Assessing Confounding in Case-Control Microbiome Analyses

M.Sc. Thesis

Presented to the Faculty of Biosciences
of the Ruprecht-Karls-Universität Heidelberg

Morgan Essex

29 August 2018

Declaration

This thesis was conducted and written at the European Molecular Biology Laboratory in Heidelberg, Germany during the period of March to September 2018 under the supervision of Dr. Georg Zeller.

1st Examiner: Prof. Dr. Ursula Kummer Institute: COS

2nd Examiner: Dr. Frederik Graw Institute: BioQuant

I hereby declare that I performed all work and wrote this thesis independently, under supervision, and that I used no sources and aids other than those indicated throughout the thesis.

Name: _____ Signature: _____

Acknowledgements

I would like to thank Georg and the entire Zeller team for making my time at EMBL transformative. My learning curve was steep and I had a great deal of self-doubt at the beginning, but the infrastructure, expertise, and camaraderie of the group truly exceeded my expectations. Thanks to each and every one of you for making it fun to go to work every day and challenge myself.

Georg, I learned so much from you. I specifically sought your knowledge of statistics and programming, but it is your general commitment to excellence that has left the strongest impression. I have the utmost respect for your critical eye and attention to detail. Thank you for your mentorship and confidence.

Konrad and Jakob, I feel especially lucky to have worked with you on SIAMCAT. In addition to best coding practices, I learned a tremendous amount about the spirit of (open) scientific research and the microbiome field. Thank you both for your kindness and patience as I learned R and especially Git.

It has been a privilege and a pleasure, and I will miss you all very much.

Abstract

The human gut microbiome has garnered enormous scientific attention in recent years due to several metagenome-wide association studies (MWAS) linking it to various disease phenotypes. While these insights have allowed the field to rapidly advance and expand, large-scale population and meta-studies have illuminated the limitations of such analyses. The growing number of genetic, lifestyle, and environmental factors known to modulate the gut microbiome are poorly understood and represent latent sources of variation which can lead to spurious correlation or biased effect sizes in MWAS. This phenomenon, known as confounding, is particularly likely to result in misleading conclusions considering that most MWAS lack statistical power. This work attempts to fill a gap and help microbiome researchers confront the shortcomings associated with MWAS by capturing the confounding potential of metadata variables associated with case-control metagenomic data. Stratified non-parametric hypothesis testing and generalized linear model approaches that enable analyses of variance and likelihood ratio testing were implemented and visualized as part of a modular R package. These methods were benchmarked using two previously analyzed datasets, one known to be confounded by drug treatment, and were able to corroborate published results while also contributing novel perspectives. It is anticipated that this work will provide researchers with some additional awareness into the heterogeneous factors determining the case-control distribution of their metagenomic data, currently an unmet need in the microbiome field. Armed with such knowledge, relevant covariates could be included in higher-order predictive models or used as stratification factors for post-hoc statistical analyses, thus helping to disentangle confounding effects from those of direct biological interest.

Contents

Introduction

In a January 2000 interview, the late Stephen Hawking remarked that the 21st century would be the century of complexity [1]. As a physicist, his comment alluded to the nearly complete understanding of the fundamental laws of matter that his discipline assembled in the 20th century, while also expressing the hope that higher-level, emergent phenomena could too be understood in a mechanistic light someday.

The biological sciences were particularly poised to adapt the paradigm shift to systems-level thinking that Hawking subtly articulated. Historically a data-poor discipline, technological advances have enabled a plethora of novel biological data types and effectively expanded the landscape of viable scientific inquiry. Just two years prior to Hawking's remark, one of those new data types, metagenomics, first appeared in the scientific literature [2].

1.1 A Brief History of Metagenomics

Metagenomics is defined as the process of collecting and sequencing genetic material from the members of a microbial community, also known as a microbiome [3]. Microbiota are found practically everywhere, from the soil and sea to the bodies of insects, animals, and humans alike. They are complex, dynamic ecosystems, comprising bacteria, archaea, fungi, and viruses; however, the bulk of scientific research is focused on human-associated bacterial communities, which predominate over others.

The beginning of metagenomics can be traced all the way back to the 1960s, when a scientist named Carl Woese discovered the archaea domain and redefined the phylogenetic tree into the rRNA-based, 3-domain version which remains today [4]. Woese sought an empirical way to relate microbial species to one another, instead of the rudimentary morphological characteristics used by microbiologists at the time, so he began systematically sequencing bacterial species' 16S rRNA gene—an ubiquitous 1.5 Kbp sequence with highly conserved regions as well as regions that vary over evolutionary time, making it an ideal phylogenetic marker gene [5].

Earlier methods to study microbial communities were culture-dependent, meaning only species that could be reliably grown in a lab, such as *E. coli*, could be studied. With Woese's 16S-based molecular phylogeny, however, a microbiome could suddenly be described by the abundances of different 16S sequences. Novel sequences belonging to previously unidentified members could at least be mapped to a location on the phylogenetic tree and understood in relation to known species. Of the roughly 100 bacterial phyla that currently exist, about 80 of them have never been cultured [6].

Today, DNA-based culture-independent methods such as high-throughput amplicon sequencing are widely used to investigate biodiversity in a variety of environmental samples. Raw 16S sequencing reads are binned into Operational Taxonomic Units (OTUs)

sharing high (97-99%) sequence identity—presumed to belong to the same genus—before further characterization using bioinformatic tools and curated databases. Though it revolutionized microbiology, amplicon sequencing is not without its shortcomings. Specifically, it cannot be used to study the functional capacity of a metagenome, resolve bacterial strain-level differences, or assemble genomes without culturing.

The advent of next generation sequencing around the year 2000 was able to address these challenges. Whole metagenome shotgun sequencing (WMS) amplifies random genetic material instead of a specific amplicon. The mixed community DNA is sheared and sequenced directly, then computationally assembled into larger genomic fragments which can be profiled by comparison to reference genes or genomes [7]. Similar to amplicon sequencing, the end result is a large table of community member units at a given resolution (OTUs, species, or genes) and their relative abundances or counts in a sample.

WMS is expensive and presents a more difficult bioinformatic task for researchers, but the ability to functionally characterize microbiota has opened exciting avenues for the nascent metagenomics field, particularly the sub-field focused on the human microbiome. Increasingly, WMS analysis is paired with other high-throughput data types to assess the actual functional and metabolic activity of a microbiome in addition to its metagenome—a strategy referred to as multi-omics [8].

Such integrative approaches are indicative of a microbiology reconfigured for the century of complexity; microbiologists have gone from understanding microbes in an isolated physiological or environmental context to understanding them in an ecological one.

1.2 The Human Gut Microbiome and Disease

Of the several microbiota in and on the human body, the largest and perhaps most medically relevant is the gastrointestinal (gut) microbiome. Large-scale efforts such as the NIH-led Human Microbiome Project (HMP) [9] and the European Metagenomics of the Human Intestinal Tract (MetaHIT) Consortium [10, 11] were initiated about a decade ago to characterize the dominant taxa present in the guts of industrialized populations.

Drawing from these efforts and others, researchers in China created an integrated gene catalog (IGC) of 9.9 million non-redundant genes from approximately 1,200 fecal shotgun metagenomic samples [11]. On average, each metagenome contained 800,000 genes; for reference, the human genome contains roughly 23,000 genes—about 34 times less. Any one metagenome had approximately 250,000 genes (31%) in common with any other metagenome. This modest overlap is consistent with previous results that the between-subject variation is higher, in both taxonomic and functional composition, than within-subject variation over time [12].

Nevertheless, some general patterns of gut biodiversity appear to be informative and reproducible. The enterotype model stratifies the population into three approximate clusters named after the dominant genera present in each: *Bacteroides, Prevotella,* and *Ruminococcus* [13]. Each enterotype exhibits significant functional differences from the other two and captures at least some inter-individual variation, positing that, by definition, the groupings represent non-random community compositions that remain relatively stable over time [14].

While enterotypes are rather coarse-grained, the drive to find informative groupings is a natural step toward understanding the clinical relevance of the gut microbiome. Indeed, a large portion of the literature consists of observational studies attempting to

correlate the community composition of the gut to various host phenotypes, most often disease states. These metagenome-wide association studies (MWAS) are analogous to genome-wide association studies (GWAS) and make use of case-control study designs in order to elucidate microbiome changes associated with human disease.

To date, there are several diverse illnesses associated with compositional changes in the gut microbiome. While it is somewhat intuitive that obesity, colorectal cancer (CRC), and inflammatory bowel disease (IBD) occur with an altered microbiome state, it is perhaps less expected that the phenomenon is also observed in neurological, metabolic, and cardiovascular diseases [15]. The increased incidence of these multifactorial diseases has happened over a relatively short time in the grand scheme of human evolution, suggesting it is unlikely due to human genetics alone [16]. Environmental factors are certainly playing a role; human populations have experienced unprecedented industrialization, migration, and urbanization the past two centuries, and gut bacteria, in contrast to their human hosts, have much shorter generation times and harbor many more genes that could adapt to such environmental pressures, perhaps at the expense of host health [17].

The open question in most cases is whether an aberrant microbiome state, broadly referred to as dysbiosis, is the cause or the result of host disease. In order for a dysbiotic microbiome to cause disease, dysbiosis may be narrowly defined as a microbial community state that is not only statistically associated with a disease, but also functionally contributes to the etiology, diagnosis, or treatment of the disease [18].

1.3 Gut Microbiome Covariates

Perhaps unsurprising in a relatively young field, MWAS investigating the same disease have reported varying results [19, 20]. Meta-analyses, which pool results in an attempt to distinguish disease-specific signals, have attempted to consolidate findings from earlier studies. Recently, a meta-analysis of 28 MWAS spanning ten diseases found that some diseases, such as CRC, consistently display an enrichment of pathogenic bacteria, while others, such as IBD, manifest as a depletion of health-associated bacteria in cases when compared to controls [21]. The authors also found that nearly 51% of the genera-level associations in a given MWAS are associated with more than one disease.

This finding hints at some of the major problems in MWAS, which meta-analyses strive to address. These include a lack of sufficient statistical power, technical or batch effects, and poor knowledge of factors influencing the microbiome, i.e. covariates. Covariates may be technical or biological in nature, and by definition are separate from the factor of interest in a given study. Some of these issues will be partially addressed as the field matures, but a deeper understanding of the host and environmental impact on microbiome variation will be a critical hurdle to overcome in order to guide future longitudinal and perturbation study designs. A 2016 population-level analysis of nearly 4,000 individuals posited that the combined effect size of known covariates on microbiome composition is only in the range of 10-15% [22].

Covariates can be empirically separated into intrinsic and extrinsic factors, the latter of which may be further sub-divided into host-intrinsic, host-extrinsic, and environmental factors (Fig. 1, adapted with permission from Schmidt *et al.* [23]). The gut microbiome undergoes pronounced shifts throughout the lifetime of an individual, such as the rapid colonization and increase in diversity during the first five years of life and the decline in microbial diversity later in life [24]. Additional longitudinal and perturbation studies will

be uniquely able to help understand these intrinsic microbiome dynamics.

The influence of microbiome-extrinsic factors will be especially challenging to tease apart, since many overlap and associate with one another. Body mass index (BMI), for example, has genetic and lifestyle components [25]. Studies in mono- and di-zygotic twins have revealed how host genetics may shape heritability of certain taxa, such as lactose-metabolizing *Bifidobacterium* in individuals who are lactase non-persisters, meaning their ancestors did not evolve the persistent ability to produce lactase in the upper GI tract after weaning [26]. This appears to be a rare occurrence, however; a more recent study estimates that the average taxa heritability of the same cohort was about 2% [27]. Environmental factors have been so far understudied, but available data point to an effect of geographic region on microbiome composition, for example, visible at the phyla, species, and strain levels [28].

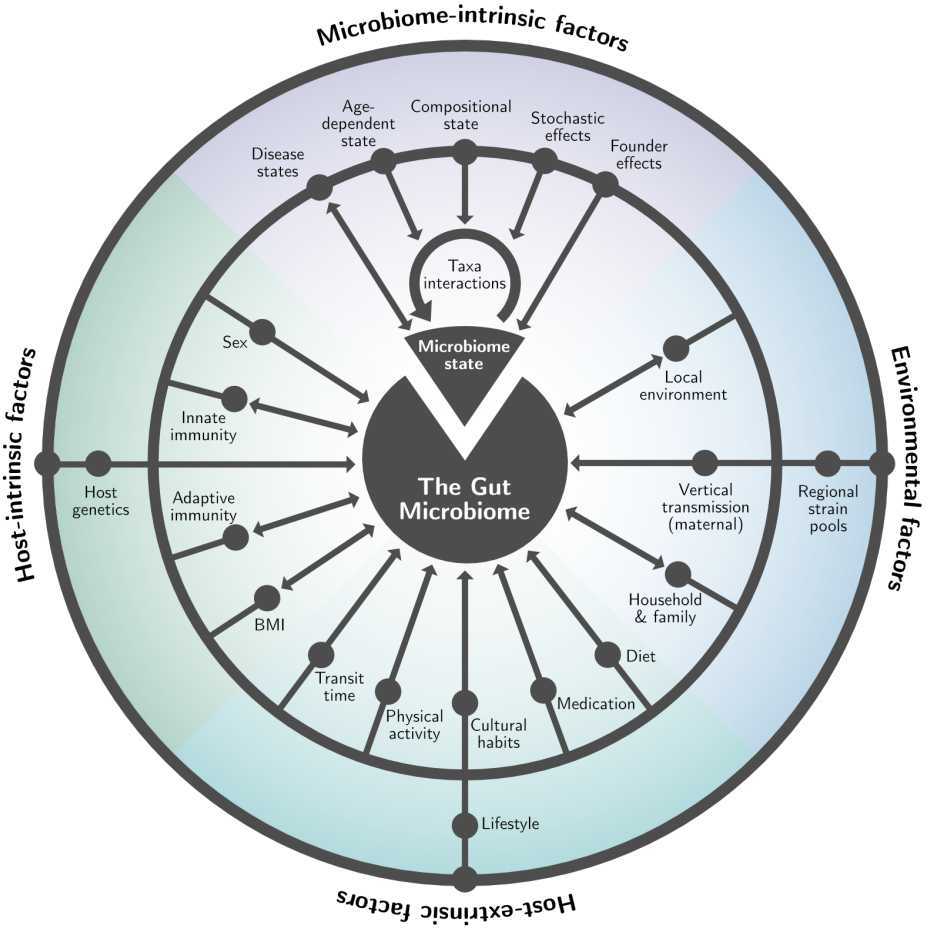

Figure 1: The complex interplay of known gut microbiome covariates, shown here empirically separated into microbiome-intrinsic and microbiome-extrinsic (host-intrinsic, host-extrinsic, environmental) factors, which overlap and associate with one another. Adapted with permission from Schmidt *et al.* [23].

Perhaps one of the most obvious covariates in the microbiome-extrinsic category is medication. It is widely known that oral antibiotics can disrupt gut microbiome stability,

prompting infection by opportunistic pathogens such as *Clostridium difficile* following a round of treatment; however, a context-dependent understanding of individualized responses and recoveries from antibiotic treatment remains elusive [29]. In addition to antibiotics, a study from earlier this year performed an *in vitro* screen of nearly 1,200 marketed drugs from a variety of non-antibiotic classes against 40 representative bacterial strains (covering 78 and 60% of the median relative abundance of the human gut microbiome at genus and species level, respectively) and found that 24% of the drugs with human targets inhibited the growth of at least one strain [30].

While this may be exciting news for novel antibiotic therapies, the immediate concern with a lack of knowledge regarding microbiome covariates is that MWAS—already statistically under-powered in most cases—are reaching erroneous conclusions as a result. This phenomenon is known as confounding, and the covariates which elicit it by exerting influences on both independent and dependent variables in a study (the microbiome composition and the disease state, respectively) are referred to as confounders.

1.4 Defining, Identifying, and Handling Confounders

To confound is to confuse, which is appropriate given that the term itself is somewhat enigmatic and eludes a simple definition in formal mathematical terms. In practice, confounding refers to spurious correlation or biased effect size that results when two groups being compared are not actually comparable [31]. For example, it is well-established that the gut microbiome becomes less taxonomically diverse and compositionally stable as an individual ages [24]. If an observational study in which the cases were significantly older than the controls were to report that a certain species was significantly associated with a disease, its results might be confounded by age.

This need not be a thought experiment; increasingly the MWAS literature is reporting such conclusions. One study investigating microbial associations with HIV-1 found a shift from *Bacteroides* to *Prevotella* predominance in patients following infection, suggesting a disease association, but this became insignificant once subjects were stratified according to sexual orientation [20]. This perceived shift instead reflected underlying lifestyle differences in the study participants which confounded its association with HIV-1.

