A randomization-based causal inference framework for uncovering environmental exposure effects on human gut microbiota

Alice J Sommer[1,2,3,*], Annette Peters[2,3,4,*], Martina Rommel[3,5], Josef Cyrys[3], Harald Grallert[5,6], Dirk Haller[7,8], Christian L Müller[9,10,11,*], and Marie-Abèle C Bind[1,12]

[1]Department of Statistics, Harvard University, Cambridge, MA, USA
[2]Institute for Medical Information Processing, Biometry, and Epidemiology, Faculty of Medicine, Ludwig-Maximilians-University München, Munich, Germany
[3]Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany
[4]Department of Environmental Health, Harvard T. H. Chan School of Public Health, Boston, MA, USA
[5]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany
[6]German Center for Diabetes Research (DZD), München-Neuherberg, Germany
[7]ZIEL - Institute for Food & Health, Technical University of Munich, Freising, Germany
[8]Chair of Nutrition and Immunology, Technical University of Munich, Freising, Germany
[9]Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany
[10]Department of Statistics, Ludwig-Maximilians-University München, Munich, Germany
[11]Center for Computational Mathematics, Flatiron Institute, New York, NY, USA
[12]Biostatistics Center, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[*]Corresponding authors: Alice J. Sommer: alice.sommer@helmholtz-muenchen.de, Annette Peters: peters@helmholtz-muenchen.de, and Chrisitan L. Müller: cmueller@flatironinstitute.org

March 2021

Abstract

Statistical analysis of microbial genomic data within epidemiological cohort studies holds the promise to assess the influence of environmental exposures on both the host and the host-associated microbiome. The observational character of prospective cohort data and the intricate characteristics of microbiome data make it, however, challenging to discover causal associations between environment and microbiome. Here, we introduce a causal inference framework based on the Rubin Causal Model that can help scientists to investigate such environment-host microbiome relationships, to capitalize on existing, possibly powerful, test statistics, and test plausible sharp null hypotheses. Using data from the German KORA cohort study, we illustrate our framework by designing two hypothetical randomized experiments with interventions of (i) air pollution reduction and (ii) smoking prevention. We study the effects of these interventions on the human gut microbiome by testing shifts in microbial diversity, changes in individual microbial abundances, and microbial network wiring between groups of matched subjects via randomization-based inference. In the smoking prevention scenario, we identify a small interconnected group of taxa worth further scrutiny, including Christensenellaceae and Ruminococcaceae genera, that have been previously associated with blood metabolite changes. These findings demonstrate that our framework may uncover potentially causal links between environmental exposure and the gut microbiome from observational data. We anticipate the present statistical framework to be a good starting point for further discoveries on the role of the gut microbiome in environmental health.

Introduction

The human microbiome plays a pivotal role in maintaining a healthy physiology via multiple interactions with the host. The gut microbiome, for instance, provides important metabolic capabilities for food digestion [1, 2] and regulates immune homeostasis [3]. Although dietary interventions [4], pathogen infections [5], and antibiotics use [6] can trigger rapid changes of gut microbial compositions and can lead to dysbiotic disruptions of host-microbiome interactions, the long-term impact of environmental exposures on the human gut microbiome remains poorly understood. In this paper, we provide a causal inference framework for assessing such epidemiological questions and analyze a prospective cohort with collected microbiome data. Recent technological advances, through culture-independent analyses, have facilitated a surge in observational studies of the human microbiome [7–9]. A common method to catalog microbial constituents is high-throughput amplicon sequencing [10], allowing the acquisition of gut microbiome survey data for large prospective cohort studies. Important examples include the Human Microbiome Project [11], the British TwinsUK study [12], the Dutch LifeLines-DEEP [13] and Rotterdam Studies [14], the Chinese Guangdong Gut Microbiome Project [15], the American Gut Project [16], and the German KORA study [17].

Thus far, these and other studies have linked alterations in gut microbial compositions to several common diseases, including rheumatoid arthritis, colorectal cancer, obesity, inflammatory bowel disease (IBD), and diabetes [18]. Although environmental exposures such as particulate matter (PM) [19] and smoking [20] are also related to these diseases, an understanding of environment-gut microbiome relationships and their implications for disease mechanisms has remained elusive. Here, we examine such environment-gut microbiome relationships within a causal inference framework [21] combined with state-of-the-art statistical methods for amplicon sequence variant (ASV) data [22]. We illustrate our analysis framework using data from the German KORA study [17] and focus on two inhaled environmental exposures previously hypothesized to be linked with gastrointestinal diseases and the gut microbiome: (i) particulate matter (PM) with diameter smaller or equal to 2.5 μm ($PM_{2.5}$) and (ii) cigarette smoking.

Air pollution exposure has been found to be associated with gastrointestinal diseases, such as appendicitis [23], inflammatory bowel disease [24], abdominal pain [25], and metabolic disorders [26]. Current research suggests that air pollution may impact the gut microbiome which, in turn, acts as a "mediator" of the association between air pollution and metabolic disorders such as obesity and type 2 diabetes [27–29]. These studies found associations between nitric oxide, nitrogen dioxyde [27], PM [28], and ozone [30] exposures and the gut microbiome. Several potential pathways explain how particles affect human health. The gut is exposed to PM through mucociliary clearance, i.e., the self-cleaning mechanism of the bronchi, inducing inhaled PM to be cleared from the lungs to the gut, and oral route exposure, when food and water is contaminated by PM [31, 32]. Results from murine studies of the effect of PM on the gut [33–37] suggest that exposure to PM changes the microbial composition and increases gut permeability, leading to higher systemic inflammation due to the unrestrained influx of microbial products from the gut into the systemic circulation [38].

Cigarette smoking, on the other hand, is an example of an inhaled exposure that has been shown to influence the susceptibility of diseases such as IBD, colorectal cancer, and systemic diseases [20, 39, 40]. Animal studies suggest that cigarette smoke may mediate its effects through alterations of intestinal microbiota [41]. In humans, shifts in the gut microbiome composition and diversity were observed after smoking cessation. These shifts were similar to previously observed shifts in obese vs. lean patients, suggesting a potential microbial link between the metabolic function of the gut and smoking cessation [42]. Comparison of the gut microbiome composition of smokers and never-smokers led to similar observations [43]. So far, the underlying mechanisms of the effect of smoking on not only gut-related, but also autoimmune diseases have not been established. It has been

hypothesizes that the gut microbiome may be the missing link between smoking and autoimmune diseases [20].

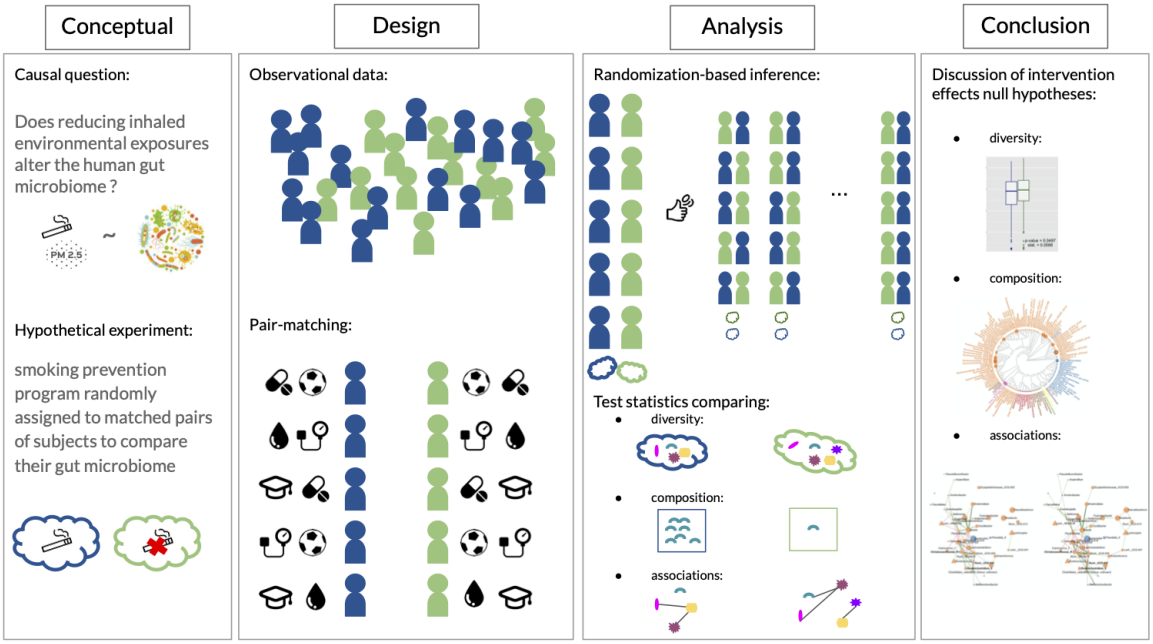

Figure 1: The four stages of the causal inference framework [21] adapted to the exploration of environment-gut microbiome relationships. Stage 1: Formulation of a plausible hypothetical intervention (e.g., decreasing inhaled environmental exposures) to examine its impacts on the gut microbiome. Stage 2: Construct a hypothetical paired-randomized experiment in which the environmental intervention been implemented randomly. Stage 3: Choose powerful test statistics comparing the gut microbiome had the subjects been hypothetically randomized to the environmental intervention vs. not and test the sharp null hypotheses of no effect of the intervention at different aggregation levels of the data. Stage 4: Interpretation of the statistical analyses and recommendations for future studies or implementation of the intervention.

Central to the present study is the investigation of the causal question: *Does reducing inhaled environmental exposures alter the human gut microbiome?* As summarized in Figure 1, we answer this question using the following four-stage analysis framework: (i) conceptualize hypothetical environmental interventions that could have resulted in the observed data at hand, (ii) design our non-randomized data, so that the unconfoundedness assumption can be assumed, (iii) choose powerful, state-of-the-art test statistics from the literature to compare human gut microbiome at different levels of taxonomic granularity between subjects assigned to the interventions vs. not, and (iv) interpret the implications of the results for recommending further studies or the studied hypothetical intervention. The Methods section elaborates on each of these steps. An essential ingredient in stage (iii) of our framework is the use of a randomization-based hypothesis testing with powerful test statistics comparing subjects under an intervention vs. not [44, 45]. We do not attempt to provide an estimate of (and uncertainty around) an estimand to avoid relying on assumptions such as the additivity of the treatment effects, asymptotic arguments, or an imputation model, which may be the case when drawing Neymanian (i.e., distribution-based) or Bayesian inferences.

The present causal inference framework relies on ideas developed in the 70s [46–49] and the Rubin Causal Model (RCM) [50, 51] to analyze observational data by reconstructing the ideal conditions of randomized experiments, the "gold standard" to draw objective causal inferences on the effects of an intervention [52]. A formidable statistical challenge is, however, to define and test these intervention effects for high-dimensional taxonomically-structured microbiome relative abundance data. Here, we adapted and advanced several state-of-the-art approaches from the statistical literature tailored to amplicon data, ranging from tests for α-diversity in networked communities [53, 54], Microbiome Regression-based Kernel Association Tests (MiRKAT) for β-diversity to randomization-based differential compositional mean tests [55]. We also applied and analyzed individual taxon differential abundance tests with taxonomic rank-dependent reference selection [56] and sparse compositionally robust taxon-taxon network estimation schemes [57] with novel differential edge tests [58], thus covering a comprehensive list of archetypical microbiome data analysis tasks.

Our framework complements recent causal inference approaches for microbiome data such as mediation methods [59, 60], graphical models [61], and Mendelian randomization [62, 63] to analyze observational gut microbiome data. In these studies, the target for interventions is the microbiome and the understanding of its effects on diseases, i.e., the microbiome is treated as the exposure and diseases as outcomes. Here, we are interested in examining the effects of environmental exposures (interventions) on the gut microbiome ("the" outcome), when only non-randomized data are available. To the best of our knowledge, no other observational study interested in environmental effects on the gut microbiome addressed their research question using causal inference methods.

