs and the error term (i.e., the first IV assumption from Section 5.2.1) is used to form a set of population moment conditions; the corresponding set of sample estimating equations is then solved to estimate the causal parameter of interest. Under homoscedasticity, 2SLS is exactly equivalent to the GMM estimator obtained from the population moment conditions: $\mathbb{E}[Z(Y - X^T\boldsymbol{\beta})] = \mathbb{E}[Zu] = \mathbf{0}$.

Instead of the orthogonality assumption $\mathbb{E}[Zu] = \mathbf{0}$ used in 2SLS, we can consider a stronger, zero conditional mean assumption: $\mathbb{E}[u|Z] = 0$. Under this assumption and homoscedasticity of the error terms, it has been shown that the optimal choice of instruments for obtaining the smallest asymptotic variance from GMM estimation is $\mathbb{E}[X|Z]$, rather than Z itself, provided that the matrix $\mathbb{E}\left[\mathbb{E}[X|Z]X^T\right]$ has full rank (Theorem 8.5 of [46]). Using $\mathbb{E}[X|Z]$ as the new instrument, the population moment condition becomes $\mathbb{E}\left[\mathbb{E}[X|Z]u\right] = \mathbf{0}$. The GMM estimator obtained by solving the corresponding sample estimating equations is equivalent to a 2SLS estimator with $\mathbb{E}[X|Z]$ as its instrument. Such an estimator allows us to gain more efficiency in causal estimation compared to the original 2SLS estimator using Z as the instrument.

As the measured covariates C can serve as their own IVs, the only unknown component of $\mathbb{E}[X|Z] = \mathbb{E}[(1, X_E, C^T)^T|Z]$ is the conditional mean for the exposure of interest, $\mathbb{E}[X_E|Z]$, which needs to be estimated in practice. In particular, when X_E is the microbial abundance, we can utilize count data models that are commonly applied to microbiome data to model the relationship between X_E and Z and estimate $\mathbb{E}[X_E|Z]$. Specific model choices are discussed in Section 5.2.4. Suppose that, based on these models, we obtain the estimated conditional mean, $\hat{\mathbb{E}}[X|Z]$. We can then construct a 2SLS estimator using $\hat{\mathbb{E}}[X|Z]$ as the instrument. Conveniently, previous work has shown that, under assumptions that are satisfied in many scenarios, whether $\mathbb{E}[X|Z]$ or $\hat{\mathbb{E}}[X|Z]$ is used as the instrument in 2SLS does not affect the resulting asymptotic distribution.

The above framework can be denoted as two-stage least squares with generated instruments (2SLS-GI), which has been discussed in detail by Wooldridge in Section 6.1.2 of [46]. We now formally introduce the 2SLS-GI estimator and present its asymptotic distribution. We consider the same notations and exposure-outcome model from (5.1) as before. Define a new instrument $Z^* := \boldsymbol{g}(Z, \boldsymbol{\lambda})$, where $\boldsymbol{g}(\cdot, \boldsymbol{\lambda})$ is a known vector-valued function but the parameter $\boldsymbol{\lambda}$ is unknown. Specifically, $Z^* \in \mathbb{R}^{k+2}$ is composed of the constant term, the measured covariates $C \in \mathbb{R}^k$ and a scalar function $g_E(Z, \boldsymbol{\lambda})$ that is assumed to represent the conditional mean model $\mathbb{E}[X_E|Z]$, i.e., $Z^* = \boldsymbol{g}(Z, \boldsymbol{\lambda}) = (1, g_E(Z, \boldsymbol{\lambda}), C^T)^T$. We further define the generated instrument: $\hat{Z}^* := \boldsymbol{g}(Z, \hat{\boldsymbol{\lambda}}) = (1, g_E(Z, \hat{\boldsymbol{\lambda}}), C^T)^T$, where $\hat{\boldsymbol{\lambda}}$ is an estimator of $\boldsymbol{\lambda}$. Let $\hat{\boldsymbol{Z}}^* \in \mathbb{R}^{n\times(k+2)}$ denote the generated instrument matrix for all n individuals. The resulting 2SLS-GI estimator is defined as:

$$\hat{\boldsymbol{\beta}}_{\text{2SLS-GI}} = (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \boldsymbol{Y}, \tag{5.3}$$

where $\tilde{\boldsymbol{X}} = \hat{\boldsymbol{Z}}^* \big[(\hat{\boldsymbol{Z}}^*)^T \hat{\boldsymbol{Z}}^*\big]^{-1} (\hat{\boldsymbol{Z}}^*)^T \boldsymbol{X}$. Assume that: (1) $\mathbb{E}[u|Z] = 0$; (2) $\hat{\boldsymbol{\lambda}}$ is $\sqrt{n}$-consistent for $\boldsymbol{\lambda}$; and (3) $\mathbb{E}[u^2|Z] = \sigma^2$. Then $\hat{\boldsymbol{\beta}}_{\text{2SLS-GI}}$ has the following asymptotic distribution:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{2SLS-GI}} - \boldsymbol{\beta}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{5.4}$$

where $\boldsymbol{\Sigma} = \sigma^2 \Big\{ \mathbb{E}[X(Z^*)^T] \big[\mathbb{E}[Z^*(Z^*)^T] \big]^{-1} \mathbb{E}[Z^* X^T] \Big\}^{-1}$. The proof for this result can be found in Appendix 6A of [46]. Based on the asymptotic normality of $\hat{\boldsymbol{\beta}}_{\text{2SLS-GI}}$, we can perform hypothesis testing regarding the causal parameter of interest, β_1. In practice, the asymptotic covariance matrix $\boldsymbol{\Sigma}$ can be estimated by the empirical covariance matrix: $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}^2 (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1}$, where $\hat{\sigma}^2 := (n-k-2)^{-1} \sum_{i=1}^n (Y_i - X_i^T \hat{\boldsymbol{\beta}}_{\text{2SLS-GI}})^2$.

The first and third assumptions for the asymptotic result are fairly standard assumptions. The second assumption on $\sqrt{n}$-consistency of $\hat{\boldsymbol{\lambda}}$ is also satisfied in many common scenarios, e.g., when $\hat{\boldsymbol{\lambda}}$ is obtained via maximum likelihood estimation (MLE) in parametric models.

When the conditional mean model for X_E given Z is correctly specified, i.e., $\mathbb{E}[X_E|Z] =$

$g_E(Z, \boldsymbol{\lambda})$, the resulting 2SLS-GI estimator provides a better efficiency than the standard 2SLS estimator. On the other hand, even if the conditional mean model for X_E is misspecified, the 2SLS-GI estimator $\hat{\boldsymbol{\beta}}_{\text{2SLS-GI}}$ is still consistent for $\boldsymbol{\beta}$ and provides valid inference, as long as the function $g_E(\cdot, \boldsymbol{\lambda})$ and the parameter $\boldsymbol{\lambda}$ are well-defined. This condition can be satisfied in many scenarios. Under model misspecification of parametric models, for example, the MLE estimator converges to a well-defined limit that minimizes the Kullback–Leibler divergence between the assumed distribution and the true distribution [176].

5.2.4 Application of 2SLS-GI to a Microbial Exposure

2SLS-GI allows us to gain more efficiency than the standard 2SLS method if the relationship between the exposure and the genetic IVs and covariates are well modeled. We now discuss how 2SLS-GI is adapted to accommodate a microbial exposure. Let the exposure X_E refer to the abundance of a microbial taxon, presumably after some transformation (e.g., scaling or rarefaction) to account for differential read depths across individuals. To capture the count nature of the microbial abundance as well as other characteristics including overdispersion and zero-inflation, here we consider two zero-inflated count models, zero-inflated Poisson (ZIP) regression and zero-inflated negative binomial (ZINB) regression, to model the relationship between X_E and the instrument Z (containing both the genetic IVs and the covariates). These two models are widely used to model microbial count data and show reasonable performance in previous microbiome studies [169, 177, 178, 179]

Zero-inflated models such as ZIP and ZINB are a mixture distribution of two components [180]. For each observation, the data generation process produces a structural zero with probability π and produces counts according to a regular count data distribution, such as a Poisson distribution or a negative binomial distribution, with probability $1 - \pi$. Formally,

the probability distribution of a zero-inflated count variable, X_E, can be written as

$$\Pr(X_E = x) = \begin{cases} \pi + (1-\pi)f(0) & \text{if } x = 0, \\ (1-\pi)f(x) & \text{if } x > 0, \end{cases} \tag{5.5}$$

where $f(x)$ is a probability density function for count data, such as the Poisson or negative binomial distribution. Therefore, the zero-inflated models account for both structural zeros (due to physical absence) and sampling zeros (due to insufficient sampling) of the microbial abundances derived from sequencing data.

In ZIP and ZINB regression, we further relate the probability of observing a structural zero, π, and the mean parameter in the Poisson or negative binomial distribution to the instrument, Z. In ZIP regression, we assume the following models on π and the mean parameter, λ_P, of the Poisson distribution:

$$\begin{aligned} \text{logit}(\pi) &= \log\left(\frac{\pi}{1-\pi}\right) = Z^T\boldsymbol{\alpha}, \\ \log(\lambda_P) &= Z^T\boldsymbol{\delta}. \end{aligned} \tag{5.6}$$

In ZINB regression, we assume similar models for π and the mean parameter, λ_{NB}, of the negative binomial distribution, again using a logit link and a log link, respectively. For ZINB regression, there is an additional dispersion parameter, ϕ, in the negative binomial distribution that accounts for the potential overdispersion of count data, where a smaller ϕ indicates a greater level of overdispersion. The ZIP distribution can be considered as a special case of the ZINB distribution as the dispersion parameter ϕ goes to infinity.

The conditional mean of X_E given Z based on ZIP regression is: $\mathbb{E}[X_E|Z] = (1-\pi)\lambda_P = [1-\text{expit}(Z^T\boldsymbol{\alpha})]\exp(Z^T\boldsymbol{\delta})$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ can be estimated via maximum likelihood. Based on ZINB regression, the conditional mean is $\mathbb{E}[X_E|Z] = (1-\pi)\lambda_{NB}$, and the unknown parameters can be estimated by maximum likelihood similarly. The resulting estimated conditional mean, $\hat{\mathbb{E}}[X_E|Z]$, can thus be used as part of the new instrument, $\hat{Z}^*$, in 2SLS-GI.

Sometimes, not all microbial taxa display zero-inflation: certain taxa might be extremely common, or the sequencing technique being used is able to produce high sequencing depths. Here we set a simple rule to prevent overfitting of zero-inflated models. When there is no zero count in the microbial data, i.e., when the microbial taxon under study is present in all individuals, we drop the zero component and directly use Poisson regression or negative binomial regression such that the conditional mean becomes $\mathbb{E}[X_E|Z] = \lambda_P$ or $\mathbb{E}[X_E|Z] = \lambda_{NB}$. The unknown parameters can be estimated similarly as before to obtain the estimated conditional mean.

In summary, the procedure for obtaining the 2SLS-GI estimator with a microbial exposure can be described as the following three steps:

1. Perform regression of X_E on genetic instruments G and covariates C using a zero-inflated count data model, either ZIP regression or ZINB regression, to obtain the estimated conditional mean, $\hat{\mathbb{E}}[X_E|Z]$.

    - If no zeros are present in the microbial data, Poisson regression or negative binomial regression can be used instead to obtain $\hat{\mathbb{E}}[X_E|Z]$.

2. Perform linear regression of X_E on $\hat{\mathbb{E}}[X_E|Z]$ and C to obtain the fitted exposure $\tilde{X}_E$.

3. Perform linear regression of Y on $\tilde{X}_E$ and C to obtain the causal estimate of X_E on Y.

To implement the 2SLS-GI approach, we use the `zeroinfl()` function in the pscl R package [181] to perform ZIP and ZINB regression; we then use the ivreg R package to perform 2SLS after obtaining the generated instruments.

5.3 Simulation Studies

We conduct simulation studies to evaluate the performance of 2SLS-GI and compare it against existing IV methods in MR analysis with a microbial exposure. We consider different data generation mechanisms for microbial abundances and assess the power gain of 2SLS-GI in detecting the causal effect of a microbial taxon on a continuous outcome.

5.3.1 Simulation Procedure

In our simulation studies, we consider two variations of the 2SLS-GI approach, where we estimate the conditional mean of the microbial exposure given IVs using ZIP regression and ZINB regression, respectively, as described in Section 5.2.4. We denote these two methods as "2SLS-GI-ZIP" and "2SLS-GI-ZINB," respectively. To compare our proposed 2SLS-GI approach to existing 1-sample MR methods, we consider two competing methods, 2SLS and limited information maximum likelihood (LIML) [182], that are commonly used in IV analysis. 2SLS is a standard IV method as we have introduced before in Section 5.2.2. LIML can be considered as a "maximum likelihood counterpart" of 2SLS [47], where a bivariate normal distribution is used to model the error terms in the exposure-outcome relationship and the genetics-exposure relationship simultaneously. In finite samples, LIML is generally less efficient than 2SLS but can be more robust to weak IVs [47]. In addition, we also apply the naive OLS estimator as a reference method in our simulation studies, which is expected to produce invalid inference in the presence of confounding.

As a general setting, we generate genotypes of five valid genetic instruments with minor allele frequencies of 0.1 according to a binomial distribution: $g_j \sim \text{Binomial}(2, 0.1)$ for $j = 1, ..., 5$ and $G := (g_1, ..., g_5)^T$. We consider two measured confounders in the exposure-outcome relationship, including one continuous variable and one binary variable: $c_1 \sim N(0, 1)$, $c_2 \sim \text{Bernoulli}(0.5)$ and $C := (c_1, c_2)^T$.

We consider three different data generation scenarios for microbial abundances. In the first and second scenarios, we generate the abundance of a microbial taxon from ZIP and ZINB regression models, respectively, so that the assumed conditional mean model for the microbial abundance is the same as or close to the true conditional mean model. In the third scenario, we generate the microbial abundance according to a beta-binomial regression model, another commonly used model for microbiome data [183], so that the assumed conditional mean model is considerably different from the true model. The detailed data generation procedure is described below.

In **Scenario 1**, for each individual i, we use the following model on the probability of observing a structural zero, π_i, and the mean parameter, $\lambda_{P,i}$, of the Poisson distribution:

$$\begin{aligned} \text{logit}(\pi_i) &= G_i^T \boldsymbol{\alpha}_G + C_i^T \boldsymbol{\alpha}_C + u_i \alpha_u, \\ \log(\lambda_{P,i}) &= G_i^T \boldsymbol{\delta}_G + C_i^T \boldsymbol{\delta}_C + u_i \delta_u, \end{aligned} \tag{5.7}$$

where $u_i \sim N(0,1)$ is an unmeasured confounder in the exposure-outcome relationship, and we let each element of $\boldsymbol{\alpha}_G$ come from Uniform$(-0.4, -0.2)$, $\boldsymbol{\alpha}_C^T = (-0.5, -0.5)$, $\alpha_u = -0.5$ and $(\boldsymbol{\delta}_G^T, \boldsymbol{\delta}_C^T, \delta_u) = -1.2(\boldsymbol{\alpha}_G^T, \boldsymbol{\alpha}_C^T, \alpha_u)$. We then generate the microbial abundance $X_{E,i}$ for individual i according to the distribution ZIP$(\pi_i, \lambda_{P,i})$. Finally, we generate a continuous outcome Y_i as:

$$Y_i = X_{E,i}\beta_1 + C_i^T \boldsymbol{\beta}_2 + u\beta_3 + \epsilon_i, \tag{5.8}$$

where we set $\beta_1 = 0.08$, $\boldsymbol{\beta}_2^T = (0.3, 0.3)$, $\beta_3 = 0.5$ and $\epsilon_i \sim N(0,1)$.

In **Scenario 2**, for each individual i, we use the following model on the probability of observing a structural zero, π_i, and the mean parameter, $\lambda_{NB,i}$, of the negative binomial distribution:

$$\begin{aligned} \text{logit}(\pi_i) &= G_i^T \boldsymbol{\alpha}_G + C_i^T \boldsymbol{\alpha}_C + u_i \alpha_u, \\ \log(\lambda_{NB,i}) &= G_i^T \boldsymbol{\delta}_G + C_i^T \boldsymbol{\delta}_C + u_i \delta_u, \end{aligned} \tag{5.9}$$

where u_i, $\boldsymbol{\delta}_G$, $\boldsymbol{\delta}_C$, δ_u, $\boldsymbol{\alpha}_G$, $\boldsymbol{\alpha}_C$, α_u are generated in the same way as in **Scenario 1**. We fix the dispersion parameter ϕ as 0.8. We then generate the microbial abundance $X_{E,i}$ for individual i according to the distribution ZINB$(\pi_i, \lambda_{P,i}, \phi)$. The outcome Y_i is again generated in the same way as in **Scenario 1** with the same parameter values.

In **Scenario 3**, we consider a beta-binomial regression model for generating the microbial abundance. A beta-binomial distribution is composed of three parameters: the total number of possible counts, N_{BB}, which corresponds to the read depth for an individual in the context of microbiome data; the probability of observing a single count, p; and a correlation parameter, ρ, which accounts for overdispersion in the count data. Here we fix $N_{BB} = 1000$

and $\rho = 0.5$. For each individual i, we use the following model on p_i, the probability of observing a count:

$$\text{logit}(p_i) = G_i^T \boldsymbol{\alpha}_G + C_i^T \boldsymbol{\alpha}_C + u_i \alpha_u - 4, \tag{5.10}$$

where $u_i \sim N(0,1)$ and we let each element of $\boldsymbol{\alpha}_G$ come from Uniform$(0.4, 0.6)$, $\boldsymbol{\alpha}_C^T = (0.5, 0.5)$, and $\alpha_u = 0.5$. We then generate the microbial abundance $X_{E,i}$ for individual i according to Beta-Binomial(N_{BB}, p_i, ρ). Finally, the outcome Y_i is generated as: $Y_i = X_{E,i}\beta_1 + C_i^T \boldsymbol{\beta}_2 + u\beta_3 + \epsilon_i$, where we set $\beta_1 = 0.004$, $\boldsymbol{\beta}_2^T = (0.3, 0.3)$, $\beta_3 = 0.5$ and $\epsilon_i \sim N(0,1)$.

In the above three scenarios, the parameters $\boldsymbol{\alpha}_G$ and $\boldsymbol{\delta}_G$ are set in a way that ensures reasonable instrument strength for the genetic IVs. The causal effect of interest, β_1, is set in each scenario to ensure a clear comparison in statistical power between different methods. When we apply the proposed 2SLS-GI methods and the competing methods to the simulated data, we only incorporate the genetic IVs, G, and measured confounders, C, in our models, but do not include the variable u, which is considered as unobserved in practice. For each scenario, we consider four different sample sizes: 400, 600, 800 and 1000, and conduct 1000 simulations for each sample size.

5.3.2 Simulation Results

For each scenario, we report both the coverage of the 95% confidence interval (CI) and the statistical power of different methods under different data generation scenarios. While we show the coverage result for the four IV methods and the OLS method, we report the power result only for the four IV methods that produce valid inference.

Figure 5.2 shows the coverage and power result for **Scenario 1**. As expected, the OLS estimator produces extremely low CI coverage, confirming that we have properly introduced confounding into the exposure-outcome relationship. On the other hand, all the IV methods including two 2SLS-GI methods, 2SLS and LIML show correct CI coverage at all sample sizes. In terms of statistical power, both 2SLS-GI-ZIP and 2SLS-GI-ZINB show a much higher power than 2SLS and LIML, demonstrating an evident efficiency gain when the true

model for the microbial abundance is ZIP.

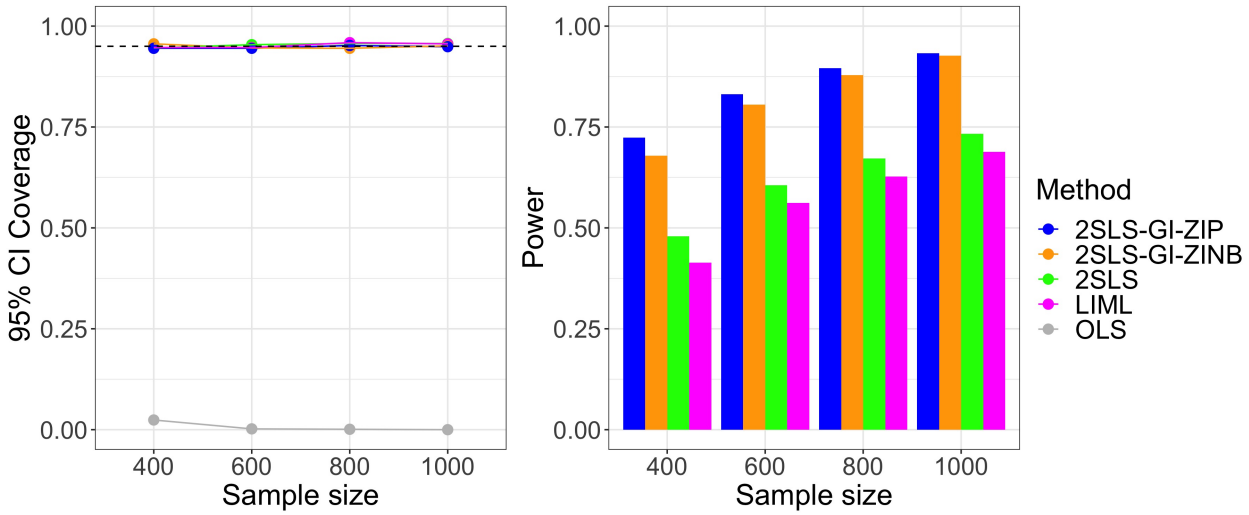

Figure 5.2: Empirical coverage of the 95% confidence interval and empirical power of different methods under **Scenario 1**, where the microbial abundance is generated from a zero-inflated Poisson model. The dashed line on the left figure indicates nominal 95% coverage.

Figure 5.3 shows the coverage and power result for **Scenario 2**, where the true model for the microbial abundance is ZINB. Under this scenario, 2SLS appears to be slightly conservative in terms of CI coverage at smaller sample sizes, whereas the two 2SLS-GI methods have their CI coverage close to the nominal 95% at all sample sizes. Again, both 2SLS-GI methods display a higher power than 2SLS and LIML.

Figure 5.4 shows the coverage and power result for **Scenario 3**, where the true model for the microbial abundance is beta-binomial. All methods are slightly conservative in terms of CI coverage when the sample size is 400, but achieve a coverage close to 95% at larger sample sizes. 2SLS-GI-ZIP has a higher power than 2SLS and LIML at all sample sizes. 2SLS-GI-ZINB has a slightly lower power than 2SLS at $n = 400$, but achieves a higher power than 2SLS and LIML at larger sample sizes. The power gain of the two 2SLS-GI methods are not as evident as in **Scenario 1** and **Scenario 2**. This is expected since the conditional mean model for the microbial abundance is largely misspecified in this case. Nevertheless, we see

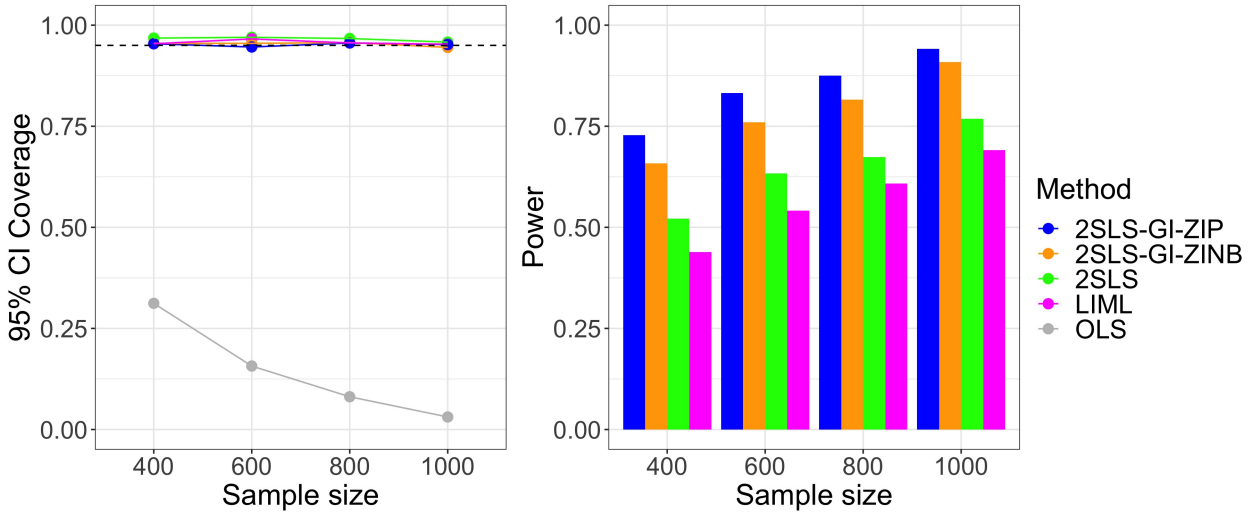

Figure 5.3: Empirical coverage of the 95% confidence interval and empirical power of different methods under **Scenario 2**, where the microbial abundance is generated from a zero-inflated negative binomial model. The dashed line on the left figure indicates nominal 95% coverage.

that both 2SLS-GI methods are able to produce correct CI coverage and can still maintain a power close to or even higher than 2SLS.

It is interesting that 2SLS-GI-ZIP appears to have a higher power than 2SLS-GI-ZINB in all scenarios, especially evident at smaller sample sizes. This is the case even in **Scenario 2**, where the true model for the microbial exposure is ZINB. One possible explanation is that ZINB regression is a more complex model than ZIP regression, requiring the estimation of a greater number of parameters, which causes a loss in efficiency at smaller sample sizes. A balance between model simplicity and goodness-of-fit to the data might be considered for model selection. Further extensions to incorporate multiple candidate conditional mean models for the microbial exposure in 2SLS-GI can also be studied.

5.4 Data Application

Hypertension, the consistent elevation of blood pressure, is one of the most common conditions worldwide and an important risk factor for cardiovascular diseases, stroke and dementia

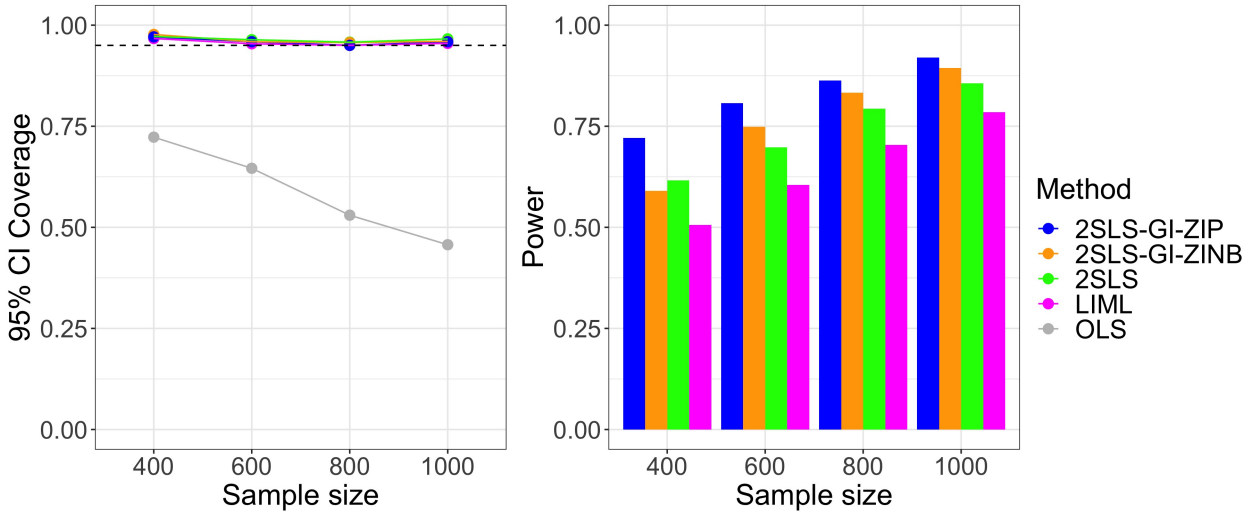

Figure 5.4: Empirical coverage of the 95% confidence interval and empirical power of different methods under **Scenario 3**, where the microbial abundance is generated from a beta-binomial model. The dashed line on the left figure indicates nominal 95% coverage.

[184, 185]. Previous studies have revealed a role of the gut microbiome in the development of hypertension and discovered specific genera associated with hypertension and blood pressure [166, 186, 187]. Fecal transplant experiments in mouse models have also demonstrated that elevated blood pressure is transferable through gut microbiota [188, 166]. However, few studies have assessed the causal effects of gut microbiota on blood pressure in the human population. Here we apply our proposed 2SLS-GI approach to an epidemiological data set, the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), where we evaluate the causal effect of different gut microbial genera on systolic and diastolic blood pressure in Hispanic/Latino populations of the United States via MR analysis.

5.4.1 Description of the HCHS/SOL Study

HCHS/SOL is a community-based prospective cohort study aimed to identify risk factors for health outcomes in Hispanic/Latino populations in the United States. The study recruited 16,415 Hispanic/Latino adults aged 18 - 74 years, representing diverse ethnic background,

at four U.S. field centers (Bronx, NY, Chicago, IL, Miami, FL, and San Diego, CA), using a two-stage probability sampling design [189, 23].

During a series of in-person clinic visits, participants answered survey questions regarding their demographics, health behaviors and medical histories. Clinical assessment and laboratory tests were also conducted. Specifically, measurement of blood pressure (BP) proceeded as follows [190]. After a 5-minute rest period, BP was measured on seated participants using their right arm, with a cuff sized to their upper right arm circumference. A trained technician used an automated sphygmomanometer (OMRON HEM-907 XL, Omron Healthcare, Inc., Lake Forest, IL) to take three BP measurements that were spaced 1 minute apart. The three measurements were then averaged. We use the resulting average systolic blood pressure (SBP) and average diastolic blood pressure (DBP) as our outcome variables.

12,803 participants consented to genetic studies. Genotyping was performed on an Illumina custom array, SOL HCHS Custom 15041502 B3, which consisted of the Illumina Omni 2.5M array (HumanOmni2.5-8v1-1) and ∼150,000 custom SNPs [65]. Quality control, genotype imputation and estimation of pairwise kinship coefficients and principal components (PCs) of genome-wide genetic variability were described in detail by Conomos et al. [65]. In addition to the quality control procedures described in [65], we also filter imputed genetic variants based on an "effective minor allele count": $N_{\text{eff}} = 2\hat{p}(1-\hat{p})Nv$, where $\hat{p}$ is the estimated minor allele frequency, N is the sample size and v is the ratio of observed variance of imputed dosages to the expected binomial variance [66]. We retain variants with sufficient minor allele counts and exclude any variants with $N_{\text{eff}} < 30$.

As an ancillary study, the HCHS/SOL Gut Origins of Latino Diabetes (GOLD) study was further conducted to investigate the role of gut microbiome in diabetes and other health outcomes in Hispanic/Latino individuals [24]. 3,035 participants from HCHS/SOL were enrolled in this ancillary study during the second in-person visit period from 2014 to 2017. Based on the collected stool samples, DNA extraction and shotgun metagenomic sequencing were performed to characterize the gut microbial profiles. Quality-controlled paired-end microbial sequences were processed using the pipelines and reference databases available in Qiita, a

web-based open-source microbiome analysis platform [191]. Sequence reads were mapped at species level and subset to bacterial species only (making up $> 99.5\%$ of all reads). Specific procedures of stool sample collection, metagenomic sequencing and subsequent bioinformatic processing are described in detail by Usyk et al. [192].

The HCHS/SOL study was approved by the Institutional Review Boards of all participating institutions, and written informed consent was obtained from all participants.