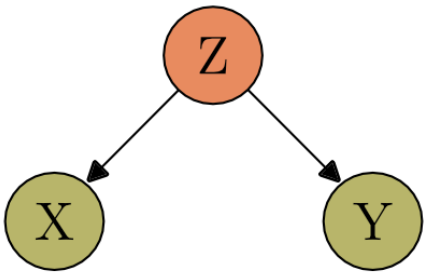

Figure 2: If X represents the independent variable (e.g. gut microbiome state) and Y the dependent variable (e.g. disease state) in a study, Z is a "lurking variable" (i.e. confounder) that influences both X and Y.

In both examples, a "lurking variable" (Fig. 2) was able to explain variation in the microbiome better than the disease state. This is the essence of confounding, and it not unique to MWAS. Epidemiology also employs observational case-control study designs, but it does so in order to explicitly model the relationship between exposure to an *a priori* known risk factor and the incidence of a disease in a subset of the population. Perfect evidence of such a relationships is, however, unattainable through observational studies; subjects can never be exposed and not exposed to the risk factor at the same time. Epidemiologists and clinicians use statistical methods to correct for this limitation—to "control for" confounding—including subject

matching, post-hoc stratification by and standardization of the confounding variable, and multiple regression [32].

Many of these tactics share the primary intent of rendering cases and controls more comparable by decreasing within-group variation. Randomized control studies rank higher than case-control studies on the evidence hierarchy because they deal with the inherent and enormous variation of biological subjects implicitly, in their randomized design. However, as MWAS necessitate a case-control framework, explicit statistical methods offer the best solution.

One of the most high-profile examples of the efficacy of these methods came from a study in 2015 that will be addressed in more detail in Section 2.3. Briefly, Forslund *et al.* re-analyzed two conflicting studies examining type II diabetes (T2D) [33, 34], neither of which stratified patients according to their drug regimens during analysis. The meta-analysis was able to reconcile the divergent results by stratifying according to metformin treatment [19]. They could differentiate the gut signatures of T2D patients that had or had not received metformin treatment; it was in fact much easier to differentiate the metformin-stratified T2D groups from one another than it was to differentiate either of them from the non-diabetic control group.

In order to correct for confounding, potential confounders must be identified. Given that so little microbiome variation is currently explained by known covariates, this is indeed the more immediate issue related to confounding. In the population-level analysis by Falony *et al.* that estimated the amount of microbiome variation explained by covariates to be between 10-15%, 63% of the genus-level abundance variation that could be explained specifically involved medication [22]. Despite this, MWAS frequently fail to record medication of subjects or report it in published metadata, making it difficult to study the effects of medication on gut microbiome variation.

Covariates that are routinely included in MWAS across diseases, however, include age, biological sex (usually denoted by gender), BMI, and library size, which refers to the metagenomic library constructed by WMS and reflects a technical process. Library size is known to influence the composition of the metagenomic OTU table; samples sequenced to a lower depth (smaller library size) are less likely to pick up lower abundance species, which could confound inferred associations [35]. Other covariates may reflect more than one source of variation or serve as a proxy for something else that is poorly defined. BMI, for example, correlates strongly with obesity, a frequently-observed comorbidity in many diseases that is itself a complex phenomenon. In a meta-analysis, the study is a coarse covariate that may reflect technical, geographic, or cultural (diet, lifestyle) factors.

The literature has plainly demonstrated that MWAS are vulnerable to confounding. The open question therefore concerns the merits of statistically controlling for perceived confounders and, more importantly, how to discern them among available covariates. So far, investigations into confounding have been conducted by individual researchers conducting informed post-hoc or meta-analyses. Thus, there is a gap between the urgent need to consider confounders in MWAS and available user-friendly tools that are tailored to the unique statistical difficulties of microbiome research, which include high-dimensionality and sparse compositional data. MaAsLin is a multivariate framework to discover associations between clinical covariates and metagenomic data [36]; however, it is standalone tool (rather than e.g. a package) that has not been published or peer-reviewed. Furthermore, it employs linear modeling approaches that make questionable assumptions about the metagenomic data.

1.5 A Package for Robust Statistical Inference

As MaAsLin and other metagenomic tools address, the inherently high-dimensional nature of metagenomic data ultimately demands multivariate analysis. Supervised machine learning approaches enable discovery of complex microbial signatures related to host phenotype, as well as construction of predictive models based on those signatures [37], and is thus well-suited to the task. To extend these capabilities to the metagenomics community at large, Zeller *et al.* [38] initially developed `SIAMCAT`[1], a full pipeline for the **S**tatistical **I**nference of **A**ssociations between **M**icrobial **C**ommunities **A**nd host pheno**T**ypes. It is implemented as a modular R package, developed using Git for version control, and available on Bioconductor as of April 2018.

SIAMCAT takes an OTU table of assembled metagenomic relative abundance or count data (features), metadata (covariates), and a binary label file as input (Fig. 3). It is thus suitable for any type of metagenomic data associated with a binary label, with case-control gut microbiome data representing just one possible example. It begins with standard preprocessing steps such as data filtering and normalization before proceeding through a suite of different modules, arranged in a branching structure, depending on the intent of the researcher.

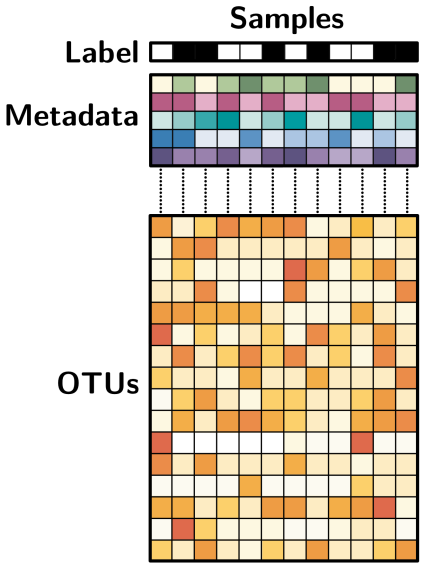

Figure 3: SIAMCAT takes three pieces of information as input: an OTU table of metagenomic data (features), metadata (covariates), and a binary label. One column represents one individual.

The main branch of SIAMCAT (Fig. 4) provides multivariable logistic regression modeling using penalized LASSO models and k-fold cross-validation to maximize interpretability and robustness, respectively. Customizable options might include different supervised learning algorithms or types of cross-validation. Separate from the main branch and having no impact on the construction of the main model is the association testing branch, which was previously a standalone module to perform univariate hypothesis testing on a subset of features. Univariate testing nicely complements multivariable machine learning approaches by providing a higher-resolution look at features that are statistically most closely associated with the binary label. Importantly, these features are not necessarily the same as those selected by the penalized regression, though there is often some overlap.

1.6 Aims

My research aim was to implement statistical measures to quantify the confounding potential of recorded covariates within the univariate framework of SIAMCAT, which is already amenable to in-depth analysis. I approached the problem from multiple angles which fell into two modules for developmental rather than conceptual reasons (Fig. 4).

[1] [39] SIAMCAT R package version 1.0.1, https://bioconductor.org/packages/SIAMCAT/

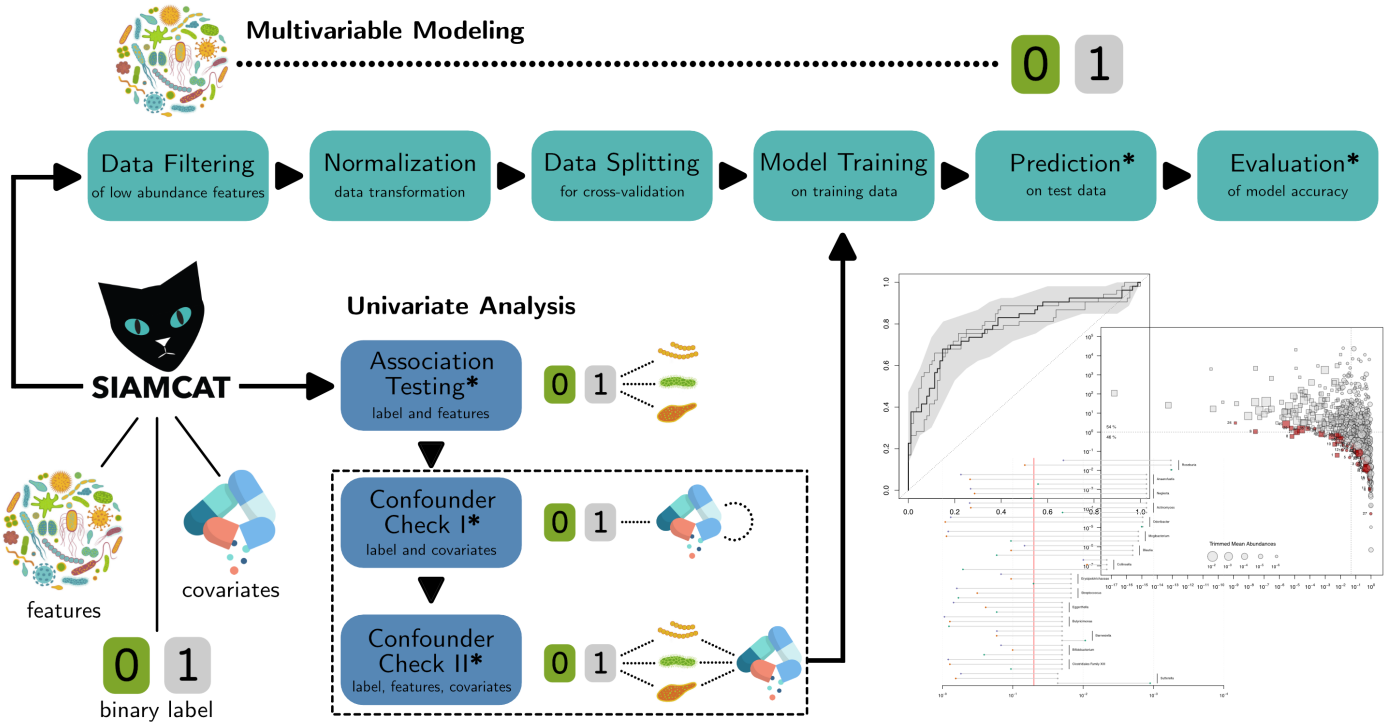

Figure 4: SIAMCAT: Statistical Inference of Associations between Microbial Communities And host phenoTypes [39]. The main branch (horizontal, top) performs multivariable logistic regression on the features and the binary label. Previously the association testing branch (vertical, blue) was a standalone module performing univariate statistical tests between single features and the label. My work (boxed) extends this branch to include two confounder check modules to additionally examine relationships with covariates. Relevant covariates may be incorporated as features along with the OTU table for further testing or construction of the main LASSO model. Modules with an asterisk produce data visualizations in PDF format (examples shown bottom right).

Conceptually, the methods chosen seek to answer whether there are associations with covariates present, what the possible sources of variation defining the case-control distribution and the species abundances are, and whether any covariates likely to impede confident inference of the label by a metagenomic classifier. Practically, the first question is addressed in Confounder Check I, which focuses on covariates and how they relate to one another as well as to the binary label, and the other two are tackled in Confounder Check II, which additionally models relationships involving species abundance data. Hence the distinction between these modules reflects differences in factors considered, i.e. input variables, more than differences in approach.

The nature of this project required primarily method development and implementation, which is summarized at the beginning of the Results section. Concepts given a more thorough treatment in the Methods section are referenced. To benchmark my methods, I applied them to two previously-analyzed case studies: a five-country colorectal cancer (CRC) meta-analysis (Wirbel *et al.*, under review), and Forslund *et al.*'s meta-analysis examining metformin status in T2D [19]. Both studies suggest the selected methods were effective at capturing confounding potential, as results were in line with expectations; the CRC studies were robust and did not appear to be confounded, while the T2D meta-analysis revealed tremendous study effects as well as a unique confounding signature for metformin, a known confounder.

Results

2.1 Extension of the Univariate Analysis Branch

Figure 5 displays the questions that guided my approach and outlines the capability of the univariate analysis framework. How features (species abundance data) are associated with the label (disease status) is addressed by the previously established Association Testing module (Fig. 4); my work specifically involved questions 2-7. The methods will be covered here with references to the motivating questions in Figure 5.

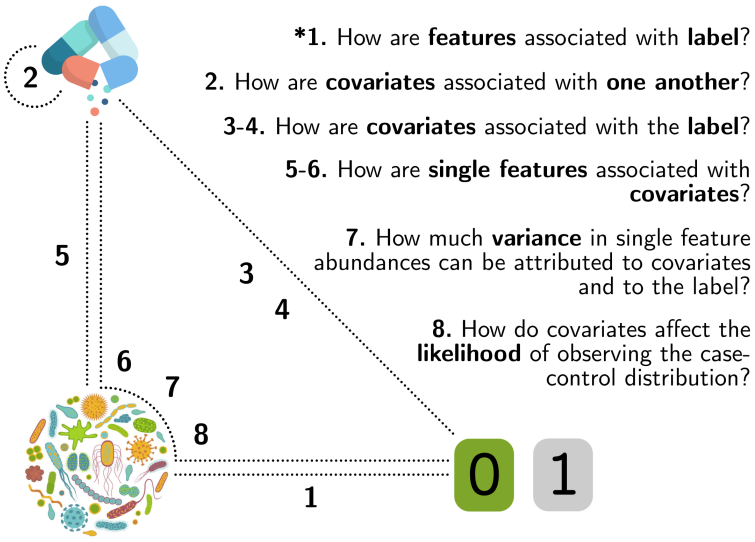

Figure 5: Research questions that guided my approach. **1** is carried out by the Association Testing module examining associations between features (microbial relative abundances) and the label (disease status), **2-4** in Confounder Check I examining associations between covariates and the label, and **5-8** in Confounder Check II which employs GLMs to examine relationships between all three and performs analyses of variance. * = not implemented as part of this work.

Associations Between Covariates (2)

As previously mentioned, some covariates such as country may be a proxy of many different factors, e.g. technical process variation that might also be reflected in library size. To examine the relationship of the covariates with one another in order to better characterize them, I sought a broadly applicable correlation measurement that does not assume a particular type of relationship, e.g. linear. The entropy of a random variable is a measure of the information it contains, and the information unique to a random variable is defined as its conditional entropy (Section 4.1). A variable completely dependent on (or derived from) another will have a conditional entropy of zero, thus in addition to quantifying the possible overlapping sources of variation captured by covariates, the conditional entropy is an appropriate criterion to identify covariates not suitable for further analysis, such as those from which the label is derived. See e.g. Figure 22 in the Appendix.

Label Association with Covariates (3-4)

A covariate that is not evenly distributed among cases and controls is more likely to confound than one that is evenly distributed. To check if factors were disproportionately associated with one group, I implemented simple descriptive statistical plots and hypothesis testing. For categorical variables, such as gender, this included stacked barplots and contingency tables, as well as the Fisher exact test. Boxplots and histograms as well as the non-parametric Wilcoxon rank-sum test (Section 4.2) were implemented for continuous variables, such as age.

The association testing module of SIAMCAT plots area under the receiver operating characteristic curve (AU-ROC, Section 4.3.2) values as another measure of association. Specifically, it quantifies the ability of individual features to predict the label. I adapted this functionality for covariates in Confounder Check I, while also going a bit further and building a generalized linear model (GLM, Sections 4.3 and 4.3.2 for logistic regression specifically) for each covariate. This enabled the logistic regression coefficient, its significance, and the AU-ROC values to be plotted for variables that do not have a conditional entropy of zero.

Single Feature Associations with Covariates (5-6)

In order for a covariate to confound, it would need to associate with the microbial relative abundances (features) as well as with the disease status (label), as previously shown in Figure 2. Intuitively, this might manifest as a feature being differentially abundant for certain values of a covariate, e.g. enriched in older samples. Thus these methods are sometimes referred to as differential abundance testing. They are non-parametric, i.e. they make less assumptions about the shape of the distributions being compared than parametric methods (Section 4.2).

To directly test for covariate-associated differential abundances, I selected the Kruskal-Wallis test (Section 4.2), sometimes called a non-parametric analysis of variance (ANOVA). Continuous variables such as age were discretized into quartiles per the requirements of ANOVA. I stratified this analysis according to disease status in order to remove variance in the feature abundances associated with the disease status and isolate the interaction with the covariate. This stratification procedure is also referred to as blocking (Section 4.3.1). As a complementary indirect method to quantify the associations between features and covariates, I adapted the Wilcoxon rank-sum test (Section 4.2) to test for differential feature abundances between cases and controls while blocking iteratively for each covariate.

The unstratified Wilcoxon rank-sum test is used by the Association Testing module, thus by comparing the significance of the results with and without stratification provides an estimation of how confounding potential. Significant Kruskal-Wallis test results were visualized using a bipartite network-type plot to provide accessible interpretation (e.g. Fig. 9 in Results Section 2.2 and Fig. 24 in the Appendix), while the stratified and unstratified Wilcoxon rank-sum test results were plotted together for the top 25 features most strongly associated with the label to highlight and quantify the effect of the covariate (e.g. Fig. 14 in Results Section 2.3 and Fig. 23 in the Appendix).

