2016

Extension of Multivariate Analyses to the Field of Microbial Ecology

Vijay Shankar
Wright State University

EXTENSION OF MULTIVARIATE ANALYSES TO THE FIELD OF MICROBIAL ECOLOGY

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

By

Vijay Shankar
B.A., Miami University, 2008

2016
Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

April 26, 2016

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Vijay Shankar ENTITLED Extension of Multivariate analyses to the field of Microbial Ecology BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Oleg Paliy, Ph.D.
Dissertation director

Mill W. Miller, Ph.D.
Director, Biomedical Sciences
Ph.D Program

Committee on
Final Examination

Robert E. W. Fyffe, Ph.D.
Vice President for Research
Dean of the Graduate School

Gerald M. Alter, Ph.D.

Jeffrey L. Peters, Ph.D.

Michael L. Raymer, Ph.D.

Nicholas V. Reo, Ph.D.

Abstract:

Shankar, Vijay. Ph.D. Biomedical Sciences PhD Program, Wright State University, 2016.
Extension of multivariate analyses to the field of microbial ecology**.**

Ground-breaking advancements in molecular and analytical techniques in the past decade have enabled researchers to accumulate data at an extraordinary rate. Especially in the field of microbial ecology, the introduction of technologies such as high-throughput sequencing, quantitative microarrays, nuclear magnetic resonance and mass spectrometry has led to the interrogation of diverse and previously unexplored microbial communities at unparalleled depth. Analysis and interpretation of patterns within datasets acquired with such high-throughput methods require powerful statistical approaches. A class of such techniques called multivariate statistical analyses is an excellent choice for analysis of complex microbiota-related datasets. This field of statistics is constantly evolving as new techniques and procedures are being developed and applied to explore and interpret the underlying patterns both statistically and visually. As a result, the decision-making process involved in the choice of the technique that best suits the scientific question and the dataset is no longer trivial. Additionally, the current trends in the use of multivariate statistics in microbial ecology indicate a strong preference toward exploratory analyses, resulting in limitations to possible biological interpretations. In order to facilitate a more extensive integration of multivariate statistics in microbial ecology, I apply a diverse set of analytical methods to human-associated microbial and metabolite datasets that allows us to draw biologically relevant inferences. Specifically, I use indirect gradient analyses to show that the largest gradients of variability correspond to the separation of samples based on sample groups. I use direct gradient analyses to explain a significant portion of the overall variability present within the response variables using independently measured environmental variables. I use classifier techniques to build highly accurate discriminant models

based on the differences in the response variables across sample groups and identify the variables that contribute the most to sample group separation. Using correlation-based bipartite analyses, I identify statistically significant associations between two different sets of response variable that were measured for the same set of samples. Finally, I integrate the analytical insights from the above approaches into a generalized protocol for the analysis of multivariate datasets in the field of microbial ecology.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

Figure **Page**

LIST OF TABLES

I. Introduction:

Analysis of microbial communities and environments:

Microbes are ubiquitous in nature and inhabit very diverse environments which include the human intestinal tract and skin, soil, roots, leaf and bark surface of plants, ocean waters, deep sea vents, and air (1-5). Microbes thrive in these environments not as individual species but as complex communities that comprise hundreds and maybe even thousands of unique members. These communities are an integral part of the systemic processes such as energy and elemental cycling, and biomass production. It is the complex metabolic interactions between the microbial community members which allows for energy and nutrients to flow through the ecosystem (6-8). The complexity associated with such communities and the fastidious nature of these microbes which leads to difficulty in culturing of individual members have made it a challenge for researchers who have attempted to study these communities and the interactions that exist within them. However, recent advancements in molecular techniques and technologies have simplified some of these challenges involved in profiling these communities by removing the need to culture the individual community members.

Many of these techniques interrogate community composition and function through direct analysis of the genetic material. In addition to removing the need to culture the microbes, these molecular techniques also tend to be high-throughput, allowing researchers to simultaneously analyze many samples and variables. Examples of such techniques include high-throughput massively parallel sequencing, phylogenetic microarrays and quantitative real-time PCR. In order to better understand the metabolic interactions between community members, it is also important to interrogate the microbial environment for metabolites and biomarkers, in addition to profiling community structure and function. Examples of high-throughput techniques that have

enabled metabolomic approaches in the field of microbial ecology include nuclear magnetic resonance spectrometry (NMR) and gas and liquid chromatography mass spectrometry (GC- and LC-MS). Recent studies that have taken advantage of the availability of these techniques continually improve our understanding of the structure, function and the dynamics of microbial communities, and the complex interactions that exist within and among the biomes. Examples of such works include the identification of definitive links between human gut microbiota and obesity (9), characterization of the impact of soil microbiomes on plant functions (10) and assessment of microbial diversity within methane seeps in deep-ocean floors (11).

High-dimensionality datasets, the typical output of high-throughput techniques, are generally represented as matrices of numerical values where each value corresponds to the measurement of a variable from a given site or sample (12). The entries within the matrix may be absolute values or relative abundances with respect to the sum of variables for each given object. The underlying distribution of values in the dataset may depend on the type of data, the method of measurement and several other factors.

Multivariate statistical analysis:

In a simple system with very few variables, the changes in the variables can be easily extracted and summarized with straightforward approaches such as visual inspection and descriptive statistics. However, in more complex systems consisting of hundreds or thousands of variables, the change in the overall dataset spans across many variables and in complicated ways with respect to various environmental gradients. For example, microbiota that live in river streams are influenced by the amount of phosphorus and nitrogen that are released from the surrounding lands into the streams as a result of rainfall. These compound responses or patterns

are much more difficult to identify in high-dimensional datasets using conventional means. Fortunately, an entire class of statistical approaches exist to tackle such problems. These methods, known as multivariate statistical analyses, attempt to deconvolve such compound responses by organizing the variability within high-dimensionality datasets into manageable and interpretable terms or "factors" (13). While the mathematical framework used to achieve this depends on the applied technique, the end result is the reduction in complexity. Aside from the reduction in complexity, there are other advantages to multivariate analyses such as enhanced statistical power due to the aggregation of responses, the ability to assign rank of importance to factors or gradients as well as the ability to partition noise out of the overall variability (14).

Different approaches to classify techniques within multivariate analysis have been considered (15). One approach, for example, is based on the objective of the investigation, which results in techniques being placed roughly into these overall categories: (i) data dimensionality reducing, (ii) sorting and grouping, (iii) building relationships between variables, (iv) machine learning (predictive) and (v) hypothesis-driving (15). Some of these categories and the specific techniques within these categories will be discussed below.

Dimensionality reduction and exploration:

Many of the dimensionality reducing techniques belong to a class of ordination methods called indirect gradient analyses or unconstrained ordination analyses. Ordination by these techniques is based solely on the matrix of response variables. They are well-suited for the exploration of structures and visualization of the most dominant gradients of variability within the dataset. These techniques reduce the dimensionality of the dataset for ease of interpretation by generating synthetic variables that represent dominant gradients from combinations of the

original response variables. The meaning behind these synthetic variables is inferred after performing the analysis to draw possible biological implications (16). Indirect gradient analyses are often used as exploratory techniques to confirm the presence of large patterns (gradients). Typical examples of unconstrained ordination techniques include Principal components analysis (PCA), Principal coordinates analysis (PCoA), and Correspondence analysis (CA). Detailed descriptions, usage, limitations and the underlying assumptions of these techniques can be found in several reports and online resources (17, 18). Unconstrained ordination techniques have successfully been used in several reports for exploratory analysis in the field of microbial ecology (19-22).