5.4.2 Procedure of MR Analysis

For our MR analysis, we consider gut microbiome and phenotype data collected during the second visit period of HCHS/SOL (2014-2017) and focus on 1,601 unrelated individuals (with pairwise kinship coefficient ≤ 0.05) where genome-wide genetic data, gut microbiome data from metagenomic sequencing and blood pressure data are available. Our exposure of interest is the abundance of individual microbial genera, and our outcomes of interest are SBP and DBP. We adjust for the following covariates: top 5 genome-wide genetic PCs, age and gender. Among these covariates, the genetic PCs are included to ensure the validity of the genetic instruments; age and gender are confounders known to affect both gut microbiome [24] and blood pressure [193, 194]. We now describe our procedure of MR analysis on HCHS/SOL data, with reference to previous microbiome MR studies [14].

We perform an initial transformation and filtering of the microbiome data to account for differential read depths across individuals and remove overly sparse taxa. We aggregate the species-level microbial abundances at the genus level, resulting in 1260 genera in total. We transform the raw abundance data to relative abundances and scale the relative abundances to the minimum read depth ($D = 116{,}776$) among all individuals. In other words, if we let r_{ij} denote the raw abundance of the j-th genus in the i-th individual, N_i denote the read depth for the i-th individual and D denote the minimum read depth among all individuals, then the scaled relative abundance for genus j in individual i is calculated as $r_{ij}D/N_i$. The scaled relative abundances are further rounded to the nearest integers. Lastly, we focus on genera that are present in $\geq 10\%$ of all individuals, leaving us 205 genera for subsequent

analysis.

The MR analysis on HCHS/SOL data includes three steps: (1) initial screening of microbial genera for associations between the genus abundance and BP; (2) identify genetic instruments for the selected microbial genera; (3) perform 2SLS-GI and competing methods to assess the causal effect of each selected microbial genus on SBP and DBP.

In the first step, we screen the 205 microbial genera to identify genera that show observed associations with the BP variables. We assess the association between SBP and each genus, via linear regression of SBP on the genus abundance while adjusting for the covariates (top 5 genetic PCs, age and gender). The same procedure is performed on DBP. We then select genera with significant associations with either SBP or DBP, with false discovery rate (FDR)-adjusted p-value < 0.2, for our subsequent analysis.

In the second step, we search for genetic instruments that are associated with the selected genera in Step 1, by performing GWAS analysis for each selected genus. We focus on common genetic variants with MAF ≥ 0.05 along the autosomes. To conduct association testing, we perform either linear regression or logistic regression depending on the prevalence of the genera, based on analysis procedures in previous microbiome GWAS studies [18, 92, 15]. Specifically, for genera present in $\geq 90\%$ of individuals, we perform rank normal transformation on the scaled relative abundance to encourage normality and use linear regression to assess the association between each rank-normal-transformed genus abundance and each genetic variant. For genera present in $\geq 10\%$ but $< 90\%$ of individuals, the presence/absence of each genus is used as the outcome and related to each genetic variant via logistic regression. The covariates in the regression models include the top 5 genetic PCs, age, gender and study site. The GWAS analysis is conducted using the GENESIS R package.

Based on the GWAS result, we identify significant variants at a relaxed genome-wide threshold (p-value $< 10^{-5}$) for each selected genus and then perform linkage disequilibrium (LD) clumping ($r^2 < 0.001$) to retain independent variants, using the ieugwasr R package. To avoid potential horizontal pleiotropy (which violates the third IV assumption from Section 5.2.1), among genetic variants identified for each genus, we remove those variants that are

also significant for other genera. The remaining variants for each genus are treated as the genetic instruments for that genus.

Finally, in the third step, we perform 2SLS-GI-ZIP and 2SLS-GI-ZINB to assess the causal effect of each selected microbial genus on SBP and DBP. We also apply three other competing methods: 2SLS, LIML and CIIV [195], a confidence-interval-based method for selecting valid instruments and performing IV analysis. 2SLS and LIML have been introduced in Sections 5.2.2 and 5.3.1. CIIV is a recently developed IV method that selects valid IVs from a larger set of potential IVs, where the selection is based on overlapping CIs of the individual-IV-based causal estimates; it then performs 2SLS to make causal inference. As the other methods (2SLS-GI, 2SLS and LIML) assume that all the genetic IVs being used are valid, including CIIV as a competing method can tell us how robust our results are to invalid IVs. We apply the first-stage thresholding for weak IVs when implementing CIIV, as suggested in [195]. In all methods, the same set of covariates (top 5 genetic PCs, age and gender) is adjusted in the IV models. We use 0.05 as the significance threshold in MR analysis.

5.4.3 Application Results

Based on initial screening of microbial genera for observed associations with blood pressure variables, we have identified five pairs of significant microbe-BP relationships which involve four genera: *Prevotella*-SBP, *Parabacteroides*-DBP, *Intestinimonas*-DBP, *Prevotella*-DBP and *Catabacter*-DBP. The number of genetic IVs discovered for each of these four genera ranges from 10 to 23. We then conduct MR analysis for the five pairs of microbe-BP association identified above. Table 5.1 shows the MR analysis results of different IV methods, where the point estimates, 95% CIs and p-values for the causal effect are reported. The point estimates and 95% CIs are with regard to one standard deviation (s.d.) increase in the scaled relative abundance of each microbial genus. For example, based on 2SLS-GI-ZIP, a 1 s.d. increase in relative abundance of *Prevotella* is estimated to generate a 2.73 (95%: [0.34, 5.12]) mmHg increase in SBP. Figure 5.5 shows the corresponding forest plot that displays a graphical comparison in causal estimates and 95% CIs between different methods.

Table 5.1: Causal effects of gut microbial genera on systolic and diastolic blood pressure (SBP and DBP), based on Mendelian randomization analysis of the HCHS/SOL data.

Association	No. of IVs	Method	Point estimate	95% CI	P-value
		2SLS-GI-ZIP	2.73	(0.34, 5.12)	0.026
		2SLS-GI-ZINB	2.55	(-0.09, 5.18)	0.058
Prevotella - SBP	15	2SLS	2.00	(-0.35, 4.36)	0.096
		LIML	2.02	(-0.37, 4.40)	0.098
		CIIV	2.17	(-0.25, 4.58)	0.078
		2SLS-GI-ZIP	-1.65	(-2.96, -0.33)	0.014
		2SLS-GI-ZINB	-1.69	(-3.04, -0.33)	0.015
Parabacteroides - DBP	17	2SLS	-1.69	(-3.04, -0.34)	0.014
		LIML	-1.74	(-3.12, -0.35)	0.014
		CIIV	-1.82	(-3.21, -0.42)	0.011
		2SLS-GI-ZIP	-0.83	(-2.48, 0.82)	0.327
		2SLS-GI-ZINB	-0.91	(-2.61, 0.79)	0.292
Intestinimonas - DBP	10	2SLS	-1.01	(-2.79, 0.77)	0.267
		LIML	-1.01	(-2.86, 0.83)	0.280
		CIIV	-1.01	(-2.78, 0.77)	0.265
		2SLS-GI-ZIP	1.08	(-0.34, 2.50)	0.136
		2SLS-GI-ZINB	1.13	(-0.43, 2.70)	0.156
Prevotella - DBP	15	2SLS	1.11	(-0.29, 2.51)	0.122
		LIML	1.12	(-0.31, 2.54)	0.124
		CIIV	1.30	(-0.14, 2.74)	0.076
		2SLS-GI-ZIP	-1.99	(-3.96, -0.01)	0.049
		2SLS-GI-ZINB	-2.44	(-4.54, -0.35)	0.022
Catabacter - DBP	23	2SLS	-1.09	(-3.26, 1.08)	0.325
		LIML	-1.23	(-3.80, 1.35)	0.350
		CIIV	0.53	(-3.28, 4.35)	0.783

The point estimates and 95% CIs are reported with regard to 1 s.d. increase in the scaled relative abundance of each microbial genus.

In general, based on the forest plot, the point estimates and 95% CIs are consistent across methods for each pair of association, confirming the validity of our proposed 2SLS-GI approach. We next discuss the specific causal relationships identified as significant by our methods.

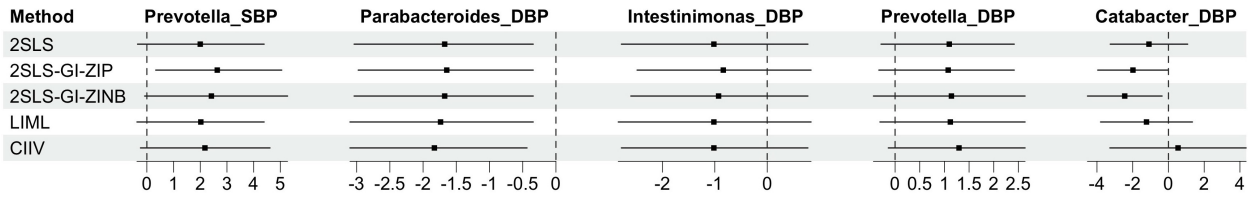

Figure 5.5: Forest plot comparing the causal effect estimate and 95% CI of different IV methods from Mendelian randomization analysis of the HCHS/SOL data.

All IV methods produce significant causal effects with similar point estimates for the *Parabacteroides*-DBP association: an increase in relative abundance of *Parabacteroides* is estimated to cause a decrease in DBP. One of the core members in the human gut microbiome [196], *Parabacteroides* is a commensal genus able to produce short chain fatty acids (SCFAs) including acetate and succinate [197]. In particular, a phase II randomized clinical trial [198] found that prebiotic SCFA supplementation reduced the 24-hour SBP in individuals with hypertension. *Parabacteroides* is also shown to alleviate obesity and acute pancreatitis through production of SCFAs in mouse models [199, 200]. Overall, these existing experimental studies corroborate the causal effect identified in our analysis.

For the *Prevotella*-SBP association, only 2SLS-GI-ZIP produces a significant causal effect (p-value = 0.026), where an increase in relative abundance of *Prevotella* is estimated to cause an increase in SBP. 2SLS-GI-ZINB appears to be borderline significant (p-value = 0.058), whereas all the other methods produce nonsignificant results. *Prevotella* is one of the most abundant genera in our current HCHS/SOL samples. A previous study [166] found that individuals with pre-hypertension and hypertension are associated with a higher abundance of *Prevotella* and more likely to have a *Prevotella*-dominated gut microbiome, compared to healthy controls. *Prevotella* has also been associated with increased inflammation [86], which is thought to play a role in the development of hypertension [201].

For the *Catabacter*-DBP association, both 2SLS-GI-ZIP (p-value = 0.049) and 2SLS-GI-ZINB (p-value = 0.022) produce significant results: an increase in relative abundance of *Catabacter* is estimated to cause a decrease in DBP. On the other hand, all the other

methods fail to produce a significant result. The role of *Catabacter* in gut microbiota is still unclear [202], but a previous study has reported negative associations between *Catabacter* and hypertension as well as between *Catabacter* and SBP [186]. Here, while most methods produce negative causal estimates, CIIV produces a positive, close to null, estimate. This is likely because CIIV only selected 3 out of 23 genetic IVs for the IV analysis under its first-stage thresholding setting. As *Catabacter* has a high level of zero-inflation (with 44% of zeros among all individuals) and CIIV performs weak IV thresholding based on a linear model between the microbial abundance and the IVs, the genetic IVs identified from nonlinear GWAS models were likely discarded by CIIV as weak IVs.

For the remaining two associations, *Intestinimonas*-DBP and *Prevotella*-DBP, all IV methods produce nonsignificant causal results, with similar point estimates and 95% CIs. Overall, our analysis of the HCHS/SOL data demonstrates a power advantage of the proposed 2SLS-GI approach: it not only produces consistent causal estimates and CIs with existing IV methods in most situations, but is able to identify a greater number of significant causal relationships than existing methods as well. The identified genus-BP causal relationships will be interesting to further investigate and verify in experimental studies.

5.5 Discussion

We have proposed a novel analysis framework to conduct MR analysis with a microbial exposure and a continuous outcome. Utilizing the 2SLS-GI method, we are able to account for characteristics specific to microbiome data such as overdispersion and zero-inflation, and incorporate nonlinear relationships between the microbial abundance and the IVs. In simulation studies, we have demonstrated the power gain of 2SLS-GI in detecting causal effects compared to existing IV methods. Through application to the HCHS/SOL data, we have shown that 2SLS-GI is able to identify a greater number of causal relationships between gut microbial genera and blood pressure variables, compared to existing IV methods.

In this work, we have considered two variations of 2SLS-GI: 2SLS-GI-ZIP and 2SLS-GI-ZINB, where we used either ZIP or ZINB to estimate the conditional mean of the microbial

abundance given genetic IVs and covariates. While a well-predicted conditional mean could improve efficiency of 2SLS-GI, in practice, we might not know which model best represents the true conditional mean relationship. One possible extension is to include multiple conditional mean estimators derived from different models as a group of instruments (which can then be plugged into Step 2 of the 2SLS-GI procedure in Section 5.2.4), instead of using one conditional mean estimator as a single instrument. Such an approach can automatically weight different estimators in terms of how well they predict the true conditional mean. Alternatively, we can also consider certain nonparametric or machine learning methods to estimate the conditional mean, which rely on fewer distributional assumptions and could potentially achieve better predictive performance. However, certain conditions might still need to satisfy in order for the asymptotic result of 2SLS-GI to hold and the theoretical validity remains to be established.

Finally, our proposed 2SLS-GI approach currently only applies to the one-sample MR setting, where individual-level data needs to be available in one study sample. In recent years, large cohort studies with data simultaneously available on genetic variation, microbiome profiles and phenotypic outcomes, such as HCHS/SOL, are increasingly popular, offering more opportunities for one-sample IV methods like our proposed approach. Nevertheless, two-sample MR analysis still has the advantage of achieving a greater statistical power compared to the one-sample setting. An extension of 2SLS-GI to two-sample analysis could be interesting to further explore: while nonlinear relationships between the microbial exposure and the IVs might be difficult to incorporate based on summary statistics, such an extension might still be feasible if individual-level data are available in the two-sample setting.

Chapter 6

DISCUSSION

6.1 Summary

Due to the importance of the human microbiome in host health, there has been a great effort to collect and analyze microbiome data. Challenges specific to microbiome data derived from high-throughput microbial sequencing call for novel statistical approaches. In this dissertation, we have developed a variety of novel statistical methods for association analyses as well as causal inference of microbiome data.

In Project 1, we have proposed a novel microbiome GWAS approach to assess the association between gene-level host genetic variation and community-level microbiome composition, which serves as an alternative powerful strategy for microbiome GWAS in addition to the existing variant-based taxon-level approaches. In Project 2, we have developed a novel multivariate independence test for clustered data, which demonstrates a superior power than competing methods in evaluating the association between the microbiome and a multivariate trait based on longitudinal data. In Project 3, we have developed a multivariate approach to construct microbial association networks, which shows an improved power in edge detection and produces networks that are consistent with existing knowledge on human microbiota. In Project 4, we have proposed a novel approach for one-sample Mendelian randomization (MR) analysis with a microbial exposure, which shows a better power in identifying causal effects of individual microbial taxa on continuous outcomes, compared to existing MR methods.

In Project 1-3, we used multivariate and kernel-based methods to account for the high-dimensional nature and intrinsic structure, such as phylogenetic relationships, of microbiome data. In Project 4, we incorporated appropriate models into MR methods to accommodate characteristics specific to microbiome data, including zero-inflation and overdispersion. We

also accounted for practical needs, such as covariate adjustment (Project 1) and analysis of clustered data (Project 2), in association analyses of microbiome data. The proposed methods in the above projects are promising tools to bring more discoveries regarding the human microbiome and help us better understand how the microbiome affects and interacts with host health.

6.2 Future Work

First, while our proposed statistical methods have addressed certain practical needs in association analyses, further extensions can be considered to make them more versatile and flexible. For example, we have accounted for covariate adjustment in Project 1 and clustered data in Project 2, which were separately incorporated into kernel-based association testing frameworks for multivariate variables. A useful extension would be a kernel-based multivariate association test that accommodates covariate adjustment and clustered data at the same time. In addition, our proposed association test in Project 2 only applies to complete and balanced clustered data, which might not be readily available in practice. A more flexible extension to incomplete clustered data (e.g., missing data due to loss of follow-up) or unbalanced clustered data (e.g., data collected from subjects of the same household or community) would also be helpful.

Second, while we have offered specific model choices in our methods to better capture microbiome-specific characteristics, we have not focused on model selection. For example, in Project 1, multiple candidate microbiome kernels can be considered to measure pairwise similarity in microbial profiles. In Project 4, multiple count data models can be considered to estimate the conditional mean of taxon abundance given instrumental variables, in order to generate a new instrument. In practice, we often do not know in advance which kernel or model best captures the data characteristics and results in the best statistical power. To ensure a practical power gain using our methods, further extensions of our methods for kernel/model selection or combining multiple candidate kernels/models would be useful.

Finally, while we have focused on the application of our methods to microbiome data in

this dissertation, our methods can be in principle extended to the analyses of other types of omics data. For example, gene expression data derived from single-cell RNA sequencing (scRNA-seq) are intrinsically high-dimensional and display sparsity and overdispersion [203, 204], so that multivariate methods and appropriate count data models can be considered. Spatial transcriptomics data further incorporates spatial relationships in gene expression among cells in a tissue, which can be modeled using kernel functions [205, 206]. Therefore, strategies similar to those proposed in our methods can be adapted to better analyze these novel omics data that are rapidly accumulating today.

BIBLIOGRAPHY

[1] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.

[2] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.

[3] HMP Integrative. The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host & Microbe*, 16(3):276–289, 2014.

[4] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, 2012.

[5] Lisa Röttjers and Karoline Faust. From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiology Reviews*, 42(6):761–780, 2018.

[6] Ran Blekhman, Julia K Goodrich, Katherine Huang, Qi Sun, Robert Bukowski, Jordana T Bell, Timothy D Spector, Alon Keinan, Ruth E Ley, Dirk Gevers, et al. Host genetic variation impacts microbiome composition across human body sites. *Genome Biology*, 16(1):1–12, 2015.

[7] Julia K Goodrich, Emily R Davenport, Andrew G Clark, and Ruth E Ley. The relationship between the human genome and microbiome comes into view. *Annual Review of Genetics*, 51:413–433, 2017.

[8] Jose C Clemente, Luke K Ursell, Laura Wegener Parfrey, and Rob Knight. The impact of the gut microbiota on human health: an integrative view. *Cell*, 148(6):1258–1270, 2012.

[9] Julian R Marchesi, David H Adams, Francesca Fava, Gerben DA Hermes, Gideon M Hirschfield, Georgina Hold, Mohammed Nabil Quraishi, James Kinross, Hauke Smidt, Kieran M Tuohy, et al. The gut microbiota and host health: a new clinical frontier. *Gut*, 65(2):330–339, 2016.

[10] Olga Castaner, Albert Goday, Yong-Moon Park, Seung-Hwan Lee, Faidon Magkos, Sue-Anne Toh Ee Shiow, and Helmut Schröder. The gut microbiome profile in obesity: a systematic review. *International Journal of Endocrinology*, 2018, 2018.

[11] Atsushi Nishida, Ryo Inoue, Osamu Inatomi, Shigeki Bamba, Yuji Naito, and Akira Andoh. Gut microbiota in the pathogenesis of inflammatory bowel disease. *Clinical Journal of Gastroenterology*, 11(1):1–10, 2018.

[12] Nadja Larsen, Finn K Vogensen, Frans WJ Van Den Berg, Dennis Sandris Nielsen, Anne Sofie Andreasen, Bente K Pedersen, Waleed Abu Al-Soud, Søren J Sørensen, Lars H Hansen, and Mogens Jakobsen. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS ONE*, 5(2):e9085, 2010.

[13] Serena Sanna, Natalie R van Zuydam, Anubha Mahajan, Alexander Kurilshikov, Arnau Vich Vila, Urmo Võsa, Zlatan Mujagic, Ad AM Masclee, Daisy MAE Jonkers, Marije Oosting, et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature Genetics*, 51(4):600–605, 2019.

[14] Xiaomin Liu, Xin Tong, Yuanqiang Zou, Xiaoqian Lin, Hui Zhao, Liu Tian, Zhuye Jie, Qi Wang, Zhe Zhang, Haorong Lu, et al. Mendelian randomization analyses support causal relationships between blood metabolites and the gut microbiome. *Nature Genetics*, pages 1–10, 2022.

[15] Alexander Kurilshikov, Carolina Medina-Gomez, Rodrigo Bacigalupe, Djawad Radjabzadeh, Jun Wang, Ayse Demirkan, Caroline I Le Roy, Juan Antonio Raygoza Garay, Casey T Finnicum, Xingrong Liu, et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nature Genetics*, 53(2):156–165, 2021.

[16] Emily R Davenport, Darren A Cusanovich, Katelyn Michelini, Luis B Barreiro, Carole Ober, and Yoav Gilad. Genome-wide association studies of the human gut microbiota. *PLoS ONE*, 10(11):e0140301, 2015.

[17] Marc Jan Bonder, Alexander Kurilshikov, Ettje F Tigchelaar, Zlatan Mujagic, Floris Imhann, Arnau Vich Vila, Patrick Deelen, Tommi Vatanen, Melanie Schirmer, Sanne P Smeekens, et al. The effect of host genetics on the gut microbiome. *Nature Genetics*, 48(11):1407–1412, 2016.

[18] David A Hughes, Rodrigo Bacigalupe, Jun Wang, Malte C Rühlemann, Raul Y Tito, Gwen Falony, Marie Joossens, Sara Vieira-Silva, Liesbet Henckaerts, Leen Rymenans, et al. Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nature Microbiology*, 5(9):1079–1087, 2020.

[19] Jun Wang, Louise B Thingholm, Jurgita Skiecevičienė, Philipp Rausch, Martin Kummen, Johannes R Hov, Frauke Degenhardt, Femke-Anouska Heinsen, Malte C Rühlemann, Silke Szymczak, et al. Genome-wide association analysis identifies variation in vitamin d receptor and other host factors influencing the gut microbiota. *Nature Genetics*, 48(11):1396–1406, 2016.

[20] Malte C Rühlemann, Frauke Degenhardt, Louise B Thingholm, Jun Wang, Jurgita Skiecevičienė, Philipp Rausch, Johannes R Hov, Wolfgang Lieb, Tom H Karlsen, Matthias Laudes, et al. Application of the distance-based F test in an mGWAS investigating β diversity of intestinal microbiota identifies variants in SLC9A8 (NHE8) and 3 other loci. *Gut Microbes*, 9(1):68–75, 2018.

[21] Xiang Zhan, Ni Zhao, Anna Plantinga, Timothy A Thornton, Karen N Conneely, Michael P Epstein, and Michael C Wu. Powerful genetic association analysis for common or rare variants with high-dimensional structured traits. *Genetics*, 206(4):1779–1790, 2017.

[22] Xiang Zhan, Anna Plantinga, Ni Zhao, and Michael C Wu. A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics*, 73(4):1453–1463, 2017.

[23] Paul D Sorlie, Larissa M Avilés-Santa, Sylvia Wassertheil-Smoller, Robert C Kaplan, Martha L Daviglus, Aida L Giachello, Neil Schneiderman, Leopoldo Raij, Gregory Talavera, Matthew Allison, et al. Design and implementation of the Hispanic Community Health Study/Study of Latinos. *Annals of Epidemiology*, 20(8):629–641, 2010.

[24] Robert C Kaplan, Zheng Wang, Mykhaylo Usyk, Daniela Sotres-Alvarez, Martha L Daviglus, Neil Schneiderman, Gregory A Talavera, Marc D Gellman, Bharat Thyagarajan, Jee-Young Moon, et al. Gut microbiome composition in the Hispanic Community Health Study/Study of Latinos is shaped by geographic relocation, environmental factors, and obesity. *Genome Biology*, 20(1):219, 2019.

[25] Hongjiao Liu, Wodan Ling, Xing Hua, Jee-Young Moon, Jessica S Williams-Nguyen, Xiang Zhan, Anna M Plantinga, Ni Zhao, Angela Zhang, Rob Knight, et al. Kernel-based genetic association analysis for microbiome phenotypes identifies host genetic drivers of beta-diversity. *Microbiome*, 11(1):1–19, 2023.

[26] Alessia Visconti, Caroline I Le Roy, Fabio Rosa, Niccolò Rossi, Tiphaine C Martin, Robert P Mohney, Weizhong Li, Emanuele de Rinaldis, Jordana T Bell, J Craig Venter, et al. Interplay between the human gut microbiome and host metabolism. *Nature Communications*, 10(1):1–10, 2019.

[27] Amy McMillan, Stephen Rulisa, Mark Sumarah, Jean M Macklaim, Justin Renaud, Jordan E Bisanz, Gregory B Gloor, and Gregor Reid. A multi-platform metabolomics approach identifies highly specific biomarkers of bacterial diversity in the vagina of pregnant and non-pregnant women. *Scientific Reports*, 5(1):1–14, 2015.

[28] Jos WR Twisk. *Applied longitudinal data analysis for epidemiology: a practical guide*. Cambridge University Press, 2013.

[29] Liam M O'Brien and Garrett M Fitzmaurice. Analysis of longitudinal multiple-source binary data using generalized estimating equations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):177–193, 2004.

[30] Laurel A Beckett, Daniel J Tancredi, and RS Wilson. Multivariate longitudinal models for complex change processes. *Statistics in Medicine*, 23(2):231–239, 2004.

[31] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005a.

[32] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, Alexander J Smola, et al. A kernel statistical test of independence. In *NIPS*, volume 20, pages 585–592. Citeseer, 2007.

[33] Xinhua Zhang, Le Song, Arthur Gretton, Alexander J Smola, et al. Kernel measures of independence for non-iid data. In *NIPS*, volume 22, pages 1937–1944, 2008.

[34] Kacper Chwialkowski and Arthur Gretton. A kernel independence test for random processes. In *International Conference on Machine Learning*, pages 1422–1430. PMLR, 2014.

[35] Seth R Flaxman, Daniel B Neill, and Alexander J Smola. Gaussian processes for independence tests with non-iid data in causal inference. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–23, 2015.

[36] Lianmin Chen, Valerie Collij, Martin Jaeger, Inge CL van den Munckhof, Arnau Vich Vila, Alexander Kurilshikov, Ranko Gacesa, Trishla Sinha, Marije Oosting, Leo AB Joosten, et al. Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nature Communications*, 11(1):1–12, 2020.

[37] Zachary D Kurtz, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology*, 11(5):e1004226, 2015.

[38] Grace Yoon, Irina Gaynanova, and Christian L Müller. Microbial networks in spring-semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Frontiers in Genetics*, page 516, 2019.

[39] Mehdi Layeghifard, David M Hwang, and David S Guttman. Disentangling interactions in the microbiome: a network perspective. *Trends in Microbiology*, 25(3):217–228, 2017.

[40] Monica Steffi Matchado, Michael Lauber, Sandra Reitmeier, Tim Kacprowski, Jan Baumbach, Dirk Haller, and Markus List. Network analysis methods for studying microbial communities: A mini review. *Computational and Structural Biotechnology Journal*, 19:2687–2698, 2021.

[41] Karoline Faust. Open challenges for microbial network construction and analysis. *The ISME Journal*, 15(11):3111–3118, 2021.

[42] Hee Cheol Chung, Irina Gaynanova, and Yang Ni. Phylogenetically informed bayesian truncated copula graphical models for microbial association networks. *The Annals of Applied Statistics*, 16(4):2437–2457, 2022.

[43] David A Savitz, Nancy Dole, James W Terry Jr, Haibo Zhou, and John M Thorp Jr. Smoking and pregnancy outcome among african-american and white women in central north carolina. *Epidemiology*, 12(6):636–642, 2001.

[44] Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163, 2008.

[45] Qian Yang, Shi Lin Lin, Man Ki Kwok, Gabriel M Leung, and C Mary Schooling. The roles of 27 genera of human gut microbiota in ischemic heart disease, type 2 diabetes mellitus, and their risk factors: a Mendelian randomization study. *American Journal of Epidemiology*, 187(9):1916–1922, 2018.

[46] Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, 2010.

[47] Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*, 26(5):2333–2355, 2017.

[48] Daphna Rothschild, Omer Weissbrod, Elad Barkan, Alexander Kurilshikov, Tal Korem, David Zeevi, Paul I Costea, Anastasia Godneva, Iris N Kalka, Noam Bar, et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature*, 555(7695):210–215, 2018.

[49] Xing Hua, Lei Song, Guoqin Yu, Emily Vogtmann, James J Goedert, Christian C Abnet, Maria Teresa Landi, and Jianxin Shi. Microbiomegwas: a tool for identifying host genetic variants associated with microbiome composition. *Genes*, 13(7):1224, 2022.

[50] Christopher Minas and Giovanni Montana. Distance-based analysis of variance: Approximate inference. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7(6):450–470, 2014.

[51] Xiaomin Liu, Shanmei Tang, Huanzi Zhong, Xin Tong, Zhuye Jie, Qiuxia Ding, Dan Wang, Ruidong Guo, Liang Xiao, Xun Xu, et al. A genome-wide association study for gut metagenome in chinese adults illuminates complex diseases. *Cell Discovery*, 7(1):1–15, 2021.

[52] Ni Zhao, Jun Chen, Ian M Carroll, Tamar Ringel-Kulka, Michael P Epstein, Hua Zhou, Jin J Zhou, Yehuda Ringel, Hongzhe Li, and Michael C Wu. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *The American Journal of Human Genetics*, 96(5):797–807, 2015.

[53] Anna Plantinga, Xiang Zhan, Ni Zhao, Jun Chen, Robert R Jenq, and Michael C Wu. MiRKAT-S: a community-level test of association between the microbiota and survival times. *Microbiome*, 5(1):17, 2017.

[54] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[55] Diptavo Dutta, Laura Scott, Michael Boehnke, and Seunggeun Lee. Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. *Genetic Epidemiology*, 43(1):4–23, 2019.

[56] K Alaine Broadaway, David J Cutler, Richard Duncan, Jacob L Moore, Erin B Ware, Min A Jhun, Lawrence F Bielak, Wei Zhao, Jennifer A Smith, Patricia A Peyser, et al. A statistical approach for testing cross-phenotype effects of rare variants. *The American Journal of Human Genetics*, 98(3):525–540, 2016.

[57] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[58] Saskia Freytag, Juliane Manitz, Martin Schlather, Thomas Kneib, Christopher I Amos, Angela Risch, Jenny Chang-Claude, Joachim Heinrich, and Heike Bickeböller. A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. *Human Heredity*, 76(2):64–75, 2013.

[59] Catherine Lozupone and Rob Knight. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, 2005.

[60] Catherine A Lozupone, Micah Hamady, Scott T Kelley, and Rob Knight. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73(5):1576–1585, 2007.

[61] Jun Chen, Kyle Bittinger, Emily S Charlson, Christian Hoffmann, James Lewis, Gary D Wu, Ronald G Collman, Frederic D Bushman, and Hongzhe Li. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28(16):2106–2113, 2012.

[62] John Aitchison. A new approach to null correlations of proportions. *Journal of the International Association for Mathematical Geology*, 13(2):175–189, 1981.

[63] Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 8:2224, 2017.

[64] Justin D Silverman, Alex D Washburne, Sayan Mukherjee, and Lawrence A David. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, 6:e21887, 2017.

[65] Matthew P Conomos, Cecelia A Laurie, Adrienne M Stilp, Stephanie M Gogarten, Caitlin P McHugh, Sarah C Nelson, Tamar Sofer, Lindsay Fernández-Rhodes, Anne E Justice, Mariaelisa Graff, et al. Genetic diversity and association studies in US Hispanic/Latino populations: applications in the Hispanic Community Health Study/Study of Latinos. *The American Journal of Human Genetics*, 98(1):165–184, 2016.

[66] Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8):816–834, 2010.

[67] Jack A Gilbert, Janet K Jansson, and Rob Knight. Earth microbiome project and global systems biology. *mSystems*, 3(3):e00217–17, 2018.

[68] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[69] 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, et al. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, 2012.

[70] Ilya Shlyakhter, Pardis C Sabeti, and Stephen F Schaffner. Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, 30(23):3427–3429, 2014.

[71] Jun Chen and Hongzhe Li. Kernel methods for regression analysis of microbiome compositional data. In *Topics in Applied Statistics*, pages 191–201. Springer, 2013.