Variance Explained by Covariates and Disease Status (7)

A species whose relative abundance variation is better explained by a given covariate than by disease status might be confounded. Through the Kruskal-Wallis and Wilcoxon rank-sum tests do employ stratification to perform some estimation of the variance that may be attributed to covariates, an ANOVA allows explicit quantification by invoking a formal model. Here I sought to quantify the amount of feature variance explained by covariates as well as by the disease status in order to compare them and estimate the confounding potential of covariates.

I converted the relative abundances to ranks in order to avoid normality assumptions that are not appropriate for metagenomic data, then used a linear model for each species to estimate the ranks using the label and a single covariate at a time. This is equivalent to a two-way ANOVA on ranks (Section 4.3.1). These models are subjected to conservative type III (adjusted) sum of squares partitioning and the F-test for each model term (label and covariate) to quantify whether the variance explained by each is statistically significant. To make these results more comparable between different models, i.e. species, I calculate the ratio of variance explained by the covariate over the variance explained by the label. This ratio of F-ratios—a "meta" F-ratio—is plotted against the FDR-adjusted significance of the F-test result for each covariate (e.g. Fig. 11 in Results Section 2.2 and Fig. 16 in Results Section 2.3).

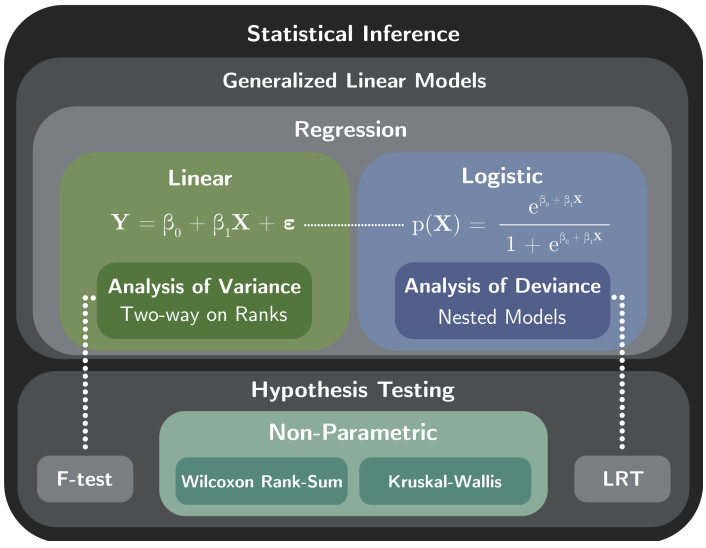

Figure 6: The two components of statistical inference used in this work include modeling and hypothesis testing. I make use of flexible generalized linear models (GLMs) and non-parametric tests to relax distribution assumptions. Logistic regression in this framework is a linear combination of predictors to which the sigmoid function is applied in order to obtain classification probabilities. Linear and logistic regression are amenable to analysis of variance and analysis of deviance, respectively, which make use of the F-test and likelihood ratio test (LRT), respectively, to evaluate the contribution of parameters in the models, i.e. covariates, feature abundances, and/or the disease status.

Logistic Regression and Analyses of Deviance (8)

Within the SIAMCAT framework, confounders are ultimately of interest if they impede the ability to confidently infer associations between metagenomic data and the case-control label. To explicity model a classification setting that might identify potential confounders, I elected to use logistic regression GLMs to predict the label using single feature abundances and the three covariates with the strongest relation to the label according to the conditional entropy analysis.

An analysis of deviance (Section 4.3.2) is analogous to an ANOVA for logistic GLMs (Fig. 6). It relies on nested models and employs the likelihood ratio test (LRT) to quantify whether

the deviance from expected probabilities is significantly less with the addition of a new parameter to the model, i.e. incorporation of species abundances or covariate information.

The decrease in model deviance per feature is plotted with the significance of the coefficients as well as the AU-ROC analysis scores for the single feature and full (feature abundances + three covariates) models (e.g. Fig. 12 in Results Section 2.2 and Fig. 18 in Results Section 2.3). The AU-ROC scores are universally comparable between very different classification models, i.e. these simple models may be compared to the multivariable LASSO model via AU-ROC scores.

2.2 Case Study I: Colorectal Cancer Meta-Analysis

CRC is a well-studied example of a diseased gut signature enriched in pathogenic species. Previous case-control studies reported enrichment of oral pathogens from the *Fusobacterium, Peptostreptococcus,* and *Porphyromonas* genera in CRC cases [38, 40], a signature that was maintained in a meta-analysis of nine CRC MWAS [41]. One of the implicated species, *Fusobacterium nucleatum*, has been studied in depth and was found to have a role in CRC tumorigenesis in mouse models [42]. *F. nucleatum* expresses FadA, a bacterial cell surface adhesion component that may bind host epithelial E-cadherin and activate β-catenin signaling, which regulates cell polarity and growth [43].

Wirbel *et al.* [44] recently submitted a meta-analysis manuscript examining 575 case-control CRC samples spanning five countries: France [38], Germany, China [45], the United States [46], and Austria [47]. Four of these are published datasets, while the German study was an unpublished collaboration with the German Cancer Research Center in Heidelberg. Table 2.1 contains the case-control distribution for each study.

Table 2.1: Samples in CRC Meta-Analysis

	FR	**DE**	**CN**	**US**	**AT**	**Total**
CTR	61	60	54	52	63	290
CRC	53	60	74	52	46	285
Total	114	120	128	104	109	**575**

Datasets were processed using the mOTU (marker gene-based operational taxonomic unit) profiler, which distinguishes reads mapped to existing reference genomes (ref mOTUs) from those clustered based on similarity of sequenced universal marker genes (meta mOTUs) [48]. Numbers found after species names are unique mOTU identifiers. Available covariates for this meta-analysis included age, gender, BMI, country of sample origin, library size, and whether the sample was taken before or after colonoscopy procedure—primarily of interest since bowel preparation for colonoscopy is known to have short-term effects on microbiome composition [49].

The median age for both cases and controls was approximately 65, although the CRC distribution was slightly left-skewed leading to a significant difference between the two populations (Wilcoxon p value = 0.0003). Gender and BMI were not significantly different between cases and controls. The shape of the library size sample distributions was nearly identical, but the CRC samples included six very deeply-sequenced outliers (Wilcoxon p value = 0.01).

The conditional entropy analysis did not show any strong relationships among covariates (Appendix Fig. 22); the strongest correlation was between colonoscopy and country, which was expected considering the studies were mostly uniform in whether samples were taken before or after (except for the US study, which was mixed).

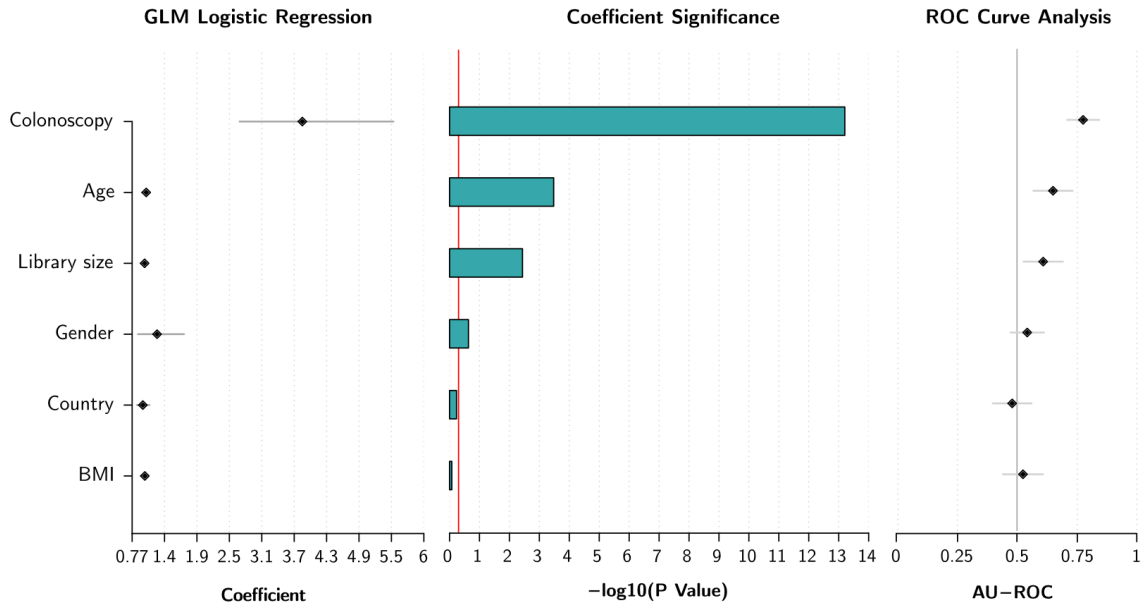

Figure 7: Output from the single covariate logistic GLM models. Colonoscopy, a categorical factor, has an AUC > 0.75, indicating a strong association with the label. Age, and library size have coefficients around 1, implying an approximately 3-fold increase in the odds of a case diagnosis for every 1-unit increase in either variable. Country and BMI were not significant; along with gender they did predict the label better than chance.

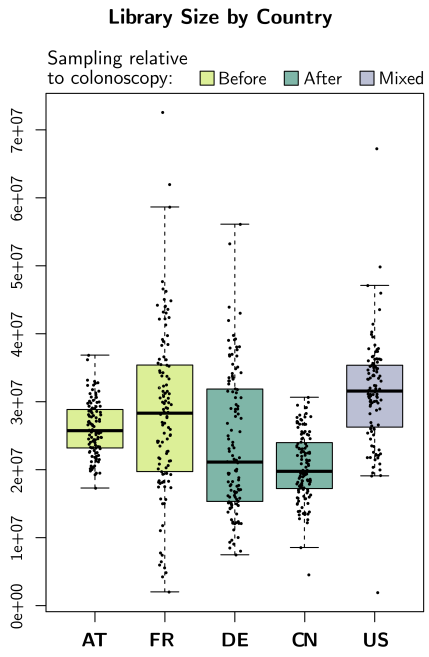

Figure 8: Sampling relative to colonoscopy and library size are both associated with country.

There was also a slight correlation between country and library size, indicating technical variation and differing protocol specifications between studies. Figure 7 shows the results from single covariate logistic regression GLMs.

Colonoscopy was the only covariate able to predict the label above 75% accuracy, though age, library size, and gender also had significant regression coefficients. Country and BMI did not predict the label better than chance (AU-ROC ≈ 0.5), indicating that these are unlikely to confound.

As a meta-study, country was of primary interest as a potential confounder. It was manually selected along with colonoscopy and library size for further analysis in Confounder Check II. Here, Figure 9 shows the significant Kruskal-Wallis test results of the top 25 species (selected for their association to the label by the Association Testing module via the scheme in Fig. 4) with each covariate while blocking for the label, as a direct measure of covariate-feature interactions.

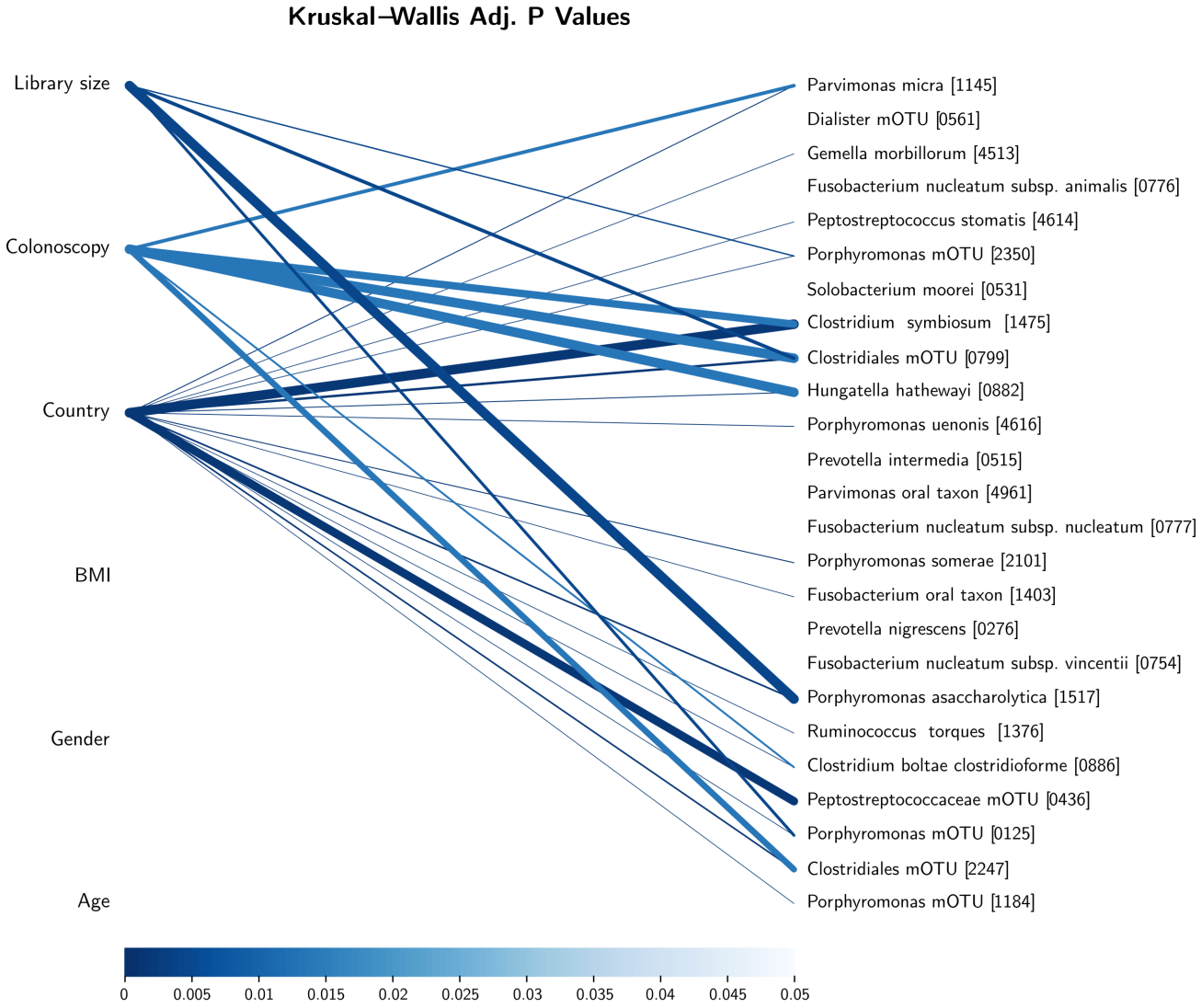

Figure 9: Significant (alpha < 0.05) non-parametric Kruskal-Wallis test results between single covariates and single species while blocking for the label. The 25 species with the strongest association with the label per the Wilcoxon rank-sum test are shown in decreasing association strength from top to bottom. Library size and colonoscopy associations overlap with country associations, indicating these variables capture similar sources of variation.

BMI, gender, and age were not significantly associated with any of the top species, and country was significantly associated with 17/25, indicating confounding potential. All six species associated with colonoscopy (five from the Clostridiaceae family and the oral pathogen *Parvimonas micra*) were also associated with country. The same overlapping pattern was also observed for the three *Porphyromonas* and single Clostridiales species associated with library size, i.e. none of the species associations with colonosocopy or library size were unique, indicating proxy effects. In general these covariates appeared to capture similar variation in differential abundances (Fig. 8).

In order to tease apart the effects of CRC status and covariates on microbial relative abundances, a two-way ANOVA on ranks was carried out. Regardless of the covariate, the amount of variance explained by the label fell between 0.08% and 17%, depending on the species being modeled, with a median around 0.3%. This range was smaller for

age, BMI, gender, and library size, and roughly the same for colonoscopy. The amount of variance explained by the country, however, was between 2% and 54% for the 849 identified species retained for analysis, with a median of 3.5%—an order of magnitude more than that explained by the label, indicating a larger confounding potential.

Figure 11 displays the country ANOVA results in detail. For 74% of the species, the influence of the country predictor was greater than that of the label in modeling the ranked relative abundances. 160 species (19%) produced significant F-test results for both country and label; however, none of the 25 species most closely associated with the label had a higher F-ratio for country than label, suggesting these species are specifically associated with the label.

Of the three additional extreme values labeled in Figure 11, only *Streptococcus thermophilus* was significantly associated with the label (Wilcoxon FDR = 0.003). It was predominately in French and Austrian samples and also one of the 160 species with significant F-test for both label and country. The Clostridiales species owes its extreme significance in this analysis to the fact that it was found almost exclusively in Chinese samples. Species such as *S. thermophilus* in the upper left quadrant of Figure 11 might be a concern if they comprise a substantial weight of the multivariable microbial signature selected by the penalized LASSO model, in which case the resulting classifier could be biased or unspecific.

