Ahn and Datta *BMC Bioinformatics* (2024) 25:117
https://doi.org/10.1186/s12859-024-05689-7

BMC Bioinformatics

RESEARCH Open Access

Differential network connectivity analysis for microbiome data adjusted for clinical covariates using jackknife pseudo-values

Seungjun Ahn[1,2,3] and Somnath Datta[1*]

*Correspondence: somnath.datta@ufl.edu

[1] Department of Biostatistics, University of Florida, Gainesville, FL, USA
[2] Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[3] Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Abstract

Background: A recent breakthrough in differential network (DN) analysis of microbiome data has been realized with the advent of next-generation sequencing technologies. The DN analysis disentangles the microbial co-abundance among taxa by comparing the network properties between two or more graphs under different biological conditions. However, the existing methods to the DN analysis for microbiome data do not adjust for other clinical differences between subjects.

Results: We propose a Statistical Approach via Pseudo-value Information and Estimation for Differential Network Analysis (SOHPIE-DNA) that incorporates additional covariates such as continuous age and categorical BMI. SOHPIE-DNA is a regression technique adopting jackknife pseudo-values that can be implemented readily for the analysis. We demonstrate through simulations that SOHPIE-DNA consistently reaches higher recall and F1-score, while maintaining similar precision and accuracy to existing methods (NetCoMi and MDiNE). Lastly, we apply SOHPIE-DNA on two real datasets from the American Gut Project and the Diet Exchange Study to showcase the utility. The analysis of the Diet Exchange Study is to showcase that SOHPIE-DNA can also be used to incorporate the temporal change of connectivity of taxa with the inclusion of additional covariates. As a result, our method has found taxa that are related to the prevention of intestinal inflammation and severity of fatigue in advanced metastatic cancer patients.

Conclusion: SOHPIE-DNA is the first attempt of introducing the regression framework for the DN analysis in microbiome data. This enables the prediction of characteristics of a connectivity of a network with the presence of additional covariate information in the regression. The R package with a vignette of our methodology is available through the CRAN repository (https://CRAN.R-project.org/package=SOHPIE), named SOHPIE (pronounced as *Sofie*). The source code and user manual can be found at https://github.com/sjahnn/SOHPIE-DNA.

Keywords: Differential network analysis, Regression modeling, Microbial co-abundance, Jackknife pseudo-values

Background

The human microbiome is the collective genomes of microbes or micro-organisms localized to the various sites of human body [1]. Recent clinical studies have shown that the microbiome has a regulatory role in a wide array of illnesses in humans, such as cancer [2], human immunodeficiency virus [3], and inflammatory bowel disease (IBD) [4]. Moreover, the human microbiome is linked to emotional well-being [5] and mental health including depression [6], autism spectrum disorders [7], and human brain diseases [8].

Following the advent of next-generation sequencing technologies, the taxonomic composition of microbial communities is better characterized by the amplification of small fragments (or amplicon) of the 16S ribosomal RNA (or 16S rRNA) gene. More recently, shotgun metagenomic sequencing has become an alternative for microbial community profiling [9]. Either sequencing platform typically employs similarity-based clustering algorithms to group 16S rRNA sequences into Operational Taxonomic Units (OTU) [10, 11] that are compositional.

The applications of network theory have been successfully utilized to better appraise the complex symbiotic (or dysbiotic) relationship between microbiome and disease states – microbial co-abundances [12]. The abundance matrix or the observed OTU table is used to infer microbial co-abundances among taxa through either correlation-based approaches or probabilistic graphical models.

The differential network (DN) analysis compares the network properties between two or more graphs under different biological conditions, such as degree centrality. Based on the recent review article [13], there are two methods that are newly available to the DN analysis for microbiome data: Microbiome Differential Network Estimation (MDiNE) [14] and Network Construction and comparison for Microbiome data (NetCoMi) [15]. These methods, however, do not assume that the association structure depends on additional binary and continuous covariates.

It has been recognized that the composition of the gut microbiome is central to the pathogenesis of IBD [4, 16]. In addition, the gut microbiome composition in patients with IBD is largely influenced by various factors including the use of antibiotics, diet, and cigarette smoking [4]. In an analogous fashion, it is not unreasonable to speculate that the structure of the microbial networks can also vary depending on these factors. Thereby, there is a need for statistical methods for DN analysis that can include additional predictor variables.

One way to accomplish this goal is to use a regression technique based on pseudo-values, a component to calculate the bias-corrected estimator of leave-one-out jackknife resampling procedure [17]. The pseudo-value technique was first postulated by Andersen and his colleagues [18, 19] in the context of multi-state survival models with right-censored data. Since then, it has been well studied in various disciplines of statistics including the interval-censored data [20, 21], clustered data [22, 23], and machine learning methods [24, 25].

The ultimate benefit of this technique is its straightforward inclusion of additional covariates in the generalized linear model [26]. An asymptotic linearity and consistency of pseudo-values given covariates are shown with the second-order von Mises expansion [27, 28]. The pseudo-values can then be used as the response variable in a regression

model with the covariates [29]. Several studies reported that the type I error is well controlled at a nominal level of 0.05 while maintaining a high statistical power under the quasi-likelihood generalized linear mixed model [30] and generalized estimating equations framework [23, 31] for pseudo-value regression approach.

Hence, we propose a regression modeling method for differential connectivity (DC) analysis that regresses the jackknife pseudo-values calculated from a degree centrality of taxa in a microbial network to directly estimate the effects of predictors. The primary focus of the methodological innovation presented in this manuscript is centered on DC analysis, a subset of the broader DN analysis. The findings of DC analysis primarily describe the DC of individual nodes (or taxa) instead of taxon-taxon co-abundance relationships [32]. In this approach, the grouping variable itself could also be included in the regression model along with additional clinical covariates while regressing the pseudo-values. We loosely refer to this as a "multivariable setting", whereas in "univariable settings" only the grouping variable is utilized in a DN analysis.

In the present study, we introduce **S**tatistical Appr**O**ac**H** via **P**seudo-value **I**nformation and **E**stimation for **D**ifferential **N**etwork **A**nalysis (SOHPIE-DNA) that can include covariate information in analyzing microbiome data. We firstly demonstrate the plausibility of the proposed method by comparing the model performances with MDiNE and NetCoMi through simulations under multivariable and univariable settings. Of note, the covariate adjustment is the main strength of our proposed method. Therefore, the findings from multivariable simulation setting corroborates the main reason for our methods paper. Furthermore, the SOHPIE-DNA is applied to illustrate its clinical utility by examining real data from the American Gut Project [33] and the Diet Exchange Study [34] to identify DC taxa with presence of covariates. All statistical analyses are performed in R version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Simulation study

The sample size $n = 20, 50, 200, 500$ are considered for each microbial network with $p = 20, 40$ taxa over 1,000 Monte Carlo replicates. Simulations are repeated to assess the effects of covariates on taxa by changing the effect size, $\delta = 0.05, 0.1, 0.2$, which is described in "Simulated data" section. A new network is generated at each simulation replicate to account for biological variability of the network structure.