In the following, we detail the characteristics of the KORA FF4 study population and highlight potential effects of the hypothetical interventions, air pollution reduction and smoking prevention, on the gut microbiome. In particular, we characterize potential effects in terms of changes in overall microbial diversity, taxon-level abundances, and microbial associations. In the smoking prevention analysis, we identified taxa, including Ruminococcaceae (UCG-005, UCG-003, UCG-002) and Christensenellaceae R-7-group, that are part of a stable sub-community in the microbial association networks and have been found to contribute to circulating blood metabolites in the LifeLines-Deep cohort [64]. The statistical workflow reproducing the present results is available at `https://github.com/AliceSommer/Pipeline_Microbiome`. A tutorial applying the workflow on the publicly available American Gut Project data [16, 65] is available at `https://github.com/AliceSommer/Causal_Microbiome_Tutorial`.

Results

To illustrate our causal inference framework, we first conceptualize two hypothetical environmental interventions that potentially influence the gut microbiome: (i) an air pollution reduction, and (ii) a smoking prevention intervention. Second, for each intervention, we construct a hypothetical matched-pair randomized experiment, aiming at satisfying the "unconfoundedness" assumption (see Methods section). Third, we analyze the "unconfounded"/"as-if randomized" data subset with randomization-based inference to test sharp null hypotheses of no effect of the interventions for each unit at different taxonomic levels of the microbial ASV data. The results presented subsequently correspond to the third stage of the framework. Fourth, causal conclusions are developed in the Discussion section. Following the American Statistical Association statement [66, 67], we avoid using the "0.05 threshold" and only describe the Fisherian p-values.

Characteristics of study population

Our study is based on data from the KORA FF4 study cohort [17]. Because we performed a design stage before analyzing the data we have two study populations, one per hypothetical experiment, which are subsets of the entire cohort (see Design stage in the Methods section). In the air pollution reduction experiment, we analyze 99 matched pairs of subjects living in highly ($PM_{2.5} \geq 13.0\ \mu g/m^3$) and less ($PM_{2.5} \leq 10.3\ \mu g/m^3$) polluted areas with similar background characteristics distributions (Table 1 and Supplementary Figures 2-4 and 8). The thresholds for the air pollution experiment intervention are based on 90^{th} and 10^{th} percentiles of the $PM_{2.5}$ distribution. We focus on the $PM_{2.5}$ pollutant, originating mainly from traffic emissions and fossil fuel combustion, for its known penetrating effects into the lung and potential implication for the gut microbiome [27]. In the smoking prevention experiment, we analyze 271 matched pairs of smokers and never-smokers (with background characteristics distributions presented in Table 1 and Supplementary Figures 5-7 and 9).

		Air pollution ($PM_{2.5}$)				**Smoking**			
		$\geq 13.0\ \mu g/m^3$		$\leq 10.3\ \mu g/m^3$		Smoker		Never-Smoker	
		Mean	St. d.	Mean	St. d.	Mean	St. d.	Mean	St. d.
Age		60.6	12.4	60.3	12.4	54.2	9.4	54.4	9.6
Body Mass Index		27.0	4.3	27.0	3.8	26.7	4.4	26.7	4.2
Alcohol intake (g/day)		11.3	14.1	11.5	13.9	13.0	15.6	11.6	14.3
Years of education		11.9	2.6	11.7	2.8	11.7	2.3	11.8	2.2
		N	%	N	%	N	%	N	%
Sex	F	130	24.0	130	24.0	41	20.7	41	20.7
	M	141	26.0	141	26.0	58	29.3	58	29.3
Smoking	Ex-S.	27	13.6	27	13.6	-	-	-	-
	Never-S.	62	31.3	62	31.3	-	-	-	-
	Smoker	10	5.1	10	5.1	-	-	-	-
Diabetes	No	95	48.0	95	48.0	264	48.7	264	48.7
	Yes	4	2.0	4	2.0	7	1.3	7	1.3
Phys. Activity	No	36	18.2	36	18.2	125	23.1	125	23.1
	Yes	63	31.8	63	31.8	146	26.9	146	26.9

Table 1: Baseline characteristics of the study population in the air pollution reduction (left table) and smoking prevention experiments (right table). Continuous variables: mean and standard deviation (St. d.). Categorical variables: number of samples per category (N) and proportion of category (%).

Statistical analysis of microbial diversity

A common first step in microbiome data analysis is estimating and assessing microbial diversity. We begin by investigating the potentially causal effects of the interventions on within-subject diversity ($\alpha-$diversity) and between-subject variation ($\beta-$diversity), respectively.

Within-subject diversity

Gut bacterial richness and Shannon diversity were estimated on the ASV level with the breakaway [68] and DivNet [54] method, respectively. Comparisons of the distributions of these estimated

variables between subject under the intervention vs. not in both hypothetical experiments are shown by boxplots in Figure 2. The small approximate Fisherian p-values based on 10,000 permutations of the intervention assignment give us ground for rejecting the null hypotheses of no effect of an air pollution reduction (p-value$_{ap,richness} \approx 0.0008$, p-value$_{ap,\alpha-div.} \approx 0.0388$) and smoking prevention (p-value$_{s,richness} \approx 0.1518$, p-value$_{s,\alpha-div.} \approx 0.0497$) on the diversity of the human gut microbiome. On average, lower diversity was observed in the subjects living in polluted areas or smokers compared to participants living in less polluted areas or non-smokers. This diversity difference motivates the more in-depth analyses of the gut microbiome composition presented subsequently.

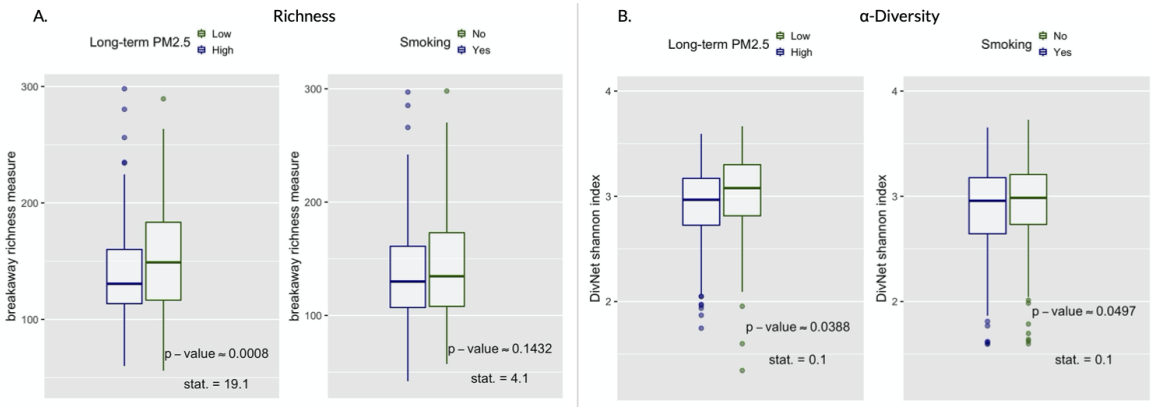

Figure 2: Richness and α-diversity. Boxplots (with median), values of the test-statistics from the `betta` regression [53], and one-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design. (A) Boxplots of the richness. (B) Boxplots of the α-diversity.

Between-subject variation

To estimate β-diversity indices, we calculated UniFrac, Aitchison, Jaccard, and Gower dissimilarities between all possible pairs of subjects. The results are shown in Table 2. To alleviate the problem of choosing the best dissimilarity metric for $\beta-$diversity estimation, we follow the Microbiome Regression-based Kernel Association Test (`MiRKAT`) of Zhao *et al.* [69] suggesting to compute several metrics and then adjust for multiple comparisons. The adjusted p-values are small, which suggests to reject the sharp null hypotheses of no effect of the intervention on between-subject variation in both experiments.

	Air pollution			**Smoking**		
distance	test-statistic	p-value	p-value$_{adj}$	test-statistic	p-value	p-value$_{adj}$
UniFrac	12.1	0.0199	0.0506	61.5	0.0024	0.0070
Aitchison	82596.0	0.1096	0.2466	356921.5	0.0001	0.0003
Jaccard	19.4	0.0884	0.2043	84.5	0.0001	0.0003
Gower	0.2	0.0089	0.0250	0.1	0.0485	0.1204

Table 2: β-diversity. Microbiome Regression-based Kernel Association Test (`MiRKAT`), unadjusted and adjusted one-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design.

Analysis of microbial compositions

We next investigated whether shifts in microbial compositions as a whole or differences in specific microbial taxa were observable in the hypothetical experiments. We illustrate this by designing and analyzing sharp null hypotheses for global compositional means and differential genus abundances.

Compositional mean differences

Testing whether two study groups have the same microbiome composition can be viewed as a two-sample testing problem for high-dimensional compositional mean equivalence. We tested sharp null hypotheses using a test statistic developed particularly for that purpose by Cao *et al.* [55]. Table 3 summarizes the results for each taxonomic level. We reject the sharp null hypotheses of gut microbiome composition equivalence for the smoking prevention experiment. At nearly all levels of taxonomic aggregation of the data, the p-values are low, except for the ASV-level. In the air pollution reduction experiment, the p-values at the Species and Genus-level are also fairly low.

		ASV	Species	Genus	Family	Order	Class	Phylum
Air Pollution	nb. of taxa (p)	4,370	414	252	74	44	29	15
	test statistic	12.8	12.9	11.9	8.8	8.4	8.4	8.1
	p-value	0.1451	0.0722	0.0733	0.1521	0.1161	0.1021	0.0591
Smoking	nb. of taxa (p)	7,409	479	271	81	48	31	16
	test statistic	13.0	14.5	13.3	11.6	8.6	9.4	10.4
	p-value	0.1607	0.0302	0.0384	0.0279	0.0859	0.0440	0.0135

Table 3: Compositional equivalence test. Test statistic for high-dimensional data suggested by [55] and one-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design.

Differential taxon abundances

For compositional microbiome data, identifying sets of potentially "differentially abundant taxa" relates to testing sharp null hypotheses of no difference in abundance of individual taxa with respect to a reference set. We conducted such an analysis on the genus level for all genera present in at least 5% of the samples. This prevalence threshold was guided by the amount of information preserved when performing filtering, i.e., microbial abundance and the number of taxa observed per sample (see Supplementary Figures 14-17). We applied the Differential abundance testing for compositional data (DACOMP) approach [56] and used two-sided tests since we lack prior knowledge on the direction of the abundance changes. Figure 3 highlights the key DACOMP results for both experiments. In the air pollution reduction experiment, we reject the sharp null hypothesis of no differential abundance only for the Marvinbryantia genus (p-value$_{adj.} = 0.0120$) (Supplementary Table 2). We do not reject the sharp null hypothesis of no effect of smoking prevention for eleven genera (see Figure 3 and Supplementary Table 3). Five belong to the Ruminococcaceae family: Ruminococcaceae-UCG-002, Ruminococcaceae-UCG-003, Ruminococcaceae-UCG-005, Ruminococcus-1, and Ruminococcaceae-NK4A214-group, three to the Lachnospiraceae family: Lachnospira, Lachnospiraceae-NK4A136-group, and Coprococcus-1, one to the Christensenellaceae family: Christensenellaceae-R-7-group, and two to the Mollicutes class, which belong to the NB1-n and Mollicutes-RF9 order.