To summarize the results of the ANOVA as well as check how the CRC microbial signature (rather than the 25 species most highly associated with the label) might be affected by the covariates, the meta F-ratios for each covariate were extracted for each species in the LASSO model and plotted in Figure 10. The meta F-ratio is a measure of the ratio of variance explained by the covariate over the ratio of variance explained by the label (y-axis of Figure 11). Values <1 indicate specificity for the label while values >1 indicate greater covariate effects and higher confounding potential.

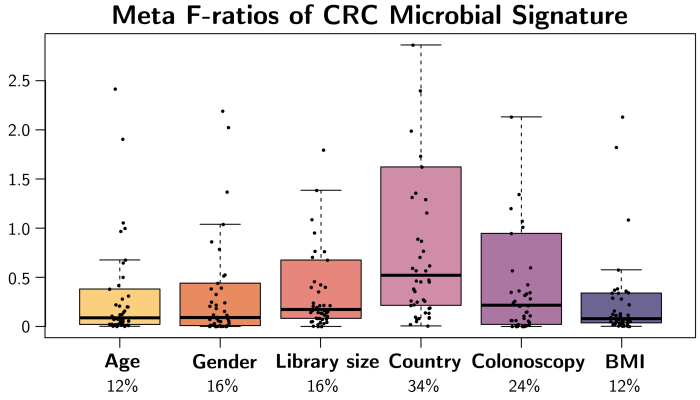

Figure 10: Ratio of variance explained by covariates over ratio of variance explained by the label for each species in the LASSO model. Each species from the LASSO model has a dot for each covariate. Extreme outliers (<5 species for any given covariate) were not plotted. The denoted percentage of species with covariate effects greater than label effects is an indicator of how the classifier might be biased or unspecific.

An analysis of deviance was carried out to assess how colonoscopy, country, and library size predicted the label in a multivariable setting, shown in Figure 12. Addition of colonoscopy always resulted in significantly less deviance from the expected probabilities, and thus a better model fit, indicating confounding potential. 75% of the controls were sampled after colonoscopy while only 40% of cases were, thus it is rather likely the logistic regression model was exploiting this discrepancy. Addition of library size was never significant, thus it is unlikely to confound. *Fusobacterium nucleatum*, *Gemella morbillorum*, *Dialister* mOTU 0561 and *Porphyromonas* mOTU 2350 were the only four models in which the species alone decreased the model deviance more than colonoscopy, indicating

high specificity for label prediction consistent with the ANOVA findings (Fig. 11).

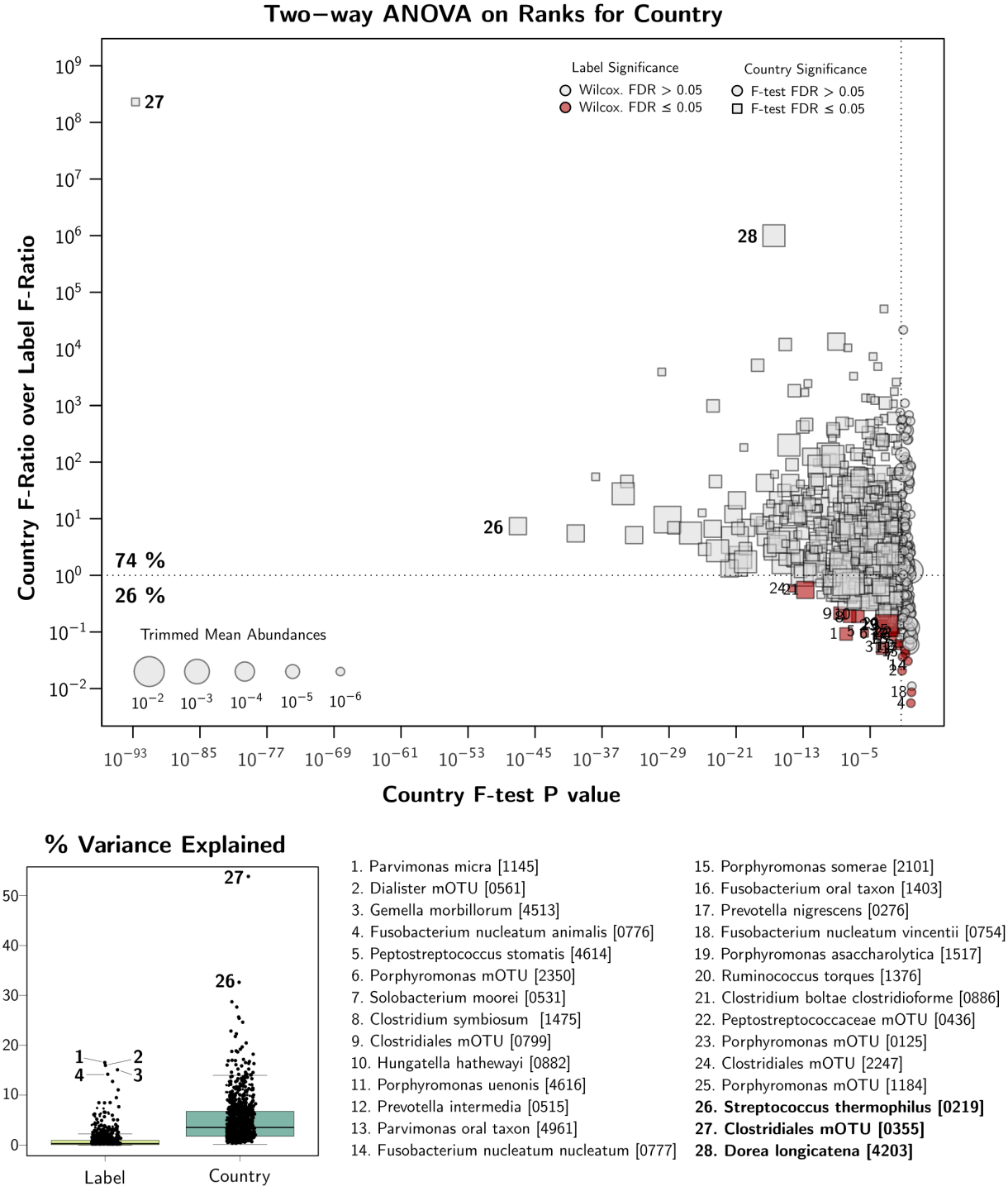

Figure 11: The horizontal line represents the case of country and label having roughly equal ability to explain the variance of ranked abundances for a given species. The boxplots on the bottom left illustrate the amount of variance explained by country and label for each species and indicate strong country effects but also strong specificity for the label for the top 25 species (1-4 shown). Species 26 was significantly associated with the label as well.

The AU-ROC values for the single species logistic GLMs were all improved upon addition of the three covariates, though all but the top five remained below 0.75. This was well below the LASSO model averages of 0.86 and primarily due to the addition of colonoscopy. Colonoscopy and country were each added individually as predictors to the LASSO model to observe the effect; both increased the AU-ROC by 0.01 on average, consistent with a robust metagenomic classifier.

Overall this meta-analysis did not appear to be confounded by any of the covariates analyzed in the univariate analysis. Country effects were present but did not predominate, and pooled findings upheld individual findings. The species most strongly associated with the label were quite specific, consistent with the literature.

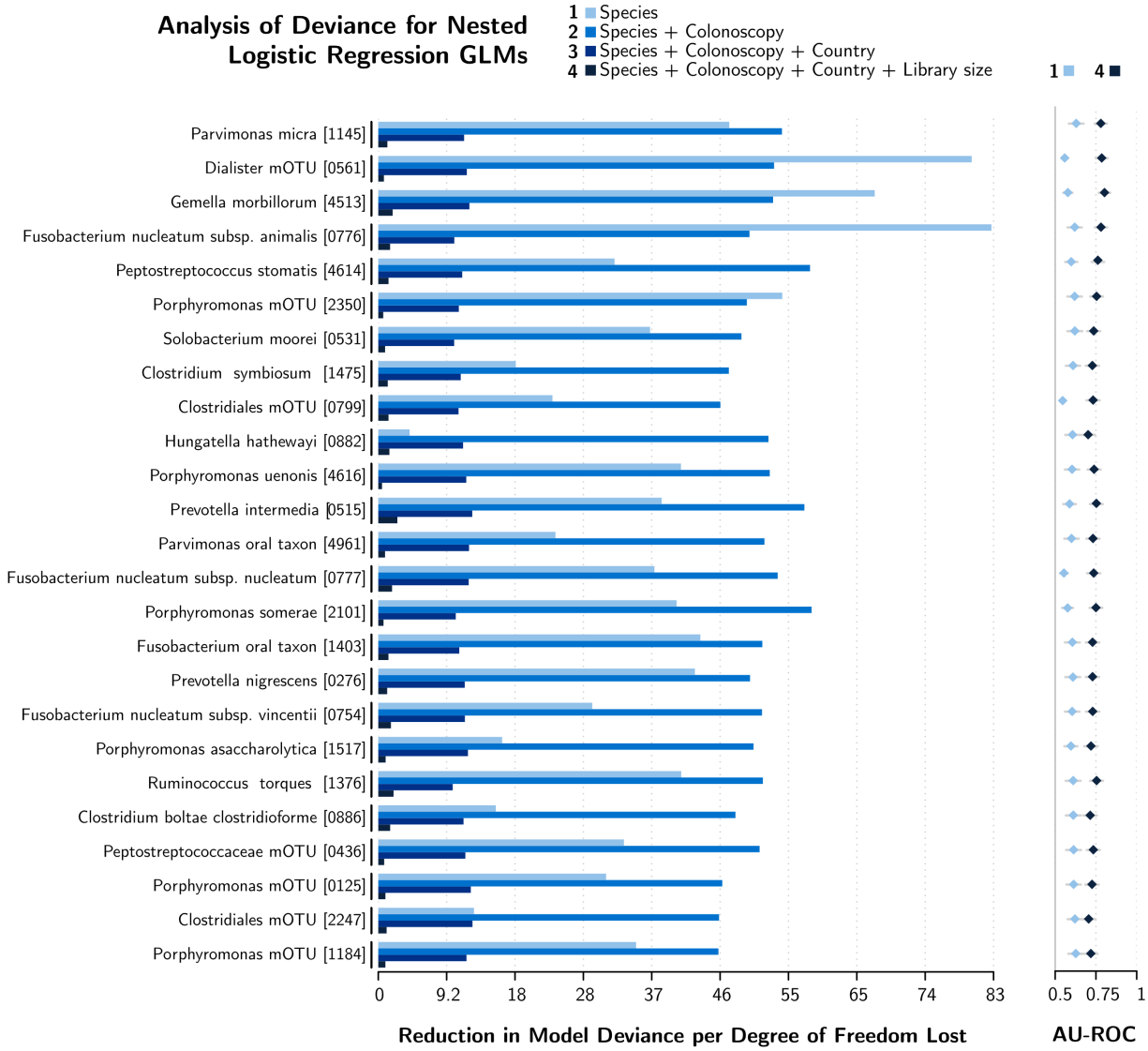

Figure 12: Horizontal bars represent the reduction in model deviance resulting from comparison to a nested model, e.g. the colonoscopy bar (2, bright blue) is a result of the comparison between model 1 (species only) and 2 (species + colonoscopy). The reduction in deviance for the species model (light blue) is calculated by comparing the model fit to the null (intercept-only) model. Coefficient significance was omitted for space, but the likelihood ratio test was significant upon addition of the species, colonoscopy, and country information for every model except *Hungatella hathewayi*; library size was never significant. At the right, the AU-ROC values of the species models (1, light blue) and the full models (4, dark blue) are shown.

2.3 Case Study II: Metformin Treatment in Type II Diabetes

Type II diabetes (T2D) was the first complex human disease to be the focus of a gut microbiome MWAS [33]. Microbiome dysbioses that have been reported include depletion of butyrate-producing taxa, specifically the genera *Blautia* and *Roseburia*, as well as a handful of unknown Clostridiales species [19, 33]. Qin *et al.* additionally reported an enrichment of *Clostridium* species in T2D [33], while Karlsson *et al.* found an increase in Lactobacilli species [34]. Forslund *et al.* combined these data sets with data from Danish subjects in the MetaHIT cohort and conducted a meta-analysis that examined metformin treatment status as a confounder to explain the divergent results. I applied my extended univariate analysis tools this data as a benchmarking measure.

The Swedish study by Karlsson *et al.* consisted of 145 women aged 68 to 72 (Fig. 13) exhibiting symptoms of T2D as well as the metabolic precursor, impaired glucose tolerance (IGT). To facilitate a cleaner binary classification, 53 IGT diagnoses were removed from my analysis, as well as samples from the Chinese dataset for which metformin treatment was unavailable, resulting in a pooled metagenome dataset as follows (MHD=MetaHIT [11], SWE=Sweden [34], CHN=Chinese [33]):

Table 2.2: Studies in Metformin Meta-Analysis

	MHD	**SWE**	**CHN**	**Total**
CTR	277	43	149	469
T2D	75	53	36	164
Total	352	95	185	**633**

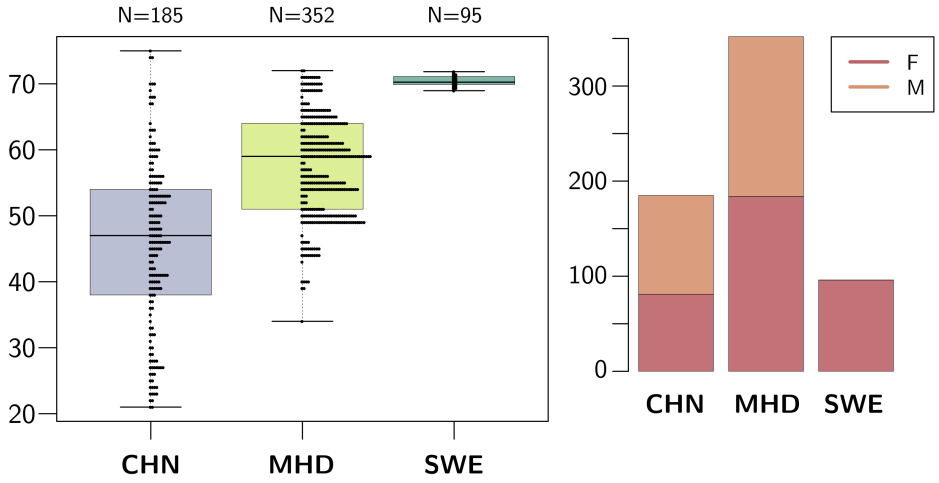

Figure 13: The meta-analysis dataset was overall poorly matched for age and gender, mainly due to the extremely homogenous Swedish study.

This meta-analysis was particularly susceptible to confounding effects considering the inclusion of datasets from very different geographic regions each exhibiting distinct age distributions, particularly the Swedish study which was also all female. Metagenomes were re-profiled using the mOTUs v2 tool [48]. Available covariates for the combined

dataset included study, age, gender, BMI, library size, and metformin treatment status (binary). The conditional entropy analysis identified significant study effects among age, gender and library size which were expected (Fig. 13), as well as an overwhelmingly strong correlation between metformin and the label (Appendix Fig. 22). Supporting this result, metformin treatment status alone had an AU-ROC of 0.76.

Figure 14 shows the results of the differential abundance testing between cases and controls while blocking for metformin, study, and age. Each covariate displayed significant confounding potential. A signature of species resembling previously reported T2D associations, including taxonomically unresolved Clostridiales species (here potentially corresponding to meta mOTUs 6134 and 7356) as well as two Lachnospiraceae species (one from genus *Roseburia*), became insignificant in this meta-analysis once adjusted for metformin treatment status, potentially explaining some of the confounding present in that study [33].