Principal components analysis:

Principal components analysis is one of the most popular and oldest dimensionality reduction multivariate tools available (23). Its popularity is attributed to the ease of performing the analysis and the simplicity in its interpretation. Briefly, PCA builds latent (compound) axes within the dataset that summarize linearly independent portions of dataset variability through matrix transformation procedures (24). Additionally, PCA ranks these axes (also known as principal components or eigen axes) based on the proportion of the overall variability captured within the dataset. Therefore, the first axis captures the largest variability, the second axis captures the second largest variability that is independent (orthogonal) to the first. This process is repeated until all of the dataset variability is organized into linearly independent components. This feature is the key behind the dimensionality reduction properties of PCA. Since each sample now has coordinates from each principal component, displaying samples as points in the first two or three axes would reveal inherent large patterns within the dataset and their effects on sample

(dis)similarity. PCA constructs eigen components using Euclidean distance. Therefore, the relationship between samples in PCA ordination space is based on this metric. PCA has been widely used in the field of microbial ecology for dataset variability exploration (17). There have been concerns regarding the use of PCA with ecological datasets because this technique assumes a linear response model (variables change in a linear fashion with respect to unknown external gradients or effects), which is rare in nature (25). However, recent considerations have shown that if the length of these gradients are short, linear techniques such as PCA can appropriately define gradients from ecological datasets (25, 26).

Principal coordinates analysis:

Principal coordinates analysis (PCoA), is similar in properties to PCA in that it too attempts to order dataset variability into independent components. The difference however, lies in PCoA's ability to use externally defined distance relationships. Therefore, PCoA can be considered a more general version of PCA, and conversely, PCA can be thought of as a special case of PCoA, where the distance used to define relationships is Euclidean (16). This is an interesting feature of PCoA, because it allows researchers to incorporate relevant information into the ordination of variable responses. There is tremendous versatility in which distance can be used, albeit, the interpretation of the ordination will be dependent on the distance used. Even distance matrices generated using qualitative, semi-qualitative or mixed variables can be analyzed using PCoA (17). A very popular application of PCoA in microbial ecology revolves around the use of a beta-diversity based phylogenetic distance called UniFrac (27), which defines the relationships between taxa within and across communities based on their evolutionary

lineage (using sequence-based comparisons). Several reports have successfully used this distance with PCoA to identify gradients within microbial community datasets (20, 28, 29).

Correspondence analysis:

Correspondence analysis is an indirect gradient technique that calculates relationships (correspondence) between samples and variables within a frequency table (cross-table or contingency table) and represents them in low-dimensional space. A graphical representation of these relationships will depict which samples are similar to one another, which variables are similar to one another based on the counts (frequency) and which variables have a higher probability of occurring in which samples (17). CA holds several advantages as an exploratory tool for microbial ecology. One of these advantages is that it is well suited to represent unimodal response variable models (variable responses change in a unimodal fashion with respect to external gradients, which is often the case in microbial ecology where microbial groups display high abundance only when specific conditions are met) within the dataset (30). Another advantage of CA is its relative insensitivity to double-zero cases (absence of a variable in the two compared samples) due to the properties of the χ^2 metric used to calculate relationships between samples and variables of the cross-table (30). Because axes of CA are not completely unrelated to one another (CA axes are only uncorrelated), occasionally, gradients that are a part of the first CA axis also appear in the second axis, usually as non-linear functions of the first. This phenomenon, called the "arch" effect, can be corrected using a post-analysis process called 'detrending' to restore the linearity of the first axis in CA. However, care should be taken with the interpretation of CA plots after 'detrending' as multi-axes distances will no longer accurately reflect but only approximate the inter-sample-variable relationships mentioned above.

Hypothesis-driven:

With the existence of hidden dataset structures confirmed with exploratory indirect gradient analyses, researchers can attempt to build hypotheses regarding the meaning behind the gradients. Testing of hypothesis-driven queries in multivariate data is typically performed using constrained ordination techniques (also referred to as direct gradient analyses). Constrained ordination can be thought of as a modification of unconstrained techniques, where the solution to the ordination is constrained in relation to an independently measured secondary variable or a set of variables. The secondary variable can be, for example, environmental variables that have been measured separately for the same set of samples as in the original dataset (pH, temperature, etc of samples or sites). As a comparison of unconstrained and constrained ordination techniques, PCA searches through a dataset to identify the largest gradients of variability, whereas Redundancy analysis, a constrained ordination technique, searches through a dataset to only find variability that is related to the changes in the constraining variables. Typical examples of constrained ordination techniques include Redundancy analysis (RDA), Canonical correspondence analysis (CCpdA), and Principal response curves (PRC) analysis. These techniques have been thoroughly described in scientific reports and online resources (17, 18). The use of hypothesis-driven multivariate analyses in the field of microbial ecology is not as widespread and popular as exploratory multivariate analyses (17). Still, a few studies have efficiently used constrained ordination analyses to answer hypothesis-driven queries in this field (31, 32).

<u>Redundancy analysis:</u>

Redundancy analysis (RDA) is thought of as the constrained extension of PCA where the ordination axes, which are linear combinations of response variables, are also linear combinations of the environmental or explanatory variables (16). Because PCA and RDA are built on the same framework, the distance metric used to define relationships is Euclidean. Because of this, RDA is best suited for use with datasets where the response variables change in a linear fashion with respect to the environmental gradients (17). The quality of how well the included explanatory variables explain the patterns in the response variables can be determined by the proportion of overall inertia (variation) due to the explanatory variables. The constraining procedure can not only be applied to a matrix of response variables, but also to a matrix of (dis)similarities. A useful application of the latter approach lies in RDA's ability to constrain a matrix of sample relationships that were generated using a non-Euclidean-based distance metric. This extension to RDA is referred to as distance-based (RDA) and has recently been applied to microbiota datasets in combination with the UniFrac beta-diversity metric (33, 34). The graphical representation of RDA is typically a bi- or a tri-plot, where arrows represent the explanatory variables (lengths are proportional to the explained variability), and dots represent sample and/or response variables (18). Care should be taken with the interpretation of RDA and should be based on the type of end-point scaling. Sample-based scaling focuses on preserving exact distances between samples in ordination space (samples with similar response and explanatory variables appear close to each other) and only the angles between response variables and explanatory variable arrows represent linear correlations. Variable-based scaling sacrifices inter-sample distance relationship to preserve relationships between all variables (angles between any two variables, response and/or explanatory, represents their linear correlation) (18). Another useful application of RDA is the ability to run multiple partial analyses (where different

environmental or explanatory variables are set as conditional variables or co-variables) to partition the overall variability explained by each explanatory variable. This type of analysis, referred to as variation partitioning (or partial RDA), lets researchers determine the relative importance of each explanatory variable based on its contribution to explaining the overall variability within the response set (18).

Principal response curves:

Principal response curves is a special case of partial RDA, where variability within the dataset is partitioned to only consider the changes in community due to time. The motivation for the development of PRC arose from the difficulty in interpreting time-dependent effects on sample and variable ordination in typical ordination plots. These effects are often masked by variability due to other environmental factors. Additionally, due to the nature of these ordination plots, time-dependent effects do not conform to a unidirectional gradient leading to a jagged arrangement of samples (35). In order to limit the interpretation to only relevant terms, the canonical coefficients derived by comparing community change to its respective control at each time point is plotted as a function of time. This process results in a response curve for each temporal data series. Very few studies in the field of microbial ecology have used PRC for the analysis of time-series data (36, 37). An interesting variant that was born out of one such application is the modified PRC where a single reference is used for every time-point comparison (36). This type of analysis depicts change in community over time, with respect to the reference.

Canonical correspondence analysis:

Canonical correspondence analysis (CCpdA) is the constrained analog of CA, just as RDA is that of PCA. Axes of CCpdA are maximally related to linear combinations of the constraining explanatory variables (30). Because the framework of CCpdA is based on that of CA, the technique is well suited for unimodal response models. Likewise, all of the advantages of CA, and its use of χ^2 metric for calculation of relationships translate over to CCpdA. The output and the interpretation of the output of CCpdA is very similar to that of RDA. One of the key differences between RDA and CCpdA, is that CCpdA is capable of utilizing categorical or nominal variables (for example, group designation) as constraining variables. Like with RDA, variation partitioning is possible with CCpdA (this is referred to as partial CCpdA) (18). The same iterative procedure used with partial RDA is also used with CCpdA. This approach is especially powerful in the context of microbiota communities because of the prevalence of unimodal response relationships (30). Use of CCpdA in microbial ecology is still somewhat rare. However, with the recent improvement in understanding and exposure, the technique has started to become popular (32).