The performance metrics provided in "Performance measures" section are computed by comparing the test results with the true network. In the true network setting, a taxon is truly DC between groups if it is connected to at least one different neighbor taxon between groups. Tables 1 and 2 summarize simulation results under the multivariable setting. That is, a continuous covariate is included with the binary group variable in the regression model. To illustrate the utility of the proposed method on covariate-dependent network, we compared the pseudo-value regression approach with the recent methods available (NetCoMi and MDiNE) that cannot incorporate the additional covariate. Results show that the SOHPIE-DNA consistently maintains high recall values in all specifications of taxa, sample sizes, and effect sizes, and outperforms NetCoMi and MDiNE in almost all cases. A higher F1 score of SOHPIE-DNA indicates that the proposed method can achieve a better overall model performance in the presence of

Table 1 The simulation results for the case when the network structure depends on age covariate

p	n	δ_1	δ_2	Precision			Recall			F1			Accuracy		
				SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE
20	20	0.05	0.05	0.26	0.28	**0.36**	**0.69**	0.25	0.01	**0.39**	0.35	0.31	0.42	0.06	**0.75**
		0.05	0.10	0.36	0.37	**0.49**	**0.69**	0.26	0.01	**0.45**	0.37	0.30	0.45	0.09	**0.65**
		0.05	0.20	**0.52**	0.51	0.39	**0.69**	0.24	0.01	**0.57**	0.36	0.24	**0.51**	0.12	0.49
		0.10	0.05	0.36	0.36	**0.44**	**0.70**	0.23	0.01	**0.46**	0.35	0.32	0.45	0.08	**0.65**
		0.10	0.10	**0.42**	**0.42**	**0.42**	**0.69**	0.25	0.01	**0.51**	0.37	0.25	0.47	0.11	**0.58**
		0.10	0.20	**0.54**	**0.54**	0.43	**0.70**	0.24	0.01	**0.59**	0.38	0.28	**0.52**	0.13	0.46
		0.20	0.05	0.50	**0.51**	0.48	**0.69**	0.25	0.01	**0.56**	0.38	0.25	**0.51**	0.12	0.50
		0.20	0.10	0.53	**0.54**	0.45	**0.70**	0.26	0.01	**0.58**	0.38	0.20	**0.52**	0.14	0.47
		0.20	0.20	**0.60**	0.58	0.42	**0.70**	0.24	0.01	**0.62**	0.38	0.18	**0.54**	0.14	0.41
	50	0.05	0.05	0.25	0.27	**0.35**	**0.82**	0.53	0.12	**0.39**	0.37	0.34	0.34	0.13	**0.72**
		0.05	0.10	0.35	0.37	**0.42**	**0.84**	0.58	0.10	**0.49**	0.44	0.30	0.41	0.20	**0.63**
		0.05	0.20	0.50	**0.52**	0.50	**0.84**	0.63	0.10	**0.61**	0.54	0.26	**0.50**	0.31	**0.50**
		0.10	0.05	0.35	0.37	**0.42**	**0.83**	0.58	0.11	**0.49**	0.44	0.30	0.40	0.20	**0.63**
		0.10	0.10	0.41	0.43	**0.44**	**0.83**	0.62	0.11	**0.54**	0.48	0.29	0.44	0.25	**0.57**
		0.10	0.20	0.53	0.55	**0.57**	**0.84**	0.64	0.12	**0.64**	0.56	0.28	**0.52**	0.34	0.47
		0.20	0.05	0.51	**0.52**	0.49	**0.84**	0.61	0.10	**0.62**	0.54	0.27	**0.51**	0.31	0.49
		0.20	0.10	0.54	**0.55**	0.54	**0.83**	0.64	0.11	**0.64**	0.57	0.27	**0.52**	0.35	0.47
		0.20	0.20	**0.59**	**0.59**	0.58	**0.84**	0.69	0.12	**0.68**	0.61	0.27	**0.56**	0.40	0.43
	200	0.05	0.05	0.26	**0.27**	0.26	**0.93**	0.63	0.45	**0.41**	0.38	0.35	0.29	0.16	**0.52**
		0.05	0.10	0.35	**0.37**	0.35	**0.94**	0.68	0.45	**0.50**	0.46	0.38	0.37	0.24	**0.51**
		0.05	0.20	0.51	**0.52**	0.50	**0.94**	0.74	0.48	**0.65**	0.58	0.47	**0.51**	0.37	0.49
		0.10	0.05	0.35	**0.36**	0.35	**0.94**	0.67	0.46	**0.50**	0.45	0.40	0.37	0.23	**0.51**
		0.10	0.10	0.41	**0.43**	0.40	**0.94**	0.74	0.48	**0.56**	0.52	0.43	0.42	0.30	**0.50**
		0.10	0.20	0.54	**0.55**	0.53	**0.93**	0.78	0.50	**0.67**	0.62	0.49	**0.53**	0.42	0.50
		0.20	0.05	0.50	**0.52**	0.48	**0.94**	0.72	0.48	**0.64**	0.58	0.46	**0.50**	0.36	0.49

Table 1 (continued)

p	n	δ_1	δ_2	Precision			Recall			F1			Accuracy		
				SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE
		0.20	0.10	0.53	**0.54**	**0.54**	**0.93**	0.76	0.51	**0.67**	0.61	0.51	**0.53**	0.41	0.51
		0.20	0.20	0.58	**0.59**	0.56	**0.93**	0.81	0.52	**0.71**	0.66	0.52	**0.57**	0.48	0.49
	500	0.05	0.05	0.25	**0.27**	0.24	**0.96**	0.73	0.69	**0.41**	0.40	0.36	0.27	0.18	**0.38**
		0.05	0.10	0.35	**0.37**	0.35	**0.97**	0.78	0.75	**0.51**	0.48	0.47	0.36	0.28	**0.42**
		0.05	0.20	0.51	**0.52**	0.48	**0.96**	0.82	0.76	**0.65**	0.61	0.57	**0.51**	0.42	0.48
		0.10	0.05	0.34	**0.35**	**0.35**	**0.96**	0.76	0.71	**0.50**	0.46	0.45	0.35	0.26	**0.43**
		0.10	0.10	0.42	**0.43**	0.42	**0.97**	0.80	0.74	**0.57**	0.53	0.52	0.42	0.33	**0.46**
		0.10	0.20	0.53	**0.54**	0.51	**0.96**	0.85	0.79	**0.67**	0.64	0.60	**0.53**	0.45	0.50
		0.20	0.05	0.50	**0.51**	0.49	**0.96**	0.82	0.78	**0.65**	0.61	0.59	**0.50**	0.41	**0.50**
		0.20	0.10	0.53	**0.54**	0.52	**0.97**	0.86	0.76	**0.67**	0.64	0.60	**0.53**	0.46	0.51
		0.20	0.20	**0.59**	**0.59**	0.58	**0.96**	0.88	0.83	**0.72**	0.68	0.66	**0.58**	0.51	0.56

The binary group variable in the multivariable regression model (continuous age and binary group) using pseudo-value approach is compared with NetCoMi and MDiNE with 1000 replicates. A random network with network size $p = 20$ is generated at each simulation replicate. The best results are highlighted in boldface. The null case is when the effect sizes for each groups are equally adjusted

Table 2 The simulation results for the case when the network structure depends on age covariate