[72] Emily S Charlson, Jun Chen, Rebecca Custers-Allen, Kyle Bittinger, Hongzhe Li, Rohini Sinha, Jennifer Hwang, Frederic D Bushman, and Ronald G Collman. Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS ONE*, 5(12):e15216, 2010.

[73] Emilie Duvallet, Luca Semerano, Eric Assier, Géraldine Falgarone, and Marie-Christophe Boissier. Interleukin-23: a key cytokine in inflammatory diseases. *Annals of Medicine*, 43(7):503–511, 2011.

[74] Richard H Duerr, Kent D Taylor, Steven R Brant, John D Rioux, Mark S Silverberg, Mark J Daly, A Hillary Steinhart, Clara Abraham, Miguel Regueiro, Anne Griffiths, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, 314(5804):1461–1463, 2006.

[75] Durga Sivanesan, Claudine Beauchamp, Christiane Quinou, Jonathan Lee, Sylvie Lesage, Sylvain Chemtob, John D Rioux, and Stephen W Michnick. IL23R (interleukin 23 receptor) variants protective against inflammatory bowel diseases (IBD) display loss of function due to impaired protein stability and intracellular trafficking. *Journal of Biological Chemistry*, 291(16):8673–8685, 2016.

[76] Martha Zakrzewski, Lisa A Simms, Allison Brown, Mark Appleyard, James Irwin, Nicola Waddell, and Graham L Radford-Smith. IL23R-protective coding variant promotes beneficial bacteria and diversity in the ileal microbiome in healthy individuals without inflammatory bowel disease. *Journal of Crohn's and Colitis*, 13(4):451–461, 2019.

[77] Konrad Aden, Ateequr Rehman, Maren Falk-Paulsen, Thomas Secher, Jan Kuiper, Florian Tran, Steffen Pfeuffer, Raheleh Sheibani-Tezerji, Alexandra Breuer, Anne Luzius, et al. Epithelial IL-23R signaling licenses protective IL-22 responses in intestinal inflammation. *Cell Reports*, 16(8):2208–2218, 2016.

[78] Shengping Hou, Liping Du, Bo Lei, Chi Pui Pang, Meifen Zhang, Wenjuan Zhuang, Minglian Zhang, Lulin Huang, Bo Gong, Meilin Wang, et al. Genome-wide association analysis of Vogt-Koyanagi-Harada syndrome identifies two new susceptibility loci at 1p31. 2 and 10q21. 3. *Nature Genetics*, 46(9):1007–1011, 2014.

[79] Takuto Sakono, Akira Meguro, Masaki Takeuchi, Takahiro Yamane, Takeshi Teshigawara, Nobuyoshi Kitaichi, Yukihiro Horie, Kenichi Namba, Shigeaki Ohno, Kumiko Nakao, et al. Variants in IL23R-C1orf141 and ADO-ZNF365-EGR2 are associated with susceptibility to Vogt-Koyanagi-Harada disease in japanese population. *PLoS ONE*, 15(5):e0233464, 2020.

[80] Nazmul Haque, Ryota Ouda, Chao Chen, Keiko Ozato, and J Robert Hogg. ZFR coordinates crosstalk between RNA decay and transcription in innate immunity. *Nature Communications*, 9(1):1–13, 2018.

[81] Vandana A Gupta, Karim Hnia, Laura L Smith, Stacey R Gundry, Jessica E McIntire, Junko Shimazu, Jessica R Bass, Ethan A Talbot, Leonela Amoasii, Nathaniel E Goldman, et al. Loss of catalytically inactive lipid phosphatase myotubularin-related protein 12 impairs myotubularin stability and promotes centronuclear myopathy in zebrafish. *PLoS Genetics*, 9(6):e1003583, 2013.

[82] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R Mende, Gabriel R Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, et al. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, 2011.

[83] Anastassia Gorvitovskaia, Susan P Holmes, and Susan M Huse. Interpreting prevotella and bacteroides as biomarkers of diet and lifestyle. *Microbiome*, 4(1):1–12, 2016.

[84] Yingting Zhou and Fachao Zhi. Lower level of bacteroides in the gut microbiota is associated with inflammatory bowel disease: a meta-analysis. *BioMed Research International*, 2016, 2016.

[85] Arnau Vich Vila, Floris Imhann, Valerie Collij, Soesma A Jankipersadsing, Thomas Gurry, Zlatan Mujagic, Alexander Kurilshikov, Marc Jan Bonder, Xiaofang Jiang, Ettje F Tigchelaar, et al. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Science Translational Medicine*, 10(472):eaap8914, 2018.

[86] Jeppe Madura Larsen. The immune response to prevotella bacteria in chronic inflammatory disease. *Immunology*, 151(4):363–374, 2017.

[87] Xiaoqiong Gu, Jean XY Sim, Wei Lin Lee, Liang Cui, Yvonne FZ Chan, Ega Danu Chang, Yii Ean Teh, An-Ni Zhang, Federica Armas, Franciscus Chandra, et al. Gut ruminococcaceae levels at baseline correlate with risk of antibiotic-associated diarrhea. *iScience*, 25(1):103644, 2022.

[88] Danping Zheng, Timur Liwinski, and Eran Elinav. Interaction between microbiota and immunity in health and disease. *Cell Research*, 30(6):492–506, 2020.

[89] Jasmohan S Bajaj, Jason M Ridlon, Phillip B Hylemon, Leroy R Thacker, Douglas M Heuman, Sean Smith, Masoumeh Sikaroodi, and Patrick M Gillevet. Linkage of gut microbiome with cognition in hepatic encephalopathy. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 302(1):G168–G175, 2012.

[90] Ana Montalban-Arques, Egle Katkeviciute, Philipp Busenhart, Anna Bircher, Jakob Wirbel, Georg Zeller, Yasser Morsy, Lubor Borsig, Jesus F Glaus Garzon, Anne Müller, et al. Commensal clostridiales strains mediate effective anti-cancer immune response against solid tumors. *Cell Host & Microbe*, 29(10):1573–1588, 2021.

[91] Nabil Sabri Enattah, Timo Sahi, Erkki Savilahti, Joseph D Terwilliger, Leena Peltonen, and Irma Järvelä. Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, 30(2):233–237, 2002.

[92] Fengzhe Xu, Yuanqing Fu, Ting-yu Sun, Zengliang Jiang, Zelei Miao, Menglei Shuai, Wanglong Gou, Chu-wen Ling, Jian Yang, Jun Wang, et al. The interplay between host genetics and the gut microbiome reveals common and distinct microbiome features for complex human diseases. *Microbiome*, 8(1):1–14, 2020.

[93] Julia K Goodrich, Emily R Davenport, Michelle Beaumont, Matthew A Jackson, Rob Knight, Carole Ober, Tim D Spector, Jordana T Bell, Andrew G Clark, and Ruth E Ley. Genetic determinants of the gut microbiome in UK twins. *Cell Host & Microbe*, 19(5):731–743, 2016.

[94] Markus Böhm and Susanne Grässel. Role of proopiomelanocortin-derived peptides and their receptors in the osteoarticular system: from basic to translational research. *Endocrine Reviews*, 33(4):623–651, 2012.

[95] Sergey V Kozyrev, Anna-Karin Abelson, Jerome Wojcik, Ammar Zaghlool, Linga Reddy, MV Prasad, Elena Sanchez, Iva Gunnarsson, Elisabet Svenungsson, Gunnar Sturfelt, et al. Functional variants in the b-cell gene bank1 are associated with systemic lupus erythematosus. *Nature Genetics*, 40(2):211–216, 2008.

[96] Catherine Labbé, Philippe Goyette, Céline Lefebvre, Christine Stevens, Todd Green, Marcela K Tello-Ruiz, Zhifang Cao, Aimee L Landry, Joanne Stempak, Vito Annese, et al. Mast3: a novel ibd risk factor that modulates tlr4 signaling. *Genes & Immunity*, 9(7):602–612, 2008.

[97] Byong Duk Ye and Dermot PB McGovern. Genetic variation in ibd: progress, clues to pathogenesis and possible clinical utility. *Expert Review of Clinical Immunology*, 12(10):1091–1107, 2016.

[98] Luke Jostins, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, Carl A Anderson, et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, 2012.

[99] Louis J Cohen, Judy H Cho, Dirk Gevers, and Hiutung Chu. Genetic factors and the intestinal microbiome guide development of microbe-based therapies for inflammatory bowel diseases. *Gastroenterology*, 156(8):2174–2189, 2019.

[100] Xochitl C Morgan, Timothy L Tickle, Harry Sokol, Dirk Gevers, Kathryn L Devaney, Doyle V Ward, Joshua A Reyes, Samir A Shah, Neal LeLeiko, Scott B Snapper, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, 13(9):1–18, 2012.

[101] Aleksandar D Kostic, Ramnik J Xavier, and Dirk Gevers. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology*, 146(6):1489–1499, 2014.

[102] Ashwin N Ananthakrishnan, Chengwei Luo, Vijay Yajnik, Hamed Khalili, John J Garber, Betsy W Stevens, Thomas Cleland, and Ramnik J Xavier. Gut microbiome function predicts response to anti-integrin biologic therapy in inflammatory bowel diseases. *Cell Host & Microbe*, 21(5):603–610, 2017.

[103] Harry Sokol, Loic Brot, Carmen Stefanescu, Claire Auzolle, Nicolas Barnich, Anthony Buisson, Mathurin Fumery, Benjamin Pariente, Lionel Le Bourhis, Xavier Treton, et al. Prominence of ileal mucosa-associated microbiota to predict postoperative endoscopic recurrence in crohn's disease. *Gut*, 69(3):462–472, 2020.

[104] Yaowu Liu, Sixing Chen, Zilin Li, Alanna C Morrison, Eric Boerwinkle, and Xihong Lin. ACAT: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104(3):410–421, 2019.

[105] Hongjiao Liu, Anna Plantinga, Yunhua Xiang, and Michael Wu. A kernel-based test of independence for cluster-correlated data. *Advances in Neural Information Processing Systems*, 34:9869–9881, 2021.

[106] Jurg Ott, Yoichiro Kamatani, and Mark Lathrop. Family-based designs for genome-wide association studies. *Nature Reviews Genetics*, 12(7):465–474, 2011.

[107] Andrzej T Galecki. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods*, 23(11):3105–3119, 1994.

[108] Gregory Reinsel. Estimation and prediction in a multivariate random effects generalized linear model. *Journal of the American Statistical Association*, 79(386):406–414, 1984.

[109] Geert Verbeke, Steffen Fieuws, Geert Molenberghs, and Marie Davidian. The analysis of multivariate longitudinal data: a review. *Statistical Methods in Medical Research*, 23(1):42–59, 2014.

[110] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

[111] Guochang Wang, Wai Keung Li, and Ke Zhu. New hsic-based tests for independence between two stationary multivariate time series. *Statistica Sinica*, 31(1):269–300, 2021.

[112] Pratyaydipta Rudra, K Alaine Broadaway, Erin B Ware, Min A Jhun, Lawrence F Bielak, Wei Zhao, Jennifer A Smith, Patricia A Peyser, Sharon LR Kardia, Michael P Epstein, et al. Testing cross-phenotype effects of rare variants in longitudinal studies of complex traits. *Genetic Epidemiology*, 42(4):320–332, 2018.

[113] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, Bernhard Schölkopf, et al. Kernel methods for measuring independence. *The Journal of Machine Learning Research*, 6:2075–2129, 2005b.

[114] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007.

[115] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, Arlington, Virginia, USA, 2011. AUAI Press.

[116] Aad W Van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

[117] Robert B Davies. The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(3):323–333, 1980.

[118] Sujatha Srinivasan, Martin T Morgan, Tina L Fiedler, Danijel Djukovic, Noah G Hoffman, Daniel Raftery, Jeanne M Marrazzo, and David N Fredricks. Metabolic signatures of bacterial vaginosis. *mBio*, 6(2), 2015.

[119] Caroline M Mitchell, Susan D Reed, Susan Diem, Joseph C Larson, Katherine M Newton, Kristine E Ensrud, Andrea Z LaCroix, Bette Caan, and Katherine A Guthrie. Efficacy of vaginal estradiol or vaginal moisturizer vs placebo for treating postmenopausal vulvovaginal symptoms: a randomized clinical trial. *JAMA Internal Medicine*, 178(5):681–690, 2018.

[120] Caroline M Mitchell, Nanxun Ma, Alissa J Mitchell, Michael C Wu, DJ Valint, Sean Proll, Susan D Reed, Katherine A Guthrie, Andrea Z Lacroix, Joseph C Larson, et al. Association between postmenopausal vulvovaginal discomfort, vaginal microbiota, and mucosal inflammation. *American Journal of Obstetrics and Gynecology*, 2021.

[121] Steven S Witkin and Iara M Linhares. Why do lactobacilli dominate the human vaginal microbiota? *BJOG: An International Journal of Obstetrics & Gynaecology*, 124(4):606–611, 2017.

[122] Camilla Ceccarani, Claudio Foschi, Carola Parolin, Antonietta D'Antuono, Valeria Gaspari, Clarissa Consolandi, Luca Laghi, Tania Camboni, Beatrice Vitali, Marco Severgnini, et al. Diversity of vaginal microbiome and metabolome during genital infections. *Scientific Reports*, 9(1):1–12, 2019.

[123] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, ESP Lung Project Team, David C Christiani, Mark M Wurfel, Xihong Lin, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012.

[124] Jacques Ravel, Pawel Gajer, Zaid Abdo, G Maria Schneider, Sara SK Koenig, Stacey L McCulle, Shara Karlebach, Reshma Gorle, Jennifer Russell, Carol O Tacket, et al. Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, 108(supplement_1):4680–4687, 2011.

[125] Mariya I Petrova, Elke Lievens, Shweta Malik, Nicole Imholz, and Sarah Lebeer. Lactobacillus species as biomarkers and agents that can promote various aspects of vaginal health. *Frontiers in Physiology*, 6:81, 2015.

[126] Rebecca M Brotman et al. Vaginal microbiome and sexually transmitted infections: an epidemiologic perspective. *The Journal of Clinical Investigation*, 121(12):4610–4617, 2011.

[127] Jennifer M Fettweis, Myrna G Serrano, J Paul Brooks, David J Edwards, Philippe H Girerd, Hardik I Parikh, Bernice Huang, Tom J Arodz, Laahirie Edupuganti, Abigail L Glascock, et al. The vaginal microbiome and preterm birth. *Nature Medicine*, 25(6):1012–1021, 2019.

[128] Xiaodi Chen, Yune Lu, Tao Chen, and Rongguo Li. The female vaginal microbiome in health and bacterial vaginosis. *Frontiers in Cellular and Infection Microbiology*, 11:631972, 2021.

[129] Myrna G Serrano, Hardik I Parikh, J Paul Brooks, David J Edwards, Tom J Arodz, Laahirie Edupuganti, Bernice Huang, Philippe H Girerd, Yahya A Bokhari, Steven P Bradley, et al. Racioethnic diversity in the dynamics of the vaginal microbiome during pregnancy. *Nature Medicine*, 25(6):1001–1011, 2019.

[130] Jennifer M Fettweis, J Paul Brooks, Myrna G Serrano, Nihar U Sheth, Philippe H Girerd, David J Edwards, Jerome F Strauss III, Kimberly K Jefferson, Gregory A Buck, Vaginal Microbiome Consortium, et al. Differences in vaginal microbiome in african american women versus women of european ancestry. *Microbiology*, 160(Pt 10):2272, 2014.

[131] Vinod K Gupta, Sandip Paul, and Chitra Dutta. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Frontiers in Microbiology*, 8:1162, 2017.

[132] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, 2012.

[133] Karoline Faust, J Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial co-occurrence relationships in the human microbiome. *PLoS Computational Biology*, 8(7):e1002606, 2012.

[134] Huaying Fang, Chengcheng Huang, Hongyu Zhao, and Minghua Deng. gcoda: conditional dependence network inference for compositional data. *Journal of Computational Biology*, 24(7):699–708, 2017.

[135] Jing Ma. Joint microbial and metabolomic network estimation with the censored gaussian graphical model. *Statistics in Biosciences*, 13(2):351–372, 2021.

[136] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[137] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[138] Jennifer M Fettweis, Myrna G Serrano, Nihar U Sheth, Carly M Mayer, Abigail L Glascock, J Paul Brooks, Kimberly K Jefferson, and Gregory A Buck. Species-level classification of the vaginal microbiome. *BMC Genomics*, 13(8):1–9, 2012.

[139] Mariya I Petrova, Gregor Reid, Mario Vaneechoutte, and Sarah Lebeer. Lactobacillus iners: friend or foe? *Trends in Microbiology*, 25(3):182–191, 2017.

[140] Paul Robert and Yves Escoufier. A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 25(3):257–265, 1976.

[141] Yunhua Xiang and Noah Simon. A flexible framework for nonparametric graphical modeling that accommodates machine learning. In *International Conference on Machine Learning*, pages 10442–10451. PMLR, 2020.

[142] Eldin Jašarević, Elizabeth M Hill, Patrick J Kane, Lindsay Rutt, Trevonn Gyles, Lillian Folts, Kylie D Rock, Christopher D Howard, Kathleen E Morrison, Jacques Ravel, et al. The composition of human vaginal microbiota transferred at birth affects offspring health in a mouse model. *Nature Communications*, 12(1):6289, 2021.

[143] Shan Sun, Myrna G Serrano, Jennifer M Fettweis, Patricia Basta, Emma Rosen, Kim Ludwig, Alicia A Sorgen, Ivory C Blakley, Michael C Wu, Nancy Dole, et al. Race, the vaginal microbiome, and spontaneous preterm birth. *mSystems*, 7(3):e00017–22, 2022.

[144] Julie Josse and Susan Holmes. Measuring multivariate association and beyond. *Statistics Surveys*, 10:132, 2016.

[145] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[146] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5:1–18, 2017.

[147] Huang Lin and Shyamal Das Peddada. Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ Biofilms and Microbiomes*, 6(1):60, 2020.

[148] Julie Josse, Jérome Pagès, and François Husson. Testing the significance of the rv coefficient. *Computational Statistics & Data Analysis*, 53(1):82–91, 2008.

[149] Daniel McDonald, Embriette Hyde, Justine W Debelius, James T Morton, Antonio Gonzalez, Gail Ackermann, Alexander A Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, et al. American gut: an open platform for citizen science microbiome research. *mSystems*, 3(3):e00031–18, 2018.

[150] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, pages 935–980, 2011.

[151] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[152] Linton C Freeman et al. Centrality in social networks: Conceptual clarification. *Social Network: Critical Concepts in Sociology. Londres: Routledge*, 1:238–263, 2002.

[153] R Poudel, Ari Jumpponen, Dan C Schlatter, TC Paulitz, BB McSpadden Gardener, Linda L Kinkel, and KA Garrett. Microbiome networks: a systems framework for identifying candidate microbial assemblages for disease management. *Phytopathology*, 106(10):1083–1096, 2016.

[154] Phillip Bonacich. Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182, 1987.

[155] Britta Ruhnau. Eigenvector-centrality—a node-centrality? *Social networks*, 22(4):357–365, 2000.

[156] Tatyana Zamkovaya, Jamie S Foster, Valérie de Crécy-Lagard, and Ana Conesa. A network approach to elucidate and prioritize microbial dark matter in microbial communities. *The ISME Journal*, 15(1):228–244, 2021.

[157] Stefanie Peschel, Christian L Müller, Erika von Mutius, Anne-Laure Boulesteix, and Martin Depner. NetCoMi: network construction and comparison for microbiome data in r. *Briefings in Bioinformatics*, 22(4):bbaa290, 2021.

[158] David N Fredricks, Tina L Fiedler, Katherine K Thomas, Brian B Oakley, and Jeanne M Marrazzo. Targeted pcr for detection of vaginal bacteria associated with bacterial vaginosis. *Journal of Clinical Microbiology*, 45(10):3270–3276, 2007.

[159] Andrew B Onderdonk, Mary L Delaney, and Raina N Fichorova. The human microbiome during bacterial vaginosis. *Clinical Microbiology Reviews*, 29(2):223–238, 2016.

[160] Laura Donati, Augusto Di Vico, Marta Nucci, Lorena Quagliozzi, Terryann Spagnuolo, Antonietta Labianca, Marina Bracaglia, Francesca Ianniello, Alessandro Caruso, and Giancarlo Paradisi. Vaginal microbial flora and outcome of pregnancy. *Archives of Gynecology and Obstetrics*, 281:589–600, 2010.

[161] Teija Ojala, Matti Kankainen, Joana Castro, Nuno Cerca, Sanna Edelman, Benita Westerlund-Wikström, Lars Paulin, Liisa Holm, and Petri Auvinen. Comparative genomics of lactobacillus crispatus suggests novel mechanisms for the competitive exclusion of gardnerella vaginalis. *BMC Genomics*, 15(1):1–21, 2014.

[162] Paola Mastromarino, Beatrice Vitali, and Luciana Mosca. Bacterial vaginosis: a review on clinical trials with probiotics. *New Microbiol*, 36(3):229–238, 2013.

[163] Abigail L Glascock, Nicole R Jimenez, Sam Boundy, Vishal N Koparde, J Paul Brooks, David J Edwards, Jerome F Strauss III, Kimberly K Jefferson, Myrna G Serrano, Gregory A Buck, et al. Unique roles of vaginal megasphaera phylotypes in reproductive health. *Microbial Genomics*, 7(12), 2021.

[164] Nengneng Zheng, Renyong Guo, Jinxi Wang, Wei Zhou, and Zongxin Ling. Contribution of lactobacillus iners to vaginal health and diseases: A systematic review. *Frontiers in Cellular and Infection Microbiology*, page 1177, 2021.

[165] Joana Castro, Ana Henriques, António Machado, Mariana Henriques, Kimberly K Jefferson, and Nuno Cerca. Reciprocal interference between lactobacillus spp. and gardnerella vaginalis on initial adherence to epithelial cells. *International Journal of Medical Sciences*, 10(9):1193, 2013.

[166] Jing Li, Fangqing Zhao, Yidan Wang, Junru Chen, Jie Tao, Gang Tian, Shouling Wu, Wenbin Liu, Qinghua Cui, Bin Geng, et al. Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome*, 5:1–19, 2017.

[167] John D Sargan. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, pages 393–415, 1958.

[168] George Davey Smith and Shah Ebrahim. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22, 2003.

[169] Lizhen Xu, Andrew D Paterson, Williams Turpin, and Wei Xu. Assessment and selection of competing models for zero-inflated microbiome data. *PLoS One*, 10(7):e0129606, 2015.

[170] Jing-Jing Ni, Xiao-Song Li, Hong Zhang, Qian Xu, Xin-Tong Wei, Gui-Juan Feng, Min Zhao, Zi-Jia Zhang, Lei Zhang, Gen-Hai Shen, et al. Mendelian randomization study of causal link from gut microbiota to colorectal cancer. *BMC Cancer*, 22(1):1371, 2022.

[171] Yiwen Long, Lanhua Tang, Yangying Zhou, Shushan Zhao, and Hong Zhu. Causal relationship between gut microbiota and cancers: a two-sample mendelian randomisation study. *BMC medicine*, 21(1):1–14, 2023.

[172] Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication, Meta analysis (DIAGRAM) Consortium, Pamela AF Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, 44(4):369–375, 2012.

[173] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525, 2015.

[174] Renée Adams, Heitor Almeida, and Daniel Ferreira. Understanding the relationship between founder–ceos and firm performance. *Journal of Empirical Finance*, 16(1):136–150, 2009.

[175] Takeshi Amemiya. The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2(2):105–110, 1974.

[176] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

[177] Yinglin Xia, Jun Sun, Ding-Geng Chen, Yinglin Xia, Jun Sun, and Ding-Geng Chen. Modeling zero-inflated microbiome data. *Statistical Analysis of Microbiome Data with R*, pages 453–496, 2018.

[178] Jun Chen, Emily King, Rebecca Deek, Zhi Wei, Yue Yu, Diane Grill, and Karla Ballman. An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics*, 34(4):643–651, 2018.

[179] Yuanjing Ma, Yuan Luo, and Hongmei Jiang. A novel normalization and differential abundance test framework for microbiome data. *Bioinformatics*, 36(13):3959–3965, 2020.

[180] A Colin Cameron and Pravin K Trivedi. *Regression analysis of count data*, volume 53. Cambridge University Press, 2013.

[181] Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression models for count data in r. *Journal of Statistical Software*, 27(8):1–25, 2008.

[182] Russell Davidson, James G MacKinnon, et al. *Estimation and inference in econometrics*, volume 63. Oxford New York, 1993.

[183] Bryan D Martin, Daniela Witten, and Amy D Willis. Modeling microbial abundances and dysbiosis with beta-binomial regression. *The Annals of Applied Statistics*, 14(1):94, 2020.

[184] Daniel T Lackland and Michael A Weber. Global burden of cardiovascular disease and stroke: hypertension at the core. *Canadian Journal of Cardiology*, 31(5):569–571, 2015.

[185] Intza Hernandorena, Emmanuelle Duron, Jean-Sébastien Vidal, and Olivier Hanon. Treatment options and considerations for hypertensive patients to prevent dementia. *Expert Opinion on Pharmacotherapy*, 18(10):989–1000, 2017.

[186] Shan Sun, Anju Lulla, Michael Sioda, Kathryn Winglee, Michael C Wu, David R Jacobs Jr, James M Shikany, Donald M Lloyd-Jones, Lenore J Launer, Anthony A Fodor, et al. Gut microbiota composition and blood pressure: the cardia study. *Hypertension*, 73(5):998–1006, 2019.

[187] Joonatan Palmu, Aaro Salosensaari, Aki S Havulinna, Susan Cheng, Michael Inouye, Mohit Jain, Rodolfo A Salido, Karenina Sanders, Caitriona Brennan, Gregory C Humphrey, et al. Association between the gut microbiota and blood pressure in a population cohort of 6953 individuals. *Journal of the American Heart Association*, 9(15):e016641, 2020.

[188] David J Durgan, Bhanu P Ganesh, Julia L Cope, Nadim J Ajami, Sharon C Phillips, Joseph F Petrosino, Emily B Hollister, and Robert M Bryan Jr. Role of the gut

microbiome in obstructive sleep apnea–induced hypertension. *Hypertension*, 67(2):469–474, 2016.

[189] Lisa M LaVange, William D Kalsbeek, Paul D Sorlie, Larissa M Avilés-Santa, Robert C Kaplan, Janice Barnhart, Kiang Liu, Aida Giachello, David J Lee, John Ryan, et al. Sample design and cohort selection in the hispanic community health study/study of latinos. *Annals of Epidemiology*, 20(8):642–649, 2010.

[190] Tali Elfassy, Adina Zeki Al Hazzouri, Jianwen Cai, Pedro L Baldoni, Maria M Llabre, Tatjana Rundek, Leopoldo Raij, James P Lash, Gregory A Talavera, Sylvia Wassertheil-Smoller, et al. Incidence of hypertension among us hispanics/latinos: the hispanic community health study/study of latinos, 2008 to 2017. *Journal of the American Heart Association*, 9(12):e015031, 2020.

[191] Antonio Gonzalez, Jose A Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, Stefan Janssen, Austin D Swafford, Stephanie B Orchanian, et al. Qiita: rapid, web-enabled microbiome meta-analysis. *Nature Methods*, 15(10):796–798, 2018.

[192] Mykhaylo Usyk, Brandilyn A Peters, Smruthi Karthikeyan, Daniel McDonald, Christopher C Sollecito, Yoshiki Vazquez-Baeza, Justin P Shaffer, Marc D Gellman, Gregory A Talavera, Martha L Daviglus, et al. Comprehensive evaluation of shotgun metagenomics, amplicon sequencing, and harmonization of these platforms for epidemiological studies. *Cell Reports Methods*, 3(1), 2023.

[193] Jane F Reckelhoff. Gender differences in the regulation of blood pressure. *Hypertension*, 37(5):1199–1208, 2001.

[194] Elisabete Pinto. Blood pressure and ageing. *Postgraduate Medical Journal*, 83(976):109–114, 2007.

[195] Frank Windmeijer, Xiaoran Liang, Fernando P Hartwig, and Jack Bowden. The confidence interval method for selecting valid instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):752–776, 2021.

[196] Gwen Falony, Marie Joossens, Sara Vieira-Silva, Jun Wang, Youssef Darzi, Karoline Faust, Alexander Kurilshikov, Marc Jan Bonder, Mireia Valles-Colomer, Doris Vandeputte, et al. Population-level analysis of gut microbiome variation. *Science*, 352(6285):560–564, 2016.

[197] Mitsuo Sakamoto and Yoshimi Benno. Reclassification of bacteroides distasonis, bacteroides goldsteinii and bacteroides merdae as parabacteroides distasonis gen. nov.,

comb. nov., parabacteroides goldsteinii comb. nov. and parabacteroides merdae comb. nov. *International Journal of Systematic and Evolutionary Microbiology*, 56(7):1599–1605, 2006.

[198] Hamdi A Jama, Dakota Rhys-Jones, Michael Nakai, Chu K Yao, Rachel E Climie, Yusuke Sata, Dovile Anderson, Darren J Creek, Geoffrey A Head, David M Kaye, et al. Prebiotic intervention with hamsab in untreated essential hypertensive patients assessed in a phase ii randomized trial. *Nature Cardiovascular Research*, pages 1–9, 2023.

[199] Kai Wang, Mingfang Liao, Nan Zhou, Li Bao, Ke Ma, Zhongyong Zheng, Yujing Wang, Chang Liu, Wenzhao Wang, Jun Wang, et al. Parabacteroides distasonis alleviates obesity and metabolic dysfunctions via production of succinate and secondary bile acids. *Cell Reports*, 26(1):222–235, 2019.

[200] Yuanyuan Lei, Li Tang, Shuang Liu, Shiping Hu, Lingyi Wu, Yaojiang Liu, Min Yang, Shengjie Huang, Xuefeng Tang, Tao Tang, et al. Parabacteroides produces acetate to alleviate heparanase-exacerbated acute pancreatitis through reducing neutrophil infiltration. *Microbiome*, 9(1):115, 2021.

[201] Carmen De Miguel, Nathan P Rudemiller, Justine M Abais, and David L Mattson. Inflammation and hypertension: new understandings and potential therapeutic targets. *Current Hypertension Reports*, 17:1–10, 2015.

[202] Marci G Crowley, Yukihiro Nakanishi, Yanina Pasikhova, John N Greene, and Avan J Armaghani. A second reported case of catabacter hongkongensis bactermia in the united states. *Infectious Diseases in Clinical Practice*, 31(1):e1212, 2023.

[203] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):1–35, 2020.

[204] Saket Choudhary and Rahul Satija. Comparison and evaluation of statistical error models for scrna-seq. *Genome biology*, 23(1):27, 2022.

[205] Shiquan Sun, Jiaqiang Zhu, and Xiang Zhou. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods*, 17(2):193–200, 2020.

[206] Ruben Dries, Jiaji Chen, Natalie Del Rossi, Mohammed Muzamil Khan, Adriana Sistig, and Guo-Cheng Yuan. Advances in spatial transcriptomic data analysis. *Genome Research*, 31(10):1706–1718, 2021.

[207] Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1), 2019.

[208] Marti J Anderson. Permutational multivariate analysis of variance (PERMANOVA). *Wiley Statsref: Statistics Reference Online*, pages 1–15, 2014.

[209] Vladimir Koltchinskii, Evarist Giné, et al. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 2000.

[210] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[211] Richard M Dudley. Fréchet differentiability, p-variation and uniform donsker classes. In *Selected Works of RM Dudley*, pages 371–385. Springer, 2010.

[212] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[213] Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux, Jean-François Paiement, Pascal Vincent, and Marie Ouimet. Learning eigenfunctions links spectral embedding and kernel pca. *Neural Computation*, 16(10):2197–2219, 2004.

[214] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[215] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *NIPS*, volume 23, pages 673–681, 2009.

[216] Zaid Harchaoui, Francis R Bach, and Eric Moulines. Testing for homogeneity with kernel fisher discriminant analysis. In *NIPS*, pages 609–616. Citeseer, 2007.