Classification, prediction and variable selection:

Extending from the hypothesis-driven approaches to multivariate analyses, if the goal is to find consistent patterns within the dataset that pertain to separation of pre-defined clusters of samples, then techniques that are designed to accomplish it are called discriminant analyses. Discriminant analyses have become more sophisticated in recent years due to the advent of powerful computers. Access to ample processing power have enabled the use of complex machine learning algorithms that can search large datasets to find strong and consistent patterns through combinations of measured variables that separate groups of samples. This process,

typically referred to as 'model training', is a critical part of discriminant analyses (38, 39). Once a model is sufficiently trained, it can now be used to predict new samples. Such a feature has tremendous application in the clinical context, where rapid identification of sample identity using pre-trained pathological models can help with appropriate treatment strategies (40). Additionally, since patterns that separate sample groups are built using combinations of the measured variables, it is possible to identify the variables that contribute the most to the modeled separation, which can help with the biological interpretation of the group separation. There are several approaches to the discriminant problem for multivariate data. Some techniques use ordination based approaches, like Orthogonal projection to latent structures (OPLS-DA), while others use decision trees, like Random forest (RF) or separation of hyperplanes in multidimensional space as in Support vector machines (SVM). Discriminant models are usually assessed using cross-validation approaches where the dataset is split into 'training set' and 'test set'. The model created using the 'training set' is then tested using the 'test set' to determine overfitting and model accuracy. Several popular cross-validation approaches have been developed and tested (41-43).

Random forest:

Random forest discriminant analysis is an ensemble classifier based on decision trees. It is referred to as an 'ensemble' classifier because it creates thousands of decision trees and the results of the decision trees are merged to generate an overall output. To briefly describe the procedure, decision trees are built using the variable values in a series of quantitative conditional statements (greater than or less than) to generate a sample group output. At each decision node, only random subsest of variables are available as choices. This process is done to ensure that

trees are truly independent and a few strong predictor variables do not dominate the decisions of all the decision trees. This 'random' selection of variable set is the difference between other decision tree-based classifiers and RF. Finally, a voting procedure is used to collect the decisions of all the trees and the mode of the group decision is selected as the algorithm output. The random selection of variables for decision nodes, and the voting procedure greatly reduce the 'over-fitting' problem often encountered with typical classifier algorithms. Studies that have tested RF's performance have reported very high classification accuracies even for datasets of modest sizes (44, 45). RF has gained tremendous popularity in microbial ecology recently due to its reported high performance with these datasets (46-48).

Orthogonal projection to latent structures:

Orthogonal projection to latent structures - discriminant analysis (OPLS-DA) tackles the discrimination problem by building synthetic axes (latent axes) which are linear combinations of the measured variables that correspond, or relate specifically, to the separation of sample groups. This is done by performing least squares regression (fitting) between the latent axes and the group designation axis, which results in the projection of these axes into a new ordination space (49). Orthogonal correction of this procedure partitions variability pertaining to group separation from unrelated variability within the dataset (50). Variability related to group separation and unrelated variability can be plotted on T and $T_{orthogonal}$ axes respectively, on ordination plots for visualization of model classification. The model predictive power and regression fit are used to assess the quality of models. Models are usually tested for over-fitting using cross validation approaches. Coefficients of variables from the latent axes can be used to determine the discriminatory strength of each variable. Use of OPLS-DA in the field of microbial ecology is

somewhat rare (51). OPLS-DA is more often used in metabolomics studies (52), but a recent surge in its usage has been reported in microbial ecology thanks to integrative studies that profile and link different aspects of the microbial environment (40, 53).

Support vector machines:

Support vector machines (SVM) discriminant analysis tackles the problem of classification using kernel methods. Kernels are transformations of data to higher dimensional spaces that enable the fitting of a simple discriminant boundary (linear plane, for example) to previous complex group separation (54). As such, a model built by SVM represents the optimal hyperplane which maximizes the margin that separates sample groups in multidimensional space (54). SVM is versatile in that it allows for linear and non-linear kernel functions. Also, because SVM does not require the calculation of feature vectors (linear combinations of variables) for discrimination, and only requires the application of the kernel function for dataset transformation, SVM calculations tend to scale very well with large and complex datasets (also referred to as kernel trick) (54). Several studies that explored the predictive performance of SVM have reported that it showed high accuracy even for datasets containing low numbers of samples (44, 55). The use of SVM in the field of microbial ecology is very rare. To date, there has been only one study that has used SVM for discriminant analysis of microbial datasets (56).

Relationships among sets of variables:

Access to different types of measured variables for the same set of samples allows for an integrative approach to analysis. Usually, biologically relevant interests in integrative analysis stem from questions regarding what type of relationships exist between sets of variables. A

straightforward approach to analyzing these datasets is to compare the pattern of changes in variables between different sets, across the samples. For example, one might be interested to look at the changes in abundances of complex polysaccharide-degrading microbiota and levels of short-chain fatty acids in the human gastrointestinal environment, to determine the metabolic interactions between these terms. Pair-wise correlation-based analyses are one of the simplest ways to uncover putative associations between different sets of variables. These analyses produce a quantitative measure (correlation coefficient) of the relationship between two numerical arrays (57). Values of the correlation coefficients usually range between -1 and 1. A positive value implies that the values change together, a negative value indicates that the values change in a reciprocal manner and zero indicates a lack of a monotonic relationship. Popular correlation analyses include Pearson product-moment, Kendall-Tau and Spearman rank correlation coefficients. Of these, Kendall-Tau and Spearman rank correlation metrics are considered non-parametric (they do not assume any specific distribution for the data) are highly suited to situations where prior information regarding the input data is unavailable. Correlation analyses have been extensively used in the microbial ecology for both integrative approaches as well as to look at relationships among variables within a single set (58, 59).

Spearman rank correlations:

Spearman rank correlation coefficient is a non-parametric measure of how well two variables change together. Specifically, it measures the strength of the monotonic (as one variable increases, the other variable also increases or as one variable increases, the other decreases) relationship between two arrays of continuous, discrete or ordinal variables (60). The flexibility in the types of variables is due to the methodology used to calculate the correlation

coefficient. Spearman coefficient is based on the comparisons of the relative ranks of the variable values within the respective arrays. This feature gives rise to two very important advantages to the use of Spearman rank correlations; (i) the non-parametric nature of the coefficient, (ii) relative insensitivity to outliers compared to other correlation coefficients (61). The statistical significance of the correlation coefficient for each pair-wise comparison can be calculated by comparing the measured coefficient to a null distribution (resulting in values around zero) generated by randomizing ranks in one or both arrays.

Extending multivariate analysis to studies in microbial ecology:

The trend in the use of multivariate statistical analyses in studies from the field of microbial ecology indicates a severe bias in what types of techniques are used and for what purposes. Most often, multivariate analyses are used for exploratory purposes with microbiota-related datasets. Researchers limit their use to indirect gradient techniques such as PCA or PCoA and hypothesis-driven techniques such as RDA and CCpdA are generally avoided (17). The reasons for such limited use of hypothesis-driven techniques are usually due to unfamiliarity with using and interpreting techniques, fear of misuse and lack of user-friendly implementations. Furthermore, there are a large number of available multivariate techniques and newer one are constantly being developed. And each technique has its own set of special conditions and assumptions that need to be satisfied it order for proper analytical implementation. As a result, the difficulty involved in determining the choice of the technique that would be appropriate for a given biological query might give rise to the observed preference for older and simpler-to-interpret exploratory techniques (12, 17). **In order to facilitate a more thorough integration of multivariate statistical analyses to studies in the field of microbial ecology, a surge in the**

knowledge and understanding of the use and interpretation of these techniques is necessary. Therefore, we have applied different types of multivariate tools to several microbiota-related datasets to demonstrate their suitability for extracting biological inferences and to develop a generalized protocol for the analysis of such datasets.