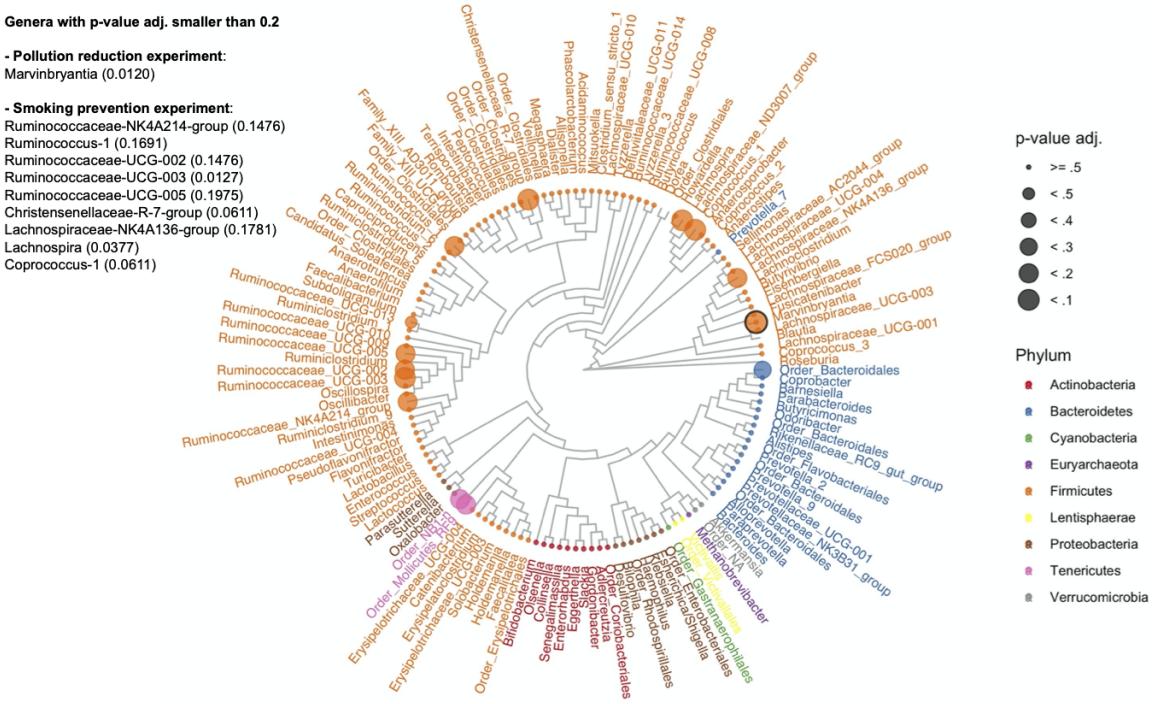

Figure 3: Differential abundance. For each genus, adjusted two-sided randomization-based p-values for 10,000 permutations of the smoking prevention intervention assignment following a matched-pair design. Genera with no tip point belong to the set of reference taxa. Black circled tip point: differentially abundant genus (Marvinbryantia) in the air pollution reduction experiment.

Microbial network analysis

To gain high-level insights into changes in the organizational structure of the underlying microbial gut ecosystem, we next calculated sparse genus-genus association networks for each exposure level and hypothetical experiment and highlight the results of our randomization-based differential association testing.

Genus-genus association networks

We used the Sparse InversE Covariance estimation for Ecological ASsociation Inference (SPIEC-EASI) framework [57] to infer genus-genus associations in our two hypothetical experiments. We used the glasso mode of SPIEC-EASI with default parameters (see Methods for details). Figure 4A shows the overall structure of the learned sparse association networks for the smoking prevention experiment (smokers (left panel) and non-smokers (right panel), respectively). Each network possesses a single large connected component consisting of 30-40 mostly Firmicutes genera (highlighted area in Figure 4A). These connected components also included the majority of the previously identified potentially differentially abundant genera, including Ruminococcaceae (UCG-005, UCG-002), Ruminococcus-1, and Christensenellaceae-R-7-group (see Figure 4B for a detailed view of the connectivity pattern). The genus-genus associations networks derived from the air pollution reduction experiment showed similar overall topological features containing one large connected component of

60 genera, including Ruminococcaceae (UCG-005, UCG-003, UCG-002) and Christensenellaceae-R-7-group among others (see also Supplementary Figure 18).

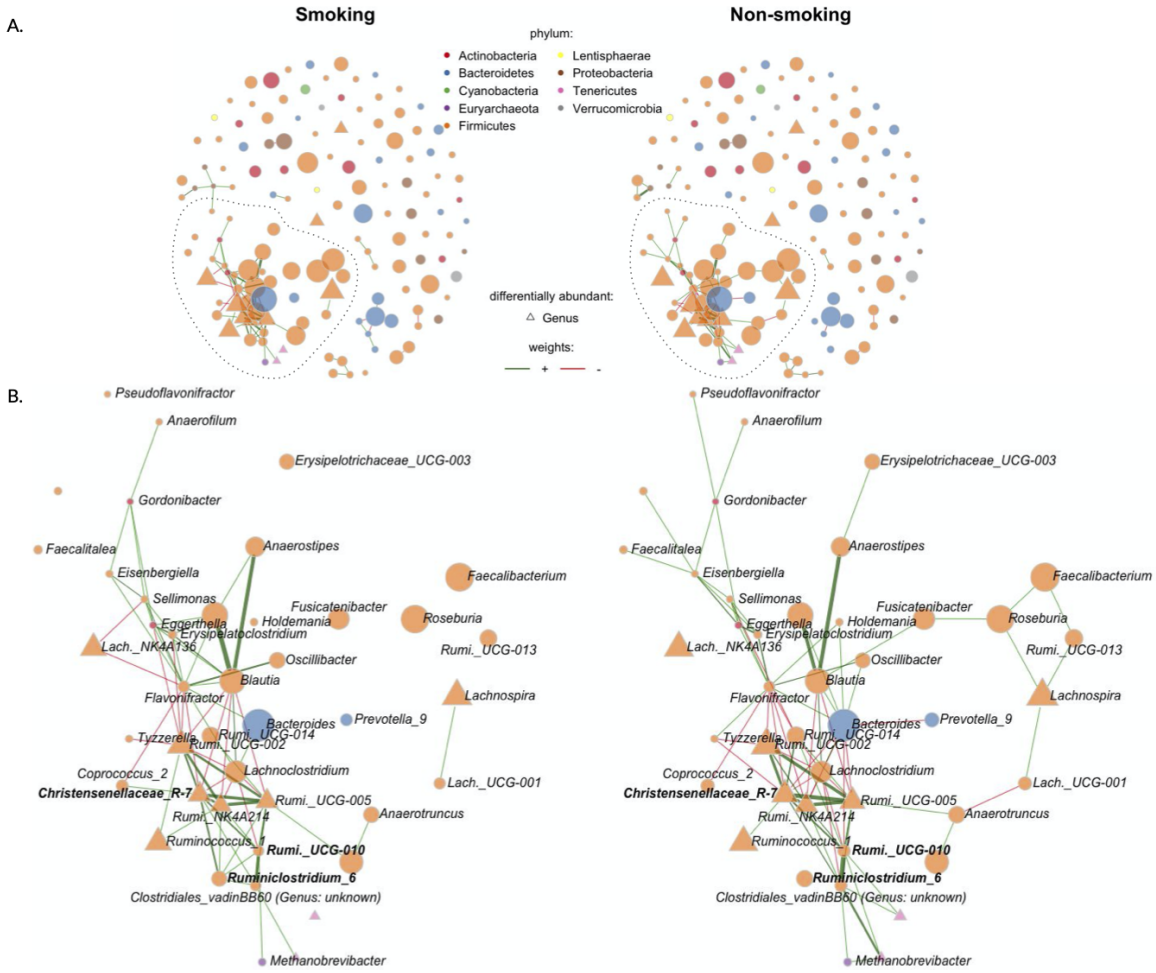

Figure 4: Genus-genus associations of smokers and never-smokers (n = 271, p = 140). (A) Visualization of the between genera partial correlations estimated with the SPIEC-EASI method. Edges thickness is proportional to partial correlation, and color to direction: red: negative partial correlation, green: positive partial correlation. Node size is proportional to the centered log ratio of the genus abundances, and color is according to phyla. Triangle shaped nodes are differentially abundant (see Figure 3). (B) Zoom in largest connected component and differential associations (bold genera).

Differential genus-genus associations

To identify potentially differential network associations in the intervention experiments, we coupled the SPIEC-EASI network estimation procedure with permutations of the intervention assignment, available in the `NetCoMi` R package [58] (see also Methods for details). For each hypothetical experiment, we list the five genus-genus associations with smallest adjusted two-sided randomization-

based p-values in Table 4 and highlight these associations in Figure 4B. In the air pollution reduction experiment, we reject the sharp null hypothesis of no differential association for two edges: the Succinivibrio/Slackia edge (p-value$_{adj.} \approx 0.0661$), and the Ruminiclostridium/Cloacibacillus edge (p-value$_{adj.} \approx 0.1063$) (see Table 4 and Supplementary Figure 18).

	Air pollution	
	Genus-genus associations (- : disappearance after intervention)	p-value$_{adj}$
	Succinivibrio/Slackia (-)	0.0661
	Ruminiclostridium/Cloacibacillus (-)	0.1063
	Cloacibacillus/Lachnospiraceae-FCS020-group	0.2795
	Megasphaera/Alistipes	0.4147
	Bacteroidales (Genus: unknown)/Prevotella-2	0.4753
	Smoking	
	Genus-genus associations (- : disappearance after intervention)	p-value$_{adj}$
	Christensenellaceae-R-7/Ruminiclostridium-6 (-)	0.1585
	Ruminococcaceae-UCG-010/Ruminiclostridium-6 (-)	0.1585
	Ruminococcaceae-UCG-014/Flavonifractor	0.2031
	Clostridiales-vadinBB60/Ruminiclostridium-6	0.2376
	Ruminococcaceae-UCG-013/Faecalibacterium	0.2492

Table 4: Differential associations of genera. Smallest five adjusted two-sided randomization-based p-values for 10,000 permutations of the intervention assignment following a matched-pair design.

In the smoking prevention experiment, we also reject the sharp null hypothesis of no differential association for two edges: the Ruminiclostridium-6/Ruminococcaceae-UCG-010 edge (p-value$_{adj.} \approx 0.1585$), and the Ruminiclostridium-6/Christensenellaceae-R-7-group edge (p-value$_{adj.} \approx 0.1585$) (see Table 4). The genera that participate in these potentially differential associations are also highlighted in Figure 4B.

Exploring associations between genera and lipid metabolites

The gut microbiome is a substantial driver of circulating lipid levels, and prior work has shown [64, 70, 71] that the relative abundance of several microbial families, including Christensenellaceae, Ruminococcaceae, and the Tenericutes phylum were negatively correlated with triglyceride and positively associated with high-density lipoproteins (HDL) cholesterol. Since our analysis identified a small interconnected group of genera, including Christensenellaceae and Ruminococcaceae, for whom we rejected the no differential abundance hypothesis, we performed an exploratory data analysis to investigate taxa-serum lipid measurements associations. Four lipids were measured in blood serum samples of our study population from the KORA cohort: total, HDL, and LDL, cholesterol, as well as triglyceride levels. Figure 5A shows the correlation between these lipids and the genera we discovered in our hypothetical experiments. Tendencies similar to those reported in previous studies can be observed in our data.

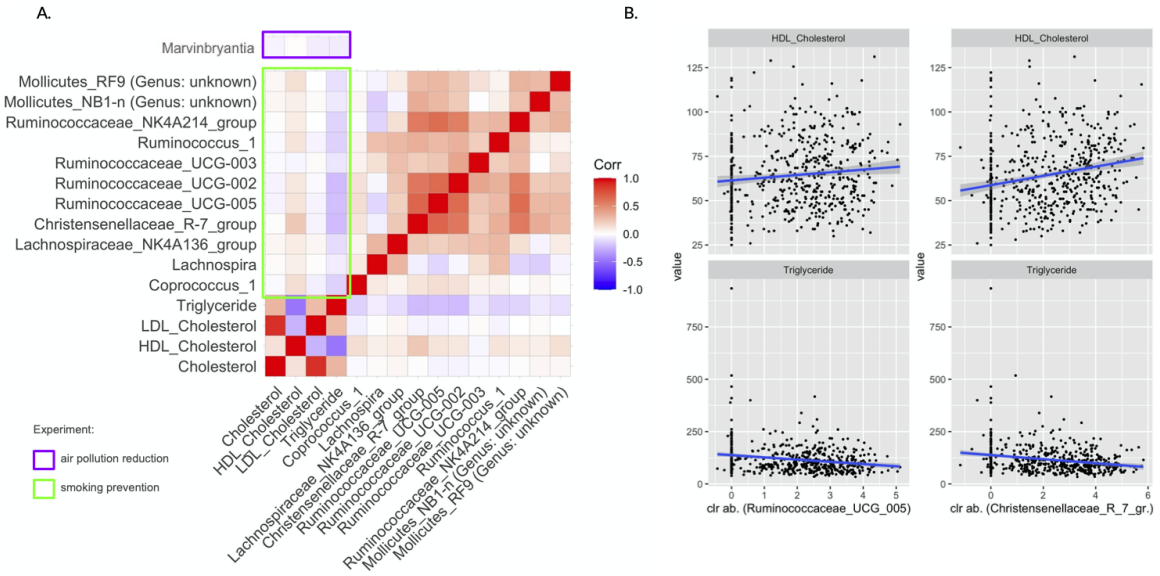

Figure 5: (A) Lipid metabolites correlation with selected genera from the smoking prevention experiment (green). (B) Scatterplots of high-density lipoprotein (HDL) cholesterol and triglycerides vs. centered log-ratio transformed relative abundances of the genera Ruminococcaceae-UCG-005 and Christensenellaceae-R-7-group.