[217] Harold Widom. Asymptotic behavior of block toeplitz matrices and determinants. ii. *Advances in Mathematics*, 21(1):1–29, 1976.

[218] Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer, 1996.

[219] P. Duchesne and P. Lafaye de Micheaux. Computing the distribution of quadratic forms: Further comparisons between the liu-tang-zhang approximation and exact methods. *Computational Statistics and Data Analysis*, 54:858–862, 2010.

[220] DA Savitz, N Dole, J Williams, JM Thorp, T McDonald, AC Carter, and B Eucker. Determinants of participation in an epidemiological study of preterm delivery. *Paediatric and Perinatal Epidemiology*, 13(1):114–125, 1999.

Appendix A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

A.1 Derivation of Covariate-adjusted KRV Coefficient

Suppose that we have a phenotype kernel matrix $\boldsymbol{L}$ and a full-rank covariates matrix $\boldsymbol{X}$ that includes a column of 1's. We first perform a kernel principal component analysis (kernel PCA; equivalent to an eigendecomposition) on the phenotype kernel matrix and obtain a matrix $\boldsymbol{\Phi}$ such that:

$$\boldsymbol{L} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T.$$

Here each column of $\boldsymbol{\Phi}$ is a kernel principal component (kernel PC) of $\boldsymbol{L}$ and has the form $\sqrt{\lambda_r}\boldsymbol{\phi}_r$ for $r = 1, \cdots, n$, where λ_r is the rth eigenvalue of $\boldsymbol{L}$ and $\boldsymbol{\phi}_r$ is the corresponding eigenvector for λ_r. We can view $\boldsymbol{\Phi}$ as a finite sample basis for the space spanned by the phenotype kernel function $\ell(\cdot, \cdot)$.

We then regress out the covariates $\boldsymbol{X}$ from each kernel PC:

$$\hat{\boldsymbol{\epsilon}} := \boldsymbol{\Phi} - \boldsymbol{P}_X\boldsymbol{\Phi},$$

where $\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the projection matrix onto the column space of $\boldsymbol{X}$. Now $\hat{\boldsymbol{\epsilon}}$ represents a sample basis that is orthogonal to the covariates $\boldsymbol{X}$. We can construct a new phenotype kernel matrix from this residual basis: $\boldsymbol{L}^* := \hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T$. Note that $\boldsymbol{L}^*$ can be expressed in terms of $\boldsymbol{L}$:

$$\boldsymbol{L}^* = (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{\Phi}\boldsymbol{\Phi}^T(\boldsymbol{I} - \boldsymbol{P}_X) = (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{L}(\boldsymbol{I} - \boldsymbol{P}_X) = \boldsymbol{P}_X^{\perp}\boldsymbol{L}\boldsymbol{P}_X^{\perp},$$

where we let $\boldsymbol{P}_X^{\perp} := \boldsymbol{I} - \boldsymbol{P}_X$. Similar procedures can be performed on the genotype kernel

matrix $\boldsymbol{K}$ to obtain the adjusted genotype kernel matrix $\boldsymbol{K}^* := \boldsymbol{P}_X^{\perp}\boldsymbol{K}\boldsymbol{P}_X^{\perp}$. Both $\boldsymbol{K}^*$ and $\boldsymbol{L}^*$ are column-centered, since the covariates matrix $\boldsymbol{X}$ includes a column of 1's, accounting for the intercept in a regression. We can then construct a KRV statistic from the adjusted kernel matrices $\boldsymbol{K}^*$ and $\boldsymbol{L}^*$:

$$\mathrm{KRV}_{adj}(G, Y|X) = \frac{\mathrm{tr}(\boldsymbol{K}^*\boldsymbol{L}^*)}{\sqrt{\mathrm{tr}(\boldsymbol{K}^*\boldsymbol{K}^*)\,\mathrm{tr}(\boldsymbol{L}^*\boldsymbol{L}^*)}} = \frac{\mathrm{tr}(\boldsymbol{P}_X^{\perp}\boldsymbol{K}\boldsymbol{P}_X^{\perp}\boldsymbol{L})}{\sqrt{\mathrm{tr}(\boldsymbol{P}_X^{\perp}\boldsymbol{K}\boldsymbol{P}_X^{\perp}\boldsymbol{K})\,\mathrm{tr}(\boldsymbol{P}_X^{\perp}\boldsymbol{L}\boldsymbol{P}_X^{\perp}\boldsymbol{L})}}.$$

Such a strategy of covariate adjustment can be seen as a special case of conditional independence (or uncorrelatedness) testing in a kernel-based framework, as proposed by Zhang et al. and Strobl et al. [115, 207]. In the context of microbiome GWAS, we are testing the correlation between genetic variants and microbiome community profiles, while conditioning on the covariates.

A.1.1 Special Case of the Linear Kernel

Suppose that we use a linear kernel $\ell(\boldsymbol{y}_i, \boldsymbol{y}_j) = \boldsymbol{y}_i^T\boldsymbol{y}_j$ for the phenotype data, where $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iq})^T$ is the set of q traits for individual i.

Let $\boldsymbol{Y}$ be the $n \times q$ matrix that stores the phenotype data for all n individuals. Then the resulting phenotype kernel matrix can be constructed as $\boldsymbol{L} = \boldsymbol{Y}\boldsymbol{Y}^T$. Note that we can rewrite the covariate-adjusted kernel matrix $\boldsymbol{L}^*$ as:

$$\boldsymbol{L}^* = \boldsymbol{P}_X^{\perp}\boldsymbol{L}\boldsymbol{P}_X^{\perp} = \boldsymbol{P}_X^{\perp}\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{P}_X^{\perp} = (\boldsymbol{P}_X^{\perp}\boldsymbol{Y})(\boldsymbol{P}_X^{\perp}\boldsymbol{Y})^T.$$

Therefore, in the case of a linear kernel, our proposed approach for covariate adjustment is equivalent to the previously proposed residual-based approach [54, 56, 21], where we first regress out the covariates from each raw phenotype and then construct the phenotype kernel matrix using the resulting residuals.

A.1.2 Connection between Euclidean Distance and Linear Kernel

When constructing a microbiome kernel matrix, we can often obtain the kernel matrix by transforming existing distance or dissimilarity matrices calculated based on microbiome data. For example, assuming that the original microbial abundance data matrix is $\boldsymbol{Y}$, we can obtain a "CLR-Euclidean" kernel matrix by first constructing the Euclidean distance matrix $\boldsymbol{D}$ based on the CLR-transformed abundance data CLR($\boldsymbol{Y}$) and then transforming $\boldsymbol{D}$ into a kernel matrix $\boldsymbol{L}$ via:

$$\boldsymbol{L} = -\frac{1}{2}\Big(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\Big)\boldsymbol{D}^2\Big(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\Big),$$

where $\boldsymbol{D}^2$ is the element-wise square of $\boldsymbol{D}$.

Now we show that, taking Euclidean distances of data CLR($\boldsymbol{Y}$) followed by kernel matrix transformation is equivalent to constructing a centered linear kernel matrix based on CLR($\boldsymbol{Y}$). For convenience, we still use $\boldsymbol{y}_i$ to represent the CLR-transformed abundances for individual i.

Let d_{ij}^2 be the (i, j)-th entry of matrix $\boldsymbol{D}^2$. Then we have

$$d_{ij}^2 = (\boldsymbol{y}_i - \boldsymbol{y}_j)^T(\boldsymbol{y}_i - \boldsymbol{y}_j) = \boldsymbol{y}_i^T\boldsymbol{y}_i - 2\boldsymbol{y}_i^T\boldsymbol{y}_j + \boldsymbol{y}_j^T\boldsymbol{y}_j.$$

As $\boldsymbol{H} := \boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}$ is a centering matrix, the (i, j)-th entry of matrix $\Big(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\Big)\boldsymbol{D}^2\Big(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\Big)$

becomes

$$\begin{aligned}
\tilde{d}_{ij}^2 &= d_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 \\
&= -2\Big[\boldsymbol{y}_i^T\boldsymbol{y}_j - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i^T\boldsymbol{y}_j - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{y}_i^T\boldsymbol{y}_j + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{y}_i^T\boldsymbol{y}_j\Big] \\
&\quad + \Big[\boldsymbol{y}_i^T\boldsymbol{y}_i + \boldsymbol{y}_j^T\boldsymbol{y}_j\Big] - \frac{1}{n}\sum_{i=1}^{n}\Big[\boldsymbol{y}_i^T\boldsymbol{y}_i + \boldsymbol{y}_j^T\boldsymbol{y}_j\Big] - \frac{1}{n}\sum_{j=1}^{n}\Big[\boldsymbol{y}_i^T\boldsymbol{y}_i + \boldsymbol{y}_j^T\boldsymbol{y}_j\Big] \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[\boldsymbol{y}_i^T\boldsymbol{y}_i + \boldsymbol{y}_j^T\boldsymbol{y}_j\Big] \\
&= -2\Big[\boldsymbol{y}_i^T\boldsymbol{y}_j - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i^T\boldsymbol{y}_j - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{y}_i^T\boldsymbol{y}_j + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{y}_i^T\boldsymbol{y}_j\Big].
\end{aligned}$$

Therefore, the (i,j)-th entry of matrix $\boldsymbol{L}$ is

$$(\boldsymbol{L})_{i,j} = -\frac{1}{2}\tilde{d}_{ij}^2 = \boldsymbol{y}_i^T\boldsymbol{y}_j - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i^T\boldsymbol{y}_j - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{y}_i^T\boldsymbol{y}_j + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{y}_i^T\boldsymbol{y}_j.$$

Consequently, the resulting kernel matrix $\boldsymbol{L}$ is a centered linear kernel matrix based on CLR($\boldsymbol{Y}$): $\boldsymbol{L} = \boldsymbol{H}\boldsymbol{L}_0\boldsymbol{H}$, where $(\boldsymbol{L}_0)_{i,j} = \boldsymbol{y}_i^T\boldsymbol{y}_j$ and $\boldsymbol{H} = \boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}$.

In light of this result, we can view the CLR-Euclidean kernel as a centered linear kernel applied to the CLR-transformed microbiome data (or denote it as the CLR-linear kernel). Applying our proposed covariate adjustment approach to the CLR-Euclidean kernel matrix is thus equivalent to using the residual-based approach on the CLR-transformed data (i.e., regressing out the covariates from the CLR-transformed data and constructing a kernel matrix based on the residuals).

A.2 Taxon-level Microbiome GWAS of the HCHS/SOL Study

As a comparison to our proposed gene-based community-level microbiome GWAS framework, we performed a traditional variant-based taxon-level microbiome GWAS based on the same

set of HCHS/SOL data ($n = 1219$) used in our main analysis, where we tested the association between individual genetic variants and individual microbial genera.

For genetic data, we applied the same quality control criteria as for the community-level analysis and focused on common genetic variants with minor allele frequency (MAF) ≥ 0.05 along the genome (including both coding and non-coding regions). For microbiome data, we focused our analysis on relatively common genera that are present in $\geq 10\%$ of all 1219 individuals under analysis.

To conduct association testing, we performed either linear regression or logistic regression depending on the prevalence of the genera, based on analysis procedures in previous microbiome GWAS studies [18, 92, 15]. Specifically, for genera present in $\geq 90\%$ of individuals, we performed rank normal transformation on the rarefied abundance data to encourage normality and used linear regression to assess the association between each rank-normal-transformed microbial abundance and each genetic variant. For genera present in $\geq 10\%$ but $< 90\%$ of individuals, the presence/absence of each genus was used as the outcome and associated with each genetic variant via logistic regression. Similar to the community-level analysis, the top 5 PCs of genome-wide genetic variability were included in the regression models as covariates. The genome-wide association testing was conducted using the GENESIS R package v2.28.0: https://bioconductor.org/packages/GENESIS.

A.3 Analyses to Assess the Robustness of the IL23R-C1orf141 Signal

Based on our main analysis of 1219 HCHS/SOL subjects, we have identified genome-widely significant associations between variants in *IL23R* and *C1orf141* and gut microbiome composition using the Bray-Curtis kernel (Table 2.1), where population structure, a major confounder captured by the top 5 PCs of genetic variability, was adjusted. However, these two associations no longer have genome-wide significance in a reduced sample ($n = 1096$) where additional covariates (age, gender and study sites) were available and adjusted. To assess the robustness of these two signals, we have conducted several additional analyses.

First, to investigate if there is additional confounding caused by age, gender and study

site, we have assessed the association between our identified loci and these covariates in the reduced sample. *IL23R* and *C1orf141* were combined into a single *IL23R-C1orf141* region due to overlapping variants. We applied the SNP-set kernel association test (SKAT) [54] to assess the association between age/gender and common variants in the *IL23R-C1orf141* region, with a linear model for age and a logistic model for gender. Since there were no available SKAT models to accommodate study site as an outcome, which is a categorical variable with four levels, we used linear regression to regress the genotype of the top variant (rs10789226) in the *IL23R-C1orf141* region on study site. In all models, the population structure captured by the top 5 genome-wide genetic PCs were adjusted. We found no significant association between the genetics and any of the covariates (p-values for age, gender and study site were 0.06, 0.08 and 0.75, respectively), thus confirming that these covariates are not likely to be confounders in the genetics-microbiome relationship in our study.

Next, to discover any systematic differences between participants with ($n = 123$) and without missing data ($n = 1096$) for the three covariates, we have compared the overall microbiome composition and genetic features of the *IL23R-C1orf141* region between these two sub-samples. We plotted the top two kernel PCs of the Bray-Curtis microbiome kernel matrix to identify any clustering by sub-samples and conducted permutational multivariate analysis of variance (PERMANOVA) [208] to test the difference between the two sub-samples in microbiome composition (see Figure A.7). Similar analysis was performed for the genetic kernel matrix, constructed based on common variants in the *IL23R-C1orf141* region using a linear kernel. These analyses revealed no significant difference between participants with and without missing covariates data (PERMANOVA p-value = 0.07 for microbiome; 0.40 for genotypes).

Based on the above analyses, we have further confirmed that there is not likely to be systematic differences between the original sample and the reduced sample, and the additionally adjusted covariates are not likely to be confounders in the genetics-microbiome relationship. These results have confirmed the robustness of our identified genetic loci based on the Bray-

Curtis kernel, and the reduced genome-wide significance in the reduced sample is likely due to sample size loss.

A.4 Supplementary Tables and Figures

Table A.1: P-values for the significant genes from Table 2.1 when additional covariates were adjusted in the first-stage KRV analysis of the HCHS/SOL data.

Microbiome kernel	Genes	Number of common variants	P-value
Bray-Curtis	*C1orf141*	484	2.5×10^{-5}
	IL23R	284	3.7×10^{-5}
Unweighted UniFrac	*MTMR12*	174	2.3×10^{-7}
	ZFR	288	3.3×10^{-8}
CLR-linear	*MTMR12*	174	3.3×10^{-6}

Adjusted covariates include the top 5 PCs of genome-wide genetic variability, age, gender and study sites. The analysis was performed on 1096 unrelated individuals where all relevant data were available.

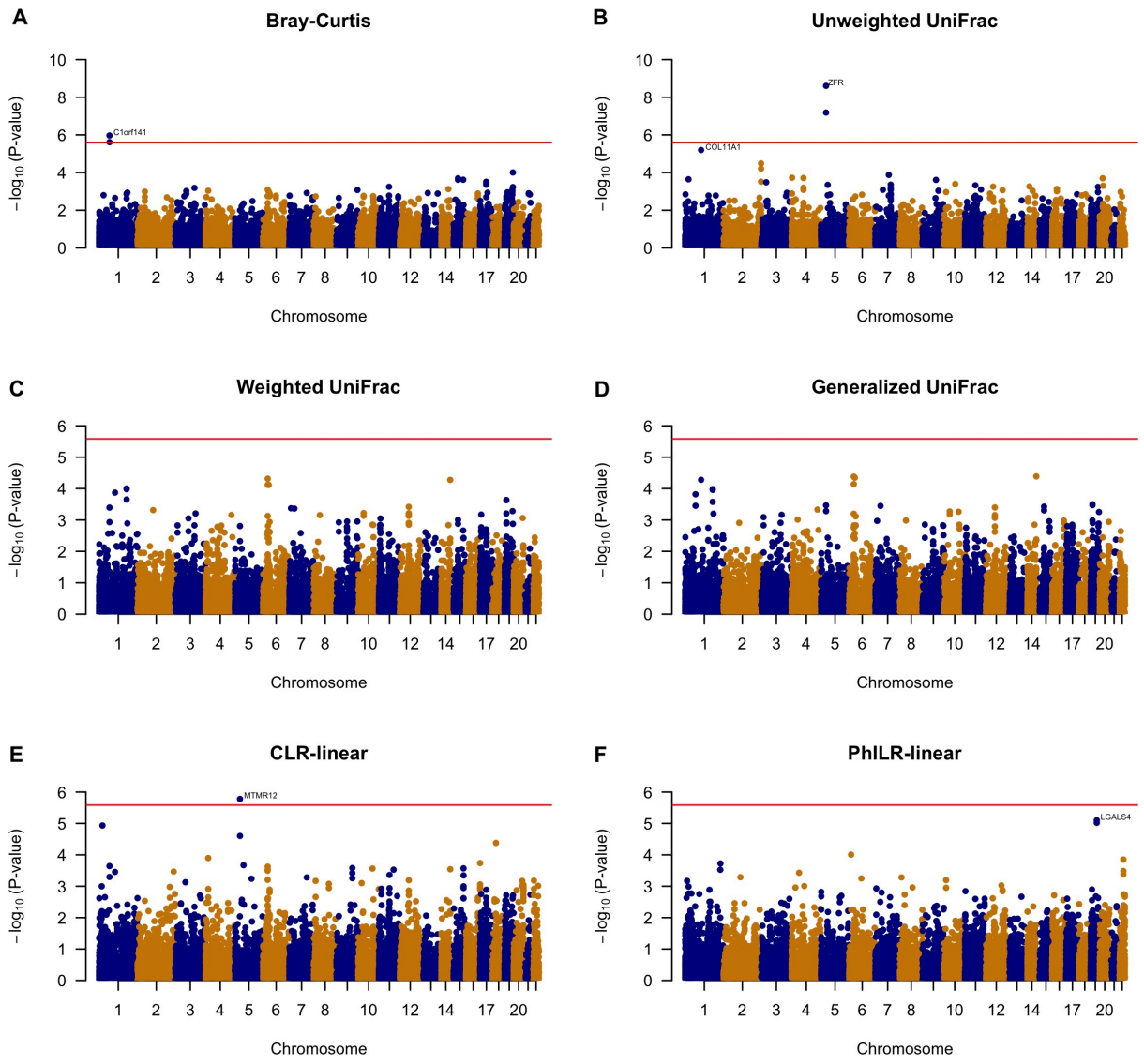

Figure A.1: Manhattan plots from the first-stage gene-level analysis of the HCHS/SOL data, using the PC-adjusted KRV. Each panel corresponds to a distinct microbiome kernel. The top 5 PCs of genome-wide genetic variability were adjusted. The red lines represent the genome-wide significance threshold ($\alpha = 2.6 \times 10^{-6}$).

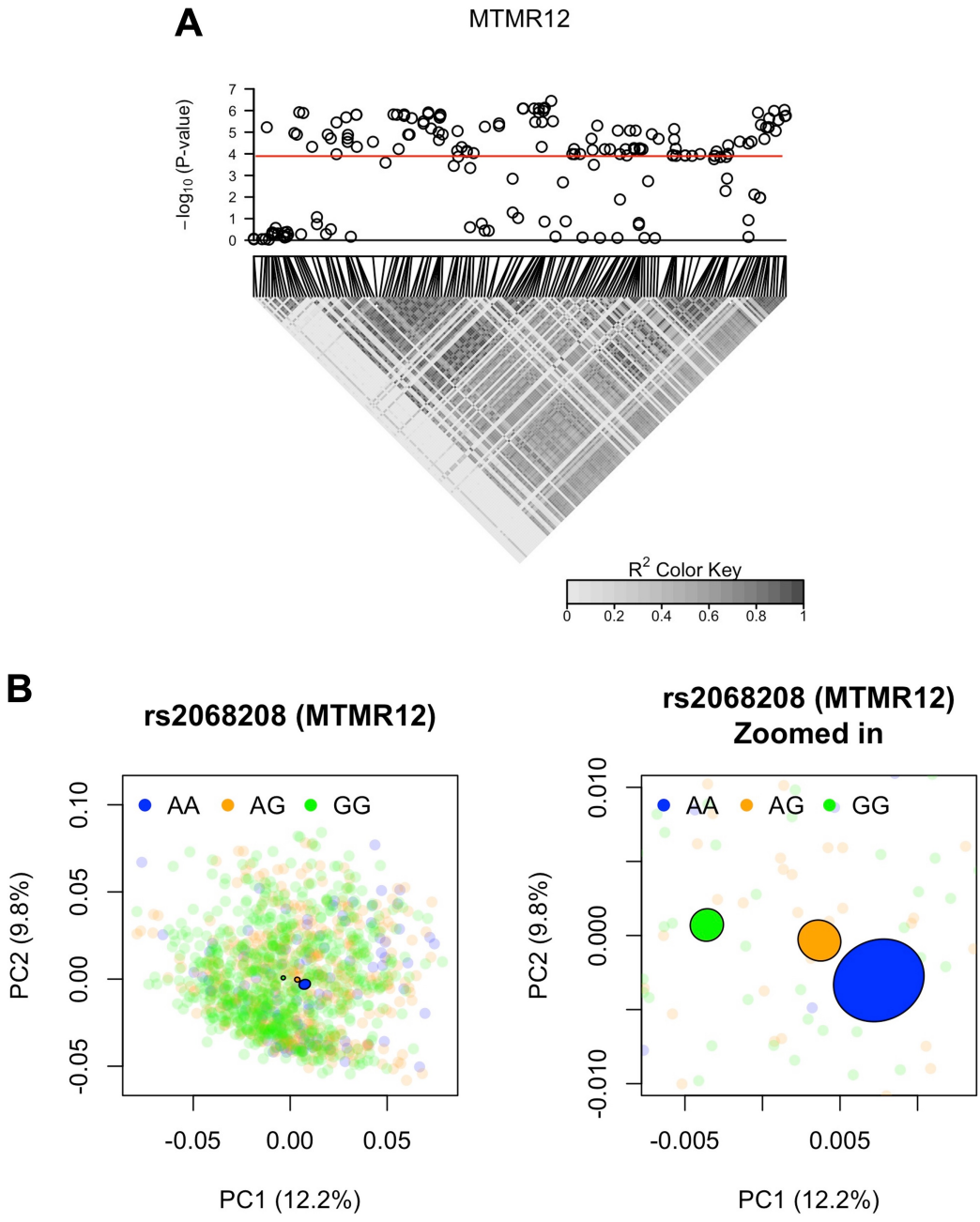

Figure A.2: Microbiome GWAS results of *MTMR12*, based on the CLR-linear kernel. Panel **A**: Manhattan plot and linkage disequilibrium (LD; R^2) heatmap from the second-stage variant-level analysis of the HCHS/SOL data, using the PC-adjusted KRV. The red line represents variant-level significance ($\alpha = 1.08 \times 10^{-4}$) used in the main analysis. Panel **B**: PC2 vs. PC1 from kernel PCA on the CLR-linear kernel, colored by genotype of the top variant from *MTMR12*. The percent of variance captured by each kernel PC was provided in the axis labels.

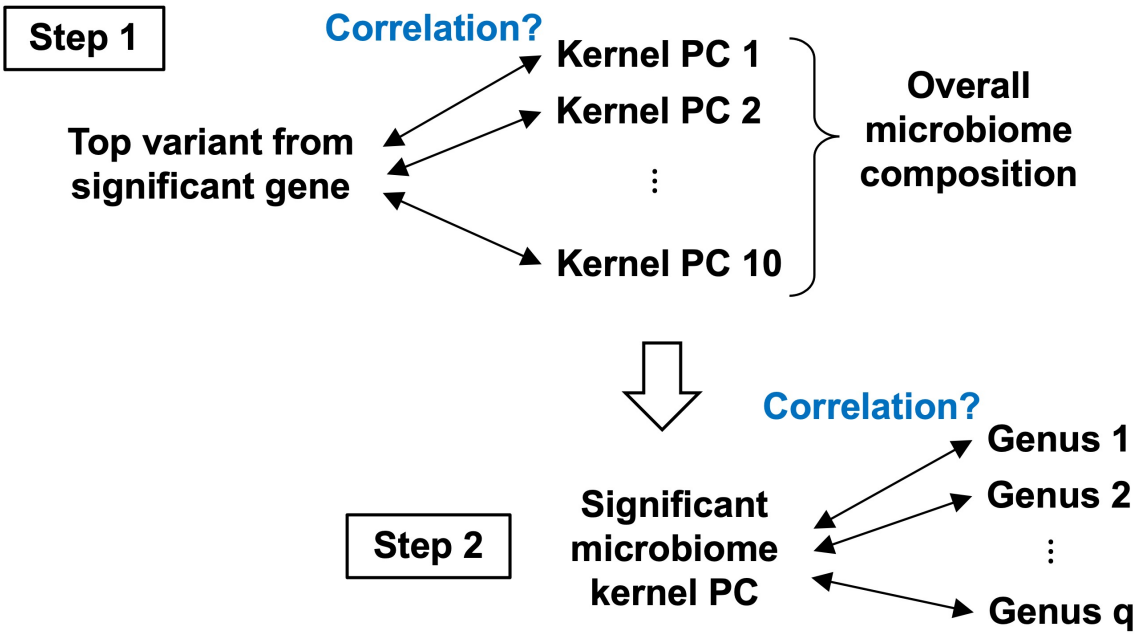

Figure A.3: Illustration of procedures to identify specific microbial taxa involved in the community-level microbiome GWAS associations.

Table A.2: Empirical type I error rate of unadjusted and covariate-adjusted KRV at nominal level α under Type I Error Scenario 2.

Method	Microbiome kernel	α		
		0.05	0.01	0.001
Unadjusted KRV	Bray-Curtis	1.0000	1.0000	1.0000
	Unweighted UniFrac	1.0000	1.0000	1.0000
	Weighted UniFrac	0.9980	0.9794	0.8312
	Generalized UniFrac	1.0000	1.0000	1.0000
	CLR-linear	1.0000	1.0000	1.0000
	PhILR-linear	1.0000	1.0000	0.9983
Adjusted KRV	Bray-Curtis	0.0489	0.0104	0.0014
	Unweighted UniFrac	0.0473	0.0079	0.0007
	Weighted UniFrac	0.0482	0.0102	0.0018
	Generalized UniFrac	0.0467	0.0096	0.0009
	CLR-linear	0.0521	0.0116	0.0010
	PhILR-linear	0.0524	0.0094	0.0018

Linear kernel was used for genetic data.

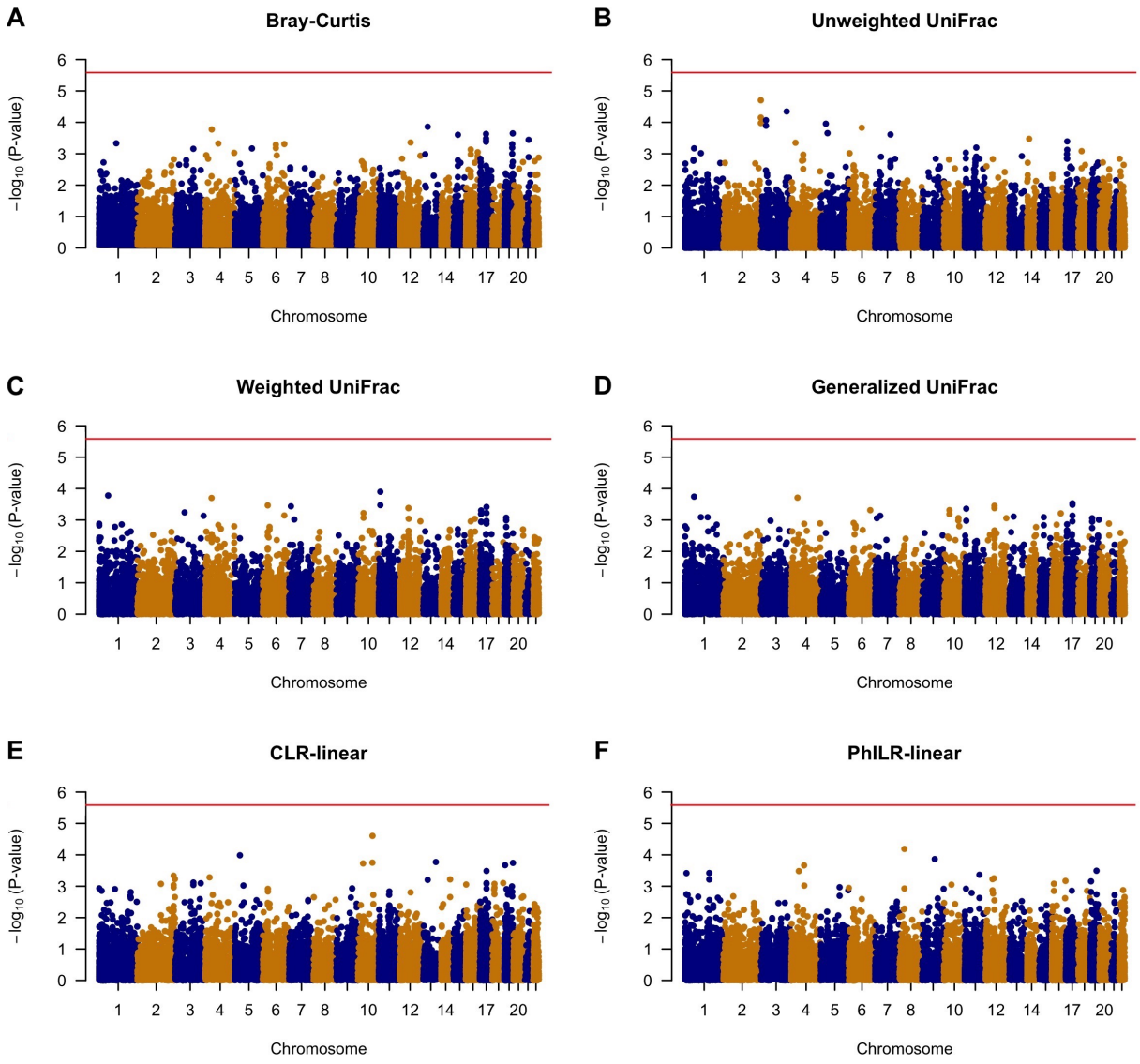

Figure A.4: Manhattan plots from alternative analysis of the HCHS/SOL data, via linear regression of the top PC of the community-level microbiome kernel matrix on the top PC of the gene-level genotype kernel matrix. Each panel corresponds to a distinct microbiome kernel. The top 5 PCs of genome-wide genetic variability were adjusted. The red lines represent the genome-wide significance threshold ($\alpha = 2.6 \times 10^{-6}$).