II. Materials and methods:

Fecal water extraction:

Similar to previous reports (62-64), fecal metabolites were analyzed in fecal water extracts prepared from each sample. A total of 250mg of homogenized stool was suspended in 1.25ml of sterile cold phosphate buffer (4.3mM $Na_2HPO_4 \bullet 7H_2O$, 1.5mM KH_2PO_4, 2.7mM KCl). The mixture was homogenized for 5 minutes and then centrifuged at 16,000g for 5 minutes. The supernatant was collected and filtered through a GDX syringe filter (10.0µm - 0.2µm pore size). The filtrate was centrifuged again at 16,000 g for 15 minutes, and the supernatant was retained and stored at -70C for subsequent analyses.

Genomic DNA isolation:

Genomic DNA isolation from human feces was performed as previously described (65). Briefly, 150mg of material was processed using the ZR Fecal DNA kit (Zymo Research Corporation) processed according to manufacturer's instructions. The genomic DNA from the procedure was eluted into 90µl of DNAse/RNAse free molecular grade H_2O. The quality and quantity of the eluted DNA was analyzed using electrophoresis on a 1% agarose gel and the 260/280 ratio of OD obtained through Nanodrop 1000. Eluted high quality DNA was stored at -70C.

Proton NMR of fecal water extracts:

A 550µl aliquot of the prepared fecal extract sample was transferred to a 5 mm NMR tube together with 150µl of 9mM trimethylsilylpropionic-2,2,3,3-d4 acid (TSP) in D2O. Proton (1H) NMR spectra was acquired at 25C using a Varian INOVA operating at 600MHz (14.1

Tesla). TSP served as a chemical shift reference and quantification standard, and D_2O provided a field-frequency lock for NMR acquisition. Water suppression was achieved using the first increment of a NOESY pulse sequence. Spectral data was pre-processed using Varian software that employs exponential multiplication (0.3Hz line-broadening), Fourier transformation, and phase correction. Spectra were then baseline corrected (flattened) in MATLAB (The Mathworks, Inc.). Further spectral processing included removal of the residual water signal, chemical shift referencing, and sum normalization. For multivariate data analyses, spectra were binned to reduce the dimensionality and mitigate peak misalignment, and signal intensities were auto-scaled. A dynamic programming-based adaptive binning technique was employed (66) using a minimum and maximum distance between peaks in a single bin of 0.001 and 0.04ppm, respectively.

Quantification of specific metabolite resonances was accomplished using an interactive spectral deconvolution algorithm in MATLAB adapted from our previously described methods (66). The deconvolution tool fits a defined spectral region using a combination of tunable baseline shapes (spline, v-shaped, linear, or constant) and a Gauss-Lorentz peak-fitting function. All metabolite peak intensities were corrected for equivalent number of protons and normalized relative to the TSP signal intensity. We used a combination of three sources to assign peaks to specific metabolites – (i) database of proton NMR peaks assigned to specific small compounds (such as Human Metabolome Database), (ii) literature that defines specific peaks to belong to specific compounds, and (ii) the above tentative assignments were confirmed by addition of the suspect compound (spiking) to a test extract sample, carrying out proton NMR spectrum acquisition, and identifying corresponding peaks.

Taxonomic analysis:

Microbiota Array:

Amplification of genomic DNA for each sample was performed using the phylogenetically conserved primers Bact-27Fv4 and Univ-1492Rv1 which target the full-length prokaryotic 16S rRNA gene as previously described (65, 67). The following conditions for PCR were used with 250ng of starting DNA template: 25 cycles of PCR amplification, 50µl reaction volume, and previously described cycle conditions (65, 67). PCR was performed in replicates of 4 reactions and pooled together prior to purification. Purified PCR products were fragmented and processed using the Affymetrix protocol as described by the manufacturer and hybridized to the custom designed Microbiota Array developed in the Paliy lab (65, 67). Post-hybridization, the chips were washed and scanned as described previously (65, 67). The analysis of microarray data was performed as previously described (65, 67). Briefly, to quantitate phylotype presence based on detection calls, the raw data were processed in GCOS using the standard MAS5 detection algorithm. To quantitate phylotype abundance data, hybridization signal estimates were first normalized using the MAS5-VSN-MAS5-MedianPolish pipeline using CARMAweb online portal (67). The acquired normalized phylotype abundance data were adjusted for 16S copy number variations and probe cross-hybridization using custom MS Excel templates.

High-throughput next generation sequencing:

The 16S rRNA gene V4 variable region PCR primers 515/806 with barcode on the forward primer were used with the extracted genomic DNA from each sample in a 30 cycle PCR using the HotStarTaq Plus Master Mix Kit (Qiagen, USA) under the following conditions: 94C for 3 minutes, followed by 28 cycles of 94C for 30 seconds, 53C for 40 seconds and 72C for 1 minute, after which a final elongation step at 72C for 5 minutes was performed. After

amplification, PCR products were checked in 2% agarose gels to determine the success of amplification and the relative intensity of bands. Multiple samples were pooled together (e.g., 100 samples) in equal proportions based on their molecular weight and DNA concentrations. Pooled samples were purified using calibrated Ampure XP beads. Then the pooled and purified PCR product were used to prepare the DNA library by following Illumina TruSeq DNA library preparation protocol. Sequencing was performed at MR DNA (www.mrdnalab.com, Shallowater, TX, USA) on a MiSeq following the manufacturer's guidelines and the 2x250bp sequencing chemistry. Sequence data was processed using a slightly modified QIIME analysis pipeline (68). Barcode and adapter sequences were removed. Reads shorter than 100bp were removed. Sequences were denoised and chimeras were removed. Sequences reads were clustered at the default 97% identity. OTUs were annotated using the GreenGenes database (69).

Statistical procedures:

Principal components analysis (PCA) was performed in MATLAB using custom developed scripts. Principal coordinates analysis (PCoA) using UniFrac distance (27) was performed using the *beta_diversity.py* script in QIIME. Correspondence analysis (CA) and detrended correspondence analyses (DCA) were performed in the PAST statistical software (70). Canonical correspondence analysis (CCpdA), distance-based Redundancy analysis (dbRDA), Principal response curves (PRC) and variation partitioning as well as variable margination were performed in R using the VEGAN package (71). Random forest (RF) was performed using the RANDOMFOREST package in R. Significant variables in RF were selected based on mean decrease in model accuracy. Orthogonal projection to latent structures – discriminant analysis (OPLS-DA) was performed using the ROPLS package in R. Significant variables were selected based on the

absolute value of weights. Support vector machines (SVM) discriminant analysis was performed using the CARET package in R. Significant variables for SVM were selected using the *varImp* command in the caret package. The CARET package was used for cross-validation and other statistical testing for all three discriminant analyses techniques. Class classification probabilities obtained from the cross-validation tests using the CARET package were used with the built-in perfcurve.m MATLAB function to build Receiver operating characteristics curves (ROC) for model comparisons. Venn diagram for the variation partitioning was performed using the EULER utility (72). The calculation of the Davies-Bouldin index, the visualization of 3x standard error of mean cloud around the centroid and the Monte Carlo Permutation Procedure for statistical significance testing were performed using custom written MATLAB code. To calculate Spearman rank correlations between the microbiota and metabolite datasets, the built-in *corr.m* MATLAB function was used. Additionally, multiple hypothesis testing correction was performed using Benjamini and Hochberg's False Discovery Rate method (FDR) in custom MS Excel templates. Visualization of bipartite networks were performed using the NAVIGATOR software (73).