For instance, in the smoking prevention dataset, we observed a positive correlation of Christensenellaceae R-7-group and Ruminococcaceae (UCG-005) genus abundances (under centered log-ratio transformation) with HDL cholesterol and negative correlation with triglyceride levels, respectively (see Figure 5B). Similar correlation patterns were also found for the other genera for whom we rejected the no differential abundance hypothesis (see second and forth column in Figure 5A). Our findings were also in line with recently reported correlation results in Vojinovic *et al.* [64] using the Dutch LifeLines-DEEP cohort [13] and the Rotterdam Study [14].

Discussion

Due to the interdisciplinary (epidemiology, microbiome research, and statistics) nature of this paper we first discuss the results presented above, then elaborate on the statistical framework we used for our analyses and suggest statistical and epidemiological extensions of our work.

In the air pollution ($PM_{2.5}$) reduction hypothetical experiment, we reject the sharp null hypotheses of no richness, α-diversity, and high-dimensional mean differences. We also reject the no differential abundance hypothesis for the Marvinbryantia genus, and the no differential association hypothesis between: the Succinivibrio and Slackia genera, as well as the Ruminiclostridium and Cloacibacillus genera. Experiments exposing mice to $PM_{2.5}$ resulted in mixed findings concerning difference in microbial richness and diversity. This might be due to regional differences in the chemical composition of $PM_{2.5}$ as well as differences in the duration of exposure [29]. Thus far, only one human study estimated associations between $PM_{2.5}$ exposure and the gut microbiome, and investigated the pathway of diabetes induction associated with PM exposure [28]. One of their key findings was that $PM_{2.5}$ exposure reduced α-diversity (measured by Chao1 and Shannon indices), which is

consistent with our observations.

In the smoking prevention hypothetical experiment, we rejected the sharp null hypotheses of no richness, no α-diversity, no β-diversity, and no high-dimensional mean differences. We also rejected the no differential abundance hypothesis for eleven genera (five of the Ruminococcaceae family, three of the Lachnospiraceae family, one of the Christensenellaceae family, and two of the Mollicutes class), and the no differential association hypothesis between the Ruminiclostridium-6 and Ruminococcaceae-UCG-010 genera, and between the Ruminiclostridium-6 and Christensenellaceae R-7-group genera. Interestingly, the associations of Ruminococcaceae-UCG-010 and Christensenellaceae R-7-group with Ruminiclostridium-6 were also found to be worth further scrutiny. Their positive associations in the genus-genus network of smokers was absent in the genus-genus network of the never-smokers. The one study comparing the gut microbiome of smokers ($n = 203$) and never-smokers ($n = 288$) with similar sample size has a men-only study population [43]. They did not find any differences in α-diversity (measured with the Shannon index), whereas we conclude that α-diversity analyses are worth further scrutiny. Lee *et al.*'s PERMANOVA analyses for β-diversity differences, measured with Jaccard and weighted UniFrac distances, suggested differences. We reject the sharp null hypothesis at the between-subject differences analysis level. In their analysis of bacterial (phylum) composition, smokers had an increased proportion of Bacteroidetes with decreased Firmicutes and Proteobacteria compared with never-smokers. When we compare these phyla, we do not depict the same differences (see Supplementary Figure 18). Also, our compositional difference analyses do not result in the same set of differentially abundant genera than reported by Lee *et al.* [43]. These conflicting findings could be due to the fact that their study was done on Korean men only. Nonetheless, it shows that there is a lack of knowledge on the effects of smoking on the human gut microbiome and that additional scientific investigations are necessary to make causal conclusions.

Throughout the extensive statistical analyses presented in this paper, we have tested sharp null hypotheses of no effect of an intervention on a wide range of gut microbiome outcomes, ranging from high-level microbial diversity estimates to differential genus-genus associations. To do so, we have performed randomization-based inference based on 10,000 permutations. This mode of inference has been motivated by two reasons: (i) it is difficult to postulate a joint model for the potential outcomes, and thereby provide an estimate of (and uncertainty around) a causal estimand, and (ii) it has been shown that using the actual randomization procedure that led to the observed data helps to report valid Fisher-exact p-values as opposed to p-values relying on approximating null randomization distributions [45]. As an example, in our mean difference analyses, we found slight differences between the null randomization distribution of the test-statistic when approximated by permuting the intervention assignment vector and when drawn from the approximating asymptotic distribution (see Supplementary Figures 10 and 11).

An important component of our randomization-based procedure is that the permutations of the intervention assignment vector conserves the matched-pair design of the hypothetical randomized experiment. This strategy has been advocated by Rubin [72] in the context of randomized trials, and more recently by Bind & Rubin [45] in the context of hypothetical randomized experiments, because assumptions on the underlying distribution of the data are not required. Only few R packages were built to perform randomization-based inference while conserving the design of the intervention assignment. Therefore, for every analysis in our study, we imported a matrix of 10,000 unique randomized intervention assignments to calculate our p-values (see `https://github.com/AliceSommer/Causal_Microbiome_Tutorial` for a reproducible example on the American Gut Data [16, 65]). Nonetheless, the `DACOMP` and `NetCoMi` R packages provide flexible functions enabling the calculation of randomization-based p-values for our study design to test sharp null hypotheses of no difference in taxa abundance and associations, respectively. We advocate for more development of

such user-friendly software functions permitting flexibility and accountability of the design stage of observational studies. P-value adjustments for multiple comparison also follow a fully randomization-based procedure, while preserving the design of the experiment. The method has proven to be more powerful while maintaining the family-wise error rate [73].

Notice that when presenting our results, we never accepted alternative hypotheses but only rejected sharp nulls when unadjusted and adjusted p-values were small, i.e., indicating the hypotheses warrants further scrutiny [67]. In the field of microbiome data analysis, the terms differential abundance and associations are frequently used. Researchers report "differential abundant" and "differentially associated" sets of taxa after testing sharp null hypotheses of no effect of an intervention. This terminology implicitly implies acceptance of the alternative hypotheses. However, when testing sharp null hypotheses we assess the amount of evidence against them in the observed data, which does not prove the alternative hypothesis to be true.

During the design stage, the outcome variable was ignored and only pre-exposure covariates were considered. The chosen balanced data is a sub-sample of units that can be used to estimate the effects of an intervention. Omitting the outcome data until the analysis avoids "model cherry-picking", because the effect of the intervention is estimated once, after a successful design stage. Nonetheless, at the design stage, we can only consider the observed pre-exposure variables but the assignment mechanism could depend on unobserved pre-exposure variables. In gut microbiome studies, diet is often an unobserved confounder. For example, in this study, dietary intake data was collected for only 1,469/2,033 (i.e., 72%) participants. We verified balance in dietary intake for our balanced data subset (see Supplementary Figures 8 and 9). Because of the lack of information, there could still be imbalances in diet and/or other unobserved covariates. In such cases, Rosenbaum [74] has recommended to consider sensitivity analyses of how the Fisher-exact p-value would change, had the intervention assignment been plausibly different (see also Bind & Rubin [45]). Subject-matter knowledge on air pollution exposure or smoking assessment should guide the plausible range of "sensitivity" p-values and the reason why they could deviate from the p-value calculated based on the assumed hypothetical intervention assignment. This idea provides material for an extension of the framework presented in this study.

The framework suggested in this paper facilitates a more transparent interpretation of results than standard approaches. First, interpretation is only valid within the range of the background covariates of the study population in the respective hypothetical experiment (see their detailed characteristics in Table 1 and Supplementary Figures 2-9). The data do not provide direct information for the "unmatched" units. Also, the assumed assignment mechanism and underlying assumptions have to be clearly stated to obtain meaningful p-values. Standard approaches usually make strong assumptions (e.g., linearity), whose discussions are often neglected. Modeling the observed data and solely adjusting for confounders by including them in a regression, without a design stage, can be unreliable, especially when the pre-exposure covariates distributions of the control and intervention units are not similar. For instance, Cochran & Rubin [46], Heckman *et al.* [75], and Rubin [76] have shown that regression models can estimate biased treatment effects when the true relationship between the covariates and the outcome is not modeled accurately.

In contrast to other studies interested in the effect of air pollution exposures on health outcomes, this study does not provide any estimation of an exposure-response curve. Instead, we examine the effect of interventions and provide results that can directly contribute to policy recommendations. Until now, relationships between inhaled environmental exposures and the human gut microbiome were not examined with causal inference methods, so a first step to make advances in the field is to test, whether air pollution and smoking have no effect on the units of our study. If so, a natural next step would be to work with a dataset adequate for balancing covariates along different doses of the exposure such as suggested in [77] and estimate a causal dose-response in order to protect

populations at risk.

In the smoking prevention experiment, the subset of genera retained at the differential abundance analysis step was linked to the serum markers triglycerides and high-density lipoprotein in previous studies [64, 70, 71]. In our data, we observe correlations between these genera and metabolites in the same direction than previously found by Vojinovic *et al.* [64] (see Figure 5). Serum triglycerides and high-density lipoprotein play a role in metabolic syndrome and associations between smoking and metabolic syndrome have also been found previously [78]. Therefore, we suggest further investigation on the pathway of cigarette smoke impacting the gut, which in turn has effects on circulating metabolites (and metabolic syndrome). A logical next step would be to apply our framework to other cohorts with similar amplicon data preprocessing and available pre-exposure covariates such as the the Dutch LifeLines-DEEP [13] and Rotterdam Studies [14], and observe whether our results replicate.

Methods

The German KORA FF4 cohort study

The data come from the German KORA FF4 cohort study, which involves participants aged 25 to 74 years old living in the city of Augsburg [17]. The participants were subject to health questionnaires and follow-up examinations. During the study, stool samples were collected and the gut microbiota data for 2,033 participants were obtained with 16S rRNA gene sequencing. For each participant we have their long-term exposure to air pollution (particulate matter). The long-term exposure variables come from the ULTRA III study, in which air pollutants were monitored several times a year at 20 locations within the Augsburg region. From this data, annual averages of air pollutants were calculated using land-use regression models. The models explain the spatial variation of the pollutants with predictor variables derived from geographic information systems (GIS). To obtain the long-term air pollution values for each participant, land-use regression models were applied to their residential address. Moreover, to elucidate relationships between health outcomes and diet, dietary intake data were collected for 1,469 participants of the KORA FF4 cohort. Dietary intake was derived using a method combining information from a food frequency questionnaire (FFQ) and repeated 24-h food lists [79]. In brief, the usual food intake (in gram/day) was calculated as the product of the probability of consumption of a food on a given day and the average amount of a food consumed on a consumption day.

Gut microbiome data sequencing and preprocessing

DNA Extraction, 16S rRNA Gene Amplification, and Amplicon Sequencing. Fecal DNA extraction was isolated by following the protocol of [80]. The samples were profiled by high-throughput amplicon sequencing with dual-index barcoding using the Illumina MiSeq platform. Based on a study providing guidelines for selecting primer pairs [81], the V3-V4 region of the gene encoding 16S ribosomal RNA was amplified using the primers 341-forward (CCTACGGGNGGCWGCAG; bacterial domain specific) and 785-reverse (GACTACHVGGGTATCTAATCC; bacterial domain specific). Amplification was undertaken using the Phusian High-Fidelity DNA Polymerase Hotstart as per manufacturer's instructions. The PCR libraries are then barcoded using a dual-index system. Following a round of purification with AMPure XP beads (Beckman Coulter), libraries were quantified and pooled to 2nM. The libraries were sequenced on an Illumina MiSeq (2 x 250 bp), using facilities provided by the Ziel NGS-Core Facility of the Technical University Muenchen (TUM).