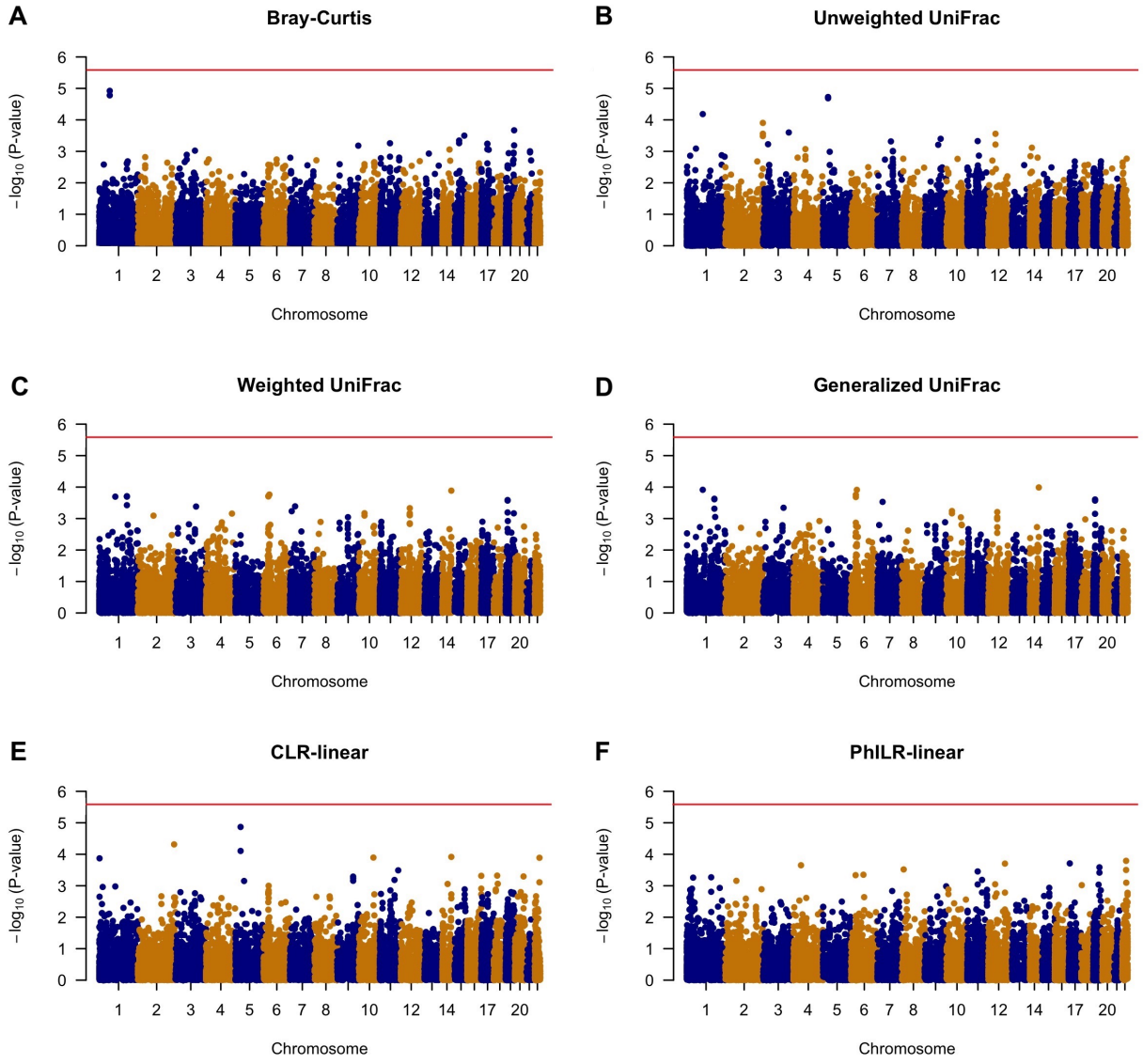

Figure A.5: Manhattan plots from alternative analysis of the HCHS/SOL data, via SKAT test of the top PC of the community-level microbiome kernel matrix on gene-level genetic variation. Each panel corresponds to a distinct microbiome kernel. The top 5 PCs of genome-wide genetic variability were adjusted. The red lines represent the genome-wide significance threshold ($\alpha = 2.6 \times 10^{-6}$).

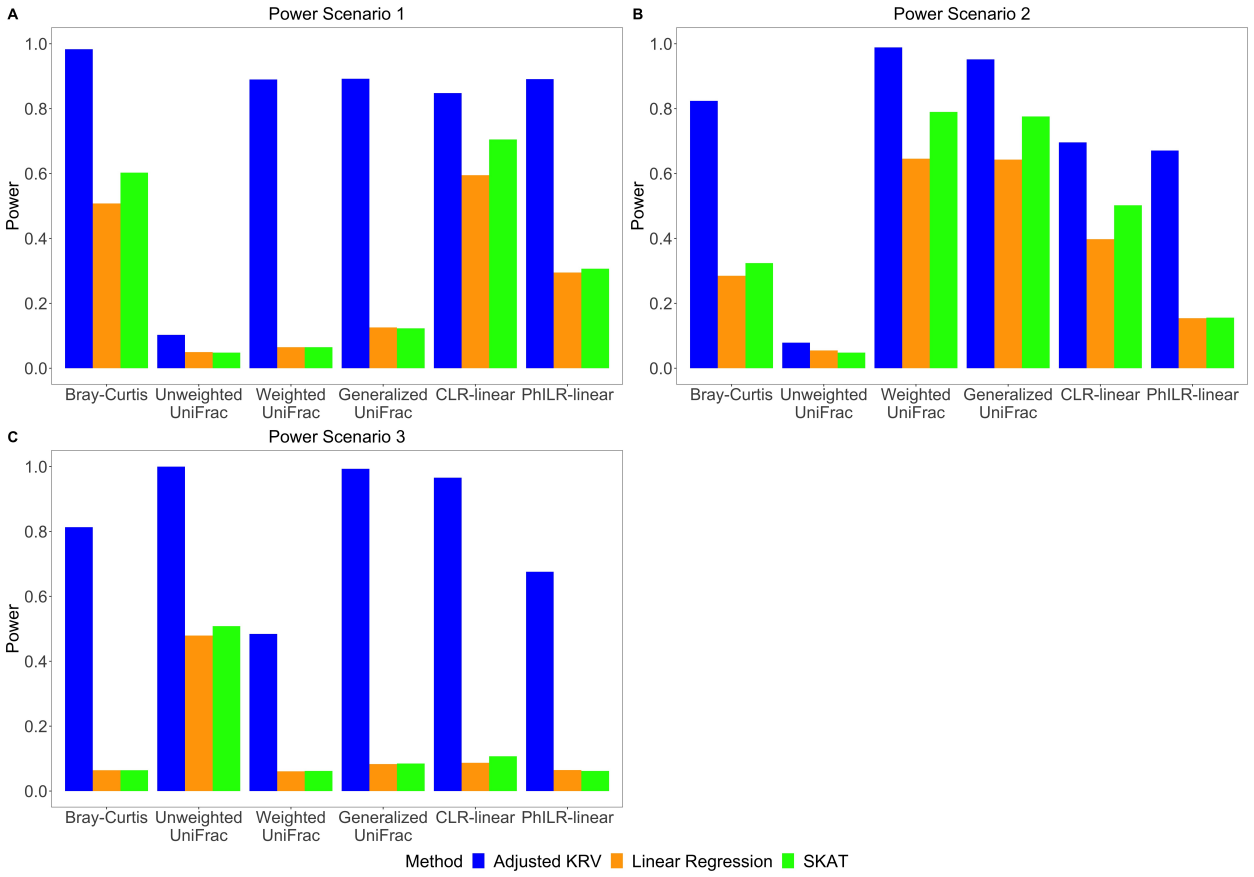

Figure A.6: Empirical power of covariate-adjusted KRV and competing methods at nominal level $\alpha = 0.05$ for different microbiome kernels under large effect sizes. Panel **A**: A single SNP affects the abundance of common OTUs. Panel **B**: A single SNP affects the abundance of OTUs from a common phylogenetic cluster. Panel **C**: A single SNP affects the abundance of rare OTUs. In each scenario, linear kernel was used for genetic data.

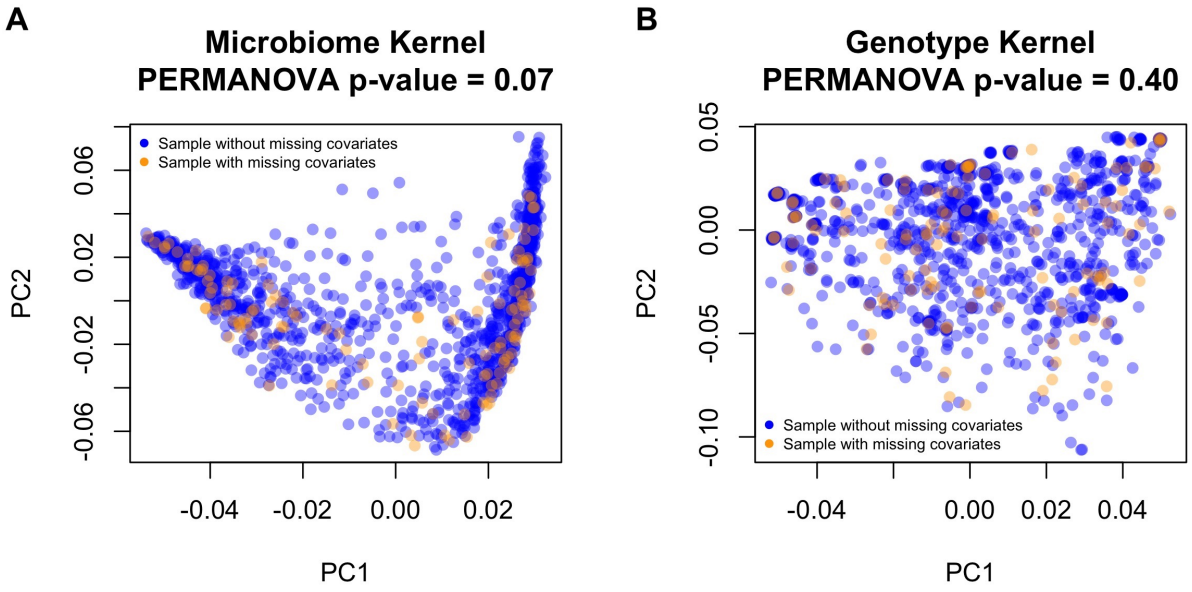

Figure A.7: PC2 vs. PC1 from kernel PCA on the Bray-Curtis microbiome kernel and the *IL23R-C1orf141* genotype kernel, colored by missing status of three covariates: age, gender and study site. Panel **A**: Kernel PCA was conducted on the Bray-Curtis microbiome kernel matrix. Panel **B**: Kernel PCA was conducted on the linear genotype kernel matrix, which was constructed based on common variants in the *IL23R-C1orf141* region.

Table A.3: Specific microbial taxa involved in identified microbiome GWAS associations from the HCHS/SOL data.

Microbiome GWAS signals				Specific microbial taxa involved	
Gene	Microbiome kernel	Top variant	Allele 1 / 2	Taxon name	Direction of association
IL23R-C1orf141	Bray-Curtis	rs10789226	A/G	Genus *Bacteroides*	+
IL23R-C1orf141	Bray-Curtis	rs10789226	A/G	Genus *Prevotella*	−
IL23R-C1orf141	Bray-Curtis	rs10789226	A/G	Genus *Blautia*	+
ZFR	Unweighted UniFrac	rs2113093	A/T	Order *Clostridiales*	−
ZFR	Unweighted UniFrac	rs2113093	A/T	Family *Ruminococcaceae*	−
MTMR12	Unweighted UniFrac	rs7729980	T/C	Order *Clostridiales*	+
MTMR12	Unweighted UniFrac	rs7729980	T/C	Family *Ruminococcaceae*	+
MTMR12	CLR-linear	rs2068208	G/A	Family *Rikenellaceae*	+
MTMR12	CLR-linear	rs2068208	G/A	Genus *Odoribacter*	+
MTMR12	CLR-linear	rs2068208	G/A	Family *Barnesiellaceae*	+
MTMR12	CLR-linear	rs2068208	G/A	Genus *Akkermansia*	+
MTMR12	CLR-linear	rs2068208	G/A	Family *Christensenellaceae*	+
MTMR12	CLR-linear	rs2068208	G/A	Family *Ruminococcaceae*	+
MTMR12	CLR-linear	rs2068208	G/A	Genus *Coprococcus*	+
MTMR12	CLR-linear	rs2068208	G/A	Genus *Ruminococcus*	+
MTMR12	CLR-linear	rs2068208	G/A	Genus *Oscillospira*	+
MTMR12	CLR-linear	rs2068208	G/A	Order *Clostridiales*	+
MTMR12	CLR-linear	rs2068208	G/A	Genus *Blautia*	+

The taxon name refers to the lowest taxonomic level available.
The direction of association is with respect to Allele 1 of the top variant.

Appendix B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

B.1 Preliminary Results

In this section, we present some preliminary results that will be useful in proving Theorem 3.3.2, Theorem 3.3.3 and Proposition 3.3.4. We draw upon existing theory on properties of random kernel matrices and extend these properties to cluster-correlated data. Specifically, we show the convergence of eigenvalues and eigenvectors of an empirical kernel matrix based on clustered data.

Let $(\mathcal{X}, \mathcal{F}, P)$ be a probability space and $\mathcal{H}$ be a Hilbert space over $(\mathcal{X}, \mathcal{F}, P)$ with a symmetric kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let H be a compact operator on $\mathcal{H}$, defined by

$$Hg(x) = \int_{\mathcal{X}} k(x, x')g(x')dP(x') \quad \text{for } x \in \mathcal{X}, g \in \mathcal{H}.$$

Let $\mathcal{H}_n$ be the Hilbert space over $(\mathcal{X}, \mathcal{F}, P_n)$, where $P_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}$ is the empirical version of P for a given $n \in \mathbb{N}$. Let $H_n : \mathcal{H}_n \to \mathcal{H}_n$ be the empirical version of the operator H, defined by

$$H_n g(x) = \int_{\mathcal{X}} k(x, x')g(x')dP_n(x') \quad \text{for } x \in \mathcal{X}, g \in \mathcal{H}_n.$$

Equivalently, H_n can be viewed as an $n \times n$ real matrix whose (i, j)-th entry is

$$\{H_n\}_{i,j} = \frac{1}{n}k(X_i, X_j).$$

This is the empirical kernel matrix scaled by a factor of $1/n$.

Here we restrict our discussion to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, where the kernel function k is positive semi-definite. We also assume that the operator H is

Hilbert–Schmidt, with $\mathbb{E}[k^2(X, X')] < \infty$.

Let $\lambda(T)$ denote the spectrum of a compact, symmetric operator T. Then $\lambda(H)$ and $\lambda(H_n)$ are the sets of eigenvalues for H and H_n, respectively. Since H_n is an operator on $\mathbb{R}^n$, we add to its spectrum an infinite number of zeros, such that $\lambda(H)$ and $\lambda(H_n)$ are comparable. For an operator with a positive semi-definite kernel, the associated eigenvalues are non-negative.

For any two compact, symmetric operators A and B with positive semi-definite kernels, let $a_1 \geq a_2 \geq \cdots \geq 0$ be the eigenvalues in $\lambda(A)$ arranged in a non-increasing order and let $b_1 \geq b_2 \geq \cdots \geq 0$ be the eigenvalues in $\lambda(B)$ arranged in a non-increasing order. Following the work by Koltchinskii et al. [209], we can define a distance measure δ_2 on $\ell_2(\mathbb{N})$ such that

$$\delta_2(\lambda(A), \lambda(B)) = \Big[\sum_i (a_i - b_i)^2 \Big]^{1/2}.$$

As shown in [209], the measure δ_2 is a well-defined distance between spectra of Hilbert–Schmidt operators or operators on $\mathbb{R}^n$. It satisfies the triangle inequality with

$$\delta_2(\lambda(A), \lambda(B)) \leq \delta_2(\lambda(A), \lambda(C)) + \delta_2(\lambda(C), \lambda(B))$$

for any operators A, B and C.

We now consider a sample $\underline{X}_n = (X_1, \cdots, X_n)$ with clustered correlation among the observations, as defined in Section 3.2 of the main text.

Assumption B.1.1. *Assume that $\underline{X}_n$ can be divided into m i.i.d. clusters of fixed size d (i.e., $n = md$). The observations $X_1, \cdots, X_n$ are identically distributed according to P, and the clusters $\{[X_{di-d+1}, X_{di-d+2}, \cdots, X_{di}]\}_{i=1}^m$ are independent from each other while having identical within-cluster correlation structure.*

B.1.1 Convergence of eigenvalues

We show that, with clustered data, the set of eigenvalues for H_n converges to the set of eigenvalues for H as the number of clusters goes to infinity. We first introduce a lemma that will be useful in proving this statement.

Lemma B.1.2. *Suppose that Assumption B.1.1 holds for the sample $\underline{X}_n$. Let V_n be a V-statistic based on $\underline{X}_n$ with a bivariate symmetric kernel function $f(x, x')$:*

$$V_n := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f(X_i, X_j).$$

Assume that V_n is non-degenerate and the class $\mathcal{C} := \{x \mapsto f(x, x') : x' \in \mathcal{X}\}$ is a P-Donsker class [116], then

$$V_n \xrightarrow{p} \mathbb{E}[f(X, X')] \text{ as } m \to \infty,$$

where X' is an independent copy of X.

The proof of Lemma B.1.2 is provided in Section B.5.

We have assumed that the operator H is Hilbert–Schmidt and the kernel k is positive semi-definite. By Mercer's Theorem (see, e.g. Theorem 3.6 of [210]), there exists an orthonormal set $\{\phi_r : r \in J\}$ of $\mathcal{H}$, where $J \subseteq \mathbb{N}$, and a sequence of descending non-negative real numbers $\{\lambda_r : r \in J\}$ with $\sum_{r \in J} \lambda_r^2 < \infty$ such that

$$k(x, x') = \sum_{r \in J} \lambda_r \phi_r(x) \phi_r(x').$$

Here the set $\lambda(H) := \{\lambda_r : r \in J\}$ is the set of eigenvalues for H, and the set $\{\phi_r : r \in J\}$ is the corresponding set of eigenfunctions. For any fixed $R \in \mathbb{N}$, we further define $k_R(x, x') := \sum_{r=1}^{R} \lambda_r \phi_r(x) \phi_r(x')$.

The following theorem states the convergence of eigenvalues of the empirical kernel matrix. The theorem and associated proof are adapted from Theorem 3.1 of Koltchinskii et al. [209], where we extend the assumption from i.i.d. data to clustered data.

Theorem B.1.3. *Assume that the kernel $k(\cdot,\cdot)$ is symmetric and positive semi-definite, and $\mathbb{E}[k^2(X,X')] < \infty$. Suppose that Assumption B.1.1 holds for the sample $\underline{X}_n$, and $\mathcal{C} := \{x \mapsto (k-k_R)^2(x,x') : x' \in \mathcal{X}\}$ is a P-Donsker class for any $R \in \mathbb{N}$. Then we have*

$$\delta_2(\lambda(H_n), \lambda(H)) \xrightarrow{p} 0 \text{ as } m \to \infty. \tag{B.1}$$

As a result, the r-th largest eigenvalue of H_n converges in probability to the r-th largest eigenvalue of H, for each r.

Proof. Symmetric kernel k and $\mathbb{E}[k^2(X,X')] < \infty$ ensure that the operator H is Hilbert–Schmidt. For the decomposition $k(x,x') = \sum_{r\in J} \lambda_r \phi_r(x)\phi_r(x')$, we consider first the basic case where $J = \{1, \cdots, R_0\}$ for some $R_0 < \infty$, and then the general case where $J = \mathbb{N}$.

In the basic case, we have $k(x,x') = \sum_{r=1}^{R_0} \lambda_r \phi_r(x)\phi_r(x')$. Following the argument of Koltchinskii et al. [209], we can prove that (B.1) holds, using tools of operator perturbation theory. The majority of the proof in the basic case would be similar to that of Koltchinskii et al. and we skip the details here. To account for clustered data, in Eq. (3.4) of [209], we apply the law of large numbers (LLN) to the clusters instead of individual observations.

We next focus on the general case where $J = \mathbb{N}$. For any fixed $R < \infty$, let H_R be the integral operator with kernel $k_R(x,x') = \sum_{r=1}^{R} \lambda_r \phi_r(x)\phi_r(x')$. Then we have

$$\lim_{R\to\infty} \delta_2(\lambda(H), \lambda(H_R)) = \lim_{R\to\infty} \Big[\sum_{r=R+1}^{\infty} \lambda_r^2 \Big]^{1/2} = 0. \tag{B.2}$$

Let $H_{R,n}$ be an $n \times n$ real matrix (i.e., an operator on $\mathbb{R}^n$) whose (i,j)-th entry is

$$\{H_{R,n}\}_{i,j} = \frac{1}{n} k_R(X_i, X_j).$$

By the result (B.1) established in the basic case, we have

$$\lim_{m\to\infty} \delta_2(\lambda(H_{R,n}), \lambda(H_R)) = 0 \text{ in probability for all } R < \infty. \tag{B.3}$$

Now, by Hoffman–Wielandt Inequality (Theorem 2.2 of [209]), we have

$$\begin{aligned}\lim_{R\to\infty}\lim_{m\to\infty}\delta_2(\lambda(H_{R,n}),\lambda(H_n)) &\le \lim_{R\to\infty}\lim_{m\to\infty}\|H_{R,n}-H_n\|_{HS}\\ &= \lim_{R\to\infty}\lim_{m\to\infty}\Big[\frac{1}{n^2}\sum_{1\le i,j\le n}(k-k_R)^2(X_i,X_j)\Big]^{1/2}.\end{aligned} \tag{B.4}$$

Here $V_n := \frac{1}{n^2}\sum_{1\le i,j\le n}(k-k_R)^2(X_i,X_j)$ is a V-statistic with kernel $f(x,x') = (k-k_R)^2(x,x')$. By Lemma B.1.2, which shows the convergence in probability of a V-statistic based on clustered data, we have

$$V_n \xrightarrow{p} \mathbb{E}[(k-k_R)^2(X,X')] \text{ as } m\to\infty,$$

where X' is an independent copy of X. Therefore, (B.4) becomes

$$\begin{aligned}\lim_{R\to\infty}\lim_{m\to\infty}\delta_2(\lambda(H_{R,n}),\lambda(H_n)) &\le \lim_{R\to\infty}\mathbb{E}[(k-k_R)^2(X,X')]^{1/2}\\ &= \lim_{R\to\infty}\left[\sum_{r=R+1}^{\infty}\lambda_r^2\right]^{1/2} = 0 \text{ in probability.}\end{aligned} \tag{B.5}$$

Finally, combining (B.2), (B.3) and (B.5), we have

$$\begin{aligned}&\lim_{m\to\infty}\delta_2(\lambda(H_n),\lambda(H))\\ &\le \lim_{R\to\infty}\limsup_{m\to\infty}\Big[\delta_2(\lambda(H_n),\lambda(H_{R,n})) + \delta_2(\lambda(H_{R,n}),\lambda(H_R)) + \delta_2(\lambda(H_R),\lambda(H))\Big]\\ &= 0 \text{ in probability.}\end{aligned}$$

Suppose γ_r is the r-th largest eigenvalue of H_n, and recall that λ_r is the r-th largest eigenvalue of H. For any $r\in\mathbb{N}$, we have that

$$|\gamma_r-\lambda_r| \le \Big[\sum_i(\gamma_i-\lambda_i)^2\Big]^{1/2} = \delta_2(\lambda(H_n),\lambda(H)) \xrightarrow{p} 0 \text{ as } m\to\infty.$$

We have thus proved the theorem. □

We now show that some common kernel functions such as linear kernels and Gaussian kernels satisfy the conditions in Theorem B.1.3.

For any $\mathcal{X} \subseteq \mathbb{R}^p$, the linear kernel $k(x, x') = x^T x'$ has exactly p non-zero eigenvalues. Therefore, the linear kernel trivially satisfies the basic case in the proof of Theorem B.1.3.

Next, we show that conditions in Theorem B.1.3 hold for a Gaussian kernel $k(x, x') = \exp(-\|x - x'\|_2^2/(2\sigma^2))$ under $\mathcal{X} = \mathbb{R}$ and a normal distribution P. The case where $\mathcal{X}$ is multi-dimensional can be generalized from the univariate case.

Proposition B.1.4. *Suppose that $\mathcal{X} = \mathbb{R}$ and $P = \mathcal{N}(0, \tau^2)$. Given a Gaussian kernel $k(x, x') = \exp(-(x - x')^2/(2\sigma^2))$, it holds that the class $\mathcal{C} := \{x \mapsto (k - k_R)^2(x, x') : x' \in \mathcal{X}\}$ is a P-Donsker class.*

Proof. We utilize the fact that, classes of functions with uniformly bounded variation are P-Donsker [211]. We will show that there exists some $M < \infty$ such that the total variation norm $\|\cdot\|_V$ of each function f in $\mathcal{C}$ satisfies: $\|f\|_V \leq M$. The total variation norm of a differentiable function f is of the form: $\|f\|_V = \int |f'(x)| dx$.

As shown in Section 4.3.1 of [212], the eigenvalues λ_r and eigenfunctions ϕ_r of the Gaussian kernel $k(x, x') = \exp(-(x - x')^2/(2\sigma^2))$, with $r = 0, 1, 2, \cdots$, are of the following form:

$$\lambda_r = \sqrt{\frac{2a}{A}} B^k, \quad \phi_r(x) = \exp(-(c - a)x^2) H_r(\sqrt{2c}x),$$

where $H_r(x) = (-1)^r \exp(x^2) \frac{d^r}{dx^r} \exp(-x^2)$ is the r-th order Hermite polynomial, $a = 1/(4\tau^2)$, $b = 1/(2\sigma^2)$ and

$$c = \sqrt{a^2 + 2ab}, \quad A = a + b + c, \quad B = b/A.$$

Therefore, we can express $k_R(x, x') = \sum_{r=0}^{R-1} \lambda_r \phi_r(x) \phi_r(x')$ in a closed form.

Both k and k_R are differentiable. Fixing $x' \in \mathcal{X}$, by definition of total variation norm,

we have

$$\|(k-k_R)^2(\cdot,x')\|_V = \int \Big|\frac{d}{dx}(k-k_R)^2(x,x')\Big|dx$$
$$= \int \Big|2(k-k_R)(x,x')\Big[-\frac{1}{\sigma^2}(x-x')k(x,x') - k_R^*(x,x')\Big]\Big|dx,$$

where $k_R^*(x,x') = \sum_{r=0}^{R-1}\lambda_r\phi_r^*(x)\phi_r(x')$, with

$$\phi_r^*(x) = \begin{cases} \exp(-(c-a)x^2)\big[-2(c-a)x\big] & \text{if } r=0, \\ \exp(-(c-a)x^2)\big[2r\sqrt{2c}H_{r-1}(\sqrt{2c}x) - 2(c-a)xH_r(\sqrt{2c}x)\big] & \text{if } r>0. \end{cases}$$

By triangle inequality and Cauchy-Schwarz inequality, $\|(k-k_R)^2(\cdot,x')\|_V$ is uniformly bounded (with respect to x') as long as the following terms are uniformly bounded:

$$\int (x-x')^2k^2(x,x')dx,\ \int k^2(x,x')dx,\ \int k_R^2(x,x')dx,\ \int (k_R^*(x,x'))^2dx.$$

We have:

$$\int (x-x')^2k^2(x,x')dx = \int (x-x')^2\exp\Big(-\frac{(x-x')^2}{\sigma^2}\Big)dx = \sqrt{\pi}\sigma\,\mathbb{E}_{P_0}[(X-x')^2] = \frac{\sqrt{\pi}\sigma^3}{2},$$

where the expectation is evaluated according to the distribution $P_0 = \mathcal{N}(x', \sigma^2/2)$. Similarly, it is easy to show that $\int k^2(x,x')dx = \sqrt{\pi}\sigma$.

To show that $\int k_R^2(x,x')dx$ is uniformly bounded, by Cauchy-Schwarz inequality, it suffices to show that $\int \phi_r^2(x)dx$ and $\phi_r(x')\phi_s(x')$ are uniformly bounded for any $r,s \in \{0,\cdots,R-1\}$ and $x' \in \mathcal{X}$. We have

$$\int \phi_r^2(x)dx = \int \exp(-2(c-a)x^2)H_r^2(\sqrt{2c}x)dx = \sqrt{\frac{\pi}{2(c-a)}}\,\mathbb{E}_{P_1}[H_r^2(\sqrt{2c}X)],$$

where the expectation is evaluated according to the distribution $P_1 = \mathcal{N}(0, 1/[4(c-a)])$. The r-th order Hermite polynomial is a polynomial of degree r. Since a normal distribution

has finite r-th moments for any non-negative integer r, the term $\mathbb{E}_{P_1}[H_r^2(\sqrt{2c}X)]$ is finite.

Let $M_1 = \max_{r\in\{0,\cdots,R-1\}} \mathbb{E}_{P_1}[H_r^2(\sqrt{2c}X)]$, then $\int \phi_r^2(x)dx \leq M_1$ for all $r \in \{0, \cdots, R-1\}$.

Note that $\phi_r(x') = \exp(-(c-a)(x')^2)H_r(\sqrt{2c}x')$ is a bounded function for any r: Intuitively, the exponential part changes at a larger rate than the polynomial part, and $\exp(-(c-a)(x')^2)$ is bounded between 0 and 1. Therefore, there exists $c_0, \cdots, c_{R-1} > 0$ such that $\sup_{x'\in\mathcal{X}} |\phi_r(x')| \leq c_r$ for each $r = 0, \cdots, R-1$.

Let $M_2 = \max_{r,s\in\{0,\cdots,R-1\}} c_r c_s$, then $|\phi_r(x')\phi_s(x')| \leq M_2$ for all $r, s \in \{0, \cdots, R-1\}$ and $x' \in \mathcal{X}$. Similarly, we can show that $\int (k_R^*(x, x'))^2 dx$ is also uniformly bounded.

As a result, for each σ^2 and τ^2, there exists $M < \infty$ such that, for any $x' \in \mathcal{X}$, the function $f(x) = (k - k_R)^2(x, x')$ has a total variation bounded by M. Thus $\mathcal{C} = \{x \mapsto (k - k_R)^2(x, x') : x' \in \mathcal{X}\}$ is a P-Donsker class. □

B.1.2 Convergence of Eigenvectors

We next show that, under sufficient conditions, the eigenvectors of the scaled empirical kernel matrix H_n would converge in probability to the corresponding eigenfunctions for the operator H, as the number of clusters goes to infinity. In particular, the i-th element of the r-th eigenvector of H_n converges in probability to the r-th eigenfunction of H evaluated at X_i, up to some scaling.

Let $\lambda(H) := \{\lambda_r : r \in \mathbb{N}\}$ be the set of eigenvalues for H, and $\{\phi_r : r \in \mathbb{N}\}$ be the corresponding set of eigenfunctions for H. Let $\lambda(H_n) := \{\gamma_r : r = 1, \cdots, n\}$ be the set of eigenvalues for H_n. Let $\boldsymbol{u}_r = (u_r(X_i), \cdots, u_r(X_n))^T$ be the r-th eigenvector for H_n.

Here (and in the proofs of the following sections) we assume that all eigenvalues in $\lambda(H)$ have multiplicity one for simplicity. In the case where certain eigenvalues have multiplicity larger than one and the corresponding eigenfunctions are not unique, we can always find an orthogonal matrix that transforms the set of eigenvectors $\{\boldsymbol{u}_r : r = 1, \cdots, n\}$ to match the components of $\{\phi_r : r \in \mathbb{N}\}$, as considered by Zhang et al. (2012) (Lemma 8 in [115]).

For each r where $\gamma_r > 0$, define the function

$$g_{r,n}(x) := \frac{1}{\sqrt{n}\gamma_r} \sum_{j=1}^{n} k(x, X_j) u_r(X_j).$$

As discussed by Bengio et al. [213], the function $g_{r,n}$ can be viewed as an eigenfunction for H_n. In particular, it is easy to show that $g_{r,n}(X_i) = \sqrt{n} u_r(X_i)$ for each i.

The following proposition states the convergence of eigenvectors of the kernel matrix. The proposition and associated proof are adapted from Proposition 2 from Bengio et al. [213], while accommodating the clustered correlation among the observations.