III. Datasets and scientific questions

Distal gut microbiota and metabolite profiles from IBS and healthy children:

Study design: Fecal samples collected from 22 healthy adolescent volunteers (designated as kHLT, age range: 11-18 years, average: 12.6 years, gender distribution: 10 males and 12 females) and 22 volunteers who were recently diagnosed with IBS-D (designated as kIBS, age range: 8-18, average: 13.2 years, gender distribution: 10 males and 12 females) were subjected to taxonomic analysis using Microbiota Array and metabolomics analysis using H^1 NMR as described in the Materials and methods section. Healthy volunteers were confirmed to not have any GI disease or disorder symptoms. All of the enrolled volunteers were confirmed to not have been on any prebiotic supplementation or antibiotic treatment for at least 6 months prior to fecal sample collection. Volunteers with indication of organic abnormalities such as persistent vomiting, dysphagia, hematemesis, rectal bleeding, fever, weight loss, fatigue and arthritis were excluded from the study. All volunteers diagnosed with IBS-D fulfilled the Rome II criteria for the syndrome (74). Specific inclusion and exclusion can be found in Rigsbee et al 2012. Fecal sample processing and taxonomic data acquisition were performed by Laura Rigsbee. Metabolite quantitation was performed by Daniel Homer.

Scientific questions:

Taxonomic:

1.) Are there differences in the genus abundance profiles in the distal gut microbiota from healthy and IBS children?
2.) Can a classification model be built based on these differences?
3.) Can these differences be identified and ranked?

Metabolomic:

1.) Are there differences in the quantitated distal gut metabolite abundance profiles from healthy and IBS children?
2.) Can an accurate classification model be built based on these differences?
3.) What are the top discriminatory metabolites that contribute to the separation between sample groups?

Associations between taxonomic groups and metabolites:

1.) Can we identify statistically significant associations between microbes and metabolite for the healthy group and IBS group?
2.) Are there differences in the microbe-metabolite associations between IBS and healthy groups?

Distal gut microbiota profiles from patients with Clostridium difficile infection before and after fecal microbiota transplantation therapy:

Study design: Three studied patients suffered from recurrent *C.difficile* infection (CDI) which was first treated using standard antibiotic therapies described previously (75). The fecal microbiota transplantation (FMT) procedure was performed by using concentrated fecal microbiota from healthy donor meeting specific criteria that have been previously described (75, 76). Although the same donor was used for the treatment of all three volunteers, the sample collection from the healthy donor was performed on different dates. Until 2 days prior to FMT, patients were treated with 125mg of Vancomycin, administered orally for four times per day. The day before the FMT procedure, patients received purgative to wash out residual antibiotics

from the intestinal environment. Fresh fecal gavage from the healthy donor were administered to the CDI patients through colonoscopy as previous described (75). Fecal samples for microbiota analysis were collected from the healthy donor on collection dates and from CDI patients before- and after-FMT on specific dates listed in Shankar et al 2014. Taxonomic data was acquired from fecal samples using Microbiota Array as described in the Materials and methods section. Fecal sample processing and taxonomic data acquisition were performed by Vijay Shankar and Amanda Kilburn.

Scientific questions:

1.) Are there significant changes in the distal gut microbiota profiles in CDI patients as a result of FMT?
2.) Is the distal gut microbiota community profile in CDI patients after FMT similar to that of the healthy donor?
3.) How does the distal gut microbiota community change in CDI patients with respect to time (from before-FMT to days after-FMT)?
4.) What are the key microbial drivers of the CDI disease state and healthy state?

Distal gut microbiota and metabolite profiles from healthy US and Egyptian children:

Study design: Fresh fecal samples were collected in sterile containers from healthy pre- and adolescent male volunteers from Giza, Egypt (designated as egkHLT; n=28, average age=13.9 years; average body mass index BMI=18.9 kg/m2) and from Dayton, OH, United States (designated as uskHLT; n=14, average age=12.9 years; average BMI=21.2 kg/m2). Fresh fecal samples were homogenized immediately after collection and frozen as described previously (65)

Healthy volunteers did not have any gastrointestinal symptoms and had not consumed antibiotics or probiotics for at least three months prior to sample collection. For each volunteer, age and BMI values were collected and used in data interpretation. Taxonomic analysis using high-throughput next generation sequencing and metabolomics analysis using H^1 NMR were acquired from fecal samples as described in Materials and methods section. Fecal samples were processed by Mostafa Gouda, Jessica Moncivaiz and Vijay Shankar. Taxonomic data was acquired by Vijay Shankar. Metabolite quantitation was performed by Jessica Moncivaiz.

Scientific questions:

Taxonomic:

1.) Are there significant differences in the genus abundance profiles from distal gut microbiota between healthy US and Egyptian children?
2.) Can a discriminant model be built based on these differences?
3.) What are the top discriminatory genera that separate distal gut microbiota profiles from these two populations?

Associations between taxonomic groups and metabolites:

1.) Are there statistically significant associations between the distal gut genera and metabolites that are common to the Egyptian and US cohorts?

Distal gut microbiota profiles of human populations from industrialized and non-industrialized countries:

Study design: High-throughput next generation-based 16S rRNA sequence data from the previous study (US-vs-Egypt study) was combined with publically available comparable data from the US-vs-Malawi-vs-Venezuela subject comparison study (46), the Tanzania-vs-Italy subject comparison study (28), and the US-vs-Peru subject comparison study (77) for taxonomic analysis. Care was taken to only include sequence data from samples from these studies that were age-matched to those of the US-vs-Egypt study. Taxonomic data from the combined dataset was acquired by Vijay Shankar as described in the Materials and methods section.

Scientific questions:

1.) Can a discriminant model be built to define the differences in the distal gut genus abundance profiles between sample from industrialized and non-industrialized countries?

2.) Which genera contribute the most to these differences?

IV. Specific aims

Current trends in the use of multivariate tools in microbial ecology have been predominantly exploratory in nature, limiting the types of observations that can be made (17). Therefore, I developed four aims to demonstrate that a multitude of biologically relevant insights can be drawn from an extensive application of different types of multivariate techniques. Additionally, a fifth aim was used to integrate these approaches into a protocol for the generalized use of multivariate statistical analyses. These aims are:

1. Use indirect gradient analyses to determine if the largest gradients of variability correspond to differences across sample groups.
2. Apply direct gradient analyses to explain variability within multivariate datasets using known independent variables.
3. Construct discriminant models, compare performances of classifier techniques, and determine variables that are relevant to separation of samples between groups.
4. Identify and evaluate associations among response variables across datasets using correlation based network analyses.
5. Construct a protocol using previous aims for the exploratory and hypothesis-driven analysis of microbiota-related multivariate datasets.

Specific aim 1: ***Use indirect gradient analyses to determine if the largest gradients of variability correspond to differences across sample groups.***

Rationale

Often, the initial question after the generation of high-throughput data from complex microbial communities is if there are large patterns within the response variables (genus, species, etc.,). Additionally, if the datasets are from different sample groups, one might be interested to test if such patterns or gradients relate to separation of samples between sample groups. In order to use these patterns for such biological interpretations, they need to be identified and extracted from the data. While, patterns involving a few variables are easily extracted from datasets through simple visualizations and descriptive statistics, large and complex patterns, which comprise many taxa, for example, are difficult to identify and visualize through conventional means (12). Such exploratory analyses of high-dimensionality data are best accomplished through the use of indirect gradient analyses which result in hypothetical variables (also referred to as latent variables) that are constructed by fitting values of response variables (taxa, metabolite, etc.,) to a specific statistical model that defines how these variables change across a gradient (78). In this aim, we use indirect gradient analyses on datasets obtained from human distal gut microbiota communities and environments to test if these hypothetical variables correspond to sample group gradients in the respective datasets.