p	n	δ_1	δ_2	Precision			Recall			F1			Accuracy		
				SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE
40	20	0.05	0.05	**0.26**	0.25	0.23	0.64	0.27	**0.68**	**0.68**	0.26	0.31	**0.44**	0.07	0.37
		0.05	0.10	**0.34**	0.33	0.27	0.64	0.28	**0.68**	0.43	0.30	0.40	**0.46**	0.09	0.41
		0.05	0.20	**0.50**	0.47	0.46	**0.66**	0.24	0.64	0.55	0.31	**0.60**	**0.50**	0.12	0.49
		0.10	0.05	**0.35**	0.34	**0.35**	**0.65**	0.28	0.59	0.44	0.30	**0.48**	0.46	0.09	**0.48**
		0.10	0.10	**0.42**	0.40	0.35	**0.65**	0.27	0.38	**0.49**	0.32	0.35	0.48	0.11	**0.49**
		0.10	0.20	0.53	0.50	**0.54**	**0.67**	0.24	0.50	**0.58**	0.32	0.51	0.51	0.13	**0.52**
		0.20	0.05	**0.50**	0.47	0.47	**0.65**	0.23	0.58	0.55	0.31	**0.58**	0.50	0.12	**0.51**
		0.20	0.10	0.53	0.51	**0.57**	**0.66**	0.24	0.38	0.57	0.32	**0.60**	**0.51**	0.13	0.49
		0.20	0.20	0.58	0.57	**0.59**	0.69	0.22	**0.84**	0.62	0.32	**0.69**	0.53	0.13	**0.58**
	50	0.05	0.05	0.25	**0.26**	**0.26**	**0.83**	0.44	0.57	**0.38**	0.31	0.35	0.34	0.11	**0.48**
		0.05	0.10	**0.34**	**0.34**	0.33	**0.84**	0.45	0.67	**0.48**	0.37	0.42	0.40	0.15	**0.44**
		0.05	0.20	**0.50**	0.48	0.49	**0.84**	0.40	0.65	**0.62**	0.41	0.53	**0.50**	0.20	**0.50**
		0.10	0.05	0.34	0.34	**0.36**	**0.83**	0.44	0.64	**0.47**	0.36	0.42	0.39	0.15	**0.45**
		0.10	0.10	0.41	0.41	**0.44**	**0.84**	0.48	0.63	**0.54**	0.41	0.48	0.44	0.19	**0.49**
		0.10	0.20	**0.53**	0.52	**0.53**	**0.84**	0.43	0.69	**0.64**	0.44	0.55	**0.52**	0.23	0.51
		0.20	0.05	**0.49**	0.48	**0.49**	**0.84**	0.40	0.65	**0.61**	0.41	0.50	**0.50**	0.20	**0.50**
		0.20	0.10	0.53	0.51	**0.55**	**0.83**	0.41	0.52	**0.64**	0.43	0.45	**0.52**	0.22	0.50
		0.20	0.20	**0.58**	0.57	0.51	**0.84**	0.40	0.55	**0.68**	0.44	0.48	**0.56**	0.23	0.49
	200	0.05	0.05	0.25	**0.26**	0.25	**0.95**	0.65	0.92	**0.39**	0.36	**0.39**	0.28	0.16	**0.29**
		0.05	0.10	0.34	**0.35**	0.33	**0.95**	0.70	0.92	**0.50**	0.45	0.48	**0.36**	0.24	0.35
		0.05	0.20	**0.50**	**0.50**	0.49	**0.95**	0.64	0.93	**0.65**	0.53	0.63	**0.50**	0.32	0.49
		0.10	0.05	0.34	**0.35**	0.34	**0.95**	0.69	0.93	**0.50**	0.44	0.49	**0.36**	0.23	**0.36**
		0.10	0.10	0.41	**0.42**	**0.42**	**0.95**	0.72	**0.95**	0.57	0.51	**0.58**	0.42	0.30	**0.43**
		0.10	0.20	0.53	0.53	**0.55**	**0.95**	0.68	**0.95**	0.68	0.57	**0.69**	0.53	0.36	**0.54**
		0.20	0.05	**0.49**	**0.49**	**0.49**	**0.95**	0.63	**0.95**	**0.64**	0.52	**0.64**	**0.49**	0.31	**0.49**

Table 2 (continued)

p	n	δ_1	δ_2	Precision			Recall			F1			Accuracy		
				SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE
		0.20	0.10	0.52	**0.53**	**0.53**	**0.95**	0.67	0.94	**0.67**	0.57	**0.67**	**0.52**	0.35	**0.52**
		0.20	0.20	**0.58**	**0.58**	0.57	**0.95**	0.63	0.93	**0.71**	0.58	0.70	**0.57**	0.37	**0.57**
	500	0.05	0.05	**0.25**	**0.25**	0.24	0.98	0.78	**0.99**	**0.39**	0.37	0.38	**0.26**	0.20	0.24
		0.05	0.10	0.34	**0.35**	**0.35**	0.97	0.82	**0.99**	0.50	0.47	**0.51**	**0.35**	0.28	**0.35**
		0.05	0.20	0.49	**0.50**	**0.50**	**0.98**	0.80	**0.98**	**0.65**	0.59	**0.65**	0.49	0.39	**0.50**
		0.10	0.05	0.34	**0.35**	0.33	0.98	0.82	**0.99**	**0.50**	0.48	0.49	**0.35**	0.28	0.34
		0.10	0.10	**0.41**	**0.41**	0.40	**0.98**	0.85	**0.98**	**0.57**	0.54	**0.56**	**0.41**	0.35	0.40
		0.10	0.20	0.52	**0.53**	**0.53**	0.98	0.83	**1.00**	**0.68**	0.63	**0.68**	0.52	0.43	**0.53**
		0.20	0.05	0.50	0.50	**0.52**	0.98	0.79	**1.00**	0.65	0.59	**0.68**	0.50	0.39	**0.52**
		0.20	0.10	**0.52**	**0.52**	0.51	0.98	0.84	**0.99**	**0.67**	0.63	**0.67**	**0.52**	0.44	0.51
		0.20	0.20	**0.58**	0.57	**0.58**	0.98	0.81	**1.00**	0.72	0.65	**0.73**	0.57	0.47	**0.58**

The binary group variable in the multivariable regression model (continuous age and binary group) using pseudo-value approach is compared with NetCoMi and MDiNE with 1000 replicates. A random network with network size $p = 40$ is generated at each simulation replicate. The best results are highlighted in boldface. The null case is when the effect sizes for each groups are equally adjusted

Table 3 The simulation results for the case when the network structure does not depend on age covariate

p	n	δ	Precision			Recall			F1			Accuracy		
			SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE	SOHPIE	NetCoMi	MDiNE
20	20	0.05	**0.15**	0.14	0.00	**0.67**	0.15	0.00	0.26	**0.33**	0.00	0.39	0.02	**0.85**
		0.10	0.27	0.29	**0.42**	**0.68**	0.16	0.00	**0.38**	0.32	0.28	0.43	0.04	**0.73**
		0.20	**0.47**	0.46	0.25	**0.67**	0.16	0.01	**0.53**	0.31	0.24	0.49	0.07	**0.53**
	50	0.05	**0.14**	**0.14**	0.11	**0.82**	0.32	0.03	0.24	0.28	**0.42**	0.27	0.04	**0.83**
		0.10	**0.27**	**0.27**	0.26	**0.82**	0.33	0.04	**0.39**	0.33	0.31	0.35	0.09	**0.72**
		0.20	**0.46**	**0.46**	0.42	**0.81**	0.33	0.04	**0.58**	0.40	0.23	0.47	0.15	**0.53**
	200	0.05	**0.14**	**0.14**	0.13	**0.94**	0.38	0.36	0.24	**0.28**	0.27	0.19	0.05	**0.58**
		0.10	**0.27**	**0.27**	0.25	**0.93**	0.39	0.36	**0.41**	0.35	0.33	0.30	0.11	**0.54**
		0.20	**0.48**	**0.48**	0.44	**0.93**	0.40	0.36	**0.62**	0.43	0.40	0.48	0.19	**0.49**
	500	0.05	**0.14**	**0.14**	**0.14**	**0.97**	0.52	0.67	0.24	**0.26**	0.25	0.17	0.07	**0.38**
		0.10	0.27	**0.28**	0.26	**0.97**	0.54	0.65	**0.41**	0.37	0.37	0.28	0.14	**0.42**
		0.20	0.47	**0.48**	0.45	**0.96**	0.54	0.64	**0.62**	0.49	0.51	**0.47**	0.25	**0.47**
40	20	0.05	0.14	0.13	**0.20**	0.61	0.18	**0.66**	0.23	0.23	**0.29**	0.44	0.02	**0.48**
		0.10	**0.26**	**0.26**	**0.26**	0.60	0.19	**0.74**	0.35	0.24	**0.40**	**0.46**	0.05	0.38
		0.20	**0.46**	0.45	0.44	0.60	0.18	**0.79**	0.51	0.26	**0.60**	**0.49**	0.08	0.46
	50	0.05	**0.14**	**0.14**	0.13	**0.84**	0.25	0.73	**0.24**	0.22	0.23	0.27	0.04	**0.33**
		0.10	**0.26**	0.25	0.24	**0.82**	0.24	0.70	**0.39**	0.25	0.36	0.34	0.06	**0.39**
		0.20	**0.46**	**0.46**	0.42	**0.83**	0.25	0.66	**0.59**	0.32	0.50	**0.48**	0.11	0.47
	200	0.05	0.14	0.14	**0.15**	**0.94**	0.32	**0.94**	0.24	0.22	**0.25**	0.18	0.05	**0.19**
		0.10	0.26	**0.27**	**0.27**	**0.95**	0.33	0.91	**0.41**	0.30	**0.41**	0.29	0.09	**0.31**
		0.20	**0.47**	**0.47**	**0.47**	**0.95**	0.32	0.91	**0.62**	0.37	0.61	**0.47**	0.15	**0.47**
	500	0.05	0.14	0.14	**0.15**	0.97	0.41	**1.00**	0.24	0.22	**0.26**	**0.16**	0.06	**0.16**
		0.10	**0.27**	0.26	0.26	0.97	0.42	**0.99**	**0.41**	0.32	**0.41**	**0.28**	0.11	0.26
		0.20	**0.47**	**0.47**	0.46	0.98	0.42	**0.99**	**0.63**	0.43	0.62	**0.47**	0.20	0.46