Proposition B.1.5. *Suppose that Assumption B.1.1 holds for the sample $\underline{X}_n$. Assume that the conditions in Theorem B.1.3 hold and that, for each r, the function $x \mapsto g_{r,n}(x)$ converges uniformly in probability to a non-random limit function $g_{r,\infty}$ as $m \to \infty$, with $\mathbb{E}[g^2_{r,\infty}(X)] < \infty$. Then for each r, i, we have*

$$g_{r,n}(X_i) = \sqrt{n} u_r(X_i) \xrightarrow{p} \phi_r(X_i) \text{ as } m \to \infty.$$

Proof. We restrict our discussion to positive γ_r and λ_r's. By algebraic manipulation, we have

$$\begin{aligned} g_{r,n}(x) &= \frac{1}{\sqrt{n}\gamma_r} \sum_{j=1}^{n} u_r(X_j) k(x, X_j) \\ &= \frac{\sqrt{n}}{\gamma_r} \sum_{j=1}^{n} \Big[\frac{1}{\sqrt{n}} g_{r,n}(X_j)\Big] \frac{1}{n} k(x, X_j) \\ &= \frac{1}{n\gamma_r} \sum_{j=1}^{n} g_{r,n}(X_j) k(x, X_j). \end{aligned}$$

The above result shows that $g_{r,n}$ is an eigenfunction of H_n with eigenvalue γ_r. Our goal is to show that $g_{r,\infty}$ is an eigenfunction of H with eigenvalue λ_r. Following the argument of

[213], by triangle inequality, for any fixed $x \in \mathcal{X}$, we can derive

$$\begin{aligned}
&\Big|g_{r,n}(x) - \frac{1}{\lambda_r}\int g_{r,\infty}(x')k(x,x')dP(x')\Big| \\
\le &\Big|\frac{1}{n\lambda_r}\sum_{j=1}^{n} g_{r,\infty}(X_j)k(x,X_j) - \frac{1}{\lambda_r}\int g_{r,\infty}(x')k(x,x')dP(x')\Big| \\
&+ \Big|\frac{\lambda_r - \gamma_r}{n\lambda_r\gamma_r}\sum_{j=1}^{n} g_{r,\infty}(X_j)k(x,X_j)\Big| \\
&+ \Big|\frac{1}{n\gamma_r}\sum_{j=1}^{n} k(x,X_j)[g_{r,n}(X_j) - g_{r,\infty}(X_j)]\Big| \\
=: &\, A_n + B_n + C_n.
\end{aligned} \tag{B.6}$$

Next we study each of the above terms.

First, by LLN applied to clusters, we have

$$\begin{aligned}
A_n &= \Big|\frac{1}{n\lambda_r}\sum_{j=1}^{n} g_{r,\infty}(X_j)k(x,X_j) - \frac{1}{\lambda_r}\int g_{r,\infty}(x')k(x,x')dP(x')\Big| \\
&= \frac{1}{d\lambda_r}\Big|\frac{1}{m}\sum_{j=1}^{m}\Big[\sum_{\ell=dj-d+1}^{dj} g_{r,\infty}(X_\ell)k(x,X_\ell)\Big] - d\int g_{r,\infty}(x')k(x,x')dP(x')\Big| \\
&\xrightarrow{p} 0 \text{ as } m \to \infty.
\end{aligned} \tag{B.7}$$

Note that, since $\mathbb{E}[k^2(X',X)] < \infty$, we have $\mathbb{E}[k^2(x,X)] = \mathbb{E}[k^2(X',X)|X' = x] < \infty$ for any fixed $x \in \mathcal{X}$.

From Theorem B.1.3, we know that $\gamma_r \xrightarrow{p} \lambda_r$ for each r, and thus γ_r is bounded in probability. Therefore, by LLN applied to clusters,

$$\begin{aligned}
B_n &= \Big|\frac{\lambda_r - \gamma_r}{n\lambda_r\gamma_r}\sum_{j=1}^{n} g_{r,\infty}(X_j)k(x,X_j)\Big| \\
&\le \Big|\frac{\lambda_r - \gamma_r}{\lambda_r\gamma_r}\Big|\Big|\frac{1}{n}\sum_{j=1}^{n} g_{r,\infty}(X_j)k(x,X_j)\Big| \\
&\xrightarrow{p} 0 \times \Big|\mathbb{E}[g_{r,\infty}(X)k(x,X)]\Big| = 0 \text{ as } m \to \infty,
\end{aligned} \tag{B.8}$$

where we use Cauchy-Schwarz inequality:

$$\Big|\,\mathbb{E}[g_{r,\infty}(X)k(x,X)]\Big| \le \mathbb{E}[\big|g_{r,\infty}(X)k(x,X)\big|] \le \Big(\,\mathbb{E}[g^2_{r,\infty}(X)]\,\mathbb{E}[k^2(x,X)]\Big)^{1/2} < \infty.$$

Finally, again by LLN applied to clusters, we have

$$\begin{aligned} C_n &= \Big|\frac{1}{n\gamma_r}\sum_{j=1}^{n} k(x,X_j)[g_{r,n}(X_j) - g_{r,\infty}(X_j)]\Big| \\ &\le \frac{1}{n\gamma_r}\sum_{j=1}^{n} \Big|k(x,X_j)[g_{r,n}(X_j) - g_{r,\infty}(X_j)]\Big| \\ &\le \frac{1}{\gamma_r}\sup_{x'\in\mathcal{X}}\Big|g_{r,n}(x') - g_{r,\infty}(x')\Big| \times \frac{1}{n}\sum_{j=1}^{n}|k(x,X_j)| \\ &\xrightarrow{p} 0 \times \mathbb{E}[|k(x,X)|] = 0 \text{ as } m\to\infty, \end{aligned} \tag{B.9}$$

where we use the uniform convergence assumption

$$\sup_{x'\in\mathcal{X}}\Big|g_{r,n}(x') - g_{r,\infty}(x')\Big| \xrightarrow{p} 0 \text{ as } m\to\infty$$

and the fact that $\mathbb{E}[|k(x,X)|] < \infty$ for any fixed x (given $\mathbb{E}[k^2(x,X)] < \infty$).

By (B.7), (B.8) and (B.9), we see that (B.6) becomes

$$\Big|g_{r,n}(x) - \frac{1}{\lambda_r}\int g_{r,\infty}(x')k(x,x')dP(x')\Big| \xrightarrow{p} 0 \text{ as } m\to\infty,$$

i.e., $g_{r,n}(x)$ converges in probability to $\frac{1}{\lambda_r}\int g_{r,\infty}(x')k(x,x')dP(x')$ pointwise for each $x\in\mathcal{X}$.

We already know that $g_{r,n} \xrightarrow{p} g_{r,\infty}$ pointwise. By uniqueness of limit, we obtain

$$\lambda_r g_{r,\infty}(x) = \int g_{r,\infty}(x')k(x,x')dP(x'),$$

indicating that $g_{r,\infty}$ is an eigenfunction of H with eigenvalue λ_r, i.e., $g_{r,\infty} = \phi_r$.

Finally, consider any observation X_i from the random sample $\underline{X}_n$. By the uniform con-

vergence assumption for $g_{r,n}(x)$, we have, for any $\epsilon > 0$,

$$\Pr\Big(|g_{r,n}(X_i) - \phi_r(X_i)| > \epsilon\Big) \leq \Pr\Big(\sup_{x\in\mathcal{X}} |g_{r,n}(x) - \phi_r(x)| > \epsilon\Big)$$
$$\to 0 \text{ as } m \to \infty.$$

Therefore, for each r, i, we have

$$g_{r,n}(X_i) = \sqrt{n}u_r(X_i) \xrightarrow{p} \phi_r(X_i) \text{ as } m \to \infty,$$

completing the proof. □

B.2 Proof of Theorem 3.3.2

We first introduce a lemma that will be useful in proving the theorem. The following lemma is adapted from Theorem 4.2 of [214] and Lemma 9 of [115].

Lemma B.2.1. *Let $\{A_{R,m}\}$ be a double sequence of random vectors indexed by R and m. Let $\{B_R\}$ and $\{C_m\}$ be sequences of random vectors and D be a random vector. Suppose that we have $A_{R,m} \xrightarrow{p} B_R$ as $m \to \infty$ for each R, and $B_R \xrightarrow{d} D$ as $R \to \infty$. Further suppose that*

$$\lim_{R\to\infty} \limsup_{m\to\infty} \Pr(\|A_{R,m} - C_m\| > \epsilon) = 0$$

for any $\epsilon > 0$. Then $C_m \xrightarrow{d} D$ as $m \to \infty$.

The proof of Lemma B.2.1 is provided in Section B.5.

B.2.1 Proof of Theorem 3.3.2

To prove Theorem 3.3.2, we adopt a strategy similar to the proof for Theorem 3 of Zhang et al. (2012) [115]. Let $\boldsymbol{u}^*_{X,r} := \sqrt{\gamma_{X,r}}\boldsymbol{u}_{X,r}$ and $\boldsymbol{u}^*_{X,r} := \sqrt{\gamma_{Y,r}}\boldsymbol{u}_{Y,r}$ for each r. Define

$$Q_{rs} := \frac{1}{\sqrt{n}}\Big(\boldsymbol{u}^*_{X,r}\Big)^T\Big(\boldsymbol{u}^*_{Y,s}\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} u^*_{X,r}(X_i)u^*_{Y,s}(Y_i),$$

where $u^*_{X,r}(X_i)$ and $u^*_{Y,r}(Y_i)$ are the i-th elements of $\boldsymbol{u}^*_{X,r}$ and $\boldsymbol{u}^*_{Y,r}$, respectively. We then note that

$$n\,\mathrm{HSIC}(P_n) = \frac{1}{n}\,\mathrm{tr}(\widetilde{\boldsymbol{K}}_X\widetilde{\boldsymbol{K}}_Y) = \frac{1}{n}\sum_{1\le r,s\le n}\Big[\Big(\boldsymbol{u}^*_{X,r}\Big)^T\Big(\boldsymbol{u}^*_{Y,s}\Big)\Big]^2 = \sum_{1\le r,s\le n} Q_{rs}^2.$$

We will break the proof of Theorem 3.3.2 into two parts:

(i) If $X \perp\!\!\!\perp Y$, then for any fixed $L \in \mathbb{N}$, we have

$$\sum_{1\le r,s\le L} Q_{rs}^2 \xrightarrow{d} \sum_{t=1}^{L^2} \ell_t z_t^2 \text{ as } m \to \infty, \tag{B.10}$$

where z_t's are i.i.d. standard normal variables, and ℓ_t's are the eigenvalues of $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]$, with $\boldsymbol{w}$ being the random vector obtained by stacking the columns in the $L \times L$ matrix $\boldsymbol{N}$, whose (r,s)-th entry is

$$N_{rs} = \frac{1}{\sqrt{d}}\sum_{i=1}^{d}\sqrt{\lambda_{X,r}\lambda_{Y,s}}\phi_{X,r}(X_i)\phi_{Y,s}(Y_i).$$

(ii) The result (B.10) still holds when $L = n \to \infty$, which is satisfied as $m \to \infty$. In other words,

$$n\,\mathrm{HSIC}(P_n) = \sum_{1\le r,s\le n} Q_{rs}^2 \xrightarrow{d} \sum_{t=1}^{\infty} \ell_t z_t^2 \text{ as } m \to \infty,$$

where ℓ_t's are now the eigenvalues of the infinite matrix $\mathbb{E}[\boldsymbol{w}_\infty\boldsymbol{w}_\infty^T]$, with $\boldsymbol{w}_\infty$ being the infinite random vector whose elements are of the form:

$$\frac{1}{\sqrt{d}}\sum_{i=1}^{d}\sqrt{\lambda_{X,r}\lambda_{Y,s}}\phi_{X,r}(X_i)\phi_{Y,s}(Y_i) \quad \text{for } r,s \in \mathbb{N}.$$

Alternatively, ℓ_t's can be viewed as the solutions to the eigenvalue problem

$$\begin{aligned}&\ell_t\psi_{t,rs}\\ &= \frac{1}{d}\sum_{p,q=1}^{\infty}\mathbb{E}\Big[\Big(\sum_{i=1}^{d}\sqrt{\lambda_{X,r}\lambda_{Y,s}}\phi_{X,r}(X_i)\phi_{Y,s}(Y_i)\Big)\Big(\sum_{i=1}^{d}\sqrt{\lambda_{X,p}\lambda_{Y,q}}\phi_{X,p}(X_i)\phi_{Y,q}(Y_i)\Big)\Big]\psi_{t,pq}\end{aligned}$$

for some double sequence $\{\psi_{t,rs}\}_{r,s=1}^{\infty} \in \mathbb{R}$.

We focus on proving part **(i)** of the theorem. Assume that the null hypothesis $H_0 : X \perp\!\!\!\perp Y$ hold, and consider a fixed $L \in \mathbb{N}$.

For $i = 1, \cdots, m$, let $\boldsymbol{v}_i$ be the random vector obtained by stacking the columns in the $L \times L$ matrix $\boldsymbol{M}_i$, whose (r, s)-th entry is

$$M_{i,rs} = \frac{1}{\sqrt{d}} \sum_{j=di-d+1}^{di} u^*_{X,r}(X_j) u^*_{Y,s}(Y_j).$$

Let $\boldsymbol{w}_i$ be the random vector obtained by stacking the columns in the $L \times L$ matrix $\boldsymbol{N}_i$, whose (r, s)-th entry is

$$N_{i,rs} = \frac{1}{\sqrt{d}} \sum_{j=di-d+1}^{di} \sqrt{\lambda_{X,r}\lambda_{Y,s}} \phi_{X,r}(X_j)\phi_{Y,s}(Y_j).$$

Since $X \perp\!\!\!\perp Y$, we have

$$\mathbb{E}[\phi_{X,r}(X_i)\phi_{Y,s}(Y_i)] = \mathbb{E}[\phi_{X,r}(X_i)]\,\mathbb{E}[\phi_{Y,s}(Y_i)] = 0 \text{ for all } i,$$

where we use the assumption that the kernels $\tilde{k}_X$ and $\tilde{k}_Y$ are centered. Therefore, we have $\mathbb{E}[\boldsymbol{w}] = \boldsymbol{0}$ and it follows that $\mathbb{C}\mathrm{ov}[\boldsymbol{w}] = \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]$.

By multivariate central limit theorem, we then have

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \boldsymbol{w}_i \xrightarrow{d} N(\boldsymbol{0}, \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]) \text{ as } m \to \infty.$$

By Theorem B.1.3 and Proposition B.1.5, for any X_i, Y_i and any fixed r, we have [1]

$$\frac{1}{n}\gamma_{X,r} \xrightarrow{p} \lambda_{X,r}, \quad \frac{1}{n}\gamma_{Y,r} \xrightarrow{p} \lambda_{Y,r} \text{ as } m \to \infty;$$
$$\sqrt{n}u_{X,r}(X_i) \xrightarrow{p} \phi_{X,r}(X_i), \quad \sqrt{n}u_{Y,r}(Y_i) \xrightarrow{p} \phi_{Y,r}(Y_i) \text{ as } m \to \infty.$$

Recall that $u^*_{X,r}(X_i) = \sqrt{\gamma_{X,r}}u_{X,r}(X_i)$ and $u^*_{Y,r}(Y_i) = \sqrt{\gamma_{Y,r}}u_{Y,r}(Y_i)$. Therefore, by continuous mapping theorem and Slutsky's theorem, for any fixed r and s,

$$u^*_{X,r}(X_i)u^*_{Y,s}(Y_i) \xrightarrow{p} \sqrt{\lambda_{X,r}\lambda_{Y,s}}\phi_{X,r}(X_i)\phi_{Y,s}(Y_i) \text{ as } m \to \infty.$$

As a consequence, we have $\boldsymbol{v}_i \xrightarrow{p} \boldsymbol{w}_i$ for each i. Using Lemma B.2.1 with $A_{R,m} = \frac{1}{\sqrt{R}}\sum_{i=1}^{R}\boldsymbol{v}_i$, $B_R = \frac{1}{\sqrt{R}}\sum_{i=1}^{R}\boldsymbol{w}_i$, $C_m = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{v}_i$ and $D \sim \mathcal{N}(\mathbf{0}, \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T])$, we can derive

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{v}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]) \text{ as } m \to \infty.$$

We perform an eigendecomposition of $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]$ such that $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T] = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]$ and $\boldsymbol{U}$ is an orthogonal matrix.

Let $\boldsymbol{Z} = \boldsymbol{U}^T\left[\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{v}_i\right]$. Then by continuous mapping theorem, we have

$$\boldsymbol{Z} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{U}^T\,\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{U}) = \mathcal{N}(\mathbf{0}, \boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{U}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}) \text{ as } m \to \infty.$$

---

[1] Here we assume that the (i,j)-th entry of the centered kernel matrix $\widetilde{\boldsymbol{K}}_X$ ($\widetilde{\boldsymbol{K}}_Y$) well approximates $\tilde{k}_X(X_i, X_j)$ ($\tilde{k}_Y(Y_i, Y_j)$). To address data dependence of the entries of $\widetilde{\boldsymbol{K}}_X$ and $\widetilde{\boldsymbol{K}}_Y$, we could show that, under regularity conditions, these empirically centered kernel functions converge uniformly in probability to $\tilde{k}_X$ and $\tilde{k}_Y$, as considered in Proposition 2 of [213].

It follows that

$$\begin{aligned}
\sum_{1\le r,s\le L} Q_{rs}^2 &= \sum_{1\le r,s\le L} \Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_{X,r}^*(X_i)u_{Y,s}^*(Y_i)\Big]^2 \\
&= \sum_{1\le r,s\le L} \Big[\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\frac{1}{\sqrt{d}}\sum_{j=di-d+1}^{di} u_{X,r}^*(X_j)u_{Y,s}^*(Y_j)\Big]^2 \\
&= \Big[\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{v}_i\Big]^T\Big[\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{v}_i\Big] \\
&= \Big[\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{v}_i\Big]^T\boldsymbol{U}\boldsymbol{U}^T\Big[\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{v}_i\Big] \\
&= \boldsymbol{Z}^T\boldsymbol{Z} \xrightarrow{d} \sum_{t=1}^{L^2}\ell_t z_t^2 \text{ as } m\to\infty,
\end{aligned}$$

where ℓ_t's are the eigenvalues of $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]$ and z_t's are i.i.d. standard normal variables.

To prove part **(ii)** of the theorem, we can use Lemma 9 from Zhang et al. (2012) [115]. The argument would be similar to that in [115] and we skip the details here.

B.3 Proof of Theorem 3.3.3

We first introduce a lemma that will help with the proof. The following lemma is in the same spirit as Lemma B.2.1, while replacing convergence in distribution with convergence in probability in assumptions and results.

Lemma B.3.1. *Let $\{A_{R,m}\}$ be a double sequence of random vectors indexed by R and m. Let $\{B_R\}$ and $\{C_m\}$ be sequences of random vectors and D be a random vector. Suppose that we have $A_{R,m} \xrightarrow{p} B_R$ as $m\to\infty$ for each R, and $B_R \xrightarrow{p} D$ as $R\to\infty$. Further suppose that*

$$\lim_{R\to\infty}\limsup_{m\to\infty}\Pr(\|A_{R,m}-C_m\|>\epsilon)=0$$

for any $\epsilon>0$. Then $C_m \xrightarrow{p} D$ as $m\to\infty$.

The proof of Lemma B.3.1 is provided in Section B.5.

B.3.1 Proof of Theorem 3.3.3

To begin the proof for Theorem 3.3.3, we assume that there exists some $r, s \in \mathbb{N}$ such that $\mathbb{E}[\phi_{X,r}(X)\phi_{Y,s}(Y)] \neq 0$. We can find a fixed $L \in \mathbb{N}$ such that $r, s \leq L$. Let Q_{rs}, $\boldsymbol{w}, \boldsymbol{w}_i, \boldsymbol{v}_i$ be defined as in the proof of Theorem 3.3.2 (Section B.2.1).

Since $\mathbb{E}[\phi_{X,r}(X)\phi_{Y,s}(Y)] \neq 0$, we have

$$\mathbb{E}\left[\frac{1}{\sqrt{d}}\sum_{i=1}^{d}\sqrt{\lambda_{X,r}\lambda_{Y,s}}\phi_{X,r}(X_i)\phi_{Y,s}(Y_i)\right] = \sqrt{d}\,\mathbb{E}\left[\sqrt{\lambda_{X,r}\lambda_{Y,s}}\phi_{X,r}(X)\phi_{Y,s}(Y)\right] \neq 0.$$

It follows that $\mathbb{E}[\boldsymbol{w}] \neq 0$. By weak law of large numbers, we have

$$\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{w}_i \xrightarrow{p} \mathbb{E}[\boldsymbol{w}] \text{ as } m \to \infty.$$

Using Theorem B.1.3, Proposition B.1.5 and Lemma B.3.1 with $A_{R,m} = \frac{1}{R}\sum_{i=1}^{R}\boldsymbol{v}_i$, $B_R = \frac{1}{R}\sum_{i=1}^{R}\boldsymbol{w}_i$, $C_m = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{v}_i$ and $D = \mathbb{E}[\boldsymbol{w}]$, we can derive

$$\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{v}_i \xrightarrow{p} \mathbb{E}[\boldsymbol{w}] \text{ as } m \to \infty.$$

Therefore, by continuous mapping theorem,

$$\begin{aligned}\frac{1}{m}\sum_{1\leq r,s\leq L}Q_{rs}^2 &= \left[\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{v}_i\right]^T\left[\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{v}_i\right] \\ &\xrightarrow{p} \mathbb{E}[\boldsymbol{w}]^T\,\mathbb{E}[\boldsymbol{w}] > 0 \text{ as } m \to \infty.\end{aligned}$$

As a consequence,

$$\sum_{1\leq r,s\leq L}Q_{rs}^2 \xrightarrow{p} \infty \text{ as } m \to \infty.$$

For a fixed L, we can always find a large enough m such that $n = md \geq L$ and thus

$$\frac{1}{n}\operatorname{tr}(\widetilde{\boldsymbol{K}}_X\widetilde{\boldsymbol{K}}_Y) = \sum_{1\leq r,s\leq n} Q_{rs}^2 \geq \sum_{1\leq r,s\leq L} Q_{rs}^2.$$

It follows that

$$n\,\mathrm{HSIC}(P_n) = \frac{1}{n}\operatorname{tr}(\widetilde{\boldsymbol{K}}_X\widetilde{\boldsymbol{K}}_Y) \xrightarrow{p} \infty \text{ as } m \to \infty.$$

Finally, we show that, when $\tilde{k}_X$ and $\tilde{k}_Y$ are characteristic kernels [110],

$$\mathbb{E}[\phi_{X,r}(X)\phi_{Y,s}(Y)] \neq 0 \text{ for some } r, s \in \mathbb{N} \;\Leftrightarrow\; X \not\perp\!\!\!\perp Y.$$

(1) Given $X \perp\!\!\!\perp Y$, then $\mathbb{E}[\phi_{X,r}(X)\phi_{Y,s}(Y)] = \mathbb{E}[\phi_{X,r}(X)]\,\mathbb{E}[\phi_{Y,s}(Y)] = 0$ for all $r, s \in \mathbb{N}$. As a contrapositive, $\mathbb{E}[\phi_{X,r}(X)\phi_{Y,s}(Y)] \neq 0$ for some $r, s \in \mathbb{N}$ implies $X \not\perp\!\!\!\perp Y$. This holds true regardless of the kernels being used.

(2) Given that $\mathbb{E}[\phi_{X,r}(X)\phi_{Y,s}(Y)] = 0$ for all $r, s \in \mathbb{N}$. Then for any $f \in \widetilde{\mathcal{H}}_X$, $g \in \widetilde{\mathcal{H}}_Y$, we have $\mathbb{C}\mathrm{ov}(f(X), g(Y)) = \mathbb{E}[f(X)g(Y)] = 0$, since any f and g can be expressed as linear combinations of $\phi_{X,r}$'s and $\phi_{Y,s}$'s, respectively. Hence, the largest singular value of C_{XY}, which is the maximized covariance between functions in $\widetilde{\mathcal{H}}_X$ and $\widetilde{\mathcal{H}}_Y$, must be zero. Consequently, the squared Hilbert-Schmidt norm of C_{XY}, $\|C_{XY}\|_{HS}^2 \equiv \mathrm{HSIC}(P_{XY})$, is also zero. When $\tilde{k}_X$ and $\tilde{k}_Y$ are characteristic kernels, $\mathrm{HSIC}(P_{XY}) = 0$ if and only if $X \perp\!\!\!\perp Y$ [110]. As a result, $\mathrm{HSIC}(P_{XY}) = 0$ implies that $X \perp\!\!\!\perp Y$.

As a contrapositive, $X \not\perp\!\!\!\perp Y$ implies $\mathbb{E}[\phi_{X,r}(X)\phi_{Y,s}(Y)] \neq 0$ for some $r, s \in \mathbb{N}$.

B.4 Proof of Proposition 3.3.4

Assume that the conditions in Theorem 3.3.2 hold. We would like to show that, under the null hypothesis $H_0 : X \perp\!\!\!\perp Y$, the statistic $n\,\mathrm{HSIC}(P_n) = \frac{1}{n}\operatorname{tr}(\widetilde{\boldsymbol{K}}_X\widetilde{\boldsymbol{K}}_Y)$ has the same asymptotic distribution as

$$\widetilde{T} = \frac{1}{m}\sum_{t=1}^{n^2} \tilde{\ell}_t z_t^2,$$

where z_t's are i.i.d. standard normal variables and $\tilde{\ell}_t$'s are eigenvalues of $\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^T$, with $\widetilde{\boldsymbol{V}} = [\tilde{\boldsymbol{v}}_1, \cdots, \tilde{\boldsymbol{v}}_m]$. Each vector $\tilde{\boldsymbol{v}}_i$ is obtained by stacking the columns in the $n \times n$ matrix $\widetilde{\boldsymbol{M}}_i$, whose (r, s)-th entry is

$$\widetilde{M}_{i,rs} = \frac{1}{\sqrt{d}} \sum_{j=di-d+1}^{di} u^*_{X,r}(X_j) u^*_{Y,s}(Y_j),$$

where $u^*_{X,r}(X_j)$ is the j-th element of $\boldsymbol{u}^*_{X,r} = \sqrt{\gamma_{X,r}}\boldsymbol{u}_{X,r}$, and $u^*_{Y,s}(Y_j)$ is the j-th element of $\boldsymbol{u}^*_{Y,s} = \sqrt{\gamma_{Y,s}}\boldsymbol{u}_{Y,s}$.

Here we present a sketch proof extended from the proof for Theorem 1 of Gretton et al. (2009) [215]. From Theorem 3.3.2, we have that

$$n\,\mathrm{HSIC}(P_n) = \frac{1}{n}\,\mathrm{tr}\left(\widetilde{\boldsymbol{K}}_X \widetilde{\boldsymbol{K}}_Y\right) \xrightarrow{d} \sum_{t=1}^{\infty} \ell_t z_t^2 \text{ as } m \to \infty,$$

where ℓ_t's are the eigenvalues of the infinite matrix $\boldsymbol{\Sigma} := \mathbb{E}[\boldsymbol{w}_\infty \boldsymbol{w}_\infty^T]$, with $\boldsymbol{w}_\infty$ being the infinite random vector whose elements are of the form:

$$\frac{1}{\sqrt{d}} \sum_{i=1}^{d} \sqrt{\lambda_{X,r}\lambda_{Y,s}}\phi_{X,r}(X_i)\phi_{Y,s}(Y_i) \quad \text{for } r, s \in \mathbb{N}.$$

A natural estimator for ℓ_t's is the set of eigenvalues for the empirical matrix $\hat{\boldsymbol{\Sigma}}$, given by

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}} &:= \frac{1}{m} \sum_{i=1}^{m} \tilde{\boldsymbol{v}}_i \tilde{\boldsymbol{v}}_i^T \\
&= \frac{1}{m} \begin{pmatrix} \tilde{\boldsymbol{v}}_1 & \cdots & \tilde{\boldsymbol{v}}_m \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{v}}_1^T \\ \vdots \\ \tilde{\boldsymbol{v}}_m^T \end{pmatrix} \\
&= \frac{1}{m} \widetilde{\boldsymbol{V}} \widetilde{\boldsymbol{V}}^T.
\end{aligned}$$

Letting $\tilde{\ell}_t$'s be the eigenvalues of $\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^T$, we would like to show that

$$\sum_{t=1}^{\infty}\Big(\frac{1}{m}\tilde{\ell}_t-\ell_t\Big)z_t^2\xrightarrow{p}0 \text{ as } m\to\infty. \tag{B.11}$$

Following the proof of Theorem 1 in [215], the key step to establish (B.11) is to show that $\sum_t|\frac{1}{m}\tilde{\ell}_t-\ell_t|\xrightarrow{p}0$ as $m\to\infty$.

By an extension of the Hoffman–Wielandt inequality, we have

$$\sum_{t=1}^{\infty}\Big|\frac{1}{m}\tilde{\ell}_t-\ell_t\Big|\le\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_1,$$

where $\|\cdot\|_1$ is the trace norm (the sum of singular values of the operator).

For $i=1,\cdots,m$, let $\tilde{\boldsymbol{w}}_i$ be the random vector obtained by stacking the columns in the $n\times n$ matrix $\widetilde{\boldsymbol{N}}_i$, whose (r,s)-th entry is

$$\widetilde{N}_{i,rs}=\frac{1}{\sqrt{d}}\sum_{j=di-d+1}^{di}\sqrt{\lambda_{X,r}\lambda_{Y,s}}\phi_{X,r}(X_j)\phi_{Y,s}(Y_j).$$

We can then write

$$\begin{aligned}\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_1&=\Big\|\frac{1}{m}\sum_{i=1}^{m}\tilde{\boldsymbol{v}}_i\tilde{\boldsymbol{v}}_i^T-\mathbb{E}[\boldsymbol{w}_\infty\boldsymbol{w}_\infty^T]\Big\|_1\\&\le\Big\|\frac{1}{m}\sum_{i=1}^{m}\tilde{\boldsymbol{v}}_i\tilde{\boldsymbol{v}}_i^T-\frac{1}{m}\sum_{i=1}^{m}\tilde{\boldsymbol{w}}_i\tilde{\boldsymbol{w}}_i^T\Big\|_1+\Big\|\frac{1}{m}\sum_{i=1}^{m}\tilde{\boldsymbol{w}}_i\tilde{\boldsymbol{w}}_i^T-\mathbb{E}[\tilde{\boldsymbol{w}}_1\tilde{\boldsymbol{w}}_1^T]\Big\|_1\\&\quad+\Big\|\mathbb{E}[\tilde{\boldsymbol{w}}_1\tilde{\boldsymbol{w}}_1^T]-\mathbb{E}[\boldsymbol{w}_\infty\boldsymbol{w}_\infty^T]\Big\|_1\\&=:A_n+B_n+C_n.\end{aligned}$$

We can show that $A_n\xrightarrow{p}0$ as $m\to\infty$ due to the convergence of the eigenvalues and eigenvector elements of $\widetilde{\boldsymbol{K}}_X$ and $\widetilde{\boldsymbol{K}}_Y$ to the eigenvalues and eigenfunctions of $\tilde{k}_X$ and $\tilde{k}_Y$, using Theorem B.1.3 and Proposition B.1.5. Furthermore, $B_n\xrightarrow{p}0$ as $m\to\infty$ due to Proposition 12 from [216]. Finally, $C_n\to0$ as $m\to\infty$ due to the convergence of the finite

truncation of a linear operator (e.g., see Proposition 2.1 from [217]). As a consequence, we have

$$\sum_{t=1}^{\infty} \left| \frac{1}{m} \tilde{\ell}_t - \ell_t \right| \leq \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_1 \xrightarrow{p} 0 \text{ as } m \to \infty,$$

which gives us

$$\widetilde{T} = \frac{1}{m} \sum_{t=1}^{n^2} \tilde{\ell}_t z_t^2 \xrightarrow{p} \sum_{t=1}^{\infty} \ell_t z_t^2 \text{ as } m \to \infty,$$

completing the proof.

B.5 Proofs of the Lemmas

Lemma A.2. *Suppose that Assumption B.1.1 holds for the sample $\underline{X}_n$. Let V_n be a V-statistic based on $\underline{X}_n$ with a bivariate symmetric kernel function $f(x, x')$:*

$$V_n := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f(X_i, X_j).$$

Assume that V_n is non-degenerate and the class $\mathcal{C} := \{x \mapsto f(x, x') : x' \in \mathcal{X}\}$ is a P-Donsker class, then

$$V_n \xrightarrow{p} \mathbb{E}[f(X, X')] \text{ as } m \to \infty,$$

where X' is an independent copy of X.