Bioinformatics. The demultiplexed, per-sample, primer-free amplicon reads were processed by the DADA2 workflow [22, 82] to infer sequence variants, remove chimeras, and assign taxonomies with the Silva v128 database [83] using the naive Bayesian classifier method [84] until the genus-level assignment and the exact matching method [85] for species-level assignment. We opted for the high-resolution DADA2 method to infer sequence variants without any fixed threshold, thereby resolving variants that differ by as little as one nucleotide. Amplicon sequence variants (ASVs) do not impose the arbitrary dissimilarity thresholds that define OTUs. They provide consistent labels because they represent a biological reality that exists outside the data being analyzed: the DNA sequence of the assayed organism, thus they remain consistent into the indefinite future [22]. The multiple genome alignment for the phylogenetic tree was built with the `DECIPHER` R package enabling a profile-to-profile method aligns a sequence set by merging profiles along a guide tree until all the input sequences are aligned [86]. The multiple genome alignment was used to construct the *de novo* phylogenetic tree using `phangorn` R package. We first construct a neighbor-joining tree [87], and then fit a maximum likelihood tree using the neighbor-joining tree as a starting point. After 16s rRNA sequencing the 2,033 stool samples from the KORA cohort and processing the sequences with the DADA2 pipeline, we observe 15,801 ASVs (see Supplementary Figure 1 and Table 1).

Causal inference framework

The four stages of the causal framework [21] that we use to construct hypothetical randomized experiments to study the environment-microbiome relationship are the following:

1. *Conceptual*: Formulation of a plausible hypothetical intervention (e.g., decreasing air pollution levels) to examine its impacts on the gut microbiome.

2. *Design*: Reconstruct the hypothetical randomized experiment had the environmental intervention been implemented randomly.

3. *Analysis*: Choose valid and powerful test statistics comparing the gut microbiome had the subjects been hypothetically randomized to the environmental intervention vs. not and test the sharp null hypotheses of no effect of the intervention at different aggregation levels of the data.

4. *Summary*: Interpretation of the statistical analyses and recommendations for future studies and interventions.

Conceptual stage: formulation of the hypothetical randomized experiment in terms of potential outcomes

To understand whether environmental interventions have an effect on the human gut microbiome, the objective is to reconstruct a hypothetical experiment that mimics a controlled randomized experiment [52], in which an environmental intervention could be believed to have been randomized. Let W_i be the indicator of the assignment for subject i $(i = 1, ..., N)$ to an environmental intervention vs. none, where:

$$W_i = \begin{cases} 1 & \text{if } i \text{ is under the intervention,} \\ 0 & \text{if } i \text{ is not.} \end{cases} \quad (1)$$

The composition of a human gut microbiome can be expressed as a B-dimensional vector of the microbial abundance. We define Y_i^b as the real abundance (count) of the b^{th} bacterium, $b = 1, ..., B$

for subject i. We define the potential outcomes of subject i as $Y_i^b(1)$, the b^{th} bacterium abundance (count) had subject i been randomized to the environmental intervention ($W_i = 1$), and $Y_i^b(0)$, had subject i not been randomized to the intervention ($W_i = 0$). Table 5 shows the potential outcomes for the N subjects.

Bacteria	**1**		**2**		**...**		**B**	
Subjects	$W_i = 0$	$W_i = 1$	$W_i = 0$	$W_i = 1$			$W_i = 0$	$W_i = 1$
1	$Y_1^1(0)$	$Y_1^1(1)$	$Y_1^2(0)$	$Y_1^2(1)$	...	...	$Y_1^B(0)$	$Y_1^B(1)$
2	$Y_2^1(0)$	$Y_2^1(1)$	$Y_2^2(0)$	$Y_2^2(1)$	...	...	$Y_2^B(0)$	$Y_2^B(1)$
...	...	...	...	...	...	...	...	...
N	$Y_N^1(0)$	$Y_N^1(1)$	$Y_N^2(0)$	$Y_N^2(1)$	...	...	$Y_N^B(0)$	$Y_N^B(1)$

Table 5: Potential outcomes for the subjects of the hypothetical experiment

Only one of the two potential outcomes can actually be observed for each subject: this is why Rubin characterizes causal inference as a *missing data problem* [51], where the observed outcome of subject-i and bacteria-b can be expressed as a function of both potential outcomes:

$$Y_i^{b,obs} = W_i Y_i^b(1) + (1 - W_i) Y_i^b(0) \tag{2}$$

Observed outcomes measurement

The human gut microbiome can be composed of trillions of bacteria. However, due to technology limitations, the exact abundance and number of all species present in a human subject cannot be measured. To tackle this limitation, we opted for the processing of Amplicon Sequence Variants (ASVs) from our sequencing data to approximate the true gut microbiome composition of our study population [22, 82]. ASVs refer to individual DNA sequences recovered from a high-throughput marker gene analysis, the 16S rRNA gene in our case. Therefore, in this study the observed outcome under investigation is a $N \times A$ matrix, for $a = 1, ..., A$ ASVs, an approximation of the $N \times B$ matrix described above. This limitation adds another layer of missing data, i.e., we are missing the true gut microbial composition of each subject. We define the ASV counts we measured for each subject-i as $C_i^{a,obs}$, which corresponds to $Y_i^{b \in A,obs}$ plus some measurement error.

Design stage: reconstruction of the conceptualized hypothetical experiment

To assess causality, randomized experiments have long been regarded as the "gold standard". We are interested in the effect of environmental interventions that are often unpractical or ethical to assign randomly to humans within an experiment [21]. Therefore, we resort to a design stage [88] with a matched-sampling strategy to construct two hypothetical randomized experiments to assess the effects of an intervention on the changes in gut microbiome composition. The aim of our pair-matching strategy is to achieve balance in background covariates distributions as it is expected, on average, in randomized experiments. This strategy attempts to create exchangeable groups as if the exposure was randomly assigned to each participant, to guaranty exposure assignment is not, on average, confounded by the measured background covariates.

Our pair-matching strategy aims to remove individual-specific confounding (e.g., years of age, sex, unit of BMI). Briefly, subject i under $W_i^{obs} = 1$ with pre-exposure covariates $\mathbf{X}_i$ is matched to subject $i^\star$, under $W_{i^\star}^{obs} = 0$ only if $\mathbf{X}_{i^\star}$ is "similar" to $\mathbf{X}_i$. For each unit, the vector of covariates is

given by $\mathbf{X}_i = (X_i^{(1)}, ..., X_i^{(k)})$. In order to ensure covariate balance, we only allow a treated unit to be matched with a control unit if the component-wise distances between their covariate vectors are less than some pre-specified thresholds $\delta_1, ..., \delta_k$. For any pair of covariate vectors X_i and $X_{i^\star}$, we define the difference between them as

$$\Delta(X_i, X_{i^\star}) = \begin{cases} 0 & \text{if } |X_i^{(k)} - X_{i^\star}^{(k)}| < \delta_k \text{ for } k = 1, ...K, \\ +\infty & \text{otherwise} \end{cases}$$

This constrained pair matching can be achieved using a maximum bipartite matching [89] on a graph such that: (i) there is one node per unit, partitioned into intervention nodes and control nodes, (ii) the edges are pairs of treated and control nodes with covariates X_i and $X_{i^\star}$, and (iii) an edge exists if and only if $\Delta(X_i, X_{i^\star}) < +\infty$. By construction, using a maximum bipartite matching algorithm on this graph as implemented in the `igraph` R package produces the largest set of matched pairs that satisfy the unit-specific proximity constraints set by our thresholds. Let $N_E = \sum_{i=1}^{N} W_i$ be the number of subjects under the environmental intervention and $N_C = \sum_{i=1}^{N} 1 - W_i$ the number of control subjects, after matching.

After excluding the participants of the cohort that take antibiotics and had a cancer of the digestive organ, the pre-matched data set consists of 1,967 participants. At this stage, the objective to create balanced data subsets for which the plausibility of the "unconfoundedness" assumption is based on a diagnostic of our choice. We choose the thresholds according to the covariates pre-matching distributions diagnostic plots (see Supplementary Figures 2-7). The thresholds are: the absolute differences between the amount of alcohol consumption is less than $\delta_1 = 25$ g/day, between the body-mass-index is less than $\delta_2 = 4$ kg/m^2, between age is less than $\delta_3 = 5$ years, the diabetes status (diabetic, non-diabetic) is identical, i.e., $\delta_4 = 0$, and so are sex (male, female), i.e., $\delta_5 = 0$, and physical activity (active, inactive), i.e., $\delta_6 = 0$. Additionally, in the air pollution reduction experiment: the smoking status (smoker, ex-smoker, never-smoker) is identical, i.e., $\delta_7 = 0$, and in the smoking cessation experiment: the absolute difference between years of education is less than $\delta_7 = 3$ years.

After matching, we obtain two subsets of the data that can be analyzed as coming from two pair-randomized experiments: (i) an air pollution (ap) reduction hypothetical experiment ($N_{ap} = 198$), and (ii) a smoking prevention hypothetical experiment ($N_s = 542$); both data sets exhibit no evidence against covariate imbalance (see Table 2 and Supplementary Figures 2-7).

	Air pollution		**Smoking**	
	N_C	N_E	N_C	N_E
Matching	$PM_{2.5} \geq 13.0$ μg/m^3	$PM_{2.5} \leq 10.3$ μg/m^3	Smoker	Never smoker
Before	206	193	302	908
After	99	99	271	271

Table 6: Before and after matching number of units. The thresholds for the air pollution experiment are based on 90^{th} and 10^{th} percentiles of the $PM_{2.5}$ distribution.

It is well known that diet has an influence on the microbiome and more studies on the gut should include dietary intake data in their analysis [90, 91]. In our study, we only have access to dietary intake data for a portion of our samples, therefore we look at balance diagnostics in usual nutrient intake after matching in order to maintain a large data set before matching. Supplementary Figures 8 and 9 show that after matching, our intervention and control units (in both hypothetical experiment) do not exhibit imbalance with respect to the following food items: potatoes/roots,

vegetables, legumes, fruits/nuts, dairy products, cereal products, meat, fish, egg products, fat, and sugar.

Statistical analysis stage: randomization-based inference

To compare the gut microbiome of subjects under the environmental intervention to control subjects, we choose to not rely on asymptotic arguments, but instead take a Fisherian perspective (i.e., randomization-based inference) [44, 92]. We test sharp null hypotheses (H_0) of no effect of the intervention for any unit by choosing test statistics that account for the complex microbiome data structure, including the additional "layer" of missing data. The ASV-count data has a challenging structure because: (i) it is high-dimensional, (ii) some ASVs have low prevalence, (iii) the ASVs are strongly correlated, and (iv) it is compositional. ASV-count data is said to be "compositional" because between units comparison of ASV counts might not be informative due to the limited sequencing depth of the machine and the total number of sequenced reads varies from unit to unit (i.e., they have no common denominator) [93].

In randomization-based inference the goal is to construct the null randomization distribution of a test statistic assuming H_0, T, by computing the values of the test statistic for all possible intervention assignments. Because the number of assignments is very large, we calculate an approximating p-value using N_{iter} iterations, i.e., the proportion of computed test statistics that are as large or larger than the observed test statistic: $\frac{1}{N_{iter}} \sum_{l=1}^{N_{iter}} \mathbb{1}_{|T_l| \geq |T^{obs}|}$, where $\mathbb{1}_{|T_l| \geq |T^{obs}|} = 1$ when $|T_l| \geq |T^{obs}|$, and 0 otherwise. A small p-value shows that the observed test statistic is a rare event when the null hypothesis is true, which indicates the results are worth further scrutiny [67]. In the following subsections, we describe the null hypotheses we test and the test statistics we use to draw randomization-based inferences with $N_{iter} = 10{,}000$ possible intervention assignments following a matched-pair design. This means that the permutations of the intervention assignment vectors needed to calculate the Fisher p-values follow the design of our hypothetical experiments. When units have varying probabilities of being treated, the analysis of experiments, even when hypothetical, should reflect their design [52, 88].

analysis level	data transformation	test statistic
richness	breakaway [68]	betta regression coefficient [53]
α-diversity	DivNet [54]	betta regression coefficient [53]
β-diversity	pairwise distance matrices	MiRKAT score statistic [69]
high-dimensional means	centered log ratios	mean abundance difference [55]
abundance	normalization by ratio [56]	LogFold mean difference
correlation	association matrices [57]	differential associations [58]

Table 7: Data transformation and choice of test statistics.

Diversity analyses

Within Subjects Diversity.