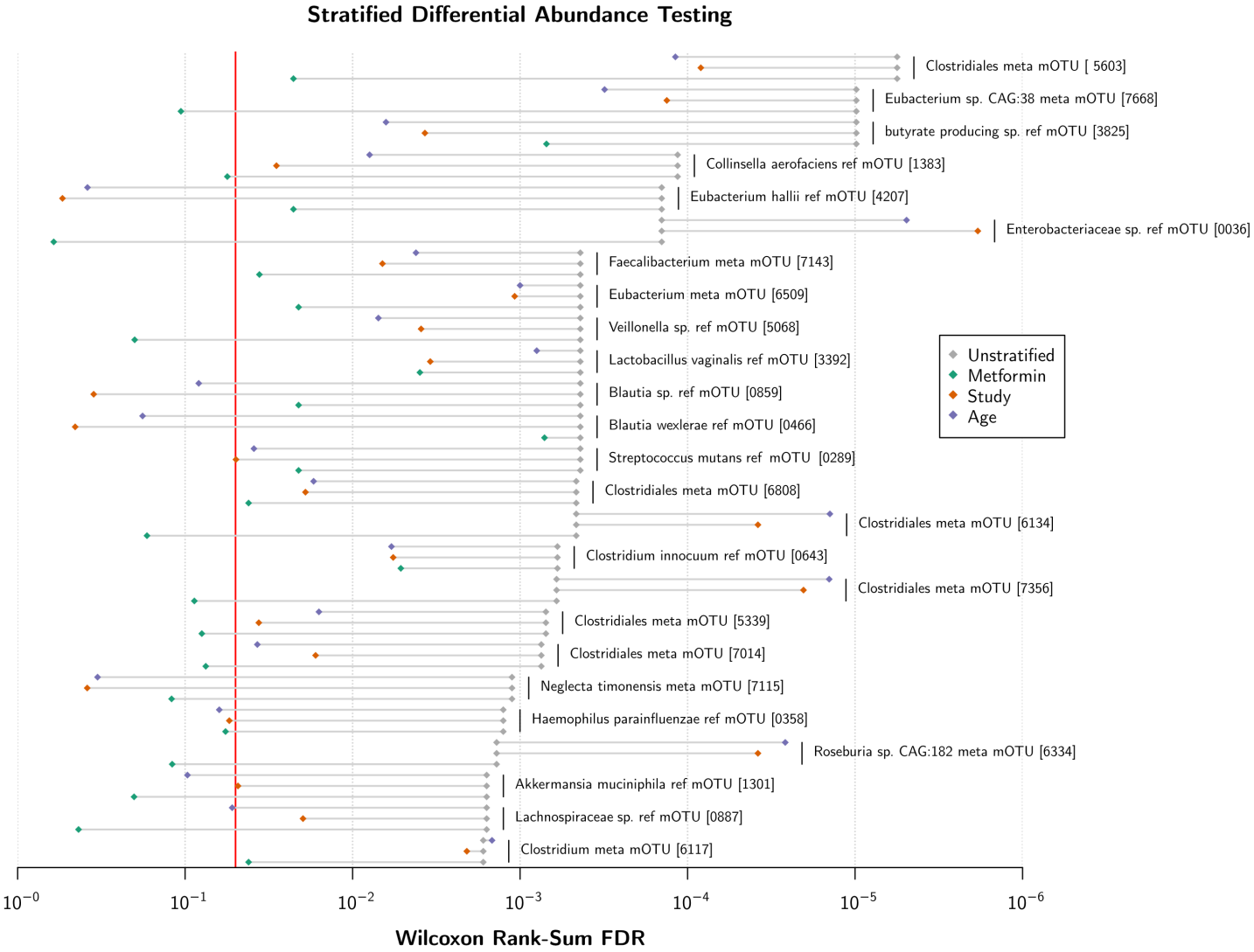

Figure 14: Differential abundance testing (Wilcoxon rank-sum tests) between cases and controls while blocking for covariates. The vertical red line is the threshold for statistical significance (alpha < 0.05). Species shown are the 25 with the strongest association to the label in decreasing order.

Forslund *et al.* focused on an increase in *Escherichia* species following metformin

treatment, which is likely reflected by Enterobacteriaceae species 0036 in this analysis due to the magnitude of change in the significance of this species when blocking for metformin (Fig. 14). These five species (four Clostridiales and one Enterobacteriaceae) also displayed a uniform increase in association strength with the label when blocked for study and age, reflecting a study effect as well as the pronounced relationship between these covariates. Generally, 13 of the top 25 species associated with the label became insignificant after blocking for metformin and 22/25 were Firmicutes, consistent with previous phyla-level associations in T2D [15].

In tandem with an enrichment of *Escherichia* species, Forslund *et al.* reported a strong depletion of *Intestinibacter* species correlating with metformin treatment, which was recovered by the ANOVA with metformin (Fig. 16, species 6 and 26). A detailed look at this signal broken down by study and metformin status is shown in Figure 15. After functional analysis revealed reduced lipid absorption and increased inflammation following metformin treatment, it was hypothesized that *Escherichia* species might thrive in this environment and outcompete *Intestinibacter* [19]. In addition to *Intestinbacter bartlettii*, the ANOVA analysis identified two other species from the Peptostreptococceae family that associated strongly with metformin status, which could support this.

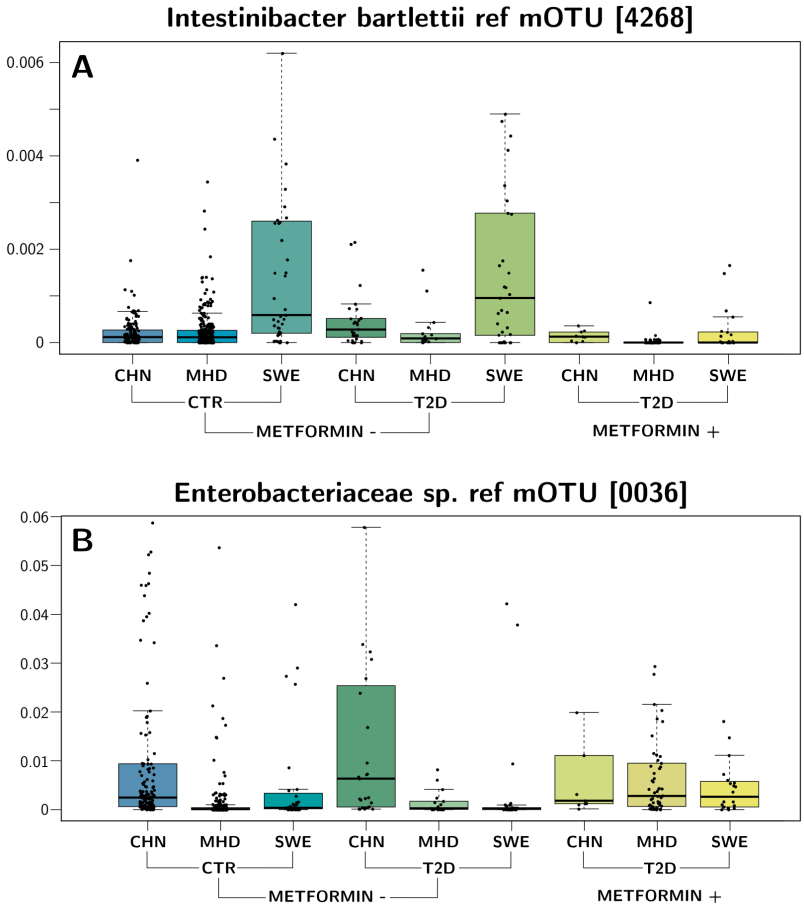

Figure 15: The main univariate effects following metformin treatment reported in Forslund *et al.*'s analysis were recovered in the ANOVA: **A)** A uniform decrease in *Intestinibacter bartlettii* and **B)** an increase in an *Enterobacteriaceae* species in Scandinavian subjects, absent from Chinese subjects which were already enriched in this species.

Strong study effects were also present in the two-way ANOVA on ranks, where study described more of the relative abundance variance than the label in 88% of the 1256 total species and 68% of the top 25 species most strongly associated with the label. Two highly abundant gut microbes that were in the top 25, *Blautia wexlerae* and *Eubacterium hallii*, were more strongly associated with age and gender than the label. These both exhibited strong study associations in Figure 14 as well, suggesting along with the conditional entropy and blocking analyses that all three variables (study, gender, and age) are capturing similar variation.

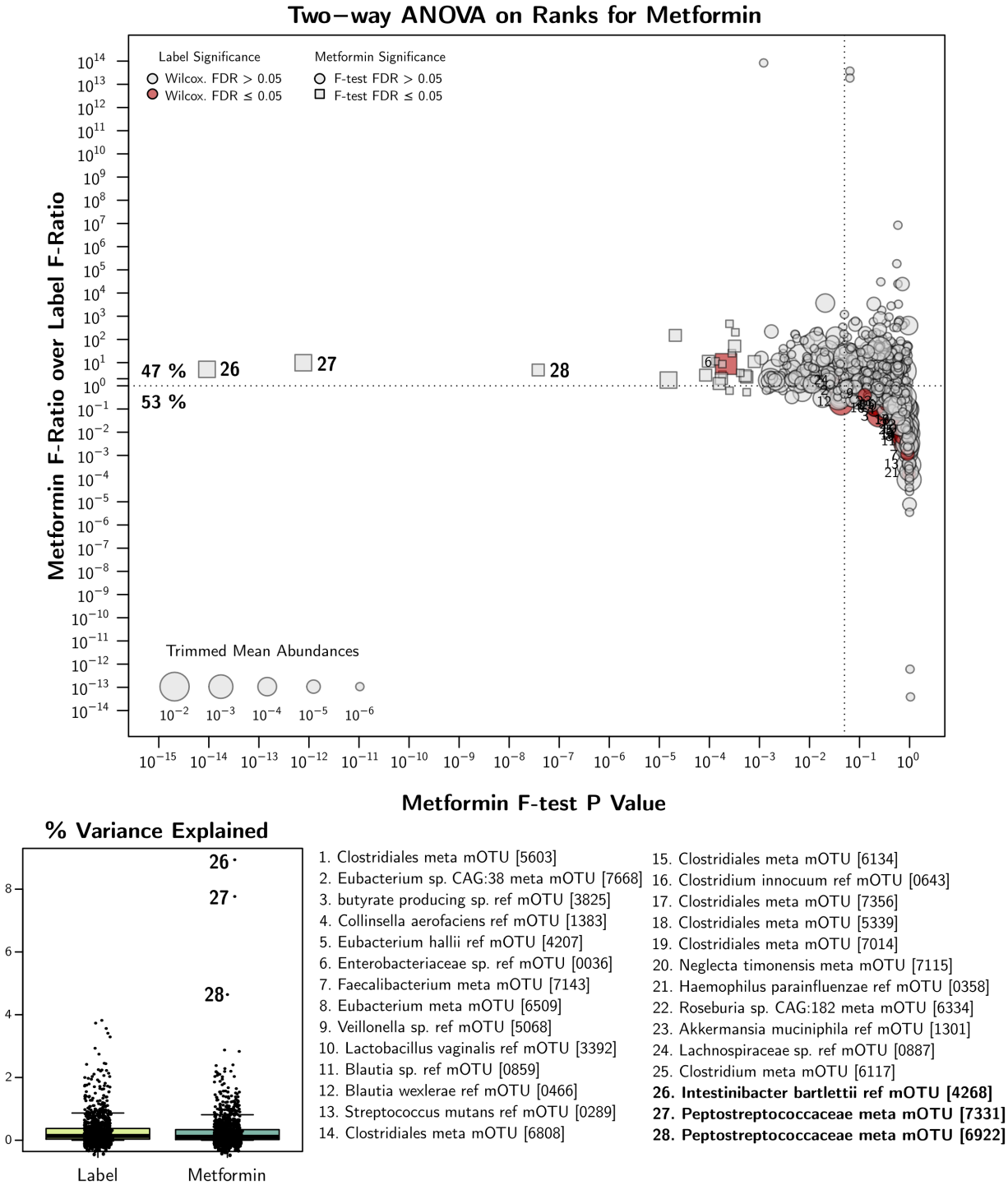

Figure 16: The horizontal line represents the case of metformin treatment and label having roughly equal ability to explain the variance of ranked abundances for a given species. The boxplots on the bottom left illustrate the amount of variance explained by metformin and the label for each species. Species uniquely and/or highly associated with either variable are numbered, except the three in the upper right and left hand corners for which these metrics were uninformative due to the prevalence of the species in a tiny, homogenous subset of samples. Notably, species 6 and 26 are in line with the signal reported by Forslund *et al.*, while 27-28 are from the same taxonomic family as 26.

Since there were no controls taking metformin, treatment status in conjunction with the label produced a unique two-way ANOVA (Fig. 16). Many rare and low abundance species had little to no variance explained by one of the predictors (either metformin or the label) in the linear models. All three species in the upper and lower right hand corners fell into this category; each was found only found in 1-2 T2D/MET+ or CTR/MET- samples, respectively, which made the meta F-ratio somewhat uninformative. This phenomenon was also reflected by the numerous species with a metformin F-test p value close to 1, likely predominating in the guts of the control group.

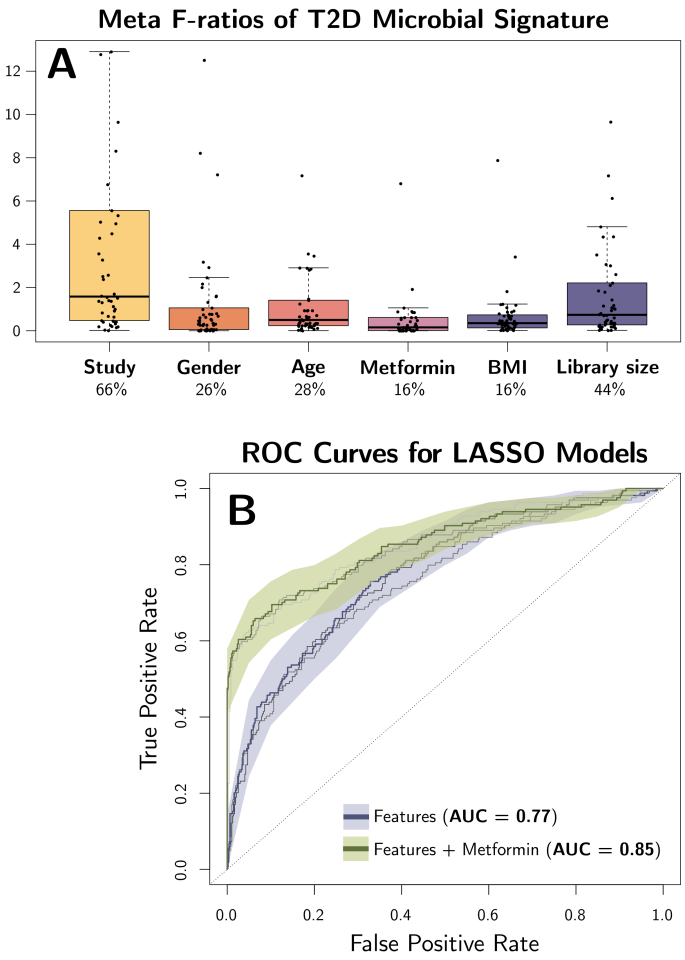

Figure 17: Examining confounding in the T2D disease signature. **A)** The meta F-ratio is the ratio of variance explained by covariates over ratio of variance explained by the label. Meta F-ratios for each species selected by the LASSO were extracted; each species has a dot for each covariate. Extreme outliers (<5 species for any given covariate) were not plotted. The denoted percentage of species with covariate effects greater than label effects is an indicator of how the classifier might be biased or unspecific. **B)** Metformin was added as a predictor to the LASSO regression, which exploited metformin's correlation with the label and substantially increased the AU-ROC score, confirming a confounding effect.

To check how relative abundances of the species in the T2D microbial signature (rather than those most closely associated with the label) could be explained by covariates, I extracted the meta F-ratios for each species selected by the LASSO model (Fig. 17A). Extreme study effects predominated, with study explaining relative abundance variation in 2/3 of the species better than disease status. Library size likely captured technical variation, some of which appears to be jointly associated with study.

Metformin's confounding potential was best captured by the analysis of deviance (Fig. 18), which, upon inclusion as a predictor, revealed a reduction in the model deviation from expected probabilities by an entire order of magnitude more than species abundances alone. Additionally, it uniformly boosted the single feature AU-ROC scores above 0.8 for each of the top 25 species associated with the label. When added as a predictor to the LASSO regression, a similar effect was observed (Fig. 17B).

Forslund *et al.* additionally performed multivariate ANOVA and distance-based dissimilarity analyses between their three cohorts (T2D with and without metformin and non-diabetic controls) and concluded that metformin's influence on the gut mi-

crobiome was poorly captured by multivariate analysis. Taken together, the univariate ANOVA and analysis of deviance support this by revealing an extraordinarily strong relationship between metformin and the label that manifests in a select few species. This supports the hypothesis originally laid out by Forslund *et al.* of specific gut microbes (potentially those of the Peptostreptococceae family, based on the triad identified in Fig. 16) at least partially mediating the therapeutic and adverse effects of metformin [19].