The binary group variable in the univariable regression model (binary group only) using pseudo-value approach is compared with NetCoMi and MDiNE with 1000 replicates. A random network is generated at each simulation replicate. The best results are highlighted in boldface

additional covariates, compared with the two competing methods. In general, all metrics improve as n increases and/or when the larger effect size is provided ($\delta = 0.2$), as expected. It is worth noting that the MDiNE poses a practical challenge associated with substantially large computational time and costs. For instance, it requires more than 9 days to complete each simulation for $p = 40$ and $n = 200$ from the University of Florida Research Computing Linux server, HiPerGator 3.0 with 32CPU cores and 4GB of RAM per node, while it takes up to 18 h to execute the same simulation tasks for both the SOHPIE-DNA and NetCoMi with 4CPU cores and 6GB of RAM per node. See Additional file 1: Table S1 for more details.

Table 3 presents results of the univariable setting, where only the binary group variable is included in the model. In other words, only the effect of group was considered when generating random networks. On the whole, a similar pattern is shown in the univariable setting that the SOHPIE-DNA reaches a higher level of recall, compared with NetCoMi and MDiNE. Overall, our method resulted in a higher F1 score when the smaller network is considered. All of the methods suffer from a low precision with a small effect size ($\delta = 0.05$), but eventually improves with a larger effect size ($\delta = 0.2$).

Analysis of the American Gut Project Data

Six out of 138 taxa are found significantly DC between migraineurs vs. non-migraineurs while adjusting for age, sex, exercise frequency, categorical alcohol consumption, oral hygiene behavior, and dog ownership. At the family-level, the DC taxa are members of *Ruminococcaceae*, *Lachnospiraceae*, *Enterobacteriaceae*, *Erysipelotrichaceae*, and *Bacteroidaceae*. Of these families, the absence of *Lachnospiraceae* has been linked to the active or severe *Clostridium difficile* infection [35]. *Erysipelotrichaceae* has been associated with dyslipidemic phenotypes and systemic inflammation [36]. Moreover, a recent study [37] reported that the species enriched among migraineurs include *Ruminococcus gnavus* and *Lachnospiraceae bacterium*. The computational time for our analysis was about 12 h on the high-performance Linux cluster, HiPerGator 3.0 with 16CPU cores and 4GB of RAM per node.

Analysis of the diet exchange study data

Out of 112 taxa, 16 are predicted to be significantly DC between AA and RA after the two-week dietary exchange intervention while accounting for their age and BMI group. A complete list of DC taxa represent *Bacillus*, *Bacteroides uniformis et rel.*, *Bacteroides vulgatus et rel.*, *Clostridium ramosum et rel.*, *Coprococcus eutactus et rel.*, *Eggerthella lenta et rel.*, *Escherichia coli et rel.*, *Eubacterium hallii et rel.*, *Eubacterium siraeum et rel.*, *Faecalibacterium prausnitzii et rel.*, *Prevotella oralis et rel.*, *Roseburia intestinalis et rel.*, *Ruminococcus gnavus et rel.*, *Staphylococcus*, *Uncultured Bacteroidetes*, and *Xanthomonadaceae*. Notably, *Roseburia intestinalis* contributes to the prevention and management of intestinal inflammation and atherosclerosis [38]. *Eubacterium hallii* has been negatively associated with the fatigue severity scores of patients with advanced metastatic cancer [39]. The analysis took about an hour and 11 min on the HiPerGator 3.0 with 16CPU cores and 4GB of RAM per node.

Discussion

In this manuscript, we introduce the SOHPIE-DNA, a pseudo-value regression approach that determines whether a microbial taxa is significantly DC between groups after adjusting for additional covariates. This study is the first of its kind in the literature to develop a regression modeling for the DN analysis in microbiome data, which includes more than one predictor (e.g., group) in the model and predicts features of connectivity of a network. A simulation study shows that, at least for the scenarios considered, the SOHPIE-DNA generally maintains higher recall and F1-score while maintaining similar precision and accuracy, when compared with the most recent state-of-the-art methods: NetCoMi and MDiNE.

In this study, the group-specific jackknife pseudo-values are calculated. Another way of calculating jackknife pseudo-values is to use the entire sample and introduce the group-level indicator as a covariate into the model. However, in our preliminary simulations, we found that doing it that way led to worse performance.

Albeit not reported, we also looked at the familywise error rate (FWER), as defined to be the probability of at least one false positive and the values were fairly high in some cases. However, our simulation results shown in this paper, still reassure the utility of our proposed method since we generally are not expecting the complete null (where none of the edges to be DC) to hold and the FWER is a stringent measure as generally accepted by many statisticians. In our opinion, the reverse engineering methods such as ours should only a used as a screening tool and any positive discovery should be experimentally validated to alleviate such concerns. Incidentally, if FWER control is deemed to be very important for some situations, our tests could be combined with a Westfall-Young type procedure [40]. The detailed performance of such a modification could be studied elsewhere.

Another issue that we encountered was the incorporation of q-values, into our procedure. Since our individual tests are not independent, the q-values may not have the classical properties. Nevertheless, our tests seem to have reasonable FDR values as can be seen from the empirical results (Tables 1, 2 and 3).

We want to highlight that the SOHPIE-DNA is theoretically feasible to accommodate categorical biological groups, in lieu of binary biological groups. To the best of our knowledge, the use of binary groups has been commonly used for the DN analysis. Further, we have presented our simulation and real-data application studies with binary groups only.

We analyzed the data from two published studies to showcase the utility of the SOHPIE-DNA. Firstly, 6 taxa are found to be significantly DC between migraineurs and non-migraineurs while accounting for covariates using the data from the American Gut Project. A slight modification to the proposed method is grafted for analyzing the Diet Exchange Study data, where the group-specific difference of the estimated association matrices between two time points are used for the pseudo-value calculation. As a result, 16 significantly DC taxa are identified between AA and RA after the two-week diet exchange intervention with the inclusion of covariates. The real-world microbiome data often includes hundreds to thousands of taxa. We recommend that the users should (1) focus on a subset of taxa that are chosen based on experts with biological or clinical knowledge or (2) utilize our method at higher taxonomic levels (such as phylum level).

The latter application demonstrates the capability of assessing the temporal variation in connectivity measures. However, the SOHPIE-DNA currently has no feature to address the within-subject correlation for repeated measurements at different time points. This opens up an avenue for future investigation of longitudinal microbiome studies. One way of handling this is to use a generalized estimating equations (GEE) type approach for the pseudo-values and utilizing a jackknife estimate of the variance-covariance matrix of the pseudo-values at different time points.

Another line of future research direction to extend our work is to consider the idea of variable selection. This will help finding the best prediction model with a subset of phenotypic variables that are more biologically relevant across more heterogeneous study samples.