One of the challenges of analyzing ASV-count data is working around the low prevalence of some ASVs that are due to the limited sequencing depth of the machine and the fact that some ASVs are not shared in the entire population (see Supplementary Figure 1). Therefore, before directly testing within-subject diversity differences with so called "plug-in" estimates, it has been recently suggested to start with estimating the diversity with statistical models [53]. We will follow this idea by estimating richness with the breakaway method [68] and estimating the Shannon index for α-diversity with the DivNet method [54].

Richness. The sharp null hypothesis of no effect of the intervention on the richness can be written as: $\mathbf{H_{0,R}} : \sum_{b=1}^{B} \mathbb{1}_{Y_i^b(0)>0} = \sum_{b=1}^{B} \mathbb{1}_{Y_i^b(1)>0}$. To estimate the richness of subject i (i.e., the number of bacteria present in subject i), we will estimate the total richness in subject i, observed and unobserved, by B_i with the `breakaway` model [68]. Let $f_{i,1}, f_{i,2}, ...$ denote the number of bacteria observed once, twice, and so on, in a subject i, and let $f_{i,0}$ denote the number of unobserved bacteria, so that $B_i = f_{i,0} + f_{i,1} + f_{i,2} + ...$ The idea behind the breakaway method is that for each subject i, it predicts the number of unobserved bacteria, $f_{i,0}$, with a nonlinear regression model to, in turn, provide an estimate of B_i.

α-diversity. The sharp null hypothesis of no effect of the intervention on α-diversity can be written as: $\mathbf{H_{0,\alpha}} : \sum_{b=1}^{B} Y_i^b(0) = \sum_{b=1}^{B} Y_i^b(1)$. To have estimates for indices of the α-diversity of subject i (i.e., its total microbial abundance) and their variance, we use the `DivNet` method, because it accounts for the co-occurrence patterns (i.e., ecological networks) of bacteria in the microbial community [54]. Let $Z_i^b = Y_i^b / \sum_{b=1}^{B} Y_i^b \in [0, 1]$ denote the unknown relative abundance of bacteria b in subject i, noting that $\sum_{b=1}^{B} Z_i^b = 1$. As a reminder, $C_i^{a,obs}$ denotes the number of times bacteria a was observed in the stool sample of subject i in our data. One of the most common α-diversity indices is the Shannon entropy [94], which is defined as: $\alpha_{i,Shannon} = -\sum_{b=1}^{B} Z_i^b \ log(Z_i^b)$. This index captures information about both the species richness (i.e., number of species) and relative abundances of the species: as the number of species in the population increases, so does the Shannon index, and as the relative abundances diverge from a uniform distribution and become more unequal, the Shannon index decreases. In the ecological literature, researchers mostly use the following maximum likelihood estimate of $\alpha_{i,Shannon}$ (often referred to as a "plug-in" estimate): $-\sum_{a=1}^{A} \frac{C_i^a}{\sum_{a=1}^{A} C_i^a} \ log\left(\frac{C_i^a}{\sum_{a=1}^{A} C_i^a}\right)$. It has been proven that this estimate is negatively biased [95]. Therefore, various corrections have been proposed and are detailed in [54]. However, most of the suggested estimates are only functions of the ASV count vectors C_i^a and do not utilize the full ASV count data matrix C and the co-occurrence pattern, i.e., ecological network, of the ASVs. Willis & Martin [54] showed that these networks can have substantial effects on estimates of diversity and proposed an approach, called DivNet, to estimating diversity in the presence of an ecological network. DivNet estimates are based on log-ratio transformations by fixing a "baseline" taxon for comparison, which are modeled by a multivariate normal distribution to incorporate the co-occurrence structure between the taxa as the covariance matrix. The main advantage of DivNet method is the use of information shared across all samples to obtain more precise and accurate estimates.

Choice of test statistic. The test statistic we use to test $\mathbf{H_{0,B}}$ and $\mathbf{H_{0,\alpha}}$ are the coefficient of the intervention indicator estimated by the regression suggested by [53]. Using the coefficient of a model as the test statistic of a Fisher test was introduced in the 70s [96]. At this stage, to achieve larger bias reductions, frequentist regression models can be used to remove residual confounding that was not accounted for, during the design stage [46, 47].

Willis *et al.* [53] suggest to test changes in richness (B_i) and α-diversity ($\hat{\alpha}_i$) with a hierarchical regression model, assuming that richness is a function of: the intervention indicator W_i, random variation that is not attributed to the covariates, and the standard error previously estimated with breakaway or DivNet (because not every bacteria in each subject was observed so we cannot not know the true richness or α-diversity for any i). The regression models are built with the `betta` function available in the `breakaway` R package [53, 68].

Between Subjects Diversity.

β-diversity. Distance-based analysis is a popular approach for evaluating the association between an exposure and microbiome diversity. The pairwise distances, d_{ii^*}, for high-dimensional data we consider are the: UniFrac (unweighted) distance [97], Jaccard index, Aitchison distance

[98] (i.e, euclidean distance on centered log-ratio transformed data), and Gower distance [99] (on centered log-ratio transformed data). We choose the unweighted paired UniFrac, because it is a proper distance metric as opposed to the generalized UniFrac. The same applies to the Jaccard distance as opposed to the commonly used Bray-Curtis. The sharp null hypothesis of no effect of the intervention on β-diversity can be written as: $\mathbf{H_{0},_{\beta}} : \mathbf{d}_{ii^{\star}}(0) = \mathbf{d}_{ii^{\star}}(1)$.

Choice of test statistic. Despite the popularity of distance-based approaches, they suffer from technical challenges, especially in selecting the best distance. Therefore, we use the suggested microbiome regression-based kernel association test (MiRKAT) [69] that uses a kernel regression and a standard variance-component score test statistic [100]. To consider different distance measures, the optimal MiRKAT: tests $\mathbf{H_{0},_{\beta}}$ for each individual kernel, obtains the p-value for each of the tests, and then adjust for multiple comparison with a p-value with an omnibus test. Instead, we will use a fully randomization-based multiple comparison adjustment method detailed subsequently.

Multiple comparison adjustments. We will follow the fully randomization-based procedure for multiple comparisons adjustments suggested by [73], which is directly motivated by the intervention assignment actually used in the experiment. In our case a hypothetical matched pair experiment. Both the unadjusted and adjusted p-values in the procedure are randomization-based, so do not require any assumptions about the underlying distribution of the data. The *adjusted* p-values are calculated following Step 1-4:

1. Calculate unadjusted p-values for the observed test statistic as explained previously. For each hypothesis h, $h = 1, .., H$, record the $T_{\beta}^{h,iter} = (T_{\beta}^{1,1}, ..., T_{\beta}^{H,N_{iter}})$, where $iter = 1, ..., N_{iter}$.

2. For each h and each $T_{\beta}^{h,iter}$, calculate an unadjusted randomization-based p-value. For each iteration $iter$, record the minimum p-value of the H p-values.

3. The repetitions of Step 2 capture the joint randomization distribution of the test statistics and thus, of the unadjusted p-values.

4. To calculate the adjusted p-values for the observed test statistics, take the proportion of "minimum p-values" (recorded in Step 2) that are less than or equal to the unadjusted p-value calculated in Step 1.

Step 2-3. essentially represent a translation of the multiple test statistics into p-values sharing a common 0-1 scale.

Composition analyses

Compositional equivalence.

The compositionality problem means that: a change in abundance (i.e., sequenced counts) of a taxon in a sample induces a change in sequenced counts across all taxa. This problem, among others, leads to many false positive discoveries when comparing taxon abundances between groups. Moreover, because the components of a composition must sum to unity, directly applying standard multivariate statistical methods intended for unconstrained data to compositional data may result in inappropriate and misleading inferences [98]. Therefore, we impose a centered log-ratio transformation of the compositions before testing the null hypothesis of no difference in average microbial abundance as suggested by [55].

For the measured microbiome data C, the centered log-ratio matrices $L = (L_i, ..., L_N)$ are defined by $L_i^a = log\big(\frac{C_i^a}{g(\mathbf{C}_i)}\big)$, where $g(\mathbf{C}_i) = (\prod_{a=1}^{A} C_i^a)^{1/a}$ denotes the geometric mean of the vector $\mathbf{C}_i = (C_i^1, ..., C_i^A)$. The sharp null hypothesis of no microbiome composition difference between the subjects under the intervention vs. not can be written as $\mathbf{H_{0,M}}$: for each subject i, $L_i(0) = L_i(1)$.

Choice of test statistic. The scale invariant test statistic suggested by [55] for testing $\mathbf{H_{0,M}}$ is based on the differences $\bar{L}_E^{a,obs} - \bar{L}_C^{a,obs}$, where $\bar{L}_E^{a,obs} = 1/N_E \sum_{i:W_i=1} L_i^a$ is the sample mean of the centered log ratios for subjects under the intervention. Because microbiome data are often sparse (i.e., only a small number of taxa may have different mean abundance), the following test statistic is considered: $T_M = \frac{N_E N_C}{N_E+N_C} \underset{1\leq a\leq A}{max} \frac{(\bar{L}_E^{a,obs} - \bar{L}_C^{a,obs})^2}{\hat{\gamma}_{aa}}$, where $\hat{\gamma}_{aa}$ are the pooled-sample centered log-ratio variances.

Differential abundance

Morton *et al.* recommended to choose reference frames for testing changes in individual bacteria abundance with compositional data [101]. Accordingly, a DACOMP (i.e., differential abundance testing for compositional data) approach has been developed [56]. It is a data-adaptive approach that: 1) identifies a subset of non-differentially abundant (reference) ASVs (R) in a testing dataset, and 2) tests the null of no differential abundance (DA) of the other ASVs (a) "normalized-by-ratio" in a training dataset. First, a bacteria enters the set $R = (r_1, ..., r_F)$ if it has low variance (< 2) and high prevalence ($> 90\%$) (see Supplementary Figures 12 and 13). For the analyses at the ASV level, we chose the variance to be < 3 and the prevalence to be $> 40\%$ as thresholds in order the have at least one reference per subject. Second, using the suggested "normalization-by-ratio" approach, the null hypothesis to be tested for ASV a is that ASV a is not differentially abundant: $\mathbf{H_{0,DA}^{(a \notin R)}} : \frac{C_i^a(0)}{C_i^a(0)+\sum_{f=1}^R C_i^{rf}(0)} = \frac{C_i^a(1)}{C_i^a(1)+\sum_{f=1}^R C_i^{rf}(1)}$,

Choice of test statistic. To test this sharp null hypothesis, we use the LogFold change available in the `dacomp` package with the `Compute.resample.test` function. This function is useful to perform randomization-based inference for differential abundance testing, because it enables to directly incorporate a matrix of hypothetically randomized intervention assignments, which is an appealing feature when researchers work with particular designs. Because we are testing $\mathbf{H_{0,DA}^{(a \notin R)}}$ $||A|| - ||R||$ times at all taxonomic rank levels, we adjust for multiple tests with the method described in the β-diversity analysis section [73].

Partial correlation structure

For our matched intervention and control subjects, we predicted microbial association networks using the Sparse InversE Covariance estimation for Ecological ASsociation Inference (SPIEC-EASI) framework [57] that uses 1) centered log-ratio transformations of the observed ASV counts, $C_i^{a,obs}$, to perform 2) Sparse Inverse Covariance selection (with the graphical lasso method [102]), and finally 3) pick a model based on edge stability (with the StARS method [103]) to obtain a sparse inverse covariance matrix. The non-zero entries of this matrix are proportional to the negative partial correlations among the taxa and form the edge set in an undirected weighted graph $G = (V, E)$. Here, the vertex (or node) set $V = v_1, ..., v_p$ represents the p genera and the edge set $E \subset V \times V$ the possible associations among them. The null hypotheses of no effect of the environmental intervention on the observed genera network associations can be expressed as: $\mathbf{H_{0,N}} : E(0) = E(1)$.

Choice of test statistic. We compare the intervention and control networks with test statistics for the difference in genera associations individually. To generate sampling distributions of the test statistics under $\mathbf{H_{0,N}}$, the intervention and control labels are reassigned 10,000 times to the samples while the matched pair structure is kept. The SPIEC-EASI framework is then re-applied to each permuted data set. This procedure is implemented with the Network Construction and Comparison for Microbiome Data, `NetCoMi`, R package [58]. To adjust for multiple differential association tests, we use the method described in the β-diversity and differential abundance analyses section [73].