Proof. We utilize empirical process theory in this proof. For a given bivariate function f, we use the notation Pf to denote the function

$$x \mapsto \int f(x, x') dP(x'),$$

and we define $P_1 P_2 f$ for any set of probability measures P_1, P_2 as the mapping

$$(x, x') \mapsto \int \int f(x, x') dP_2(x') dP_1(x).$$

Using these notations, we can write $V_n = P_n^2 f$ and $\mathbb{E}[f(X, X')] = P^2 f$.

By symmetry of f, we have

$$\begin{aligned} V_n = P_n^2 f &= P^2 f + P_n^2 f - P^2 f \\ &= P^2 f + P_n(P_n - P)f + (P_n - P)Pf \\ &= P^2 f + (P_n - P)(P_n - P)f + P(P_n - P)f + (P_n - P)Pf \\ &= P^2 f + 2(P_n - P)Pf + (P_n - P)^2 f. \end{aligned} \tag{B.12}$$

Letting $f_1 := Pf$, note that

$$\begin{aligned} (P_n - P)Pf &= \frac{1}{n}\sum_{i=1}^{n}\Big(f_1(X_i) - \mathbb{E}[f_1(X)]\Big) \\ &= \frac{1}{md}\sum_{i=1}^{m}\sum_{j=di-d+1}^{di}\Big(f_1(X_j) - \mathbb{E}[f_1(X)]\Big) \\ &= \frac{1}{m}\sum_{i=1}^{m}\Big(\frac{1}{d}\sum_{j=di-d+1}^{di} f_1(X_j) - \mathbb{E}[f_1(X)]\Big) \\ &\xrightarrow{p} \mathbb{E}\left[\frac{1}{d}\sum_{j=1}^{d} f_1(X_j) - \mathbb{E}[f_1(X)]\right] = \mathbb{E}[f_1(X)] - \mathbb{E}[f_1(X)] = 0 \text{ as } m \to \infty. \end{aligned} \tag{B.13}$$

We now show that $(P_n - P)^2 f$ is also asymptotically negligible.

Letting $f_{1n} := (P_n - P)f$, we have

$$\begin{aligned} \sup_{x\in\mathcal{X}}|f_{1n}(x)| &= \sup_{x\in\mathcal{X}}\left|\int f(x, x')d(P_n - P)(x')\right| = \sup_{x\in\mathcal{X}}\left|\int f(x', x)d(P_n - P)(x')\right| \\ &\leq \frac{1}{d}\sup_{x\in\mathcal{X}}\left|\frac{1}{m}\sum_{i=1}^{m} f(X_{di-d+1}, x) - Pf\right| + \cdots + \frac{1}{d}\sup_{x\in\mathcal{X}}\left|\frac{1}{m}\sum_{i=1}^{m} f(X_{di}, x) - Pf\right|. \end{aligned} \tag{B.14}$$

Since $\mathcal{C} = \{x \mapsto f(x, x') : x' \in \mathcal{X}\}$ is a P-Donsker class, it is also a P-Glivenko-Cantelli class (see Chapter 19.2 of [116] for definition of these classes). Therefore, by definition of a

Glivenko-Cantelli class, each element in (B.14) would converge to 0 in probability as $m \to \infty$. As a result,

$$\sup_{x \in \mathcal{X}} |f_{1n}(x)| \xrightarrow{p} 0 \text{ as } m \to \infty.$$

Since $Pf_{1n}^2 \leq \left[\sup_{x \in \mathcal{X}} |f_{1n}(x)|\right]^2$, it follows that

$$Pf_{1n}^2 \xrightarrow{p} 0 \text{ as } m \to \infty. \tag{B.15}$$

Next, note that $x \mapsto \int f(x, x')dP_n(x')$ is in the closure of the convex hull of $\mathcal{C}$, and $x \mapsto \int f(x, x')dP(x')$ is a fixed function. By Theorems 2.10.2 and 2.10.3 of [218], $f_{1n}(x) = \int f(x, x')d(P_n - P)(x')$ also falls in a P-Donsker class.

Finally, combining the above result (f_{1n} belongs to a P-Donsker class) with (B.15), by Lemma 19.24 of [116], we have

$$\begin{aligned}(P_n - P)^2 f &= (P_n - P)f_{1n} \\ &= \frac{1}{d}\Big[\frac{1}{m}\sum_{i=1}^{m} f_{1n}(X_{di-d+1}) - Pf_{1n}\Big] + \cdots + \frac{1}{d}\Big[\frac{1}{m}\sum_{i=1}^{m} f_{1n}(X_{di}) - Pf_{1n}\Big] \\ &= o_P(m^{-1/2}).\end{aligned} \tag{B.16}$$

By (B.12), (B.13) and (B.16), we have that

$$V_n \xrightarrow{p} \mathbb{E}[f(X, X')] \text{ as } m \to \infty,$$

thus completing the proof. □

Lemma B.1. *Let $\{A_{R,m}\}$ be a double sequence of random vectors indexed by R and m. Let $\{B_R\}$ and $\{C_m\}$ be sequences of random vectors and D be a random vector. Suppose that we have $A_{R,m} \xrightarrow{p} B_R$ as $m \to \infty$ for each R, and $B_R \xrightarrow{d} D$ as $R \to \infty$. Further suppose that*

$$\lim_{R \to \infty} \limsup_{m \to \infty} \Pr(\|A_{R,m} - C_m\| > \epsilon) = 0$$

for any $\epsilon > 0$. *Then* $C_m \xrightarrow{d} D$ *as* $m \to \infty$.

Proof. Since $A_{R,m} \xrightarrow{p} B_R$ as $m \to \infty$ for each R, we have $A_{R,m} \xrightarrow{d} B_R$ as $m \to \infty$ for each R. The rest follows from the proof for Theorem 4.2 of [214]. □

Lemma C.1. *Let* $\{A_{R,m}\}$ *be a double sequence of random vectors indexed by* R *and* m. *Let* $\{B_R\}$ *and* $\{C_m\}$ *be sequences of random vectors and* D *be a random vector. Suppose that we have* $A_{R,m} \xrightarrow{p} B_R$ *as* $m \to \infty$ *for each* R, *and* $B_R \xrightarrow{p} D$ *as* $R \to \infty$. *Further suppose that*

$$\lim_{R\to\infty} \limsup_{m\to\infty} \Pr(\|A_{R,m} - C_m\| > \epsilon) = 0$$

for any $\epsilon > 0$. *Then* $C_m \xrightarrow{p} D$ *as* $m \to \infty$.

Proof. Given $\epsilon > 0$, we have that

$$\begin{aligned}\Pr(\|C_m - D\| > \epsilon) \le \Pr(\|C_m - A_{R,m}\| > \epsilon/3) + \Pr(\|A_{R,m} - B_R\| > \epsilon/3) \\ + \Pr(\|B_R - D\| > \epsilon/3).\end{aligned}$$

Fixing R and letting $m \to \infty$, we have

$$\begin{aligned}&\lim_{m\to\infty} \Pr(\|C_m - D\| > \epsilon) \\ &\le \limsup_{m\to\infty} \Pr(\|C_m - A_{R,m}\| > \epsilon/3) + \lim_{m\to\infty} \Pr(\|A_{R,m} - B_R\| > \epsilon/3) + \Pr(\|B_R - D\| > \epsilon/3) \\ &= \limsup_{m\to\infty} \Pr(\|C_m - A_{R,m}\| > \epsilon/3) + \Pr(\|B_R - D\| > \epsilon/3),\end{aligned}$$

where we use the fact that $A_{R,m} \xrightarrow{p} B_R$ as $m \to \infty$ for fixed R. Now letting $R \to \infty$, we have

$$\begin{aligned}&\lim_{m\to\infty} \Pr(\|C_m - D\| > \epsilon) \\ &\le \lim_{R\to\infty} \limsup_{m\to\infty} \Pr(\|C_m - A_{R,m}\| > \epsilon/3) + \lim_{R\to\infty} \Pr(\|B_R - D\| > \epsilon/3) \\ &= 0,\end{aligned}$$

where we use $B_R \xrightarrow{p} D$ as $R \to \infty$ and $\lim_{R\to\infty} \limsup_{m\to\infty} \Pr(\|A_{R,m} - C_m\| > \epsilon) = 0$. Therefore, we have shown that $C_m \xrightarrow{p} D$ as $m \to \infty$.

□

B.6 Additional Simulations

B.6.1 Type I Error Simulation in Non-normal Data

To examine the type I error control of $\mathbf{HSIC_{cl}}$ in the presence of non-normal data, we modify Model (3) in the main text such that the variable Y has a non-normal distribution under the null hypothesis.

Following the general simulation setting in Section 3.4.1, we let

$$\boldsymbol{y}_0 = (\beta_1 f(x_{r1}), \beta_1 f(x_{r2}), \beta_1 f(x_{r3}), \cdots, \beta_q f(x_{r1}), \beta_q f(x_{r2}), \beta_q f(x_{r3}))^T + \boldsymbol{\epsilon}.$$

We then generate $\boldsymbol{y} := (y_{11}, y_{12}, y_{13}, \cdots, y_{q1}, y_{q2}, y_{q3})^T$ by nonlinear transformations of $\boldsymbol{y}_0$. We consider three types of transformation functions: **Scenario A**: $f_A(y) = I\{y > 0\}$, **Scenario B**: $f_B(y) = \exp(y)$ and **Scenario C**: $f_C(y) = \sin(y)$. In each scenario, $\boldsymbol{y}$ is generated by applying the transformation function to each element of $\boldsymbol{y}_0$ such that the within-cluster correlation can be preserved to some extent. We fix $m = 500$ and $\rho_c = 0.5$.

Figure B.1 shows the p-value QQ-plots for $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{orig}}$ under Type I error simulation for all three scenarios.

B.6.2 Effect of Cluster Size on Performance of $\boldsymbol{HSIC_{cl}}$

We investigate the effect of cluster size on the performance of $\mathbf{HSIC_{cl}}$. We use the same simulation setting as in Section 3.4.1, where we fix $m = 500$ and $\rho_c = 0.5$ and consider different cluster sizes: d =2, 3, 4 or 5. We evaluate both type I error control and power of $\mathbf{HSIC_{cl}}$. In the power simulation, we consider **Power Scenario 1** and set $\eta = 20\%$.

Figure B.2 shows the p-value QQ-plots for $\mathbf{HSIC_{cl}}$ under Type I error simulation with different cluster sizes, when the Gaussian kernel is used. Figure B.3 shows the p-value QQ-

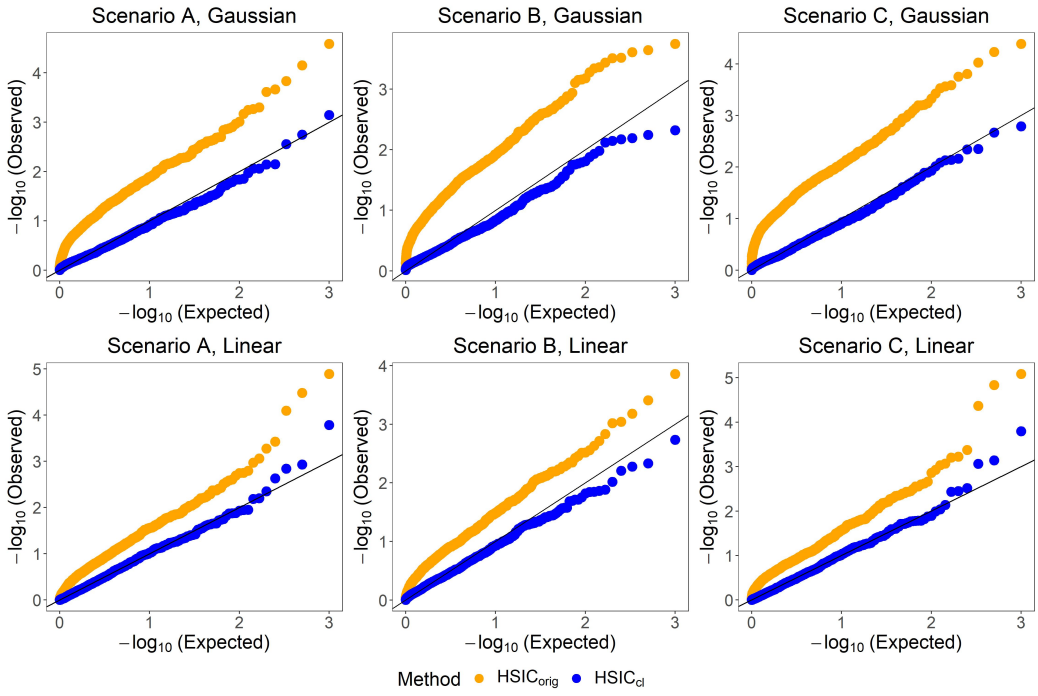

Figure B.1: P-value QQ-plots for $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{orig}}$ under Type I error simulation for three non-normal data scenarios. Simulation parameters are set as: $m = 500$, $d = 3$ and $\rho_c = 0.5$. The top row shows results based on the Gaussian kernel, and the bottom row shows results based on the linear kernel.

plots for $\mathbf{HSIC_{cl}}$ under Type I error simulation with different cluster sizes, when the linear kernel is used.

Figure B.4 shows the empirical power of $\mathbf{HSIC_{cl}}$ under different cluster sizes, for both the Gaussian kernel and the linear kernel.

B.6.3 Comparison of $\boldsymbol{HSIC_{cl}}$ against $\boldsymbol{HSIC_{perm}}$

An alternative way to assess the significance of the HSIC statistic is to compare the observed statistic against its permutation distribution, which could approximate the sampling distribution of the test statistic under the null hypothesis. This approach does not rely on asymptotic results and is suitable for small sample sizes. We implement a permutation-based HSIC test for clustered data, $\mathbf{HSIC_{perm}}$. We construct the empirical permutation distribu-

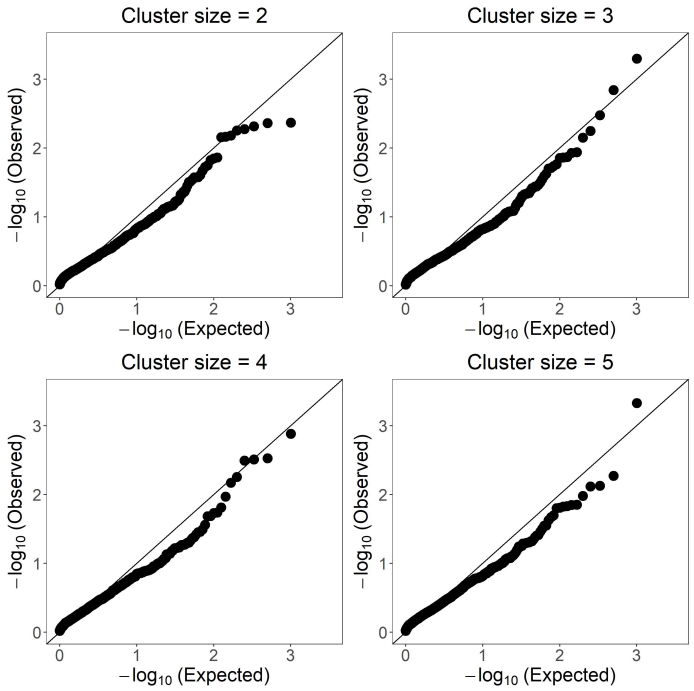

Figure B.2: P-value QQ-plots for $\mathbf{HSIC_{cl}}$ under Type I error simulation with different cluster sizes. The Gaussian kernel is used. Simulation parameters are set as: $m = 500$, $\rho_c = 0.5$.

tion of HSIC in the following way: in each permutation, we randomly shuffle the clusters for one variable and then re-construct the HSIC statistic using the shuffled observations; this procedure is repeated many times to obtain an empirical permutation distribution. The p-value is calculated as

$$p_{\text{perm}} = \frac{\sum_{j=1}^{n_{\text{perm}}} I\{\text{HSIC}_{\text{perm j}} \geq \text{HSIC}_{\text{obs}}\}}{n_{\text{perm}}}, \tag{B.17}$$

where n_{perm} is the number of permutations, $\text{HSIC}_{\text{perm j}}$ is the HSIC statistic at the jth permutation and HSIC_{obs} is the original observed HSIC statistic.

We use the same simulation setting as in Section 3.4.1, where we consider a small sample size $m = 100$ and set $d = 3$ and $\rho_c = 0.5$. We compare $\mathbf{HSIC_{cl}}$ against $\mathbf{HSIC_{perm}}$ in both type I error control and power. In the power simulation, we consider **Power Scenario 1**

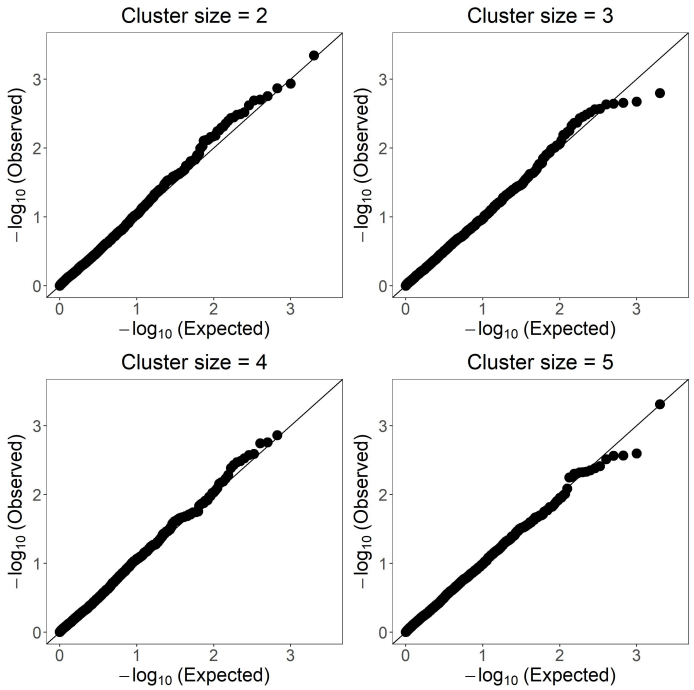

Figure B.3: P-value QQ-plots for $\mathbf{HSIC_{cl}}$ under Type I error simulation with different cluster sizes. The linear kernel is used. Simulation parameters are set as: $m = 500$, $\rho_c = 0.5$.

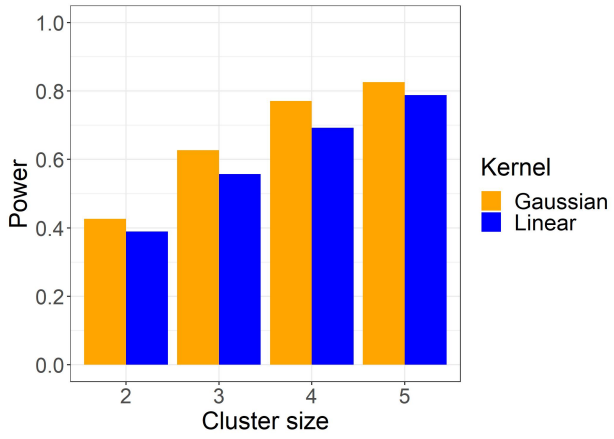

Figure B.4: Empirical power of $\mathbf{HSIC_{cl}}$ at nominal level $\alpha = 0.05$ under different cluster sizes. Simulation parameters are set as: $m = 500$, $\rho_c = 0.5$ and $\eta = 20\%$. **Power Scenario 1** is considered.

and set $\eta = 20\%$. In each simulation run, 1000 permutations are conducted in $\mathbf{HSIC_{perm}}$.

Figure B.5 shows the p-value QQ-plots for $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{perm}}$ under Type I error simulation for both the Gaussian kernel and the linear kernel. Table B.1 shows the empirical power of $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{perm}}$ for the two kernels.

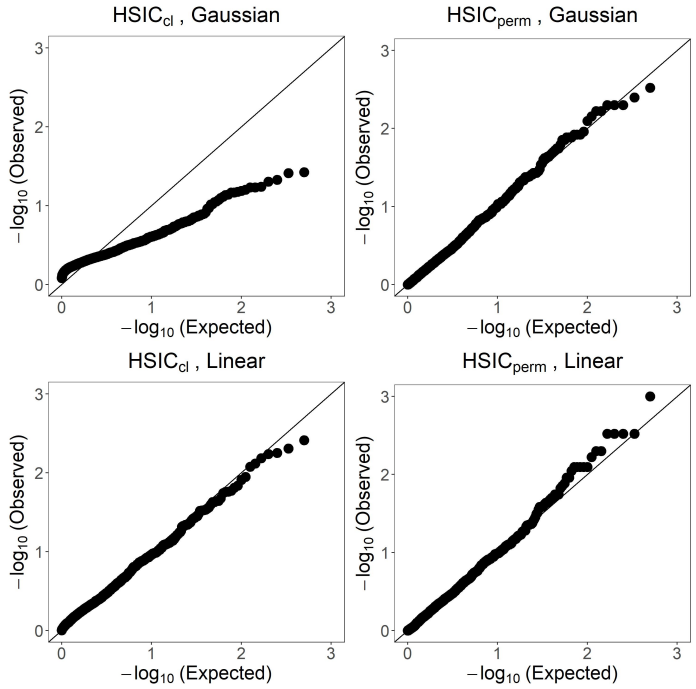

Figure B.5: P-value QQ-plots for $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{perm}}$ under Type I error simulation. Simulation parameters are set as: $m = 100$, $d = 3$ and $\rho_c = 0.5$. The top row shows results based on the Gaussian kernel, and the bottom row shows results based on the linear kernel.

Based on the Gaussian kernel, while $\mathbf{HSIC_{cl}}$ is over-conservative, $\mathbf{HSIC_{perm}}$ produces a valid type I error rate; as a result, $\mathbf{HSIC_{perm}}$ has a higher power than $\mathbf{HSIC_{cl}}$ in this case. Based on the linear kernel, $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{perm}}$ have similar performances. These results show that $\mathbf{HSIC_{perm}}$ could be a useful surrogate for $\mathbf{HSIC_{cl}}$ at small sample sizes. However, the computational burden of $\mathbf{HSIC_{perm}}$ is large compared to $\mathbf{HSIC_{cl}}$, especially as sample sizes increase (Table B.2) or as we require a more stringent significance level (which requires a larger number of permutations).

Table B.1: Empirical power of $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{perm}}$ at nominal level $\alpha = 0.05$ under simulation ($m = 100$).

Kernel	$\mathbf{HSIC_{cl}}$	$\mathbf{HSIC_{perm}}$
Gaussian	0.304	0.726
Linear	0.534	0.559

B.7 Additional Details on Implementation

B.7.1 Code Availability

In our simulations, the $\mathbf{HSIC_{orig}}$ method is implemented according to Broadaway et al. [56] (called GAMuT in their work), where Davies' exact method [117] is used to approximate the mixture of chi-square variables in the asymptotic null distribution of HSIC. The specific code is adapted from https://github.com/epstein-software/GAMuT (license: GPL-3.0).

Both $\mathbf{HSIC_{orig}}$ and $\mathbf{HSIC_{cl}}$ use the CompQuadForm R package [219] v1.4.3: https://cran.r-project.org/web/packages/CompQuadForm (license: GPL $\geq$ 2), which implements Davies' exact method.

$\mathbf{HSIC_{cl}}$ is implemented as the `HSIC_cl()` function in R environment. In the Github page: https://github.com/pearl-liu/HSIC_cl, we provide code and instructions for using the `HSIC_cl()` function and for reproducing the simulation results in Section 3.4.2 of the main text.

B.7.2 Computation Time

We have estimated the computation time of $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{perm}}$ (with 1000 permutations; see Section B.6.3 for details) for different number of clusters (m). For each m, we simulate 10 data sets according to Section 3.4.1 of the main text and report the average computation time. Given constructed kernel matrices for X and Y, the average computation times on a 12-core computer with 2.40 GHz CPUs and 256 GB memory are shown in

Table B.2.

Table B.2: Average computation time (in seconds) of $\mathbf{HSIC_{cl}}$ and $\mathbf{HSIC_{perm}}$ (with 1000 permutations) for different number of clusters, with cluster size 3.

Method	Kernel	Number of clusters, m					
		50	100	200	500	800	1000
$\mathbf{HSIC_{cl}}$	Gaussian	0.168	0.531	3.857	83.689	485.141	1109.086
	Linear	0.013	0.040	0.241	2.572	10.654	21.069
$\mathbf{HSIC_{perm}}$	Gaussian	1.327	7.628	50.740	705.860	2802.158	5385.214
	Linear	1.392	7.158	49.335	692.443	2782.553	5353.857

First, we note that $\mathbf{HSIC_{cl}}$ has a much shorter computation time than $\mathbf{HSIC_{perm}}$ for each sample size, either based on the Gaussian kernel or based on the linear kernel.

Next we compare computation times of $\mathbf{HSIC_{cl}}$ between the Gaussian kernel and the linear kernel. For any $X \in \mathbb{R}^p$ with $p < n$, the associated linear kernel matrix has p non-zero eigenvalues. In contrast, the Gaussian kernel matrix always has n non-zero eigenvalues, due to the infinite dimension of its feature space. Therefore, based on the way the asymptotic null distribution of the HSIC statistic is estimated in Proposition 3.3.4, using a linear kernel can take much shorter computation times than using a Gaussian kernel. Specifically, the computational complexity of $\mathbf{HSIC_{cl}}$ based on a Gaussian kernel is $O(m^2n^2)$. While Gaussian kernels have the advantage of capturing more general dependence patterns, linear kernels could be preferable in certain situations (e.g., when $p < n$) as a computationally efficient choice.

On a high-performance computing cluster (each node with 20 cores, 2.20 GHz CPUs and ∼100 GB memory), with divided computing jobs, it took ∼70 hours to complete the simulations in Section 3.4 using the Gaussian kernel, and ∼4 hours to complete the simulations using the linear kernel. Using the same resources, the analysis of the MsFLASH data set in Section 3.5 took ∼13 minutes.

B.8 MsFLASH Study

B.8.1 Description of the MsFLASH Study

The Menopause Strategies: Finding Lasting Answers for Symptoms and Health (MsFLASH) Vaginal Health Trial was a randomized, double-blind and placebo-controlled clinical trial conducted at 2 centers in the U.S.: Kaiser Permanente Washington Health Research Institute in Seattle and University of Minnesota in Minneapolis [119]. The trial compared the treatment efficacy for moderate-to-severe vulvovaginal discomfort between 0.01 mg vaginal estradiol tablets or vaginal moisturizer and placebo in 302 postmenopausal women. Vaginal swabs were collected from the participants at baseline, and 4 and 12 weeks after randomization. In a secondary study [120], based on the samples collected from each follow-up, the vaginal microbiota was characterized via 16S ribosomal RNA (rRNA) gene sequencing, and the vaginal metabolome was profiled using liquid chromatography-mass spectrometry. The abundance data of microbial taxa were center log-ratio transformed to address differential read depth and compositionality, and the abundance data of metabolites were quantile-normalized. More details on the MsFLASH trial are described by Mitchell et al. (2018, 2021) [119, 120].

The MsFLASH trial was approved by institutional review boards of the participating institutions, and all participants provided written informed consent. Inspection of the data set reveals no personally identifiable information or offensive content. The data used in this study is available upon request from the MsFLASH Data Coordinating Center.

B.8.2 Additional Analysis Results

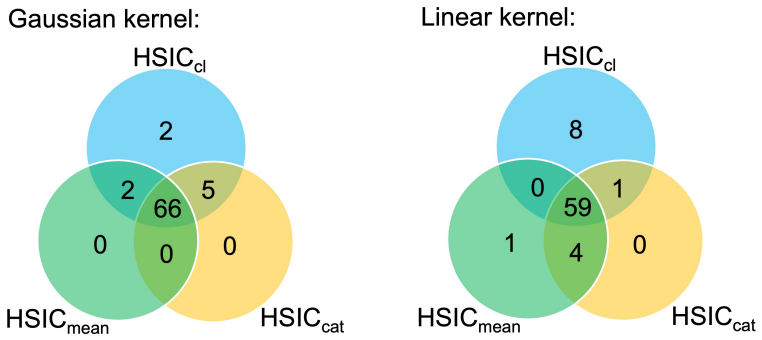

Figure B.6: Venn diagrams for the number of metabolic pathways identified to be associated with the vaginal microbiome composition based on the MsFLASH data set ($\alpha = 5.3 \times 10^{-4}$). The results are separated by kernel.

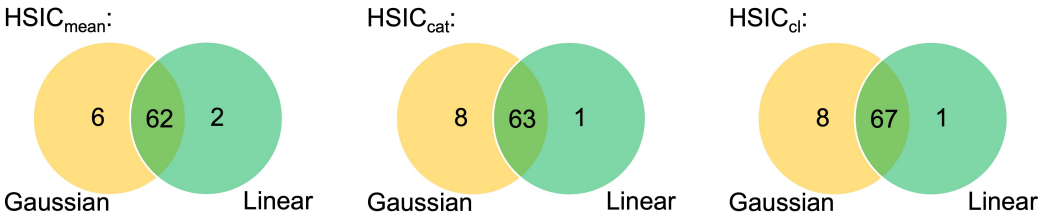

Figure B.7: Venn diagrams for the number of metabolic pathways identified to be associated with the vaginal microbiome composition based on the MsFLASH data set ($\alpha = 5.3 \times 10^{-4}$). The results are separated by method for HSIC test.

Appendix C

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

C.1 Connection between CRV and SEcov

The scaled expected conditional covariance (SEcov) is a nonparametric measure proposed by Xiang and Simon [141] to assess the conditional dependence between two univariate variables given a third variable. Consider three random variables: $Y \in \mathbb{R}$, $Z \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^m$ with a joint distribution P_{YZX}. Suppose that we observe a sample $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P_{YZX}$.

The population SEcov parameter is defined as

$$\mathrm{SEcov}_{P_{YZX}}(Y, Z|\boldsymbol{X}) = \frac{\mathbb{E}[\mathbb{C}\mathrm{ov}(Y, Z|\boldsymbol{X})]}{\sqrt{\mathbb{E}[\mathbb{V}\mathrm{ar}(Y|\boldsymbol{X})]}\sqrt{\mathbb{E}[\mathbb{V}\mathrm{ar}(Z|\boldsymbol{X})]}}, \tag{C.1}$$

where $\mathbb{C}\mathrm{ov}(Y, Z|\boldsymbol{X})$ is the conditional covariance of Y and Z given $\boldsymbol{X}$ and $\mathbb{V}\mathrm{ar}(Y|\boldsymbol{X})$ is the conditional variance of Y given $\boldsymbol{X}$. The SEcov parameter measures the average degree of association between Y and Z conditional on $\boldsymbol{X}$ without assumptions on the joint distribution, P_{YZX}, of the variables. When P_{YZX} is a multivariate normal distribution, the SEcov parameter is equivalent to the partial correlation between Y and Z.

The corresponding sample SEcov statistic based on $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{i=1}^n$ is defined as

$$\mathrm{SEcov}_{P_n}(Y, Z|\boldsymbol{X}) = \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\mu}_Y(\boldsymbol{X}_i))(Z_i - \hat{\mu}_Z(\boldsymbol{X}_i))}{\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\mu}_Y(\boldsymbol{X}_i))^2}\sqrt{\frac{1}{n}\sum_{i=1}^n (Z_i - \hat{\mu}_Z(\boldsymbol{X}_i))^2}}, \tag{C.2}$$

where $\hat{\mu}_Y$ and $\hat{\mu}_Z$ are predictive models for the conditional means $\mathbb{E}[Y|\boldsymbol{X}]$ and $\mathbb{E}[Z|\boldsymbol{X}]$. These conditional mean estimators can again be derived using common machine learning approaches, similar to CRV in the main text.