Additionally, we made an attempt of fitting a model under the generalized linear model for binary outcomes: logistic regression with or without the Firth's correction, in case of small sample size. It was challenging to appropriately dichotomize the matrices with jackknife pseudo-values. Further studies will be needed to devise an adaptive algorithm to find a threshold value that better classify the jackknife pseudo-values.

As a last remark, it should be emphasized that methods other than SparCC were also considered for network estimation, which includes the CCLasso [41] and SPIEC-EASI [42] with graphical lasso or neighborhood selection algorithms. However, these were not favorable in terms of runtime or due to not being able to run under certain simulation scenarios. For instance, the computational time to complete the re-estimation step for the SPIEC-EASI took more than 200 min for $p = 20$ with $n = 200$ for a single simulation replicate. The CCLasso could not estimate the association matrix with small sample size for a smaller network ($p = 20$ for $n = 20, 40, 60$).

Conclusions

There has been limited research to date that discusses how to adjust for additional covariate information in DN analysis for microbiome data. Herewith, we propose SOHPIE-DNA, a novel pseudo-value regression approach for the DN analysis, which can include additional clinical covariate in the model.

Methods

Compositional correlation-based methods for network estimation

The correlation is a useful proxy measure for identifying co-abundances or dependencies among taxa (or OTUs) in a microbial network. The Sparse Correlations for Compositional Data (SparCC) [43] estimates the pairwise correlations of the log-ratio transformed OTU abundances. Of note, a recent method, namely a Pseudo-value Regression Approach for Network Analysis (PRANA) [44], operates on gene expression data only, which therefore does not use a correlation measure that preserves the compositional profiling.

The co-abundance among taxa is described by a covariance matrix $T \in \mathbb{R}^{p \times p}$ where the non-diagonal elements t_{jk} are expressed by

$$\begin{aligned} t_{jk} \equiv & \mathrm{Var}\left(\log \frac{u_j}{u_k}\right) \\ = & \mathrm{Var}\left(\log u_j\right) + \mathrm{Var}\left(\log u_k\right) - 2\mathrm{Cov}\left(\log u_j, \log u_k\right) \\ = & \sigma_j^2 + \sigma_k^2 - 2\rho_{jk}\sigma_j\sigma_k, \end{aligned} \tag{1}$$

where u_j and u_k are the fraction of OTU abundances, σ_j^2 and σ_k^2 are the variances of the log-transformed abundances, and ρ_{jk} is the correlation of taxa j and k, respectively. Moreover, the variance t_{jj} is approximated by

$$t_{jj} \cong (p-1)\sigma_j^2 + \sum_{k \neq j} \sigma_k^2, \tag{2}$$

where $j, k \in \{1, \ldots p\}$. Then the correlation can be estimated by solving Eqs. 1 and 2:

$$\hat{\rho}_{jk} = \frac{\hat{\sigma}_j^2 + \hat{\sigma}_k^2 - \hat{t}_{jk}}{2\hat{\sigma}_j\hat{\sigma}_k}, \tag{3}$$

where $\hat{\sigma}_j$, $\hat{\sigma}_k$, and $\hat{t}_{jk}$ are the sample estimates of σ_j, σ_k, and t_{jk}, respectively.

Furthermore, SparCC takes an iterative approach under the assumption ("sparsity of correlations" as in the original paper) that a small number of strong correlations exists in a true network, which hinders the detection of spurious correlations among taxa.

Besides SparCC, we have attempted to use other compositional correlation measures for our differential network analysis. See the "Discussion" section for further details.

Pseudo-value approach

Consider undirected network estimated from n individuals. It can then be represented by the $p \times p$ association matrix that encodes the pairwise correlations $\hat{\rho}_{jk}$ between a pair of taxa $j, k \in \{1, \ldots, p\}$. The association matrix is symmetric ($\hat{\rho}_{jk} = \hat{\rho}_{kj}$) where the non-diagonal entries are either non-zero (i.e., some association between two taxa) or zero (i.e., no association between two taxa). The diagonal entries are all equal to one, because the network is assumed that there is no self-loop (i.e., a node cannot redirect to itself).

The network centrality has been studied to measure the extent of biological or topological importance that a node has in a network [45, 46]. For each taxa k, the network centrality is calculated as the marginal sum of the association matrix.

$$\hat{\theta}_k = \sum_{j=1}^{p} \hat{\rho}_{jk},$$

where $k = 1, \ldots, p$.

The jackknife pseudo-values [17] for the ith individual and k^{th} taxon are defined by:

$$\tilde{\theta}_{ik} = n\hat{\theta}_k - (n-1)\hat{\theta}_{k(i)}, \tag{4}$$

where $\hat{\theta}_{k(i)}$ is the marginal sum of a taxon calculated based on the re-estimated association matrix using the microbiome data eliminating the ith subject.

The computational cost of the re-estimation process is dependent on the sample size, as for each taxa k requires n such calculations with the data size of $n-1$. A solution to speed up the processing time is the use of parallel computing such as mclapply function in *parallel* R package.

Let $Z \in \{1,2\}$ be a binary group indicator and denote $\mathcal{G}_1 = \{i : Z_i = 1\}$ and $\mathcal{G}_2 = \{i : Z_i = 2\}$. Each group has the same set of p taxa, but group-specific sample size $n_z = |\mathcal{G}_z|$ for the two groups $z = 1,2$. Total sample size is $n = \sum_z n_z$. The Eq. 4 is used to calculate the group-specific jackknife pseudo-values. That is, for taxon k and group z, we define $\hat{\theta}_k^z$ and $\hat{\theta}_{k(i)}^z$, where $i = 1,\ldots,n_z$. Then for each $i \in \mathcal{G}_z$, the kth taxon jackknife pseudo-values are calculated from $\tilde{\theta}_{ik} = n_z\hat{\theta}_k^z - (n_z-1)\hat{\theta}_{k(i)}^z$.

Let $\mathbf{X} = (X_1,\ldots,X_q)$ denote q vector of covariates, such as age at diagnosis, current smoking status, and etc. The pseudo-value regression model for the ith individual and k^{th} taxon is

$$\mu_i = E[\tilde{\theta}_{ik} \mid Z_i, \mathbf{X}_i] = \alpha_k + \beta_k Z_i + \sum_{m=1}^{q} \gamma_{km} X_{im}, \tag{5}$$

where μ_i is the k-dimensional mean vector of pseudo-value $\tilde{\theta}_{ik}$ for the ith individual, α_k is the intercept, β_k is the regression coefficient for Z, and $\gamma_{k1},\ldots,\gamma_{kq}$ is the set of regression coefficients to be estimated for $\mathbf{X}$. In our setting, the main parameter of interest is given by β_k, the change in network centrality measure of the kth taxon between two groups.

The least trimmed squares (LTS), also known as least trimmed sum of squares [47], is then implemented to carry out a robust regression. The main advantages of the LTS estimator over other robust estimators including the M-estimator and least median of squares estimator are its computational efficiency and robustness to outliers in both the response and predictor variables [48, 49].

The LTS estimator is defined by

$$\min_{\alpha_k,\beta_k,\gamma_{k1},\ldots,\gamma_{kq}} \sum_{i=1}^{h} r_{(i)}(\alpha_k,\beta_k,\gamma_{k1},\ldots,\gamma_{kq})^2,$$

where $r_{(i)}$ is the set of ordered absolute values of the residuals sorted in increasing order of absolute value and h may depend on a pre-determined trimming proportion $c \in [0.5,1]$ [50]. For example, one can take $h = [n(1-c)] + 1$.

Hypothesis testing

We construct the null hypothesis of $H_0 : \beta_k = 0$ against the research hypothesis $H_1 : \beta_k \neq 0$ to test if there is a true difference between groups in the network centrality measure of the kth taxon. The t-statistic is defined by $U_k = \hat{\beta}_k / SE(\hat{\beta}_k)$ for $k = 1,\ldots,p$, where $\hat{\beta}_k$ is the least-squares estimator from the robust regression described in the above Eq. 5 and $SE(\hat{\beta}_k)$ is the standard error of $\hat{\beta}_k$. As far as the decision-making process, the asymptotically α-level test rejects H_0 if $|U_k| > t_{\alpha/2}$. p values are calculated using a t-distribution as in *robustbase* R package [51, 52].