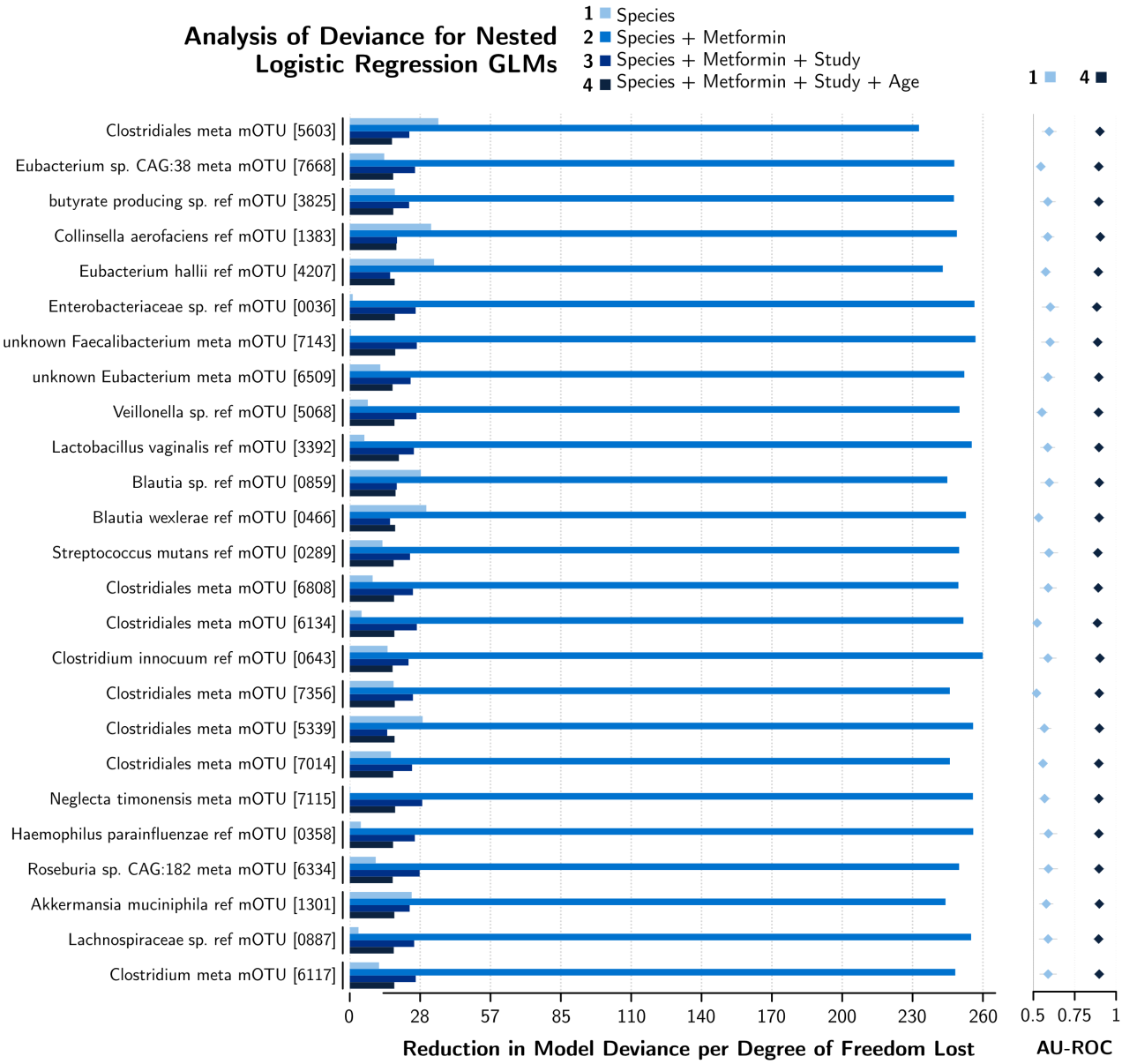

Figure 18: Horizontal bars represent the reduction in model deviance resulting from comparison to a nested model, e.g. the metformin bar (2, bright blue) is a result of the comparison between model 1 (species only) and 2 (species + metformin). The reduction in deviance for the species model (light blue) is calculated by comparing the model fit to the null (intercept-only) model. Metformin treatment status dominates for the 25 showing the strongest association with the label, consistent with a strong confounding effect. At the right, the AU-ROC values of the species models (1, light blue) and the full models (4, dark blue) are shown.

Discussion

3.1 A General Framework for Confounder Analysis

Addressing confounding first and foremost requires identifying potential confounders, a largely unmet need in the microbiome field [22]. Previous work in this area was undertaken at the level of individual analyses, such as the one mentioned in the introduction in which stratification by sexual orientation abolished a metagenomic association with HIV-1 [20], and the one recapitulated in this work initially by Forslund *et al.* that investigated metformin treatment in T2D. The major goal of my efforts was to generalize this level of detailed confounder analysis by extending the univariate branch of SIAMCAT, making it the first openly available metagenomic statistical analysis pipeline to do so [39].

This work laid a strong foundation to that effect and was benchmarked with two meta-analysis datasets [19, 44]. Colorectal cancer (CRC) is a classically strong example of robust metagenomic associations, observed as an enrichment in pathogenic species, many of which have been experimentally investigated and found to support the hypothesis of a microbial role in CRC pathogenesis [38, 41–43]. Type II diabetes (T2D), on the other hand, represents the case of reported associations involving a more vague dysbiosis that has been harder to disentangle from confounding effects or link to clear mechanistic hypotheses [18, 19, 33, 34, 50].

The general approach I took involved stratified non-parametric hypothesis tests, analyses of variance, and generalized linear models, which were paired with appropriate visualizations to capture potentially confounding patterns in the 25 species most closely associated with the disease phenotype. In this manner I was able to closely examine associations of covariates with both the gut microbiome state and the disease status per the simple confounding scheme in Figure 2. Vulnerability to confounding was most succinctly captured by the blocked differential abundance testing and related plots, in which all 25 of the CRC associations remained very significantly associated with the label (Appendix Fig. 23) and nearly all 25 of the T2D associations displayed substantial interaction with covariates (Fig. 14 in Section 2.3). The ANOVA revealed that the expected study effects were dominant in both datasets, explaining 74% and 88% of the variance in single species relative abundances in the CRC and T2D analyses, respectively. The 25 species most specific to CRC maintained their specificity when study effects were adjusted for, however, while the 25 most closely associated with T2D did not; more than half showed stronger associations with study than with the disease status.

The two-way ANOVA was able to capture confounding potential that could be due to heterogeneous datasets (observed for e.g. age and gender in the T2D meta-analysis) as well as that which could be due to a distinct effect (e.g. metformin). The former was expressed by the meta F-ratio, i.e. the ratio of variances explained by covariate and label, and the latter by calculating a significance for the variance explained by the covariate, plotted on the y- and x-axis in the visualizations, respectively.

The association between the covariate and the label was more intuitively captured by the conditional entropy and AU-ROC analyses than the analysis of deviance, which was somewhat less quantitatively meaningful. This is probably due to the fact that the nested models necessitated by likelihood ratio testing in an analysis of deviance require sequential sums of squares partitioning, i.e. the order in which the terms of the model are specified matters. Thus, by always entering the species first, something like metformin which has an extremely strong relationship to the disease status (by nature of the fact that no individuals in the control group were taking metformin) will hugely increase the likelihood when added as a predictor to a classification model.

3.2 Theoretical Limitations of Statistical Methods

Elegant ways to identify and correct for confounders are elusive for a variety of reasons, but the core confusion reflects a conceptual mismatch between the intuitive statistical language and formal mathematical expression. As a relatively young field investigating a particularly complex topic, microbiome research is mostly confined to associational statistical concepts, which are those that may be described by the joint distribution of observed variables, such as correlation, regression, and independence. Concepts of spurious correlation and effect size, on the other hand, are causal; their existence is predicated on assumed relationships between variables that cannot be defined from a distribution alone [51]. Confounding too is a causal concept; it is only of interest insofar as it is an impediment to learning about genuine causal effects [52]. Thus it does not belong with observational studies, which are associational in nature, and yet, it is an intuitive way to describe distinct phenomena that have been observed in case-control settings, which is why it has crept into MWAS despite lacking a formal mathematical basis for doing so.

These semantics mostly matter as they inform the different types of statistical inference with which each is compatible; associational concepts allow inference of beliefs under static conditions, and causal concepts allow it under changing conditions, such as those introduced by perturbations [53]. By definition the latter represents a more complete understanding of microbiome dynamics, and will be necessary in order for the microbiome to eventually have any real translational value. Schmidt *et al.* wrote a comprehensive review this year titled "The Human Gut Microbiome: From Association to Modulation" [23], which broadly outlined how the ability to understand microbiome covariation (particularly those detailed in Fig. 1) will be critical to advance the field in biologically meaningful ways. To reach this understanding necessitates the incorporation of causal concepts, i.e. confounding, into experimental designs and post-hoc analyses. The meta-analysis by Forslund *et al.* is perhaps an ideal example of this trajectory. In response to that initial groundwork, Wu *et al.* recently conducted a paired, double-blind, multi-omics study investigating the effect of metformin on the microbiome composition of treatment-naive T2D patients, and found the same previously-reported effects observable in *Intestinibacter* and *Escherichia* species, in addition to novel functional associations [50]. They also found that transplantation of metformin-treated human fecal matter into germ-free mice improved glucose tolerance, further supporting the hypothesis of microbial mediation of metformin's therapeutic effects.

The question of what to do with perceived confounders, once identified, remains less straightforward, as statistical correction methods are not without limitations. A practical issue common in biology is the fact that the myriad factors influencing both

the microbiome and a disease phenotype evade measurement. The meta-analyses I used for benchmarking included six covariates each and displayed strong proxy effects, e.g. country, age, and gender were capturing similar sources of variation in the T2D analysis. Whenever this is true, "lurking variables" cannot be measured in order to be adjusted for. An intuitive example of this is diet [54], which is known to have a strong influence on the microbiome but does not readily translate into a quantitative value.

This is a real issue faced by regression approaches, which use these noisy measurements to infer that the unobservable latent constructs they represent (i.e. the true sources of variation) are strongly associated with an outcome of interest. As an example, it is known that stool samples are not an ideal readout of the microenvironment of the gastrointestinal tract, which contains distinct macro- and micro-ecosystems [24, 55]. The risk of inference based on poor measurement is a phenomenon known as residual confounding, in which a perceived confounder that has been controlled for nonetheless still exerts an effect [56]. This is conceivable for metformin, which was used to stratify diabetics into just two additional groups based on whether or not an individual received treatment, rather than something more fine-grained that might better capture realistic biological variance or actually attenuate a confounding influence. Regression models do not explicitly account for the unreliability of covariates, which has been demonstrated to potentiate residual confounding and increase Type I error, i.e. spurious correlation [57].

3.3 Practical Solutions and Outlook

Theoretically good solutions such as structural equation modeling (SEM), which does explicitly account for measurement unreliability [57], as well as network-based approaches, which employ advanced probability theory concepts to measure uncertainty in high dimensional data [58], have yet to readily extend to microbiome research due to constraints on data size and complexity [23]; however, one can only imagine that will inevitably change. Cues might instead be taken from genome-wide association studies (GWAS), which have been dealing with confounding influences for much longer. For the common GWAS problem of population stratification, i.e. differential allele frequencies due to ancestry differences, the EIGENSTRAT method was developed, which uses a principal component analysis to infer axes of genetic variation directly from the data, which are then used to statistically control for ancestry [59]. Another GWAS solution generalizes this and statistically learns an implicit confounding framework that is then used for correction [60]. Like MaAsLin [36], the only current alternative to this work, these approaches make modeling assumptions that are better avoided if at all possible given the statistical difficulties present in microbiome data. However, the effect of adjusting for some broadly informative variance measurement such as an enterotype could provide an improvement to existing models [14].

The aim of this work was to generalize a detailed analysis so as to provide accessible insights into the confounding potential of microbiome covariates. The main message this work seeks to convey is that it is better to be aware of potential confounders than not. I provide evidence that univariate non-parametric statistical methods are sufficient to detect potential confounding effects by quantifying relationships of covariates with single species microbiome abundances and case-control distributions. Fraught with challenges as it may be, examining these relationships at all is already an improvement over analyses which choose to ignore them entirely.

Methods

4.1 Conditional Entropy

A correlation coefficient is a symmetric, scale-invariant measure of association between two random variables [61]. If two variables are correlated, they are generally said to be statistically dependent and associated with one another. Conversely, the presence of an association does not necessarily imply correlation since the latter implies more specific type of relationship. There are different types of correlation measurements which may or may not specify more information about the type of association. Pearson's R^2, for example, ranges from -1 to 1 and is used when examining a linear relationship between variables; a 0 indicates that there is no linear relationship while the sign indicates the direction if present.

The concept of entropy as defined in information theory is an additive quantity referring to the average amount of information contained in a single random variable [62]. When looking at the entropies of two different variables, a number of terms are needed. These are summarized in figure 19. Two variables (X and Y) that are completely independent will have a joint entropy *H(X,Y)* that is the sum of their individual marginal entropies (*H(X)* and *H(Y)*), while conditional entropy is used to denote the uncertainty remaining about the value of one variable given the value of another. For independent variables this will be non-zero, while for completely dependent variables the conditional entropy is exactly zero. In SIAMCAT, users may create a label object from the metadata, which would obviously not need to be included in a confounder analysis.

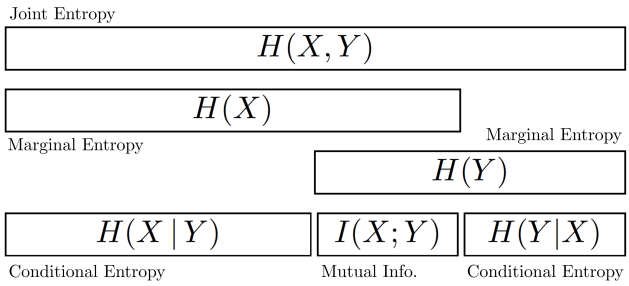

Figure 19: The relationship between joint entropy, marginal entropy, conditional entropy and mutual information, adapted from MacKay [62].

This was implemented in Confounder Check I using the `infotheo`[1] package and visualized using the `corrplot`[2] package.

[1] `infotheo` R package version 1.2.0, https://cran.r-project.org/web/packages/infotheo/index.html
[2] `corrplot` R package version 0.84, https://cran.r-project.org/web/packages/corrplot/index.html

4.2 Non-parametric Hypothesis Testing

The process of analyzing data that is sampled from a larger population in order to learn about the underlying distribution is referred to as statistical inference. One arm of statistical inference involves construction of models believed to capture some of the real world data-generating process, and the other is hypothesis testing, which assesses how likely it is that the sample at hand is representative of the population or whether two samples come from the same population [63].

Parametric tests assume that the underlying distribution of the sample is approximately normally-distributed, while non-parametric tests do not. They avoid this assumption by pooling the data being compared and transforming it into ranks before computing a test statistic; the highest value in the sample will become a 1, the second highest value a 2, and so on. The procedure for dealing with ties depends on the test. Generally, non-parametric tests are preferable for metagenomic and specifically microbiome data due to its high-dimensional, sparse, and compositional nature.

Wilcoxon Rank-Sum Test

The Wilcoxon rank-sum test is also called the Mann-Whitney or Wilcoxon-Mann-Whitney test. It is analogous to the parametric t-test and is suitable to test whether two independent samples came from the same population, but completed on ranks to make less assumptions about the data distributions. In Confounder Check II, the sample relative abundances for each feature are converted to ranks, and then grouped according to their label as well as an additional covariate before testing.

Kruskal-Wallis Test

The Kruskal-Wallis test is often described as one-way non-parametric analysis of variance (ANOVA). It extends both the Wilcoxon rank-sum test to multiple groups (for continuous data) as well as the hypothesis test portion of traditional ANOVA to non-normal data (for categorical data). In Confounder Check Module II, the sample relative abundances for each feature are converted to ranks, and then partitioned according to their discretized covariate values and according to their labels. The means of these groupings are compared and tested against the null hypothesis that they are more or less the same, and the significant associations are visualized in a bipartite network-type plot.

4.3 Generalized Linear Models

A generalized linear model (GLM) is a class of flexible models which fit data according to the maximum likelihood estimate (MLE). GLMs derive their name not by assuming a linear relationship between the predictors and response variable (as in linear models), but by assuming one between the transformed response variable and the predictors. This is encoded in the GLM via the link function and the response distribution. Table 4.1 includes all types of GLMs used in this work.

Table 4.1: Types and Components of Different GLMs

Model	Predictor Variables	Response Distribution	Link Function
Linear Regression	Continuous	Gaussian	Identity
ANOVA	Categorical	Gaussian	Identity
Logistic Regression	Mixed	Binomial	Logit

4.3.1 Analysis of Variance

Although ANOVA generally takes the form of a linear model with categorical predictors, in practice it refers to a method of analysis whose aims are slightly different than regression. Rather than estimating a response variable, ANOVA seeks to examine the explanatory power of the predictor(s) in determining the distribution of the response variable, which it does by partitioning the sums of squares.

Given a simple linear model with n observations and a single categorical predictor x_i in the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the sum of deviations of each individual observation from the single "grand" mean, $\overline{y}$, represents the total variability of the data, called the total sum of squares (TSS):

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{4.1}$$

The intent of ANOVA is to split this value into the explained sum of squares (ESS) and the residual sum of squares (RSS). Sums of squares (SS) are the variability estimators for ANOVA instead of standard variance because they are additive:

$$\text{TSS} = \text{ESS} + \text{RSS} \tag{4.2}$$

ESS is calculated by fitting a group mean $\hat{y}_i$ for each unique value of the predictor x_i and summing the deviations of each group mean from the grand mean. RSS is then the sum of deviations of individual observations from their respective group means:

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4.3}$$

The explained sum of squares represents the variability between different levels of the predictor, and thus, as the name suggests, the portion of the TSS which can be attributed to the predictor. The RSS represents the variability within each level of the predictor, which is due to noise or error.