Analysis methods:

Multiple indirect gradient analysis techniques were used on the datasets described in the **Datasets and scientific questions** section. The specific tools used for each dataset differ based on the match between the assumptions of the techniques and the overall structures of the respective datasets. The ordination output, which were sample coordinates from the first two latent variables (eigen axes, principal components, canonical axes, etc.,) were visualized as two dimensional ordination graphs. In order to define the distinction of sample clusters, the sample

values from the latent variables were used to calculate the Davies-Bouldin index (79). The Davies-Bouldin index (DB) is defined as a function of the ratio of within-cluster distance (spread of a cluster) to between-cluster centroid (separation of clusters):

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{j \neq i}(D_{i,j})$$

Where k is the number of clusters and $D_{i,j}$ is the within-to-between cluster ratio of clusters i and j and defined as:

$$D_{i,j} = \frac{(\bar{d}_i + \bar{d}_j)}{d_{i,j}}$$

where $\bar{d}_i$ and $\bar{d}_j$ are average distances between each sample point and its respective group centroid, and $d_{i,j}$ is the Euclidean distance between the group centroids (79). Smaller DB index values imply better clustering and cluster separation. In the two-dimensional ordination plots, three standards errors of the mean (3SE) were calculated for each group around the group centroid using custom Matlab code (80). Statistical significance of the DB index was calculated by the Monte Carlo Permutation Procedure (81), which involves comparison of DB index obtained from the analysis to a null distribution generated using random swapping of sample IDs between groups and calculating DB index for each iteration. 10,000 permutations were performed to generate the reported DB index p-values.

Sub aim 1a: Determine if the ordination of genus abundances from fecal microbiota communities can distribute samples from healthy and IBS patients into distinct clusters in ordination space.

Principal components analysis (PCA) and abundance-weighted phylogenetic principal coordinated analysis (PCoA) were performed on the genus abundances obtained using the

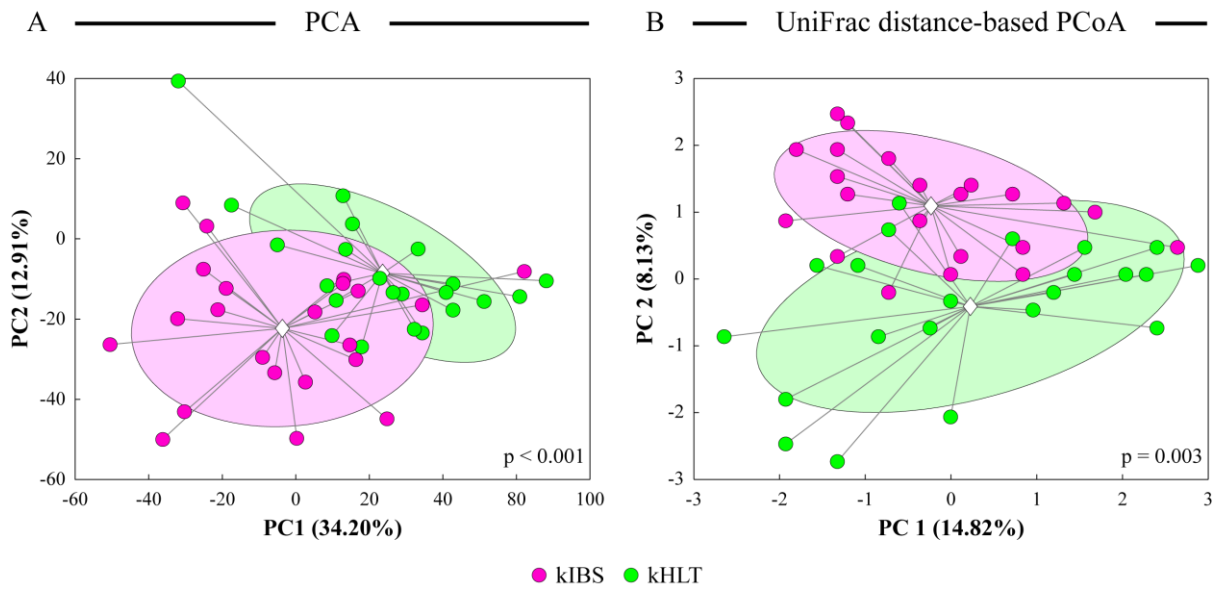

Figure 1. Ordination of genus abundance profiles from distal gut microbiota of healthy and IBS children using Principal components analysis (A) and UniFrac distance based Principal coordinates analysis (B).

phylogenetic Microbiota Array from fecal microbiota of healthy and IBS children. PCA uses Euclidean distances to define relationships between samples, while PCoA uses UniFrac distances, which is a phylogenetic beta diversity metric that takes into account the lineage information of the community membership when calculating similarities between samples (27).

Results:

Both PCA (Figure 1A) and PCoA (Figure 1B) separated samples into distinct clusters based on their health state (healthy - kHLT or IBS - kIBS). For both PCA and PCoA the separation between sample clouds occurred mostly in the second ordination axis (Figure 1). In PCA, the first principal component captured 34.20% of the overall variability and the second principal component captured 12.91% of the overall variability within the dataset. Likewise, in PCoA, the first and the second ordination axes captured 14.82% and 8.13% of the overall variability within the dataset. The DB index values for group separation and their corresponding p-values for PCA and PCoA were 3.296 with a $p<0.001$ and 4.067 with a $p=0.003$ respectively.

Discussion:

In both plots, sample groups separated on the second ordination dimension. This observation might be due to the contribution of large inter-individual variations in taxonomic abundances to the overall variability. Since both PCA and PCoA order eigen axes based on the variability captured (i.e., the axis that captures the most variability in the dataset is ranked as the first ordination axis and so on), in these plots, the first ordination axis likely corresponds to these inter-personal differences. This phenomenon, where inter-individual differences in gut microbiota composition contributes more to overall variability than other consistent changes due to co-factors such as host health state, diet, etc., has previously been reported in several studies (82-84). Little difference was found in the quality of sample group separation between PCA and PCoA (DB index and p-values for each); however, less of the overall variability was captured by the first two dimensions of PCoA when compared to PCA. This observation implies that the use of the phylogenetic distance (UniFrac), in this particular analysis, changes how the overall variability is distributed among the eigen axes and that it does not significantly enhance the latent variable that corresponds to sample group gradient (IBS vs healthy).

Sub aim 1b: Determine if the ordination of quantitative fecal metabolite levels from healthy and IBS patients can distribute samples into distinct clusters in ordination space.

Principal components analysis (PCA) and Correspondence analysis (CA) were performed on quantitated abundances of individual fecal metabolites from healthy and IBS children. PCA was performed with Mahalanobis scaling of input data to reduce the effects of inter-individual variability on the ordination results. We chose to utilize CA on this dataset because we suspected