Multiple hypothesis testing is a common feature in the DN analysis, and therefore it is crucial to appropriately control the false discovery rate (FDR). The FDR measures the proportion of false discoveries incurred among a set of DC taxa from the test. Most

classically, the concept of FDR was pioneered by Benjamini and Hochberg [53], shown to achieve the FDR control, whilst maintaining the adequate statistical power [54]. However, the q-value [55] offers a less conservative FDR estimation over the conventional Benjamini-Hochberg procedure [56]. The q-value is estimated from the empirical distribution of the observed p values, and keeps the balance between true positives and false positives [57]. Accordingly, the q-value is applied to adjust for the multiplicity control in the present paper using *fdrtool* R package.

Algorithm

The SOHPIE-DNA algorithm is described below in Algorithm 1.

Algorithm 1 SOHPIE-DNA

Require: $n_z \times p$ OTU table and metadata for each group $z = 1, 2$.
Ensure: The set of p-values of the group variable for each taxa k.

1: Estimate $p \times p$ association matrix with SparCC from the $n_z \times p$ OTU table for each group $z = 1, 2$. See the Simulated Data in Materials subsection for the data generation.
2: Calculate the group-specific marginal sums of association matrix of each taxa $k \in \{1, \cdots, p\}$, denoted by $\hat{\theta}_k^z$.
3: Calculate $\hat{\theta}_{k(i)}^z$ for each taxa k and individual $i \in \mathcal{G}_z$ from the association matrix that is re-estimated from the OTU table without the ith individual of $n_z \times p$ data.
4: Calculate the jackknife pseudo-value $\tilde{\theta}_{ik}$ using equation 4.
5: Fit a robust regression for each taxa k to obtain the p-values of the group variable, computed from the t-test.
   (i) Multivariable: A binary group variable Z and a continuous covariate X are included in the model.
   (ii) Univariable: A binary group Z is only included in the model.
6: The q-values are calculated based on the observed p-values for the multiplicity control. This will be used to compute the performance measures of the Monte Carlo simulation (see the next section under Performance Evaluations).

Performance evaluations

Construction of adjacency matrices

Generate the scale-free random network (or Barabási-Albert network) [58] with p nodes using the *igraph* R package [59]. A network is scale-free if its degree distribution follows a power-law distribution. In other words, a small portion of "hub" nodes has the highest degree centrality, while most nodes have lower degree centrality.

The two identical $p \times p$ adjacency matrices, where the diagonal entries are 0 and non-diagonal entries are either $\{0, 1\}$, are obtained from this random network. At the end of the data generation phase using *SparseDOSSA2* in Simulated Data in Materials subsection, we are able to identify which taxa are spike-in associated with the covariate for each $z = 1, 2$. In order to distinguish networks representative of $z = 1$ (e.g., healthy control) from that of $z = 2$ (e.g., disease group), we keep track of the indices of these covariate-dependent taxa. Each index with a value of 1 indicates that the corresponding covariate-dependent taxon is connected with at least one of the neighboring taxa. We use these indices to perturb the random networks by changing a value from 1 to 0 (i.e.

Fig. 1 Network plots visualizing the microbial network ($p = 20$) with a covariate dependence structure that depends on continuous age and binary group information ($\delta_1 = 0.05$ (left), $\delta_2 = 0.2$ (right)). This represents the multivariable setting

Fig. 2 Network plots visualizing the microbial network ($p = 20$) without a covariate dependence structure that depends on binary group only ($\delta_1 = 0$ (left), $\delta_2 = 0.2$ (right)). This represents the univariable setting

synthetically removing all the connected edges) around the covariate-dependent taxa for each group. This perturbation is further explained and borrowed from the recent paper [60]. The network plots are provided to visually demonstrate the perturbed adjacency matrices (see Figs. 1 and 2). The figures represent two single networks for two particular realizations corresponding to covariate profiles. The effect sizes (i.e. pre-specified proportion of taxa that are associated with the covariate; denoted as δ_1 and δ_2) control the amount of perturbation. If the effect sizes are different ($\delta_1 \neq \delta_2$), then the covariates are affecting the networks differentially (Fig. 2). See Simulated Data in Materials subsection for further details.

Performance measures

Four performance metrics are adopted to evaluate our proposed method: precision, recall, F1-score, and accuracy. Let $\Omega^z \in \mathbb{R}^{p \times p}$ be the group-specific adjacency matrix, where

$$\Omega^z_{jk} = \begin{cases} 1 & \text{if the two nodes } j \text{ and } k \text{ are connected} \\ 0 & \text{otherwise,} \end{cases}$$

for $z = 1, 2$. Next, a node-specific true connection is calculated

$$\eta_k = I\left(\sum_{j=1}^{p} |\Omega_{jk}^1 - \Omega_{jk}^2| > 0\right),$$

indicating that taxa k has differential connectivity (DC).

In terms of notation, we use q_{ks} to denote a q-value [55] of taxa k at the simulation replicate s. An error rate control of $\alpha = 0.05$ is used throughout the simulation. In the following, we present the details of each performance metric.

Precision is the fraction of taxa which are declared to be significantly DC from the test that are confirmed as true:

$$\text{Precision} = \frac{\sum_{k=1}^{p} \eta_k \, I(q_{ks} < \alpha)}{\sum_{k=1}^{p} I(q_{ks} < \alpha)}.$$

Recall is the fraction of truly DC taxa which are correctly declared to be significant between two comparing groups from the test:

$$\text{Recall} = \frac{\sum_{k=1}^{p} \eta_k \, I(q_{ks} < \alpha)}{\sum_{k=1}^{p} \eta_k}.$$

The F1 score is the harmonic mean of precision and recall values. A higher F1 score indicates a better overall performance with lower false negative and false positive predictions:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Accuracy is defined as the fraction of total number of taxa that are correctly predicted to be DC. The accuracy ranges from 0 (no correct predictions) to 1 (perfect predictions):

$$\text{Accuracy} = \frac{\sum_{k=1}^{p} I(\eta_k = 1)\, I(q_{ks} < \alpha) + \sum_{k=1}^{p} I(\eta_k = 0)\, I(q_{ks} \geq \alpha)}{\sum_{k=1}^{p} I(\eta_k = 1) + \sum_{k=1}^{p} I(\eta_k = 0)}.$$

Materials

Simulated data

The synthetic microbiome dataset are structured with p taxa and n sample size. In the simulation, binary group indicators 1 and 2 are generated from a Bernoulli distribution with equal probabilities and a single continuous covariate $X \sim N(55, 10)$ (e.g., age at diagnosis). We test our proposed method on datasets under two different simulation scenarios: taxa are impacted by the effect of (1) Z and X or (2) Z only, which each corresponds to "multivariable" and "univariable" settings, respectively.

The actual microbial data generation (e.g., OTU counts) given the covariates is described next. In this context, it is perhaps worth mentioning that this part is completely different from generating gene expression data as in PRANA [44]. For each simulation scenario, we generate an OTU table that resembles the dependence structure of covariates Z and/or X on the microbial community (or the network) using the *SparseDOSSA2* (Sparse Data Observations for the Simulation of Synthetic Abundances)

R package [61]. *SparseDOSSA2* adopts a Bayesian Gaussian copula model with zero-inflated, truncated log-normal distributions to capture the marginal distributions of each microbial taxa and to account for the correlation between taxa.

The package has a feature to indicate a user-specified percentage of taxa to be "spiked-in" association with the clinical information (or metadata). This is referred to as the "effect size" of differential abundance δ. To evaluate the effect size of Z under the univariable setting, we generate the data that half of the samples have taxa with no spike-in association, whereas the other half of the samples have spike-in association on 5%, 10%, or 20% of taxa. The distributions of age in the two groups are different. Therefore, under the multivariable setting, 5%, 10%, or 20% of taxa have spike-in association with X for each group $z = 1, 2$. In both scenarios, $n_z \times p$ matrices for each group $z = 1, 2$ will be available for use.