On the other hand, based on Eq. 4.1 in the main text, the CRV parameter for assessing

the conditional dependence between Y and Z given $\boldsymbol{X}$ can be written as:

$$\begin{aligned}
\mathrm{CRV}_{P_{YZX}}(Y, Z|\boldsymbol{X}) &= \frac{\mathbb{E}\left[(Y - \mathbb{E}[Y|\boldsymbol{X}])(Z - \mathbb{E}[Z|\boldsymbol{X}])\right]^2}{\mathbb{E}\left[(Y - \mathbb{E}[Y|\boldsymbol{X}])^2\right]\mathbb{E}\left[(Z - \mathbb{E}[Z|\boldsymbol{X}])^2\right]} \\
&= \frac{\mathbb{E}\left\{\mathbb{E}\left[(Y - \mathbb{E}[Y|\boldsymbol{X}])(Z - \mathbb{E}[Z|\boldsymbol{X}])\middle|\boldsymbol{X}\right]\right\}^2}{\mathbb{E}\left\{\mathbb{E}\left[(Y - \mathbb{E}[Y|\boldsymbol{X}])^2\middle|\boldsymbol{X}\right]\right\}\mathbb{E}\left\{\mathbb{E}\left[(Z - \mathbb{E}[Z|\boldsymbol{X}])^2\middle|\boldsymbol{X}\right]\right\}} \\
&= \frac{\mathbb{E}[\mathbb{C}\mathrm{ov}(Y, Z|\boldsymbol{X})]^2}{\mathbb{E}[\mathbb{V}\mathrm{ar}(Y|\boldsymbol{X})]\,\mathbb{E}[\mathbb{V}\mathrm{ar}(Z|\boldsymbol{X})]} \\
&= \left[\mathrm{SEcov}_{P_{YZX}}(Y, Z|\boldsymbol{X})\right]^2.
\end{aligned} \tag{C.3}$$

Similarly, based on Eq. 4.2 in the main text, we can derive that:

$$\mathrm{CRV}_{P_n}(Y, Z|\boldsymbol{X}) = \left[\mathrm{SEcov}_{P_n}(Y, Z|\boldsymbol{X})\right]^2. \tag{C.4}$$

Therefore, for univariate Y and Z, the population CRV parameter is exactly equal to the squared population SEcov parameter, and the sample CRV statistic is equal to the squared sample SEcov statistic. The same conditional independence testing procedure as described in Section 4.3.2 of the main text can be used to conduct hypothesis testing for SEcov.

C.2 Proof of Theorem 4.3.2

Following Section 4.3.1 and 4.3.2, let Ψ be the numerator of $\mathrm{CRV}_{P_{YZX}}(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$ and $\hat{\Psi}$ be the numerator of $\mathrm{CRV}_{P_n}(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$. We also note that $\mu_{Y_j}(\boldsymbol{x}) = \mathbb{E}[Y_j|\boldsymbol{X} = \boldsymbol{x}]$ and $\mu_{Z_k}(\boldsymbol{x}) = \mathbb{E}[Z_k|\boldsymbol{X} = \boldsymbol{x}]$ for $j = 1, \cdots, p$ and $k = 1, \cdots, q$. Then we have

$$\begin{aligned}
\Psi &= \mathrm{tr}(\Sigma_{\boldsymbol{YZ}|\boldsymbol{X}}\Sigma_{\boldsymbol{ZY}|\boldsymbol{X}}) = \sum_{j=1}^{p}\sum_{k=1}^{q}\mathbb{E}^2\left[(Y_j - \mu_{Y_j}(\boldsymbol{X}))(Z_k - \mu_{Z_k}(\boldsymbol{X}))\right] =: \sum_{j=1}^{p}\sum_{k=1}^{q}\varphi_{jk}^2, \\
\hat{\Psi} &= \mathrm{tr}(S_{\boldsymbol{YZ}|\boldsymbol{X}}S_{\boldsymbol{ZY}|\boldsymbol{X}}) = \sum_{j=1}^{p}\sum_{k=1}^{q}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_{ji} - \hat{\mu}_{Y_j}(\boldsymbol{X}_i))(Z_{ki} - \hat{\mu}_{Z_k}(\boldsymbol{X}_i))\right]^2 =: \sum_{j=1}^{p}\sum_{k=1}^{q}\hat{\varphi}_{jk}^2,
\end{aligned} \tag{C.5}$$

where we define $\varphi_{jk} = \mathbb{E}\left[(Y_j - \mu_{Y_j}(\boldsymbol{X}))(Z_k - \mu_{Z_k}(\boldsymbol{X}))\right]$ and $\hat{\varphi}_{jk} = \frac{1}{n}\sum_{i=1}^{n}(Y_{ji} - \hat{\mu}_{Y_j}(\boldsymbol{X}_i))(Z_{ki} - \hat{\mu}_{Z_k}(\boldsymbol{X}_i))$.

Let $\boldsymbol{O} = (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, so that we can write $P_{YZX} = P_O$. It has been shown in the work of Xiang et al. [141] that, under Assumption 4.3.1, we can derive:

1. The efficient influence function [116] of $\hat{\varphi}_{jk}$ (as an estimator of φ_{jk}) under sampling from P_O is

$$D_{jk}(\boldsymbol{o}) = [y_j - \mu_{Y_j}(\boldsymbol{x})][z_k - \mu_{Z_k}(\boldsymbol{x})] - \varphi_{jk}. \tag{C.6}$$

   In other words, we have $\hat{\varphi}_{jk} - \varphi_{jk} = \frac{1}{n}\sum_{i=1}^{n} D_{jk}(\boldsymbol{O}_i) + o_P(n^{-1/2})$.

2. $\hat{\varphi}_{jk}$ is asymptotically normal:

$$\sqrt{n}(\hat{\varphi}_{jk} - \varphi_{jk}) \xrightarrow{d} N(0, \sigma_{jk}^2), \tag{C.7}$$

   where $\sigma_{jk}^2 = \int [D_{jk}(\boldsymbol{o})]^2 dP_O(\boldsymbol{o})$.

Let

$$\boldsymbol{w}(\boldsymbol{o}) = \begin{pmatrix} [y_1 - \mu_{Y_1}(\boldsymbol{x})][z_1 - \mu_{Z_1}(\boldsymbol{x})] \\ [y_1 - \mu_{Y_1}(\boldsymbol{x})][z_2 - \mu_{Z_2}(\boldsymbol{x})] \\ \vdots \\ [y_p - \mu_{Y_p}(\boldsymbol{x})][z_q - \mu_{Z_q}(\boldsymbol{x})] \end{pmatrix}_{pq\times 1}, \boldsymbol{v}(\boldsymbol{o}_i) = \begin{pmatrix} [y_{1i} - \hat{\mu}_{Y_1}(\boldsymbol{x}_i)][z_{1i} - \hat{\mu}_{Z_1}(\boldsymbol{x}_i)] \\ [y_{1i} - \hat{\mu}_{Y_1}(\boldsymbol{x}_i)][z_{2i} - \hat{\mu}_{Z_2}(\boldsymbol{x}_i)] \\ \vdots \\ [y_{pi} - \hat{\mu}_{Y_p}(\boldsymbol{x}_i)][z_{qi} - \hat{\mu}_{Z_q}(\boldsymbol{x}_i)] \end{pmatrix}_{pq\times 1}, \tag{C.8}$$

so that $\boldsymbol{w}(\boldsymbol{o})$ is the vector of products between $y_j - \mu_{Y_j}(\boldsymbol{x})$ and $z_k - \mu_{Z_k}(\boldsymbol{x})$, and $\boldsymbol{v}(\boldsymbol{o}_i)$ is the vector of products between $y_{ji} - \hat{\mu}_{Y_j}(\boldsymbol{x}_i)$ and $z_{ki} - \hat{\mu}_{Z_k}(\boldsymbol{x}_i)$ for all j and k. Further let

$$\boldsymbol{D}(\boldsymbol{o}) = \begin{pmatrix} D_{11}(\boldsymbol{o}) \\ D_{12}(\boldsymbol{o}) \\ \vdots \\ D_{pq}(\boldsymbol{o}) \end{pmatrix}_{pq\times 1}. \tag{C.9}$$

Then $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}(\boldsymbol{O}_i)$ is an asymptotically linear estimator of $\mathbb{E}[\boldsymbol{w}(\boldsymbol{O})]$ with influence function $\boldsymbol{D}(\boldsymbol{o})$.

Under $H_0 : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}$, we have $\varphi_{jk} = 0$ for all j, k, so $\mathbb{E}[\boldsymbol{w}(\boldsymbol{O})] = \boldsymbol{0}$ and $\boldsymbol{D}(\boldsymbol{o}) = \boldsymbol{w}(\boldsymbol{o})$. Therefore, by multivariate central limit theorem and property of influence functions, under H_0, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{v}(\boldsymbol{O}_i) = \sqrt{n}\begin{pmatrix}\hat{\varphi}_{11}\\ \vdots \\ \hat{\varphi}_{pq}\end{pmatrix} \xrightarrow{d} N(\boldsymbol{0}, \mathbb{E}[\boldsymbol{w}(\boldsymbol{O})\boldsymbol{w}(\boldsymbol{O})^T]). \tag{C.10}$$

By continuous mapping theorem, it then follows that, under H_0,

$$n\hat{\Psi} = \Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{v}(\boldsymbol{O}_i)\Big]^T\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{v}(\boldsymbol{O}_i)\Big] \xrightarrow{d} \sum_{t=1}^{pq}\ell_t\tau_t, \tag{C.11}$$

where τ_t's are i.i.d. χ^2_1 variables and ℓ_t's are eigenvalues of the matrix $\mathbb{E}[\boldsymbol{w}(\boldsymbol{O})\boldsymbol{w}(\boldsymbol{O})^T]$. In practice, ℓ_t's are unknown and can be estimated using eigenvalues of the matrix $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}(\boldsymbol{O}_i)\boldsymbol{v}(\boldsymbol{O}_i)^T$.

C.3 Additional Details on Simulation

C.3.1 Synthetic Microbial Network Recovery

Simulation Procedure According to the original network simulation procedure from Kurtz et al. [37], a desired network topology is first generated and stored in a precision matrix, where its non-zero pattern corresponds to the adjacency matrix for the network under normality assumptions. Multivariate normal data are generated according to the specified precision matrix and then quantile-transformed to zero-inflated negative binomial (ZINB) data marginally, which serve as the simulated microbial counts. The parameters of the ZINB distributions are fitted based on real microbiome data from the American Gut Project [149]. Here we extend this framework to further include a multivariate aspect.

We first construct an initial network at the genus level according to the strategy of [37],

where each genus only contains one species (Figure C.1, left), and generate a positive-definite precision matrix for the network by assigning weights to non-zero entries of the adjacency matrix. We further take the inverse of the precision matrix to get the covariance matrix.

We then add an arbitrary number of additional species to each genus/node (Figure C.1, right). For example, if Genus A initially only contains Species A_1, we add another species, Species A_2, by generating a new variable: $A_2 = \beta_{A_2} A_1 + \epsilon_{A_2}$, where $\beta_{A_2} \sim Unif(0.5, 0.8)$ and $\epsilon_{A_2} \sim N(0, 1)$ is a term independent from any existing species in the graph. Additional species can be added similarly. It is easy to derive the covariance between A_2 and other existing species in the graph, since A_2 is constructed as a linear combination of A_1 and an independent error term. Therefore, an augmented species-level covariance matrix can be constructed, which contains entries that correspond to the additional species. This augmented covariance matrix will be used to generate multivariate normal data, which is then quantile-transformed into count data marginally.

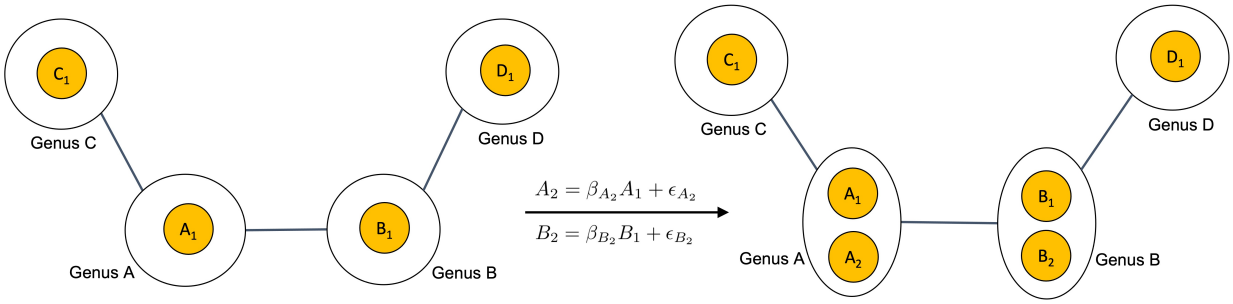

Figure C.1: Example procedure for generating a genus-level network with species-level data available.

We can show that the original genus-level network topology is preserved after adding additional species using the above strategy. Using Figure C.1 as an example, where additional

species are added to Genus A and Genus B, we see that

$$\begin{aligned}\mathbb{C}\text{ov}(A_2, C_1 \mid B_1, B_2, D_1) &= \mathbb{C}\text{ov}(A_2, C_1 \mid B_1, \beta_{B_2} B_1 + \epsilon_{B_2}, D_1) \\ &= \mathbb{C}\text{ov}(A_2, C_1 \mid B_1, \epsilon_{B_2}, D_1) \\ &= \mathbb{C}\text{ov}(A_2, C_1 \mid B_1, D_1) \\ &= \beta_{A_2}\, \mathbb{C}\text{ov}(A_1, C_1 \mid B_1, D_1) \neq 0.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{C}\text{ov}(A_2, D_1 \mid B_1, B_2, C_1) &= \mathbb{C}\text{ov}(A_2, D_1 \mid B_1, C_1) \\ &= \beta_{A_2}\, \mathbb{C}\text{ov}(A_1, D_1 \mid B_1, C_1) = 0.\end{aligned}$$

Therefore, any additional species added to a genus will have the same zero/non-zero status as the initial species in that genus, in terms of their conditional covariance with species in other genera.

Note that, in the above setting, we let the β's be positive, so all species in the same genus would share the same direction of conditional correlations with species in other genera. To introduce heterogeneous relationships among the species, we randomly add negative signs to the β's in our data generation process. For convenience, we denote a genus as containing relationships of the "same directions" if all species in that genus have positive β's, and denote a genus as containing relationships of "opposite directions" if around half of the species in that genus have positive β's and the other half have negative β's. This pattern is illustrated in Figure C.2.

Additional Results Table C.1 and C.2 show the average Hamming distance between the constructed network and the true graph for different methods under different topologies, where heterogeneous conditional correlations are present among species within 10% and 30% of all genera, respectively. Figure C.3 and C.4 show the ROC curves of different methods in recovering microbial networks of different topologies, where heterogeneous conditional

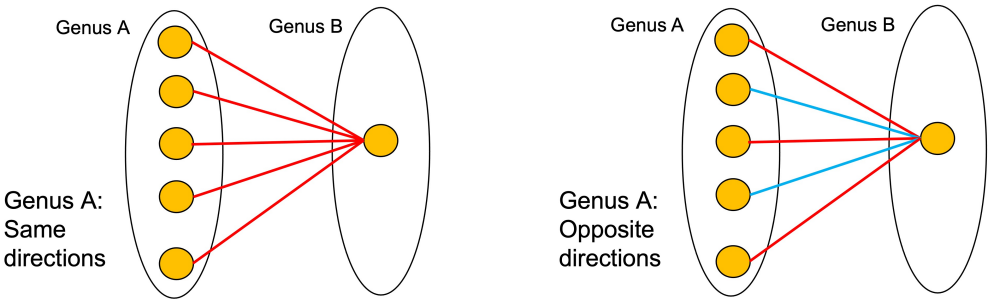

Figure C.2: Patterns of heterogeneous relationships among species within different genera. Red and blue edges represent positive and negative conditional correlations, respectively.

correlations are present among species within 10% and 30% of all genera, respectively.

Table C.1: Average Hamming distance between the constructed network and the true graph for different methods under different topologies, when conditional correlations of opposite directions are present among species within 10% of all genera.

Method	Band	Cluster	Scale-free
CRV	20.2	23.4	34.6
SEcov	29.1	32.2	40.6
SPIEC-EASI (MB)	34.7	34.9	45.1
SPIEC-EASI (glasso)	36.6	35.9	45.2
SPRING	35.0	35.5	37.1

The false discovery rate is controlled at 0.2.

C.4 Additional Details on Real Data Application

C.4.1 Description of the Study Sample from the PIN Study

The Pregnancy, Infection and Nutrition (PIN) study is a prospective cohort study of pregnant women with singleton pregnancies conducted in central North Carolina in the United States, with a goal of identifying risk factors for preterm birth (PTB) [220]. The original study recruited 3,163 pregnant women with baseline gestational age between 24 and 29 weeks

Table C.2: Average Hamming distance between the constructed network and the true graph for different methods under different topologies, when conditional correlations of opposite directions are present among species within 30% of all genera.

Method	Band	Cluster	Scale-free
CRV	19.8	23.5	35.4
SEcov	32.1	32.9	41.6
SPIEC-EASI (MB)	34.9	34.5	42.9
SPIEC-EASI (glasso)	37.0	36.7	50.0
SPRING	36.4	36.0	41.7

The false discovery rate is controlled at 0.2.

from prenatal clinics, who were followed through 12 months postpartum. At enrollment, the participants provided blood, urine and genital tract samples. Demographic information, medical histories and health behavior were collected via telephone interviews in the 2 weeks following enrollment.

Sun et al. [143] conducted a vaginal microbial analysis based on a nested case-control subset of the PIN data to study the associations between race, the vaginal microbiome and spontaneous PTB. Term birth was defined as reaching $\geq$ 37 weeks of gestation. Sun et al. selected all 402 women in the PIN population who went on to have a PTB as the case group and randomly sampled 799 women from those who went on to have a term birth as the control group. They further restricted the control group to women who self-identified as Black or White race, resulting in 652 women (375 White women and 277 Black women) with term birth. Vaginal swabs collected at enrollment were previously stored at -70 °C and processed by Sun et al. to extract DNA and perform 16S rRNA sequencing on their selected samples. A detailed description of microbial sequencing and bioinformatic processing of the 16S data can be found in [143].

For our current network analysis, we utilize the available vaginal microbiome data from the 652 women of the control group from Sun et al.'s study. These women can represent a

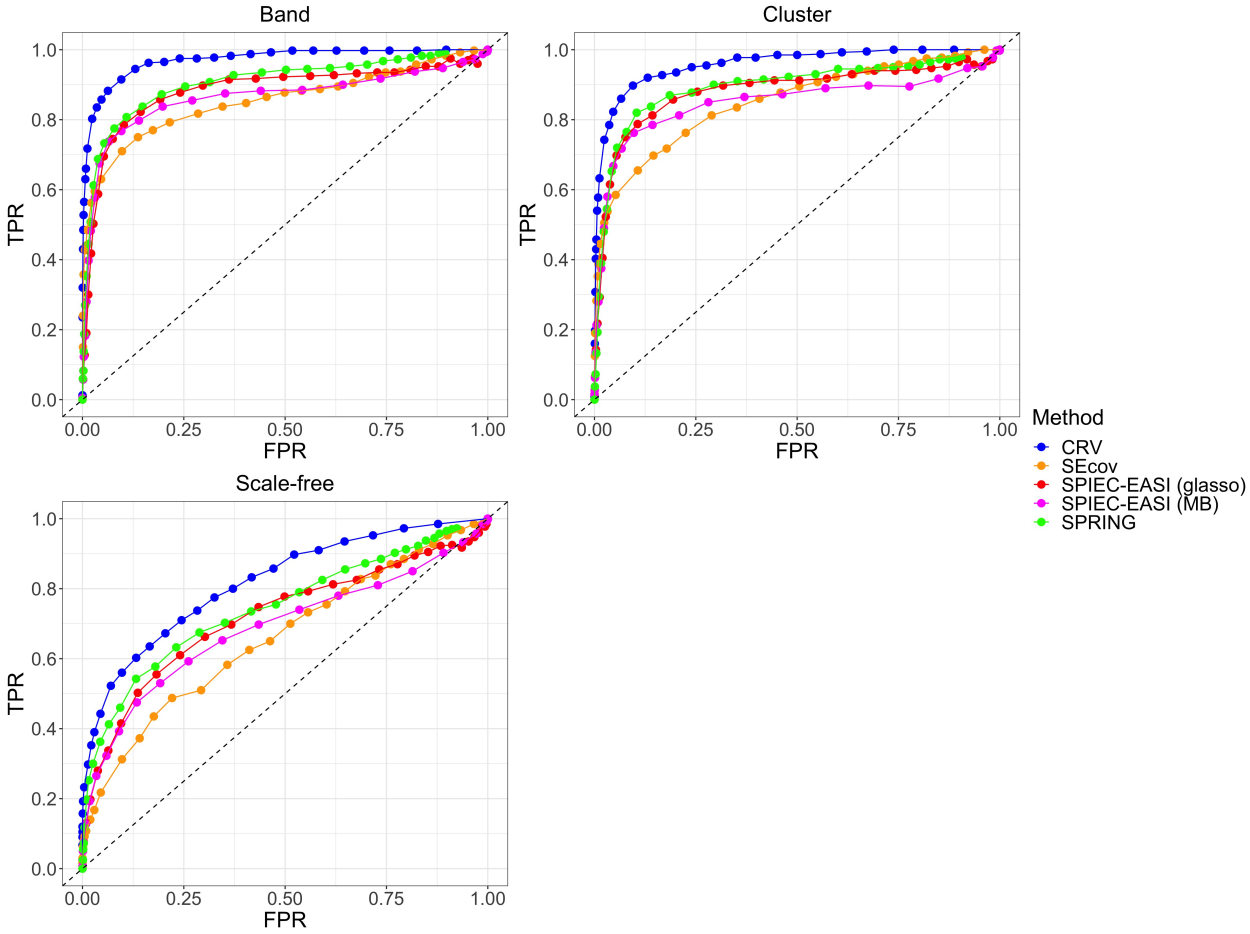

Figure C.3: ROC curves of different methods in recovering synthetic microbial networks of different topologies, averaged over 10 simulations. Conditional correlations of opposite directions are present among species within 10% of all genera.

random subsample of PIN participants who experienced a term birth and allow us to compare the vaginal network structure between Black and White women, due to the reasonable sample sizes of both races.

C.4.2 Additional Results

Figure C.5 shows the comparison of genus-level vaginal microbial networks based on a random subset of White women ($n = 274$) and based on all Black women ($n = 274$) of the PIN study.

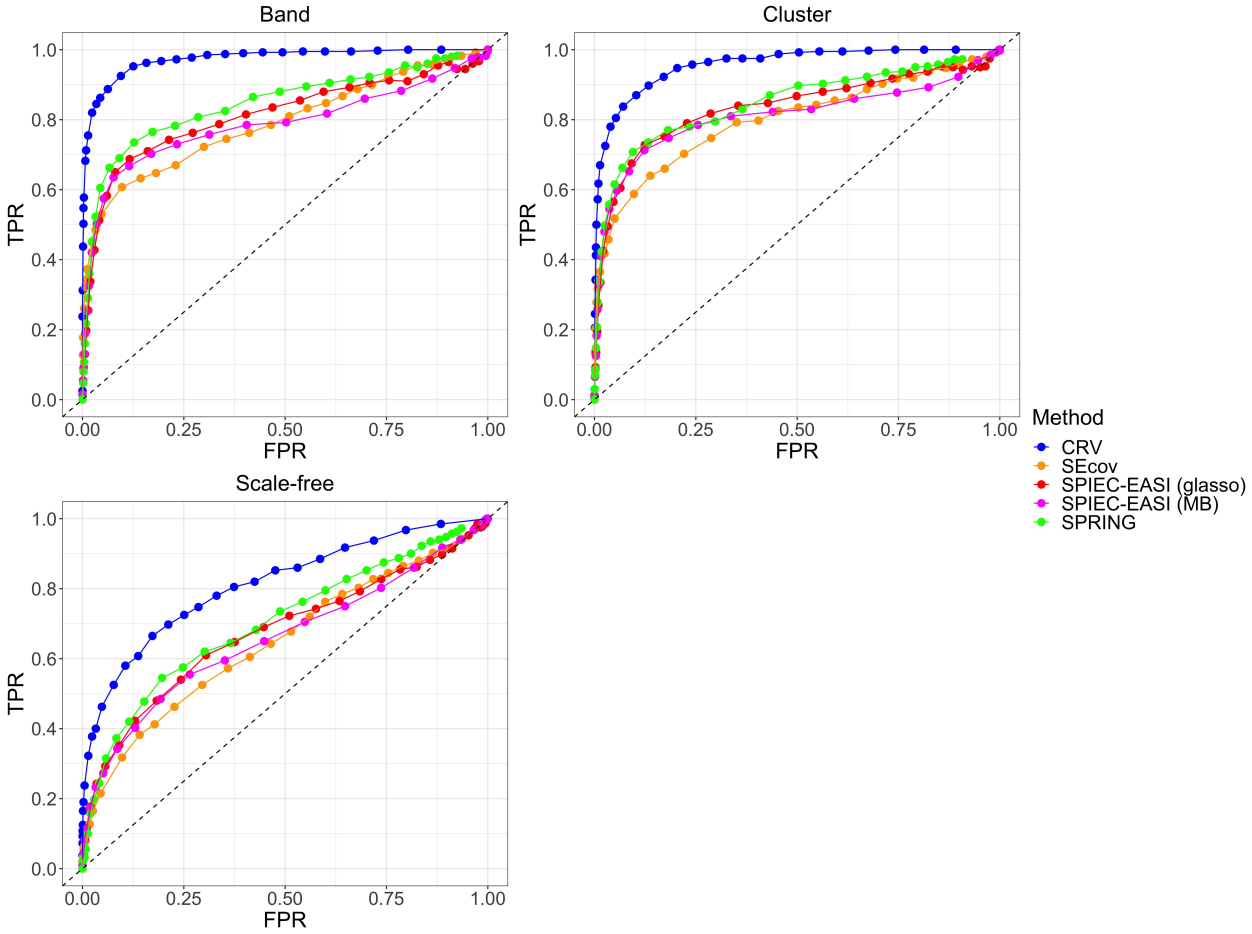

Figure C.4: ROC curves of different methods in recovering synthetic microbial networks of different topologies, averaged over 10 simulations. Conditional correlations of opposite directions are present among species within 30% of all genera.

Figure C.6 shows the corresponding heatmaps of network centrality measures for each genus in the networks from Figure C.5.

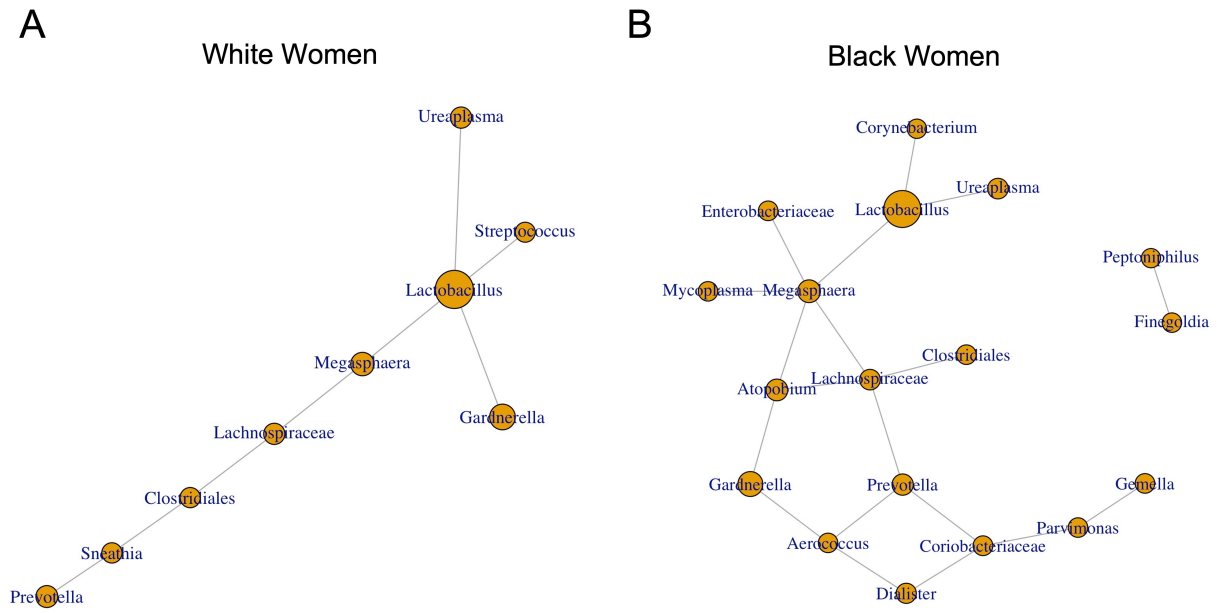

Figure C.5: Genus-level vaginal microbial networks based on a random subset of White women (Panel **A**; $n = 274$) vs. based on all Black women (Panel **B**; $n = 274$) of PIN study, with false discovery rate controlled at 0.2. The sizes of the nodes are proportional to their genus-level microbial abundance in the combined sample.

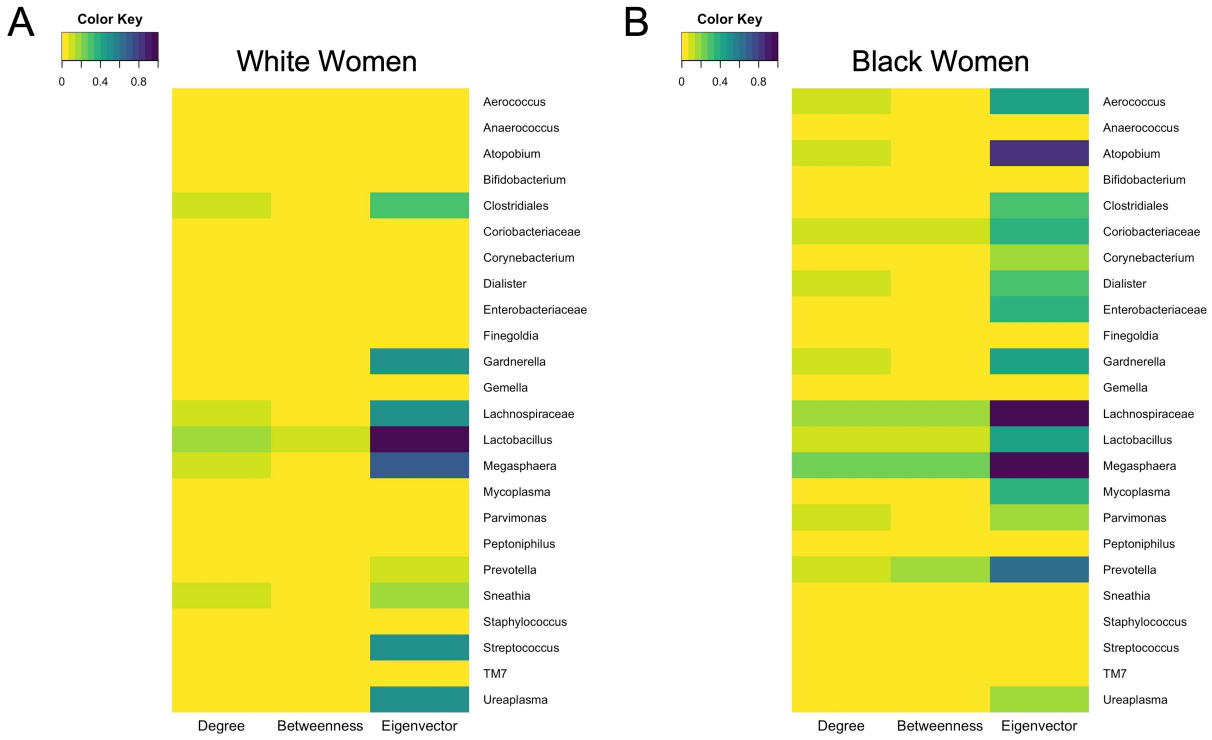

Figure C.6: Heatmaps of network centrality measures for each genus in genus-level vaginal microbial networks based on a random subset of White women (Panel **A**; $n = 274$) vs. based on all Black women (Panel **B**; $n = 274$) of PIN study.