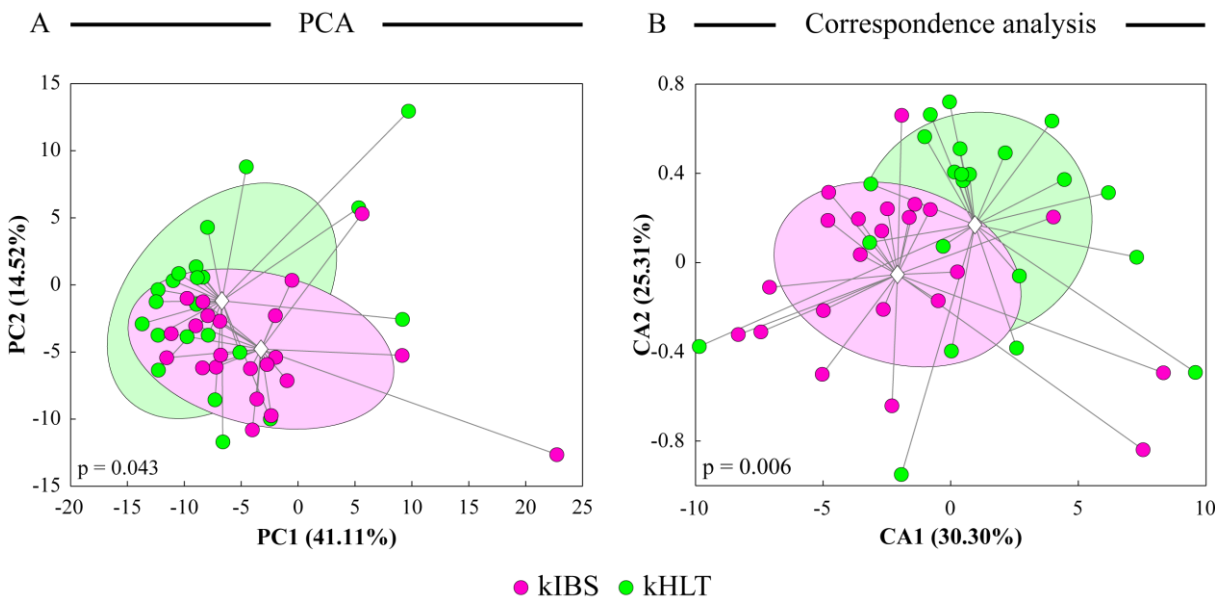

Figure 2. Ordination of metabolite abundance profiles from distal gut environment of healthy and IBS children using Principal components analysis (A) and Correspondence analysis (B).

that fecal metabolite levels changed unimodally in response to known and unknown environmental gradients (explanatory variables) such as disease state, fecal pH, % water content, host age and BMI (see **Introduction** for description of CA).

Results:

Both PCA (Figure 2A) and CA (Figure 2B) partially separated samples in their respective ordination spaces based on the hosts' health state (IBS or healthy). For both PCA and CA, the separation between sample groups occurred on both axes (Figure 2). The proportion of the overall variability captured by the first two principal components in PCA are 41.11% and 14.52% respectively. Likewise, in CA, the percent variability (also referred to as inertia) captured by the first two canonical axes are 30.30% and 25.31% respectively. The DB index values for group separation and their corresponding p-values from PCA and CA were 5.639 with a $p = 0.043$ and 5.193 with a $p = 0.006$, respectively.

Discussion:

While both PCA and CA only partially separated samples between healthy and IBS groups, the separation was nevertheless statistically significant, based on p-values of the DB indices generated through the Monte Carlo Permutation Procedure (p values < 0.05). The loss of a clear separation and tight group clustering can be explained by the following two reasons. Firstly, only a small number of metabolites were chosen for deconvolution and quantitation from the full NMR spectra of the samples. The reduction in the number of variables can result in the representation of only a small fraction of the overall variability within the full NMR spectra. Secondly, the selection of metabolites for quantitation was based on generalized biological responses associated with IBS and previously published reports and may not accurately reflect specific differences in this particular study. While this selection strategy could be effective for most pathological conditions, it may not be suitable for IBS because of the large degree of heterogeneity in symptoms and biomarker responses observed with this syndrome (85, 86).

Sub aim 1c: Determine if the ordination of genus abundance profiles from fecal microbiota communities can distribute samples from healthy US and Egyptian children into distinct clusters in ordination space.

Abundance-weighted principal coordinated analysis (PCoA) with UniFrac phylogenetic distance and Detrended correspondence analysis (DCA) were performed on chord transformed genus abundances from distal gut microbiota of healthy US and Egyptian children. Chord transformation of the input data was performed to correct for the large number of zeroes present within the genus abundance dataset which can lead to false patterns after ordination (87). DCA, which is a variant of CA, was used instead of CA because of the presence of rare genera within

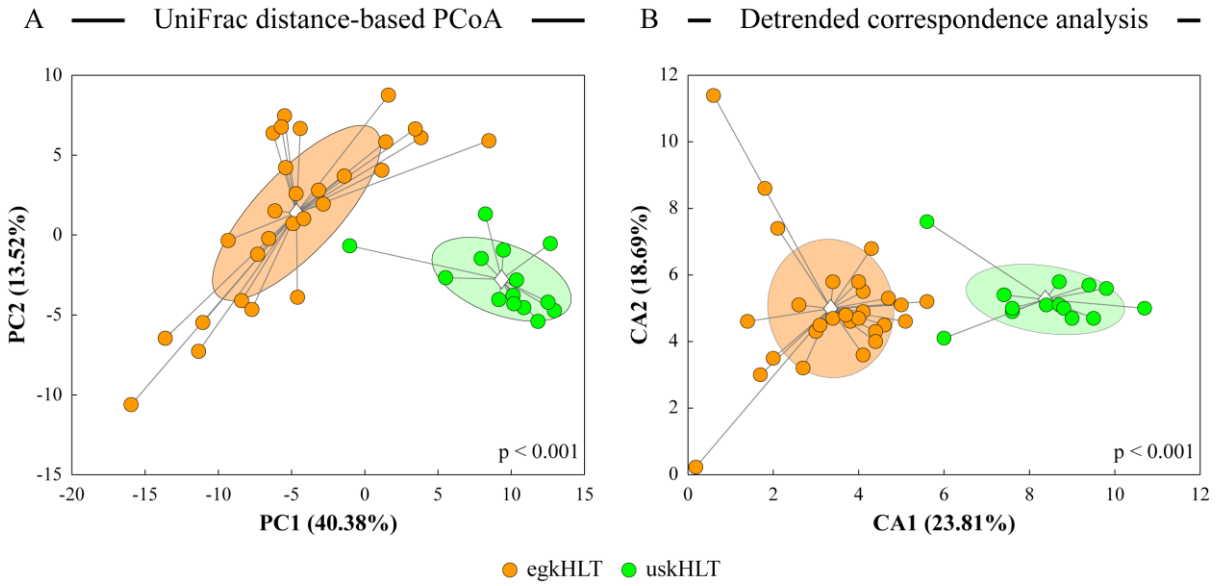

Figure 3. Ordination of chord transformed genus abundance profiles from distal gut microbiota of healthy US and Egyptian children using abundance-weighted UniFrac distance-based Principal coordinates analysis (A) and Detrended correspondence analysis (B).

the dataset (hence the presence of many zeros). Rare variables can lead to shortening of the distances between sample positions at the ends of the ordination axes when using ordination techniques that assume a unimodal variable response (CA, CcpdA, DCA, etc) (25). The detrending procedure, part of DCA, was used to correct for this phenomenon and to preserve the ordination of samples in the first dimension (12).

Results:

A clear separation between the sample groups was observed in both PCoA (Figure 3A) and DCA (Figure 3B). The separation between sample groups in PCoA occurred mostly in the first dimension, which captured 40.38% of the overall variability within the dataset. The second dimension captured 13.52% of the overall variability. In DCA, the separation between sample groups was entirely in the first dimension, which captured 23.81% of the total inertia. The second dimension in DCA contributed to 18.69% of the total inertia. It is important to note that

the reported percent variation captured for this analysis is from CA because the detrending procedure in DCA alters the total inertia, and in addition, only produces a small number output axes (30). The DB index and their corresponding p-values for PCoA and DCA are 1.286 with a p-value <0.001 and 1.217 with a p-value<0.001 respectively.

Discussion:

One of the striking features of these ordination analyses is the clear separation between the Egyptian and US sample groups in ordination space. This observation implies that there are large variations within the genus abundance profiles that correspond to the "country" gradient (i.e., the Egyptian gut microbiota, at the genus level, is very different from the US gut microbiota). In agreement with our findings, several previous reports have shown that gut microbiota from geographically distinct populations differ greatly (28, 46, 77). Also important to note is the difference in the spread of samples within each group in both analyses. The Egyptian sample group had a much greater spread compared to the tightly packed US sample group. The second dimensions in both PCoA and CDA predominantly capture the variability within the egkHLT sample group. One possible reason for this phenomenon might be that there is a greater degree of inter-individual variation in the gut microbiota composition of the Egyptian group compared to the US group. It is also possible that the unequal number of samples within the groups (egkHLT, n = 28, uskHLT, n = 14) contributes somewhat to the sample placements we see in these ordination plots.