Summary stage: interpretation of the results

If the null hypothesis of no difference in the gut microbiome between the matched groups of treated and control units is rejected, that difference warrants further scrutiny to assess whether it can be attributed to the different treatments, assuming the assignment "unconfoundness" assumption holds. We can then report that the gut microbiome composition was or was not altered by the introduction of the environmental intervention. It is important to note that interpretation should be restricted to units that remain in the finite sample after matching (see their detailed characteristics in Supplementary Figures 2-9). The data do not provide direct information for "unmatched" units. Cautiousness regarding extrapolation to units with covariate values beyond values observed in the balanced subset of the data is necessary.

Supplementary materials

The following are available online.

Data availability

The KORA cohort data discussed in the paper is available upon request via the kora.passt portal: `https://epi.helmholtz-muenchen.de`. The code for analysis and visualisation of the data are accessible on the following GitHub public repository: `https://github.com/AliceSommer/Pipeline_Microbiome`. A tutorial to get acquainted with the framework and an open source data is accessible on the following GitHub public repository: `https://github.com/AliceSommer/Causal_Microbiome_Tutorial`.

Author contribution

Conceptualization, A.J.S., M.A.B., A.P., and C.L.M.; methodology, A.J.S., M.A.B., and C.L.M.; writing—original draft preparation, A.J.S.; writing—review and editing, A.J.S., M.A.B., A.P., C.L.M, M.R., D.H., J.C., and H.G.; visualization, A.J.S.; supervision, M.A.B., A.P., and C.L.M; funding acquisition, M.A.B. and A.P. All authors have read and agreed to the published version of the manuscript.

Funding

Research reported in this publication was supported by the Office of the Director, National Institutes of Health under Award Number DP5OD021412 and the John Harvard Distinguished Science Fellows Program within the FAS Division of Science of Harvard University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The KORA study was initiated and financed by the Helmholtz Zentrum München—German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The Technical University of Munich provided funding for the ZIEL Institute for Food & Health and Caroline Ziegler provided technical support for

sample preparation and 16S rRNA gene amplicon sequencing. Microbiota profiling of KORA samples was supported by enable Kompetenzcluster der Ernährungsforschung (No. 01EA1409A) and the European Union Joint Programming Initiative DINAMIC (No. 2815ERA04E, 2815ERA11E).

Acknowledgements

We thank all KORA participants and technical assistants without whose contributions this study could not have been realized. We also thank Stefanie Peschel and Viet Tran for testing the code for the tutorial with the American Gut Data as well as Barak Brill for his support in the DACOMP implementation. The computations in this paper were run on the FASRC Odyssey cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

Conflicts of interests

The authors declare no conflict of interest.

References

1. Wikoff, W. R., Anfora, A. T., Liu, J., Schultz, P. G., Lesley, S. A., Peters, E. C. & Siuzdak, G. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 3698–3703 (2009).
2. Visconti, A., Le Roy, C. I., Rosa, F., Rossi, N., Martin, T. C., Mohney, R. P., Li, W., de Rinaldis, E., Bell, J. T., Venter, J. C., Nelson, K. E., Spector, T. D. & Falchi, M. Interplay between the human gut microbiome and host metabolism. *Nature Communications* **10** (2019).
3. Belkaid, Y. & Hand, T. Role of the Microbiota in Immunity and inflammation Yasmine. *Cell* **157,** 121–141 (2015).
4. David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., Biddinger, S. B., Dutton, R. J. & Turnbaugh, P. J. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505,** 559–563 (2014).
5. David, L. a., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., Erdman, S. E. & Alm, E. J. Host lifestyle affects human microbiota on daily timescales. *Genome Biology* **15,** R89 (2014).
6. Langdon, A., Crook, N. & Dantas, G. The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Medicine* **8** (2016).
7. Thursby, E. & Juge, N. Introduction to the human gut microbiota. *Biochemical Journal* **474,** 1823–1836 (2017).
8. Marchesi, J. R., Adams, D. H., Fava, F., Hermes, G. D. A., Hirschfield, G. M., Hold, G., Quraishi, M. N., Kinross, J., Smidt, H., Tuohy, K. M., Thomas, L. V., Zoetendal, E. G. & Hart, A. The gut microbiota and host health: a new clinical frontier. **65,** 330–339 (2016).
9. Young, V. B. The role of the microbiome in human health and disease: an introduction for clinicians. *BMJ* **356** (2017).
10. Pace, N., Stahl, D., Lane, D. & Olsen, G. in *Advances in Microbial Ecology* (ed K.C., M.) 1–55 (Springer, Boston, MA, 1986).
11. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. & Gordon, J. I. The Human Microbiome Project. *Nature* **449,** 804–810 (2007).
12. Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J. T., Spector, T. D., Clark, A. G. & Ley, R. E. Human genetics shape the gut microbiome. *Cell* **159,** 789–799 (2014).
13. Scholtens, S., Smidt, N., Swertz, M. A., Bakker, S. J., Dotinga, A., Vonk, J. M., van Dijk, F., van Zon, S. K., Wijmenga, C., Wolffenbuttel, B. H. & Stolk, R. P. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *International Journal of Epidemiology* **44,** 1172–1180 (2015).
14. Ikram, M. A., Brusselle, G. G. O., Murad, S. D., van Duijn, C. M., Franco, O. H., Goedegebure, A., Klaver, C. C. W., Nijsten, T. E. C., Peeters, R. P., Stricker, B. H., Tiemeier, H., Uitterlinden, A. G., Vernooij, M. W. & Hofman, A. The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol* **32,** 807–850 (2017).

15. He, Y., Wu, W., Zheng, H.-M., Li, P., McDonald, D., Sheng, H.-F., Chen, M.-X., Chen, Z.-H., Ji, G.-Y., Zheng, Z.-D.-X., Mujagond, P., Chen, X.-J., Rong, Z.-H., Chen, P., Lyu, L.-Y., Wang, X., Wu, C.-B., Yu, N., Xu, Y.-J., Yin, J., Raes, J., Knight, R., Ma, W.-J. & Zhou, H.-W. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nature Medicine* **24,** 1532–1535 (2018).

16. McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y., DeRight Goldasich, L., Dorrestein, P. C., Dunn, R. R., Fahimipour, A. K., Gaffney, J., Gilbert, J. A., Gogul, G., Green, J. L., Hugenholtz, P., Humphrey, G., Huttenhower, C., Jackson, M. A., Janssen, S., Jeste, D. V., Jiang, L., Kelley, S. T., Knights, D., Kosciolek, T., Ladau, J., Leach, J., Marotz, C., Meleshko, D., Melnik, A. V., Metcalf, J. L., Mohimani, H., Montassier, E., Navas-Molina, J., Nguyen, T. T., Peddada, S., Pevzner, P., Pollard, K. S., Rahnavard, G., Robbins-Pianka, A., Sangwan, N., Shorenstein, J., Smarr, L., Song, S. J., Spector, T., Swafford, A. D., Thackray, V. G., Thompson, L. R., Tripathi, A., Vázquez-Baeza, Y., Vrbanac, A., Wischmeyer, P., Wolfe, E., Zhu, Q. & Knight, R. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3** (2018).

17. Holle, R., Happich, M., Löwel, H. & Wichmann, H. KORA - A Research Platform for Population Based Health Research. *Gesundheitswesen (Bundesverband der Ärzte des Öffentlichen Gesundheitsdienstes (Germany))* **67,** 19–25 (2005).

18. Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease. *Curr Opin Gastroenterol* **31,** 69–75 (2015).

19. Rückerl, R., Schneider, A., Breitner, S., Cyrys, J. & Peters, A. Health effects of particulate air pollution: A review of epidemiological evidence. *Inhalation Toxicology* **23,** 555–592 (2011).

20. Huang, C. & Shi, G. Smoking and microbiome in oral, airway, gut and some systemic diseases. *Journal of translational medicine* **17,** 225–225 (2019).

21. Bind, M. C. & Rubin, D. B. Bridging observational studies and randomized experiments by embedding the former in the latter. *Statistical Methods in Medical Research* **28,** 1958–1978 (2019).

22. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **11** (2017).

23. Kaplan, G. G., Dixon, E., Panaccione, R., Fong, A., Chen, L., Szyszkowicz, M., Wheeler, A., MacLean, A., Buie, W. D., Leung, T., Heitman, S. J. & Villeneuve, P. J. Effect of ambient air pollution on the incidence of appendicitis. *Canadian Medical Association Journal* **181,** 591–597 (2009).

24. Ananthakrishnan, A., McGinley, E., Binion, D. & Saeian, K. Ambient air pollution correlates with hospitalizations for inflammatory bowel disease: an ecologic analysis. *Inflamm Bowel Dis* **17,** 1138–45 (2011).

25. Kaplan, G. G., Szyszkowicz, M., Fichna, J., Rowe, B. H., Porada, E., Vincent, R., Madsen, K., Ghosh, S. & Storr, M. Non-specific abdominal pain and air pollution: a novel association. *PLoS One* **7,** 1–8 (2012).

26. Peters, A. Epidemiology: Air pollution and mortality from diabetes mellitus. *Nature Reviews Endocrinology* **8,** 706 (2012).

27. Alderete, T. L., Jones, R. B., Chen, Z., Kim, J. S., Habre, R., Lurmann, F., Gilliland, F. D. & Goran, M. I. Exposure to traffic-related air pollution and the composition of the gut microbiota in overweight and obese adolescents. *Environmental Research* **161,** 472–478 (2018).

28. Liu, T., Chen, X., Xu, Y., Wu, W., Tang, W., Chen, Z., Ji, G., Peng, J., Jiang, Q., Xiao, J., Li, X., Zeng, W., Xu, X., Hu, J., Guo, Y., Zou, F., Du, Q., Zhou, H., He, Y. & Ma, W. Gut microbiota partially mediates the effects of fine particulate matter on type 2 diabetes: Evidence from a population-based epidemiological study. *Environment International* **130** (2019).

29. Bailey, M. J., Naik, N. N., Wild, L. E., Patterson, W. B. & Alderete, T. L. Exposure to air pollutants and the gut microbiota: a potential link between exposure, obesity, and type 2 diabetes. *Gut Microbes* **11,** 1188–1202 (2020).

30. Fouladi, F., Bailey, M. J., Patterson, W. B., Sioda, M., Blakley, I. C., Fodor, A. A., Jones, R. B., Chen, Z., Kim, J. S., Lurmann, F., Martino, C., Knight, R., Gilliland, F. D. & Alderete, T. L. Air pollution exposure is associated with the gut microbiome as revealed by shotgun metagenomic sequencing. *Environment International* **138,** 105604 (2020).

31. Möller, W., Häußinger, K., Winkler-Heil, R., Stahlhofen, W., Meyer, T., Hofmann, W. & Heyder, J. Mucociliary and long-term particle clearance in the airways of healthy nonsmoker subjects. *Journal of Applied Physiology* **97,** 2200–2206 (2004).

32. Beamish, L. A., Osornio-Vargas, A. R. & Wine, E. Air pollution: An environmental factor contributing to intestinal disease. *Journal of Crohn's and Colitis* **5,** 279–286 (2011).

33. Mutlu, E. A., Engen, P. A., Soberanes, S., Urich, D., Forsyth, C. B., Nigdelioglu, R., Chiarella, S. E., Radigan, K. A., Gonzalez, A., Jakate, S., Keshavarzian, A., Budinger, G. S. & Mutlu, G. M. Particulate matter air pollution causes oxidant-mediated increase in gut permeability in mice. *Particle and Fibre Technology* **8,** 19 (2011).

34. Kish, L., Hotte, N., Kaplan, G. G., Vincent, R., Tso, R., Gänzle, M., Rioux, K. P., Thiesen, A., Barkema, H. W., Wine, E. & Madsen, K. L. Environmental particulate matter induces murine intestinal inflammatory responses and alters the gut microbiome. *PLoS One* **8,** 1–15 (2013).

35. Li, R., Navab, K., Hough, G., Daher, N., Zhang, M., Mittelstein, D., Lee, K., Pakbin, P., Saffari, A., Bhetraratana, M., Sulaiman, D., Beebe, T., Wu, L., Jen, N., Wine, E., Tseng, C., Araujo, J., Fogelman, A., Sioutas, C., Navab, M. & Hsiai, T. Effect of exposure to atmospheric ultrafine particles on production of free fatty acids and lipid metabolites in the mouse small intestine. *Environ. Health Perspectives* **123,** 34–41 (2015).