The total variance of a dataset is calculated by dividing the TSS by the degrees of freedom, $n-1$. Likewise, dividing the ESS and RSS by their respective degrees of freedom produces variance equivalents, called mean squares in the context of ANOVA. The ratio of the ESS over the RSS once both values have been normalized to mean squares is the F-ratio; larger F-ratios imply a dominant numerator, i.e. a larger amount of variability explained by the predictor. The F-ratio is also a test statistic for the hypothesis test implicit in ANOVA—the F-test—that there is no difference between the group means.

In Confounder Check II, this analysis is complicated by the addition of a second factor, which follows the same basic premises shown above plus consideration of an interaction term of the two factors. Continuous covariates were discretized into quartiles and

the response variable was converted to ranks to avoid assumptions about data distributions.

Sequential and Adjusted Sums of Squares

The sequential sum of squares (Type I) is the amount of variation explained by a variable when the preceding terms in the model have been accounted for, which is to say that the order in which the terms are specified in the model matters. In a two-way ANOVA examining factors A and B, the sequential sum of squares would be calculated in the following order:

$$\begin{aligned} &\text{SS(A)} \\ &\text{SS(B | A)} \\ &\text{SS(A*B | A, B)} \end{aligned}$$

In contrast, an adjusted sum of squares (Type III) is the amount of variation explained by a variable when all other explanatory variables in the model have been accounted for. For the same model in the example above, the adjusted sum of squares would be calculated in the following order:

$$\begin{aligned} &\text{SS(A | B, A*B)} \\ &\text{SS(B | A, A*B)} \\ &\text{SS(A*B | A, B)} \end{aligned}$$

The adjusted sum of squares thus provides a more conservative estimate of the variance explained by each factor, as it will only attribute sums of squares to a factor which are uniquely attributable to that factor. Confounder Check II employs adjusted sums of squares for the two-way ANOVA on ranks calculations and sequential sums of squares for the nested analysis of deviance calculations, for reasons which will be explained in Section 4.3.2.

Stratification and Blocking

Stratification and blocking are two terms for the same concept which, like ANOVA, serves to eliminate unnecessary variation in order to more precisely estimate a parameter. Specifically, each sample being compared on the basis of some difference (e.g. case-control) is further divided up into discrete blocks and tested within each block to eliminate variation due to the blocking variable. Stratification usually refers to the idea when it is part of an experimental study design, and blocking when it is part of a post-hoc analysis, hence why the latter is used more frequently in this work. In Confounder Check II both the Kruskal-Wallis and Wilcoxon rank-sum tests are implemented using the `coin`[3] package in R to enable blocking.

4.3.2 Logistic Regression

Rather than estimating the response variable directly, as in linear regression, logistic regression predicts the probability that the response variable belongs to a particular

[3] `coin` R package version 1.2.2, https://cran.r-project.org/web/packages/coin/index.html

category, i.e. $p(X) = \Pr(Y = 1|X)$, where $Y \in \{-1, 1\}$. Using a linear equation such as $p(X) = \beta_0 + \beta_1 X$ does not suffice, but modeling a linear combination of predictors with the sigmoid function, which takes any real input and outputs a value between 0 and 1, does:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{4.4}$$

Another way of expressing the probability of an event is through the odds, the equation for which can be derived from 4.4:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X} \tag{4.5}$$

The odds function is important for the interpretation of logistic regression model coefficients. In linear regression, interpretation of the fitted coefficients is somewhat straightforward; β_1 is the average change in the response variable associated with a single unit increase in the predictor. Interpretation requires an additional step in logistic regression. Taking the natural log of Equation 4.5 yields the log-odds, also referred to as the logit (logistic unit), which is the inverse of the sigmoid function:

$$\ln\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \tag{4.6}$$

In the context of GLMs (Section 4.1), it is the logit which links the transformed response variable to the linear combination of predictors, hence the choice of nomenclature. With Equation 4.6 above, it becomes clear that increasing X by one unit increases the log-odds by β_1, or, equivalently, it multiplies the odds that $Y = 1$ by e^{β_1}.

The logistic GLMs of Confounder Check I are in the form of the example above since each model contains a single covariate as the predictor. In Confounder Check II, each model contains a single feature and three covariates.

Receiver Operating Characteristic

The binary output of a logistic regression equation enables it to classify observations, the accuracy of which is commonly assessed using a receiver operating characteristic (ROC) curve (fig. 20). This curve plots the false positive rate (FPR, also known as the Type I Error) against the true positive rate (TPR, or sensitivity) of a classifier as a function of decision thresholds varying across the range of predicted values (e.g. from 0 to 1). The default threshold for e.g. a binary logistic regression classifier is usually 0.5; probabilities less than this value are assigned to one class and probabilities greater than 0.5 to the other.

The diagonal represents the accuracy for a model assigning classifications randomly, and thus anything to the upper left of this is said to predict better than chance. Given a classification task for groups labeled 0 and 1, the area under the curve can be integrated to calculate the probability that a given classifier will assign a higher score to a random example from group 1 than to a random example from group 0. The area under the receiver operating characteristic (AU-ROC) is a frequently-used metric for model comparison. It is implemented in SIAMCAT using the `pROC`[4] package in R.

[4]pROC R package version 1.12.1, https://cran.r-project.org/web/packages/pROC/index.html

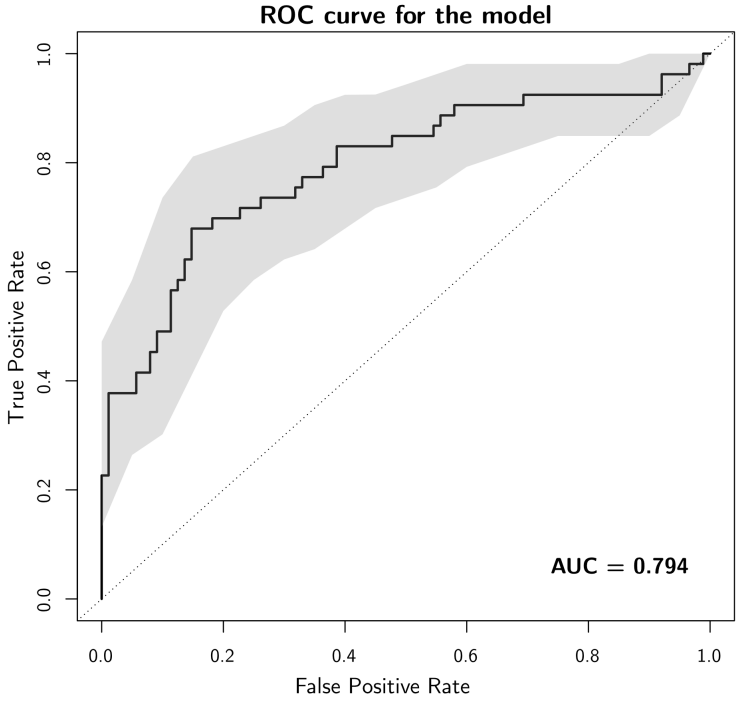

Figure 20: An example ROC curve taken from the SIAM-CAT vignette [39]. The false positive and true positive rates are plotted as functions of the classification threshold of a binary classifier. Values to the left of the diagonal line represent classifiers that perform better than chance. A perfect classifier would reach to the upper left hand corner at (0,1). The confidence interval indicates uncertainty in the predictions.

Maximum Likelihood

For any model such as the one specified by Equation 4.4, the likelihood function is a way of summarizing how well the observed data can be explained by the model. For an independent sample of multiple observations $x_i...x_n$ the likelihood function is the product of the individual likelihoods:

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} L(\theta|x_i) \tag{4.7}$$

The likelihood function is shorthanded as $L(\theta)$ and frequently converted to the log format for ease of computation:

$$\ell(\theta|x) = \log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta) = \sum_{i=1}^{n} \ell(\theta|x_i) \tag{4.8}$$

The MLE is simply the value that maximizes the likelihood function, representing the optimal θ for observing the data in the independent sample $x_i...x_n$.

Analysis of Deviance

The deviance is a goodness-of-fit statistic analogous the sum of squares when a model is fit using the MLE rather than ordinary least squares (OLS), as with logistic GLMs. Similarly, an analysis of deviance (from the expected probabilities in a logistic model) is akin to an analysis of variance. Instead of the F-test, an analysis of deviance uses the likelihood ratio test (LRT) to compare model fits and determine whether a parameter is significant or not. A significant test result for a parameter corresponds to a model in which it is included producing a significantly better fit for the data than a model without.

Given a model with three predictors $x_1...x_3$ extending the sigmoid function in 4.4, an analysis of deviance would consider models with the following linear combination of predictors (right hand side of 4.6):

$$\begin{aligned} \beta_0 \quad &\text{(null)} \\ \beta_0 + \beta_1 x_1 \quad &(1) \\ \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad &(2) \\ \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad &\text{(saturated)} \end{aligned}$$

The null model contains no predictors, only an intercept, and the saturated model contains a parameter for each observation. The models above are nested, meaning the parameters of e.g. model 2 are a subset of the parameters in the saturated model. An analysis of deviance requires nested models to perform the LRT. If $L(\theta_1)$ and $L(\theta_2)$ represent the likelihood functions for model 1 and 2, respectively, the likelihood ratio (LR) is as follows:

$$\text{LR} = 2\log\frac{L(\theta_2)}{L(\theta_1)} = 2[\log L(\theta_2) - \log L(\theta_1)] \tag{4.9}$$

Analogous to the F-ratio for ANOVA, the LR is a test statistic. Model 2 will always fit at least as well as model 1, and under the null hypothesis the LR will follow a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters estimated. From this expectation it is possible to calculate a significance of the LR.

Appendix

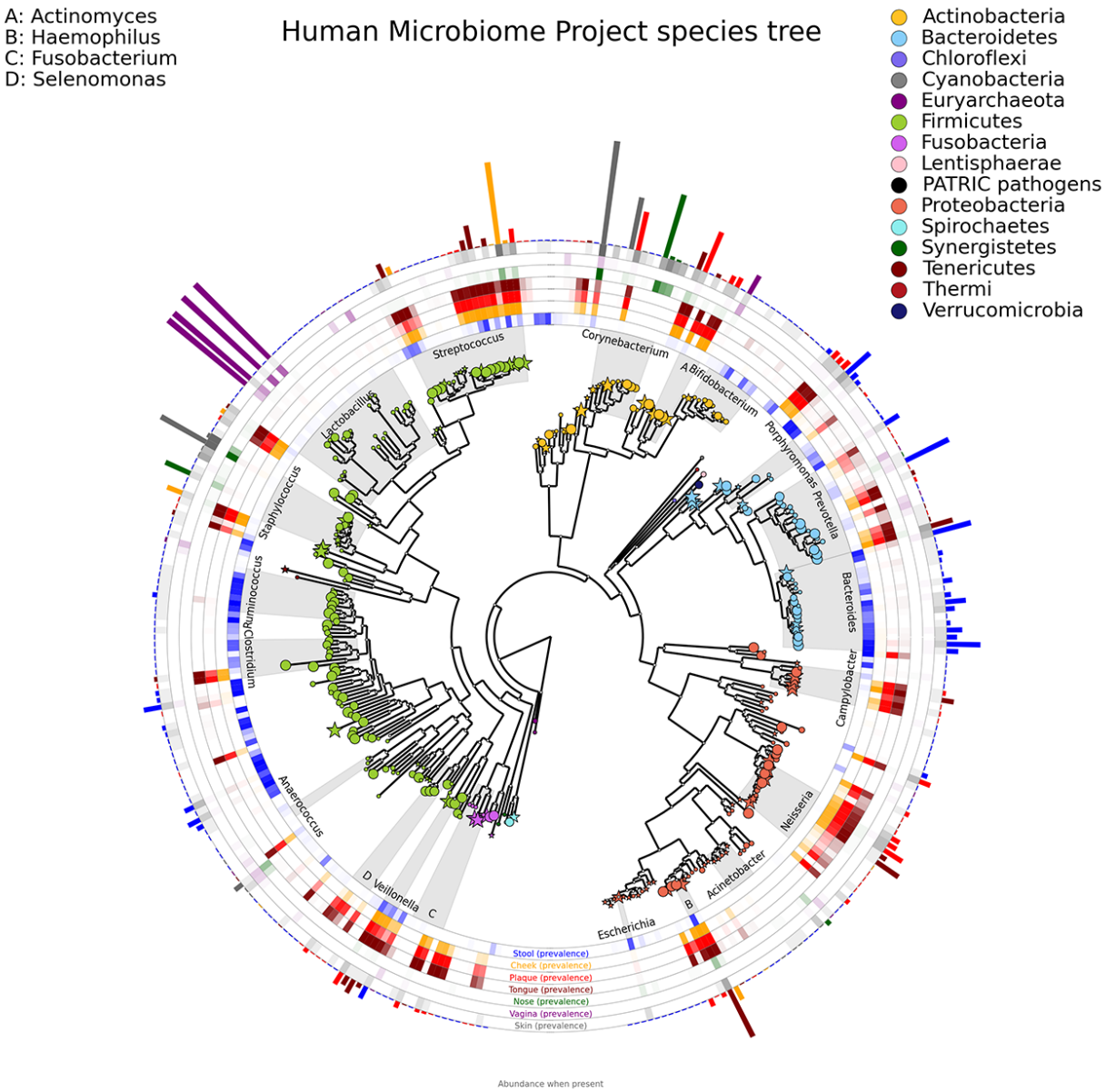

Figure 21: Most abundant species in the human microbiome (seven bodysites including the gut), using circular heatmaps and barplots. The colors intensity corresponds to species prevalence in each body site; the bar heights on the outside of the circle are proportional to taxa abundance. Reprinted from [64].

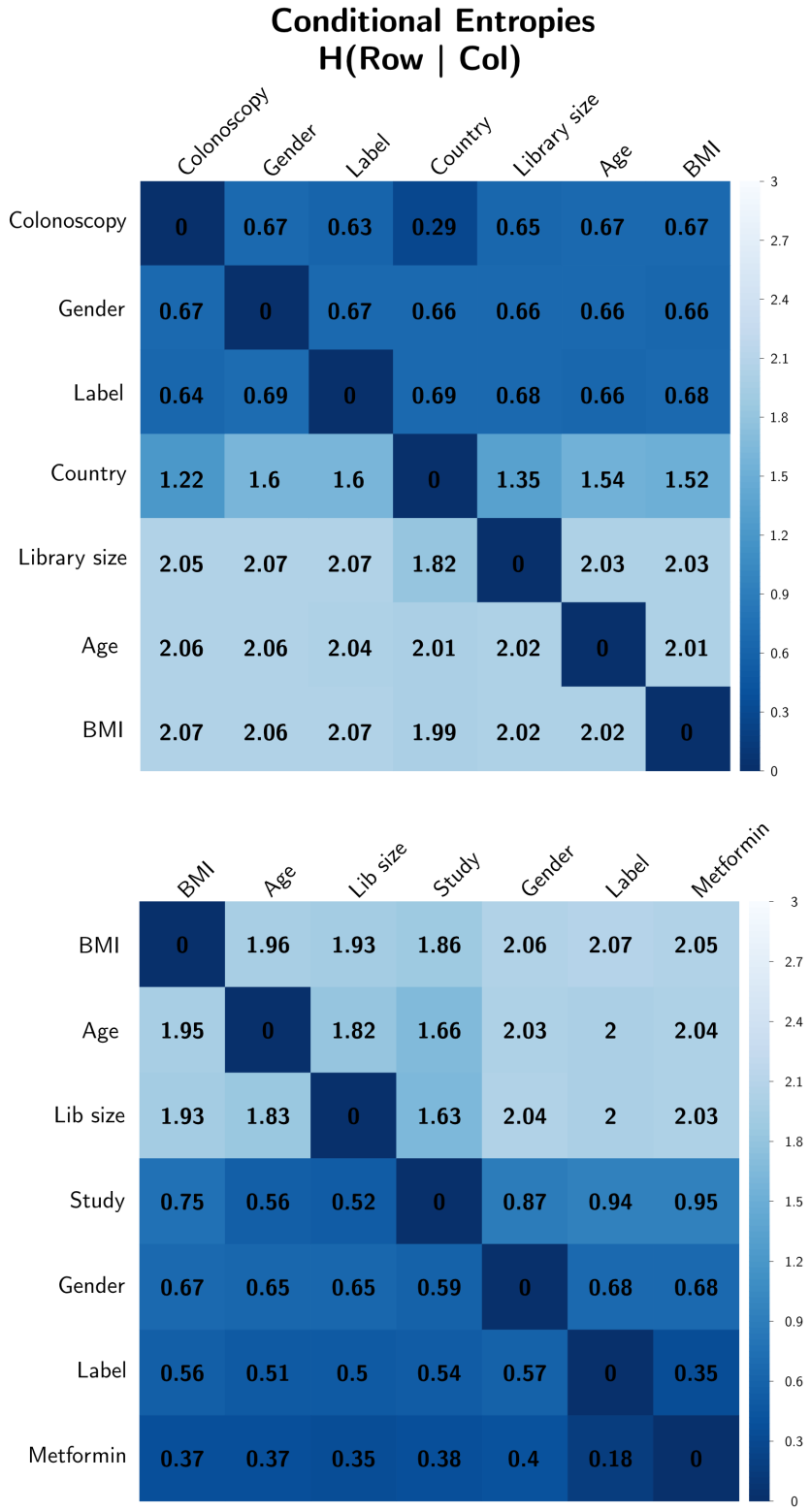

Figure 22: Conditional entropies for CRC (top) and T2D (bottom) meta-analyses. It is read as the uncertainty remaining about each row variable considering individual column variables. A conditional entropy of 0 implies no uncertainty remaining, i.e. statistical dependence.