Application study

The American Gut Project Data A pre-processed OTU table of the human stool microbiome samples from the American Gut Project [33] is available in the *SpiecEasi* R package, along with the corresponding metadata information. The gut microbiome is involved with the bidirectional relationship between the gastrointestinal system and central nervous system (i.e. gut-brain axis) that impacts on the migraine inflammation [62].

In the analysis, the main variable of interest is a binary variable indicating the migraine headache (yes or no). Age [63], sex [63], exercise frequency ($\geq$ 3 days per week or otherwise) [64], and categorical alcohol consumption (heavy, moderate, or non-drinking) [65] are covariates that are included in the multivariable model. Additionally, migraine has been associated with the periodontal inflammation [66] and pet ownership [67], and therefore the oral hygiene behavior such as dental floss frequency ($\geq$ 3 times per week or otherwise) and living with a dog (yes or no) were included in the model.

The initial OTU table consists of 138 taxa with 296 subjects. No taxa were removed, however, 28 subjects were excluded due to unidentified sampling body site and missing age or sex information. Hence, 138 taxa and 268 subjects were used for the analysis.

The Diet Exchange Study Data A pre-processed data of the geographical epidemiology study [34] is available in *microbiome* [68] R package. The aim of the study was to assess the effect of fat and fiber intake of the diet on the composition of the colonic microbiota by switching the diet in study populations with high (African-Americans from Pittsburgh area of Pennsylvania; AA) and low (rural South Africans from KwaZulu region; RA) colon cancer risk for two weeks.

An initial OTU table contains 130 taxa with 38 subjects. After the exclusion of a subject with missing post-dietary intervention data and 18 rare taxa that appear in fewer than 10% of the samples, 112 taxa with 37 subjects (20 AA and 17 RA) are used for the analysis.

The main predictor variable is binary geographic location (AA or RA). Additional covariates considered in a multivariable model were sex and BMI groups (obese, overweight, or lean).

For each groups separately, we take the difference of the estimated association matrices (as well as the re-estimated association matrices) between two time points, that is,

the endoscopy before and after two weeks of dietary change. The differences are then used to calculate the jackknife pseudo-values as in the previous sections. This additional step is intended to incorporate the temporal change of connectivity of each taxa after dietary interventions.

Abbreviations

AA	African-Americans from Pittsburgh area of Pennsylvania
RA	Rural South Africans from KwaZulu region
IBD	Inflammatory bowel disease
OTU	Operational taxonomic units
DC	Differentially connected
DE	Differential expression
DN	Differential network
MDiNE	Microbiome differential network estimation
NetCoMi	Network construction and comparison for microbiome data
SOHPIE-DNA	Statistical approach via pseudo-value information and estimation for differential network analysis
SparCC	Sparse correlations for compositional data
SparseDOSSA2	Sparse data observations for the simulation of synthetic abundances
CCLasso	Correlation inference for compositional data through lasso
SPIEC-EASI	Sparse inverse covariance estimation for ecological association inference
LASSO	Least absolute shrinkage and selection operator
LTS	Least trimmed squares

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05689-7.

Additional file 1. Comparison of computational time of SOHPIE-DNA with that of NetCoMi and MDiNE.

Acknowledgements
The authors are exceptionally thankful for the investigators involved with the American Gut Project and the Diet Exchange Study for sharing their data publicly.

Author contributions
SD conceived the use of pseudo-value in the study. SA developed the methodology, performed simulations and data analyses of the study. SA drafted the manuscript and SD provided suggestions when writing the manuscript. All authors have reviewed and edited the manuscript.

Funding
SA was supported by the National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health under grants number [NIH T32AA025877] during his time at the University of Florida for his doctoral research. Research reported in this publication was supported in part by the National Cancer Institute Cancer Center Support Grant [NIH P30CA196521-01] awarded to the Tisch Cancer Institute of the Icahn School of Medicine at Mount Sinai and used the Biostatistics Shared Resource Facility. The content is solely the responsibility of S.A. and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials
The original study data of the American Gut Project and the Diet Exchange Study are available in SpiecEasi (https://github.com/zdk123/SpiecEasi) and microbiome (https://bioconductor.org/packages/release/bioc/html/microbiome.html) R packages, respectively. The ***SOHPIE*** [69] R package can be downloaded from the CRAN repository https://CRAN.R-project.org/package=SOHPIE. The source code for SOHPIE-DNA is available at (https://github.com/sjahnn/SOHPIE-DNA). Please reach out to the authors (Seungjun Ahn, seungjun.ahn@mountsinai.org; Somnath Datta, somnath.datta@ufl.edu) if you have any further inquiries on the data or code.

Declarations

Ethical approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests. S.D. is a member of editorial board (Associate Editor), but this did not influence the scientific peer-review process nor did he discuss the submission with other editorial members of the journal.

Received: 6 May 2023 Accepted: 2 February 2024
Published online: 18 March 2024

References

1. Weinstock G. Genomic approaches to studying the human microbiota. Nature. 2012;489:250–6.
2. Bhatt AP, Redinbo MR, Bultman SJ. The role of the microbiome in cancer development and therapy. CA Cancer J Clin. 2017;67:326–44.
3. Vujkovic-Cvijin I, Sortino O, Verheij E, Sklar J, Wit FW, Kootstra NA, et al. HIV-associated gut dysbiosis is independent of sexual practice and correlates with noncommunicable diseases. Nat Commun. 2020;11:2448.
4. Glassner KL, Abraham BP, Quigley E. The microbiome and inflammatory bowel disease. J Allergy Clin Immunol. 2020;145:16–27.
5. Lee SH, Yoon SH, Jung Y, Kim N, Min U, Chun J, et al. Emotional well-being and gut microbiome profiles by enterotype. Sci Rep. 2020;10:20736.
6. Valles-Colomer M, Falony G, Darzi Y, Tigchelaar EF, Wang J, Tito RY, et al. The neuroactive potential of the human gut microbiota in quality of life and depression. Nat Microbiol. 2019;4:623–32.
7. Krajmalnik-Brown R, Lozupone C, Kang DW, Adams JB. Gut bacteria in children with autism spectrum disorders: challenges and promise of studying how a complex community influences a complex disease. Microb Ecol Health Dis. 2015;26:26914.
8. Mayer EA, Knight R, Mazmanian SK, Cryan JF, Tillisch K. Gut microbes and the brain: paradigm shift in neuroscience. J Neurosci. 2014;34:
9. Durazzi F, Sala C, Castellani G, Manfreda G, Remondini D, De Cesare A. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. Sci Rep. 2021;11:3030.
10. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019;10:5029.
11. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58:586–97.
12. Layeghifard M, Hwang DM, Guttman DS. Disentangling interactions in the microbiome: A network perspective. Trends Microbiol. 2017;25:217–28.
13. Matchado MS, Lauber M, Reitmeier S, Kacprowski T, Baumbach J, Haller D, et al. Network analysis methods for studying microbial communities: A mini review. Comput Struct Biotechnol J. 2021;9:2687–98.
14. McGregor K, Labbe A, Greenwood C. MDiNE: a model to estimate differential co-occurrence networks in microbiome studies. Bioinformatics. 2020;36:1840–7.
15. Peschel S, Muller C, von Mutius E, Boulesteix A, Depner M. NetCoMi: network construction and comparison for microbiome data in R. Brief Bioinform. 2021;22:290.
16. Lee M, Chang E. Inflammatory bowel diseases (IBD) and the microbiome-searching the crime scene for clues. Gastroenterology. 2021;160:524–37.
17. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Philadelphia: Chapman & Hall/CRC; 1993.
18. Andersen PK, Klein JP, Rosthøj S. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika. 2003;90:15–27.
19. Andersen P, Klein J. Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies. Scand J Statist. 2007;34:3–16.
20. Sabathé C, Andersen PK, Helmer C, Gerds TA, Jacqmin-Gadda H, Joly P. Regression analysis in an illness-death model with interval-censored data: A pseudo-value approach. Stat Methods Med Res. 2020;29:752–64.
21. Johansen MN, Lundbye-Christensen S, Larsen JM, Parner ET. Regression models for interval censored data using parametric pseudo-observations. BMC Med Res Methodol. 2021;21:36.
22. Logan BR, Zhang MJ, Klein JP. Marginal models for clustered time-to-event data with competing risks using pseudovalues. Biometrics. 2011;67:1–7.
23. Ahn KW, Logan BR. Pseudo-value approach for conditional quantile residual lifetime analysis for clustered survival and competing risks data with applications to bone marrow transplant data. Ann Appl Stat. 2016;10:618–37.
24. Zhao L, Feng D. Deep neural networks for survival analysis using pseudo values. IEEE J Biomed Health Inform. 2020;24:3308–14.
25. Ginestet PG, Gabriel EE, Sachs MC. Survival stacking with multiple data types using pseudo-observation-based-AUC loss. J Biopharm Stat. 2022. https://doi.org/10.1080/10543406.2022.2041655.
26. Logan BR, Klein JP, Zhang MJ. Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. Biometrics. 2008;64:733–40.
27. Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. Lifetime Data Anal. 2009;15:241–55.
28. Overgaard M, Parner ET, Pedersen J. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. Ann Stat. 2017;45:1988–2015.
29. Klein JP, Gerster M, Andersen PK, Tarima S, Perme MP. SAS and R functions to compute pseudo-values for censored data regression. Comput Methods Programs Biomed. 2008;89:289–300.
30. Wang Y, Logan B. Testing for center effects on survival and competing risks outcomes using pseudo-value regression. Lifetime Data Anal. 2019;25:206–28.
31. Ahn K, Mendolia F. Pseudo-value approach for comparing survival medians for dependent data. Stat Med. 2014;33:1531–8.
32. Zhao S, Shojaie A. Network differential connectivity analysis. Ann Appl Stat. 2022;16:2166–82.
33. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al. American gut: an open platform for citizen science microbiome research. mSystems. 2018;3:00031–18.