36. Mutlu, E. A., Comba, I. Y., Cho, T., Engen, P. A., Yazıcı, C., Soberanes, S., Hamanaka, R. B., Niğdelioğlu, R., Meliton, A. Y., Ghio, A. J., Budinger, G. S. & Mutlu, G. M. Inhalational exposure to particulate matter air pollution alters the composition of the gut microbiome. *Environmental Pollution* **240,** 817–830 (2018).

37. Wang, W., Zhou, J., Chen, M., Huang, X., Xie, X., Li, W., Cao, Q., Kan, H., Xu, Y. & Ying, Z. Exposure to concentrated ambient PM2.5 alters the composition of gut microbiota in a murine model. *Particle and Fibre Toxicology* **15,** 1–13 (2018).

38. Salim, S. Y., Kaplan, G. G. & Madsen, K. L. Air pollution effects on the gut microbiota. *Gut Microbes* **5,** 215–219 (2014).

39. Calkins, B. M. A meta-analysis of the role of smoking in inflammatory bowel disease. *Digestive Diseases and Sciences* **34,** 1841–1854 (1989).

40. Cosnes, J., Beaugerie, L., Carbonnel, F. & Gendre, J.-P. Smoking cessation and the course of Crohn's disease: An intervention study. *Gastroenterology* **120,** 1093–1099 (2001).

41. Benjamin, J. L., Hedin, C. R., Koutsoumpas, A., Ng, S. C., McCarthy, N. E., Prescott, N. J., Pessoa-Lopes, P., Mathew, C. G., Sanderson, J., Hart, A. L., Kamm, M. A., Knight, S. C., Forbes, A., Stagg, A. J., Lindsay, J. O. & Whelan, K. Smokers with Active Crohn's Disease Have a Clinically Relevant Dysbiosis of the Gastrointestinal Microbiota. *Inflammatory Bowel Diseases* **18,** 1092–1100 (2011).

42. Biedermann, L., Zeitz, J., Mwinyi, J., Sutter-Minder, E., Rehman, A., Ott, S. J., Steurer-Stey, C., Frei, A., Frei, P., Scharl, M., Loessner, M. J., Vavricka, S. R., Fried, M., Schreiber, S., Schuppler, M. & Rogler, G. Smoking cessation induces profound changes in the composition of the intestinal microbiota in humans. *PloS one* **8,** e59260–e59260 (2013).

43. Lee, S. H., Yun, Y., Kim, S. J., Lee, E.-J., Chang, Y., Ryu, S., Shin, H., Kim, H.-L., Kim, H.-N. & Lee, J. H. Association between Cigarette Smoking Status and Composition of Gut Microbiota: Population-Based Cross-Sectional Study. *Journal of clinical medicine* **7,** 282 (2018).

44. Fisher, R. A. *The Design of Experiments* (Edinburgh: Oliver and Boyd, 1935).

45. Bind, M.-A. C. & Rubin, D. B. When possible, report a Fisher-exact P value and display its underlying null randomization distribution. *Proceedings of the National Academy of Sciences* **117,** 19151–19158 (2020).

46. Cochran, W. G. & Rubin, D. B. Controlling Bias in Observational Studies: A Review. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **35,** 417–446 (1973).

47. Rubin, D. B. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics* **29,** 185–203 (1973).

48. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66,** 688–701 (1974).

49. Rubin, D. B. Inference and Missing Data. *Biometrika* **63,** 581–592 (1976).

50. Holland, P. W. Statistics and Causal Inference. *Journal of the American Statistical Association* **81,** 945–960 (1986).

51. Imbens, G. W. & Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press, New York, NY, USA, 2015).

52. Rubin, D. B. For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics* **2,** 808–840 (2008).

53. Willis, A., Bunge, J. & Whitman, T. Improved detection of changes in species richness in high diversity microbial communities. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **66,** 963–977 (2017).

54. Willis, A. D. & Martin, B. D. Estimating diversity in networked ecological communities. *Biostatistics* (2020).

55. Cao, Y., Lin, W. & Li, H. Two-sample tests of high-dimensional means for compositional data. *Biometrika* **105,** 115–132 (2018).

56. Brill, B., Amir, A. & Heller, R. Testing for differential abundance in compositional counts data, with application to microbiome studies. *arXiv.org.* `http://search.proquest.com/docview/2212340160/` (2019).

57. Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J. & Bonneau, R. A. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. **11,** e1004226 (2015).

58. Peschel, S., Müller, C. L., von Mutius, E., Boulesteix, A.-L. & Depner, M. NetCoMi: network construction and comparison for microbiome data in R. *Briefings in Bioinformatics* (2020).

59. Sohn, M. B. & Li, H. Compositional mediation analysis for microbiome studies. *The Annals of Applied Statistics* **13,** 661–681 (2019).

60. Wang, C., Hu, J., Blaser, M. J. & Li, H. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics* **36,** 347–355 (2019).

61. Sazal, M. R., Stebliankin, V., Mathee, K. & Narasimhan, G. Causal Inference in Microbiomes Using Intervention Calculus. *bioRxiv*. `https://www.biorxiv.org/content/early/2020/03/03/2020.02.28.970624.full.pdf` (2020).

62. Wade, K. H. & Hall, L. J. Improving causality in microbiome research: can human genetic epidemiology help? *Wellcome open research* **4,** 199–199 (2020).

63. Hughes, D., Bacigalupe, R., Wang, J., Rühlemann, M., Falony, G., Joossens, M., Vieira-Silva, S., Henckaerts, L., Rymenans, L., Verspecht, C., Ring, S., Franke, A., Wade, K., Timpson, N. & Raes, J. Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nature Microbiology* **5** (2020).

64. Vojinovic, D., Radjabzadeh, D., Kurilshikov, A., Amin, N., Wijmenga, C., Franke, L., Ikram, M., Uitterlinden, A., Zhernakova, A., Fu, J., Kraaij, R. & van Duijn, C. Relationship between gut microbiota and circulating metabolites in population-based cohorts. *Nature Communications* **10,** Article: 5813 (2019).

65. Mishra, A. K. & Müller, C. L. Negative binomial factor regression with application to microbiome data analysis. *arXiv [stat.ML]* (2021).

66. Wasserstein, R. & Lazar, N. The ASA's Statement on p-Values: Context, Process, and Purpose. *American Statistician* **70,** 129–131 (2016).

67. Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. Moving to a World Beyond "p ¡ 0.05". *The American Statistician* **73,** 1–19 (2019).

68. Willis, A. & Bunge, J. Estimating diversity via frequency ratios. *Biometrics* **71,** 1042–1049 (2015).

69. Zhao, N., Chen, J., Carroll, I. m., Ringel-Kulka, T., Epstein, M. p., Zhou, H., Zhou, J. j., Ringel, Y., Li, H. & Wu, M. c. Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *The American Journal of Human Genetics* **96,** 797–807 (2015).

70. Fu, J., Bonder, M. J., Cenit, M. C., Tigchelaar, E. F., Maatman, A., Dekens, J. A. M., Brandsma, E., Marczynska, J., Imhann, F., Weersma, R. K., Franke, L., Poon, T. W., Xavier, R. J., Gevers, D., Hofker, M. H., Wijmenga, C. & Zhernakova, A. The Gut Microbiome Contributes to a Substantial Proportion of the Variation in Blood Lipids. *Circulation research* **117,** 817–824 (2015).

71. He, Y., Wu, W., Wu, S., Zheng, H.-M., Li, P., Sheng, H.-F., Chen, M.-X., Chen, Z.-H., Ji, G.-Y., Zheng, Z.-D.-X., Mujagond, P., Chen, X.-J., Rong, Z.-H., Chen, P., Lyu, L.-Y., Wang, X., Xu, J.-B., Wu, C.-B., Yu, N., Xu, Y.-J., Yin, J., Raes, J., Ma, W.-J. & Zhou, H.-W. Linking gut microbiota, metabolic syndrome and economic status based on a population-level analysis. *Microbiome* **6,** 172 (2018).

72. Rubin, D. B. More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine* **17,** 371–385 (1998).

73. Lee, J. J., Forastiere, L., Miratrix, L. & Pillai, N. S. More powerful multiple testing in randomized experiments with non-compliance. *Statistica Sinica* **27,** 1319–1345 (2017).

74. Rosenbaum, P. R. *Design of Observational Studies* (Springer, New-York, 2010).

75. Heckman, J. J., Ichimura, H. & Todd, P. Matching as an econometric evaluation estimator. *Review of Economic Studies* **65,** 261–294 (1998).

76. Rubin, D. B. Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology* **2,** 169–188 (2001).

77. Wu, X., Braun, D., Schwartz, J., Kioumourtzoglou, M. A. & Dominici, F. Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. *Science advances* **6,** eaba5692 (2020).

78. Sun, K., Liu, J. & Ning, G. Active Smoking and Risk of Metabolic Syndrome: A Meta-Analysis of Prospective Studies. *PLoS ONE* **7,** e47791 (2012).

79. Breuninger, T. A., Riedl, A., Wawro, N., Rathmann, W., Strauch, K., Quante, A., Peters, A., Thorand, B., Meisinger, C., Linseisen, J. & et al. Differential associations between diet and prediabetes or diabetes in the KORA FF4 study. *Journal of Nutritional Science* **7,** e34 (2018).

80. Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Applied and environmental microbiology* **63** (1997).

81. Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M. & Glöckner, F. O. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research* **41** (2013).

82. Callahan, B. J., Sankaran, K., Fukuyama, J. A., Mcmurdie, P. J. & Holmes, S. P. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses [version 1; referees: 2 approved]. *F1000Research* **5** (2016).

83. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glöckner, F. O. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research* **41,** D590 (2013).

84. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology* **73,** 5261 (2007).

85. Edgar, R. C. & Valencia, A. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34,** 2371–2375 (2018).

86. Wright, E. S. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *The R Journal* **8,** 352–359 (2016).

87. Studier, J. A. & Keppler, K. J. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular biology and evolution* **5,** 729 (1988).

88. Rubin, D. B. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine* **26,** 20–36 (2007).

89. Micali, S. & Vazirani, V. V. *An Algoithm for Finding Maximum Matching in General Graphs* in *Proceedings of the 21st Annual Symposium on Foundations of Computer Science* (IEEE Computer Society, Washington, DC, USA, 1980), 17–27.

90. Singh, R. K., Chang, H.-W., Yan, D., Lee, K. M., Ucmak, D., Wong, K., Abrouk, M., Farahnik, B., Nakamura, M., Zhu, T. H., Bhutani, T. & Liao, W. Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine* **15,** 73 (2017).

91. Johnson, A. J., Zheng, J. J., Kang, J. W., Saboe, A., Knights, D. & Zivkovic, A. M. A Guide to Diet-Microbiome Study Design. *Frontiers in Nutrition* **7,** 79 (2020).

92. Rubin, D. B. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association* **75,** 591–593 (1980).

93. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* **8** (2017).

94. Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **27,** 379–423 (1948).

95. Basharin, G. P. On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables. *Theory of Probability and its Applications* **4,** 333 (1959).

96. Brillinger, D. R., Jones, L. V. & Tukey, J. W. *The Role of Statistics in Weather Resources Management* in *The Management of Weather Resources* **2** (U.S. Government Printing Office, Washington D.C., USA, 1978), 25.

97. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology* **71,** 8228–8235 (2005).

98. Aitchison, J. (*The statistical analysis of compositional data* (Blackburn Press, Caldwell, N.J., 2003).

99. Gower, J. C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **27,** 857–871 (1971).

100. Lin, X. Variance Component Testing in Generalised Linear Models with Random Effects. *Biometrika* **84,** 309–326 (1997).

101. Morton, J., Marotz, C., Washburne, A., Silverman, J., Zaramela, L., Edlund, A., Zengler, K. & Knight, R. Establishing microbial composition measurement standards with reference frames. *Nature Communications* **10,** 1–11 (2019).

102. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9,** 432–441 (2008).

103. Liu, H., Roeder, K. & Wasserman, L. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. *Adv Neural Inf Process Syst* **24,** 1432–1440 (2010).