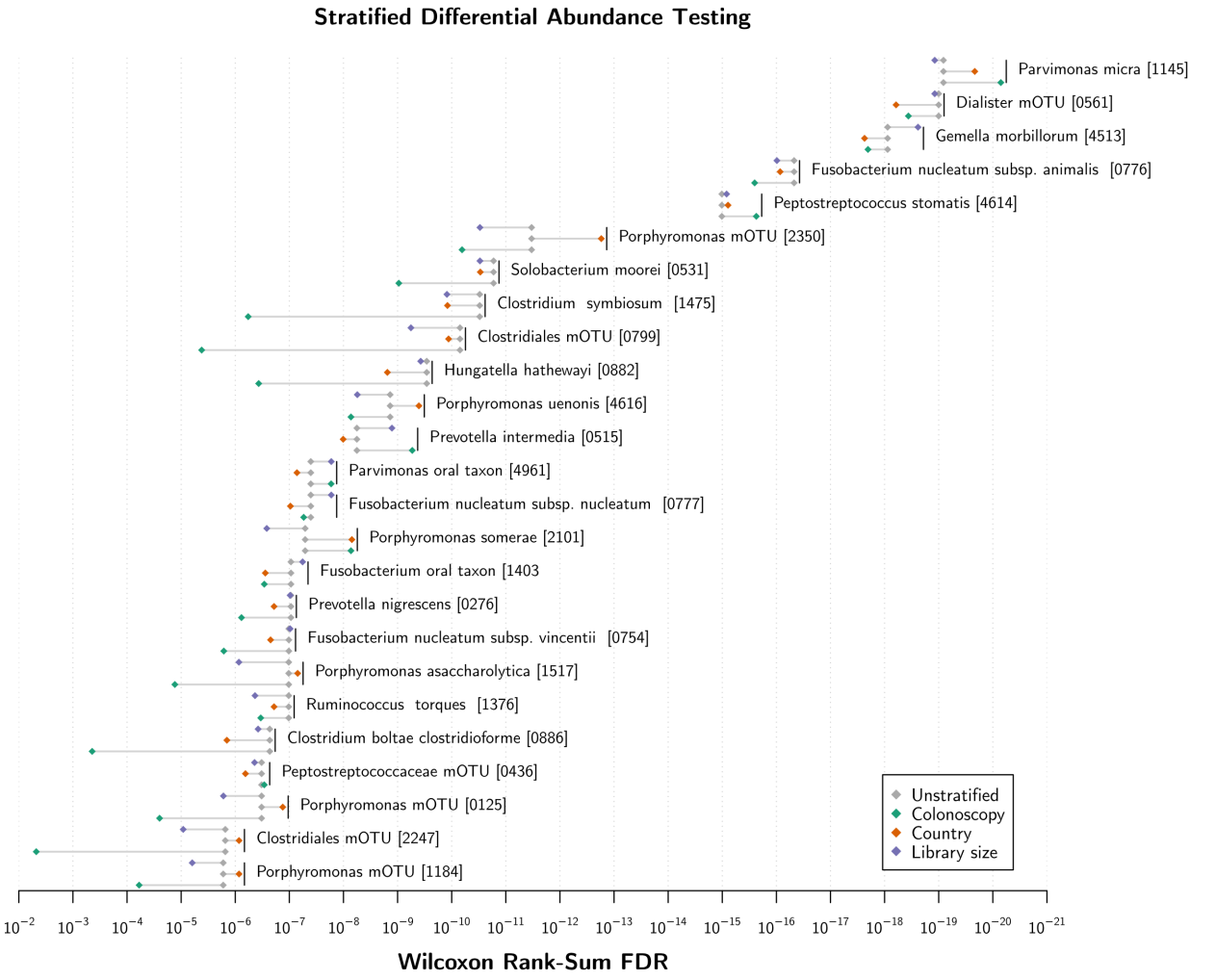

Figure 23: Differential abundance testing (Wilcoxon rank-sum tests) between cases and controls while blocking for covariates in the CRC meta-analysis. The vertical red line is the threshold for statistical significance (alpha < 0.05). Species shown are the 25 with the strongest association to the label in decreasing order.

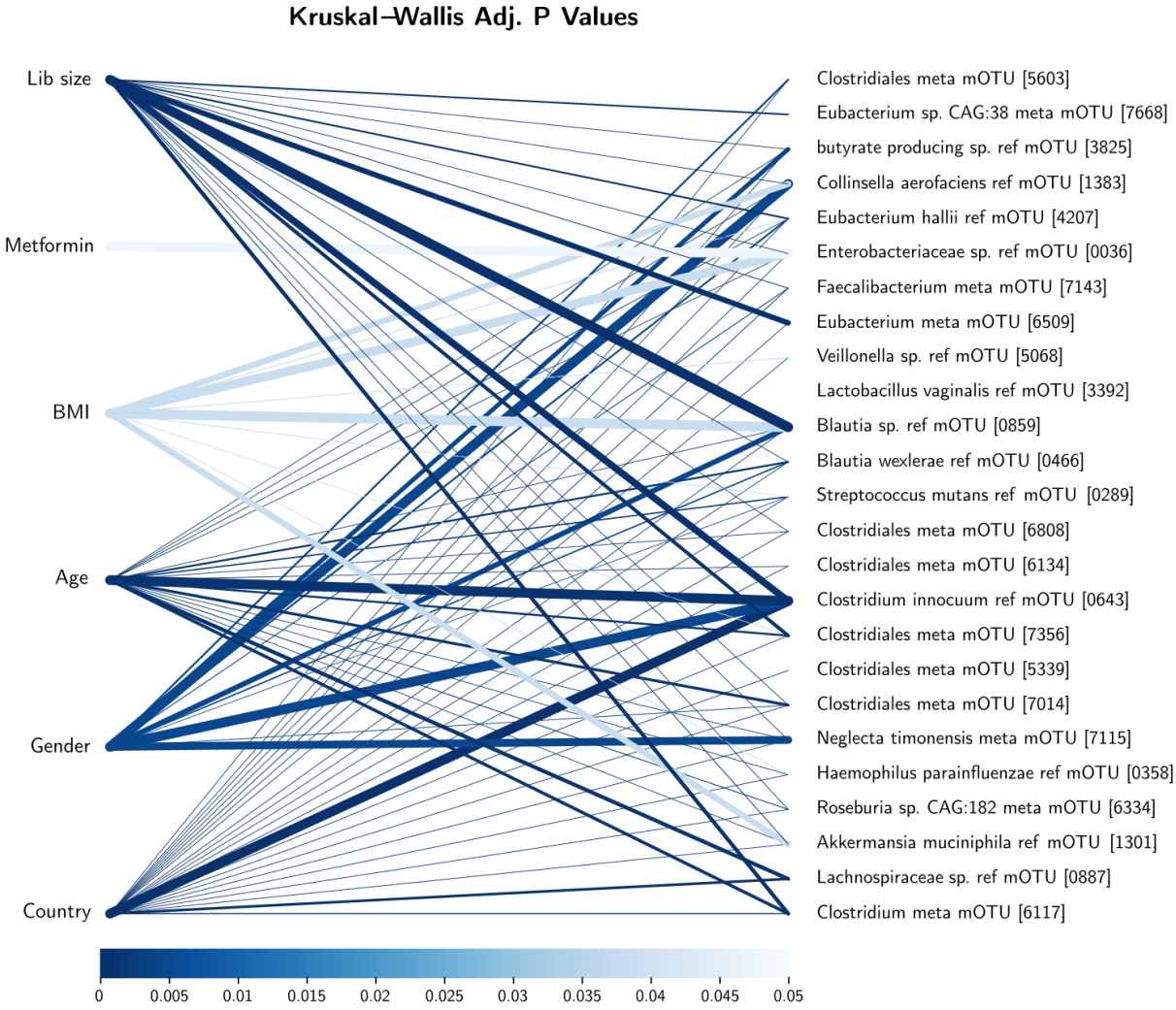

Figure 24: Significant (alpha < 0.05) non-parametric Kruskal-Wallis test results between single covariates and single species while blocking for the label from the T2D meta-analysis. The 25 species with the strongest association with the label per the Wilcoxon rank-sum test are shown in decreasing association strength from top to bottom. Multiple overlapping associations are present indicating proxy effects and/or lack of specificity. The only significant association with metformin is an *Enterobacteriaceae* species.

References

1. Jogalekar, A. *Stephen Hawking's advice for twenty-first century grads: Embrace complexity* Scientific American, Apr. 2013 (see page 1).
2. Handelsman, J. *et al.* Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology* **5,** R245–R249 (1998) (see page 1).
3. Marchesi, J. R. & Ravel, J. The vocabulary of microbiome research: a proposal. *Microbiome.* doi:10.1186/s40168-015-0094-5 (2015) (see page 1).
4. Arnold, C. *The Man Who Rewrote the Tree of Life* NOVA Next, Apr. 2014 (see page 1).
5. Morgan, X. C. & Huttenhower, C. Human microbiome analysis. *PLoS Computational Biology* **8,** e1002808 (2012) (see page 1).
6. Yong, E. *I Contain Multitudes: The Microbes Within Us and a Grander View of Life* isbn: 9780062368621 (HarperCollins, 2016) (see page 1).
7. Segata, N. *et al.* Computational meta'omics for microbial community studies. *Molecular Systems Biology* **9,** 666 (2013) (see page 2).
8. Franzosa, E. A. *et al.* Sequencing and beyond: integrating molecular'omics' for microbial community profiling. *Nature Reviews Microbiology* **13,** 360 (2015) (see page 2).
9. Methé, B. A. *et al.* A framework for human microbiome research. *Nature* **486,** 215 (2012) (see page 2).
10. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *nature* **464,** 59 (2010) (see page 2).
11. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology* **32,** 834 (2014) (see pages 2, 18).
12. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207 (2012) (see page 2).
13. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174 (2011) (see page 2).
14. Costea, P. I. *et al.* Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology* **3,** 8 (2018) (see pages 2, 26).

15. Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. *Nature Reviews Microbiology* **14,** 508 (2016) (see pages 3, 20).

16. Egger, G. & Dixon, J. Beyond obesity and lifestyle: a review of 21st century chronic disease determinants. *BioMed research international* **2014.** doi:0.1155/2014/731685 (2014) (see page 3).

17. Zuo, T. *et al.* Urbanization and the gut microbiota in health and inflammatory bowel disease. *Nature Reviews Gastroenterology & Hepatology,* 1 (2018) (see page 3).

18. Levy, M. *et al.* Dysbiosis and the immune system. *Nature Reviews Immunology* **17,** 219 (2017) (see pages 3, 24).

19. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528,** 262 (2015) (see pages 3, 6, 8, 18, 20, 23, 24).

20. Noguera-Julian, M. *et al.* Gut microbiota linked to sexual preference and HIV infection. *EBioMedicine* **5,** 135–146 (2016) (see pages 3, 5, 24).

21. Duvallet, C. *et al.* Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications* **8,** 1784 (2017) (see page 3).

22. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352,** 560–564 (2016) (see pages 3, 6, 24).

23. Schmidt, T. S., Raes, J. & Bork, P. The human hut microbiome: from association to modulation. *Cell* **172,** 1198–1215 (2018) (see pages 3, 4, 25, 26).

24. Lynch, S. V. & Pedersen, O. The human intestinal microbiome in health and disease. *New England Journal of Medicine* **375,** 2369–2379 (2016) (see pages 3, 5, 26).

25. Maes, H. H., Neale, M. C. & Eaves, L. J. Genetic and environmental factors in relative body weight and human adiposity. *Behavior genetics* **27,** 325–351 (1997) (see page 4).

26. Goodrich, J. K. *et al.* Genetic determinants of the gut microbiome in UK twins. *Cell Host & Microbe* **19,** 731–743 (2016) (see page 4).

27. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555,** 210 (2018) (see page 4).

28. Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Molecular Systems Biology* **13,** 960 (2017) (see page 4).

29. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences* **108,** 4554–4561 (2011) (see page 5).

30. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555,** 623 (2018) (see page 5).

31. Morabia, A. History of the modern epidemiological concept of confounding. *Journal of Epidemiology & Community Health* **65,** 297–300 (2011) (see page 5).

32. McNamee, R. Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine* **62,** 500–506 (2005) (see page 6).

33. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55 (2012) (see pages 6, 18, 19, 24).

34. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498,** 99 (2013) (see pages 6, 18, 24).

35. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5,** 27 (2017) (see page 6).

36. Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology* **13,** R79 (2012) (see pages 6, 26).

37. Knights, D. *et al.* Human-associated microbial signatures: examining their predictive value. *Cell Host & Microbe* **10,** 292–296 (2011) (see page 7).

38. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology* **10,** 766 (2014) (see pages 7, 12, 24).

39. Zych, K. *et al. SIAMCAT: Statistical Inference of Associations between Microbial Communities And host phenoTypes* R package version 1.0.1 (2018). doi:10.18129/B9.bioc.SIAMCAT (see pages 7, 8, 24, 32).

40. Baxter, N. T. *et al.* DNA from fecal immunochemical test can replace stool for detection of colonic lesions using a microbiota-based model. *Microbiome* **4,** 59 (2016) (see page 12).

41. Shah, M. S. *et al.* Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut.* doi:10.1136/gutjnl-2016-313189 (2017) (see pages 12, 24).

42. Kostic, A. D. *et al.* Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host & Microbe* **14,** 207–215 (2013) (see pages 12, 24).

43. Garrett, W. S. Cancer and the microbiota. *Science* **348,** 80–86 (2015) (see pages 12, 24).

44. Wirbel, J. & G, Z. (submitted) (see pages 12, 24).

45. Yu, J. *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut.* doi:10.1136/gutjnl-2015-309800 (2015) (see page 12).

46. Vogtmann, E. *et al.* Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PloS one* **11,** e0155362 (2016) (see page 12).

47. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature Communications* **6,** 6528 (2015) (see page 12).

48. Milanese, A. *et al.* Microbial abundance, activity, and population genomic profiling with mOTUs (under review) (see pages 12, 18).

49. Jalanka, J. *et al.* Effects of bowel cleansing on the intestinal microbiota. *Gut* **64,** 1562–1568 (2015) (see page 12).

50. Wu, H. *et al.* Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nature Medicine* **23,** 850 (2017) (see pages 24, 25).

51. Pearl, J. *et al.* Causal Inference in Statistics: An Overview. *Statistics Surveys* **3,** 96–146. issn: 1935-7516 (2009) (see page 25).

52. McNamee, R. Confounding and confounders. *Occupational and Environmental Medicine* **60,** 227–234 (2003) (see page 25).

53. Pearl, J. *Causality* isbn: 978-0521895606 (Cambridge University Press, 2009) (see page 25).

54. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences* **107,** 14691–14696 (2010) (see page 26).

55. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host & Microbe* **15,** 382–392 (2014) (see page 26).

56. Christenfeld, N. J. *et al.* Risk factors, confounding, and the illusion of statistical control. *Psychosomatic medicine* **66,** 868–875 (2004) (see page 26).

57. Westfall, J. & Yarkoni, T. Statistically controlling for confounding constructs is harder than you think. *PloS one* **11,** e0152719 (2016) (see page 26).

58. Layeghifard, M., Hwang, D. M. & Guttman, D. S. Disentangling interactions in the microbiome: a network perspective. *Trends in Microbiology* **25,** 217–228 (2017) (see page 26).

59. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38,** 904 (2006) (see page 26).

60. Fusi, N., Stegle, O. & Lawrence, N. D. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Computational Biology* **8,** e1002330 (2012) (see page 26).

61. Dalgaard, P. *Introductory statistics with R* isbn: 978-0-387-79053-4. doi:10.1007/978-0-387-79054-1 (Springer Science & Business Media, 2008) (see page 27).

62. MacKay, D. J. *Information theory, inference and learning algorithms* isbn: 9780521642989. doi:10.2277/0521642981 (Cambridge University Press, 2003) (see page 27).

63. Holmes, S. & Huber, W. *Modern Statistics for Modern Biology* isbn: 9781108427029 (Cambridge University Press, 2018) (see page 28).

64. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9,** 811 (2012) (see page 34).

65. James, G. *et al. An introduction to statistical learning* isbn: 978-1-4614-7137-0. doi:10.1007/978-1-4614-7138-7 (Springer, 2013).