34. O'Keefe SJ, Li JV, Lahti L, Ou J, Carbonero F, Mohammed K, et al. Fat, fibre and cancer risk in african americans and rural africans. Nat Commun. 2015;6:6342.
35. Taur Y, Pamer E. Harnessing microbiota to kill a pathogen: Fixing the microbiota to treat clostridium difficile infections. Nat Med. 2014;20:246–7.
36. Nolan-Kenney R, Wu F, Hu J, Yang L, Kelly D, Li H, et al. The association between smoking and gut microbiome in bangladesh. Nicotine Tob Res. 2020;22:1339–46.
37. Chen J, Wang Q, Wang A, Lin Z. Structural and functional characterization of the gut microbiota in elderly women with migraine. Front Cell Infect Microbiol. 2020;9:470.
38. Nie K, Ma K, Luo W, Shen Z, Yang Z, Xiao M, et al. Roseburia intestinalis: A beneficial gut organism from the discoveries in genus and species. Front Cell Infect Microbiol. 2021;11: 757718.
39. Hajjar J, Mendoza T, Zhang L, Fu S, Piha-Paul SA, Hong DS, et al. Associations between the gut microbiome and fatigue in cancer patients. Sci Rep. 2021;11:5847.
40. Westfall P, Young SS. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. New York: Wiley; 1993.
41. Fang H, Huang C, Zhao H, Deng M. CCLasso: correlation inference for compositional data through lasso. Bioinformatics. 2015;31:3172–80.
42. Kurtz ZD, Muller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol. 2015;11:1004226.
43. Friedman J, Alm E. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8:1002687.
44. Ahn S, Grimes T, Datta S. A pseudo-value regression approach for differential network analysis of co-expression data. BMC Bioinformatics. 2023;24:8.
45. Ashtiani M, Salehzadeh-Yazdi A, Razaghi-Moghadam Z, Hennig H, Wolkenhauer O, Mirzaie M, et al. A systematic survey of centrality measures for protein-protein interaction networks. BMC Syst Biol. 2018;12:80.
46. Ozgür A, Vu T, Erkan G, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics. 2008;24:277–85.
47. Rousseeuw P. Least median of squares regression. J Am Stat Assoc. 1984;79:871–80.
48. Ahdesmäki M, Lähdesmäki H, Gracey A, Shmulevich L, Yli-Harja O. Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data. BMC Bioinformatics. 2007;8:233.
49. Alfons A, Croux C, Gelper S. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. Ann Appl Stat. 2013;7:226–48.
50. Pison G, Van Aelst S, Willems G. Small sample corrections for LTS and MCD. Metrika. 2002;55:111–23.
51. Todorov V, Filzmoser P. An object-oriented framework for robust multivariate analysis. J Stat Soft. 2009;32:1–47.
52. Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, et al.: Robustbase: Basic Robust Statistics. (2022). R package version 0.95-0. http://robustbase.r-forge.r-project.org/
53. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Statist Soc B. 1995;57:289–300.
54. Benjamini Y. Discovering the false discovery rate. J R Statist Soc B. 2010;72:405–16.
55. Storey J. A direct approach to false discovery rates. J R Statist Soc B. 2002;64:479–98.
56. Strimmer K. A unified approach to false discovery rate estimation. BMC Bioinformatics. 2008;9:303.
57. Storey J, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100:9440–5.
58. Barabási A, Albert R. Emergence of scaling in random networks. Science. 1999;286:509–12.
59. Csárdi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems. 2006;1695:16–27.
60. Grimes T, Datta S. SeqNet: An R Package for Generating Gene-Gene Networks and Simulating RNA-Seq Data. J Stat Softw. 2021;98:10–1863709812.
61. Ma S, Ren B, Mallick H, Moon YS, Schwager E, Maharjan S, et al. A statistical model for describing and simulating microbial community profiles. PLoS Comput Biol. 2021;17:1008913.
62. Arzani M, Jahromi SR, Ghorbani Z, Vahabizad F, Martelletti P, Ghaemi A, et al. Gut-brain axis and migraine headache: a comprehensive review. J Headache Pain. 2020;21:1.
63. Stewart WF, Linet MS, Celentano DD, Van Natta M, Ziegler D, et al. Age- and sex-specific incidence rates of migraine with and without visual aura. Am J Epidemiol. 1991;134:1111–20.
64. Amin FM, Aristeidou S, Baraldi C, Czapinska-Ciepiela EK, Ariadni DD, Di Lenola D, et al. The association between migraine and physical exercise. J Headache Pain. 2018;19:83.
65. Mostofsky E, Bertisch SM, Vgontzas A, Buettner C, Li W, Rueschman M, et al. Prospective cohort study of daily alcoholic beverage intake as a potential trigger of headaches among adults with episodic migraine. Ann Med. 2020;52:386–92.
66. Leira Y, Ameijeira P, Domínguez C, López-Arias E, Ávila-Gómez P, Pérez-Mato M, et al. Periodontal inflammation is related to increased serum calcitonin gene-related peptide levels in patients with chronic migraine. J Periodontol. 2019;90:1088–95.
67. Koivusilta L, Ojanlatva A. To have or not to have a pet for better health? PLoS One. 2006;1:109.
68. Lahti L, Shetty S. Microbiome R Package. (2017). Bioconductor. https://doi.org/10.18129/B9.bioc.microbiome
69. Ahn S, Datta S. SOHPIE: statistical approach via pseudo-value information and estimation for differential network analysis of microbiome data. Bioinformatics. 2024;40(1):btad766. https://doi.org/10.1093/bioinformatics/btad766

Publisher's Note