Sub aim 1d: Determine if the differences in the microbial phylotype abundances from fecal samples collected from patients with *Clostridium difficile* infection before and after fecal transplantation therapy lead to separate clusters in ordination space.

Phylotype abundance data obtained using the Microbiota Array from distal gut microbiota of patients with *Clostridium difficile* infection (CDI) before and after fecal microbiota transplantation (FMT) were analyzed using Principal components analysis (PCA) and abundance-weighted UniFrac-based Principal coordinated analysis (PCoA). Multiple after-FMT samples indicate fecal samples collected at various time points after the therapy (please refer to **Datasets and scientific questions** section for description of the study).

Results:

Both PCA (Figure 4A) and PCoA (Figure 4B) showed clear separation of before-FMT samples from after-FMT samples for all three CDI patients. The donor samples clustered with the after-FMT CDI samples in both PCA and PCoA. The separation between before-FMT and after-FMT samples was entirely in the first dimension in both ordination analyses, while the second dimension captured the variability within the after-FMT samples. In PCA, the first dimension captured 38.11% of the overall variability, and the second dimension captured 13.02% of the overall variability. Likewise, in PCoA, PC1 captured 65.14% of the overall variability and PC2 captured 10.58% of the overall variability. The DB index for separation of before-FMT sample cluster from the donors and after-FMT sample cluster in PCA and PCoA ordination spaces were 0.849 with a $p<0.001$ and 0.467 with a $p<0.001$, respectively. The overlay of the collection time points against the after-FMT samples in both ordination plots

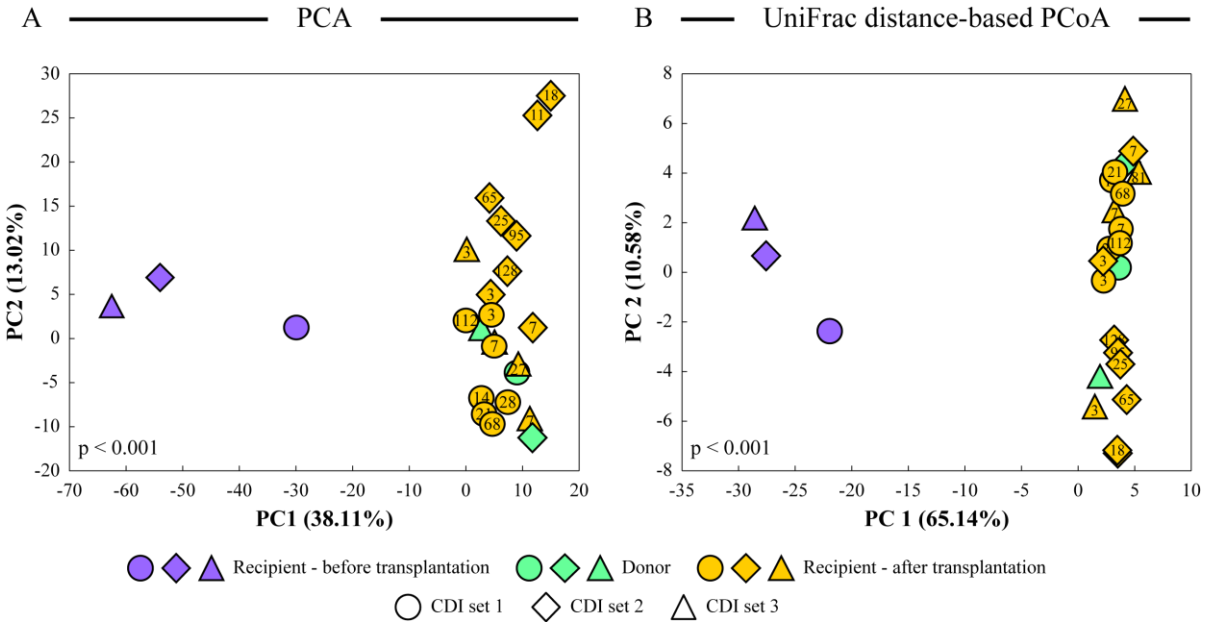

Figure 4. Principal components analysis (A) and abundance-weighted UniFrac-based Principal coordinates analysis (B) of phylotype abundance data from distal gut microbiota of patients with *Clostridium difficile* infection before FMT, after FMT and their respective donors.

indicated that the spread of samples on the second dimension mostly follow a chronological pattern.

Discussion:

The clear separation between before-FMT and after-FMT samples in both ordination analyses imply that the gut microbiota composition between these two groups are starkly different. Also, the fact that donor samples clustered with after-FMT samples indicates that the gut microbiota composition of CDI patients after the therapy closely resemble that of their respective donors. In both ordination analyses, the separation between before-FMT and after-FMT samples occurred in the first dimension, which signifies that this shift in the microbiota composition due to fecal transplantation therapy contributed to the greatest variability within the dataset. It is interesting to note that the before-FMT samples from CDI patients showed considerable variability as indicated by their positions across the first dimension. This implies

that even among the pathology of CDI, inter-individual differences in the gut microbiota composition exist. The unusual, nearly vertical arrangement of after-FMT samples along the second dimension is likely due to the variability in microbiota composition that corresponds to the "time" gradient, as indicated by the collection time points overlayed in the ordination plots.

Specific aim 2: ***Apply direct gradient analyses to explain variability within multivariate datasets using known independent variables.***

Rationale:

Application of indirect gradient analyses distributes samples onto "hypothetical" axes or gradients. The meaning of these synthetic axes and what they correspond to are only implied within these techniques. While they are useful for detecting and extracting these patterns, if the goal of the analyses is to explain the reasons behind the ordering of samples along such gradients, it is important to "constrain" our results from the indirect gradient analyses to independently measured explanatory variables of specific interests. Such a class of techniques, often referred to as constrained ordination analyses or direct gradient analyses, are well suited for this purpose because they are designed to maximize the relationships between the explanatory variables (BMI, age, treatment, etc.,) and the dependent response variables (taxa, metabolite, etc.,) (12). The ability to test associative hypotheses between these sets of variables greatly enhances the interpretation of ordination results and extraction of biological inferences. In this aim, we use direct gradient analyses to identify and quantify the magnitude of independent gradients that explain the ordination of taxonomic and metabolite response variables from various human distal gut microbiota-related datasets.

Sub aim 2a: Determine the effects of fecal pH, fecal percent water content, host age and health state on the variance of fecal metabolite profiles acquired from IBS and healthy patients.

Analysis method:

Canonical correspondence analysis (CcpdA) was used to build relationships between fecal pH, fecal percent water content, host gender and age, and the ordination of fecal metabolite profiles from healthy and IBS children. CcpdA was used because it is well suited to deal with categorical explanatory variables (health state). The quality of the direct gradient model was tested using the pseudo F-statistic, which is the ratio of rank-adjusted constrained and unconstrained total inertia of the model:

$$F = \frac{SS(\hat{\boldsymbol{Y}})/m}{RSS/(n-m-1)}$$

where *n* is the number of samples, *m* represents the degrees of freedom within the model (also represents the number of canonical eigenvalues), $SS(\hat{\boldsymbol{Y}})$ represents the explained variation and *RSS* is the total variation minus the constrained or explained variation (residual sum of squares) (39). Pseudo F-statistic therefore measures how well the constraining variables cumulatively explain the overall variation (inertia) of the response variable ordination. Statistical significance of the analysis is generated through the comparison of the pseudo F-statistic calculated for the original model to a null distribution of the metric calculated from random permutation of fecal metabolite profiles.

Results:

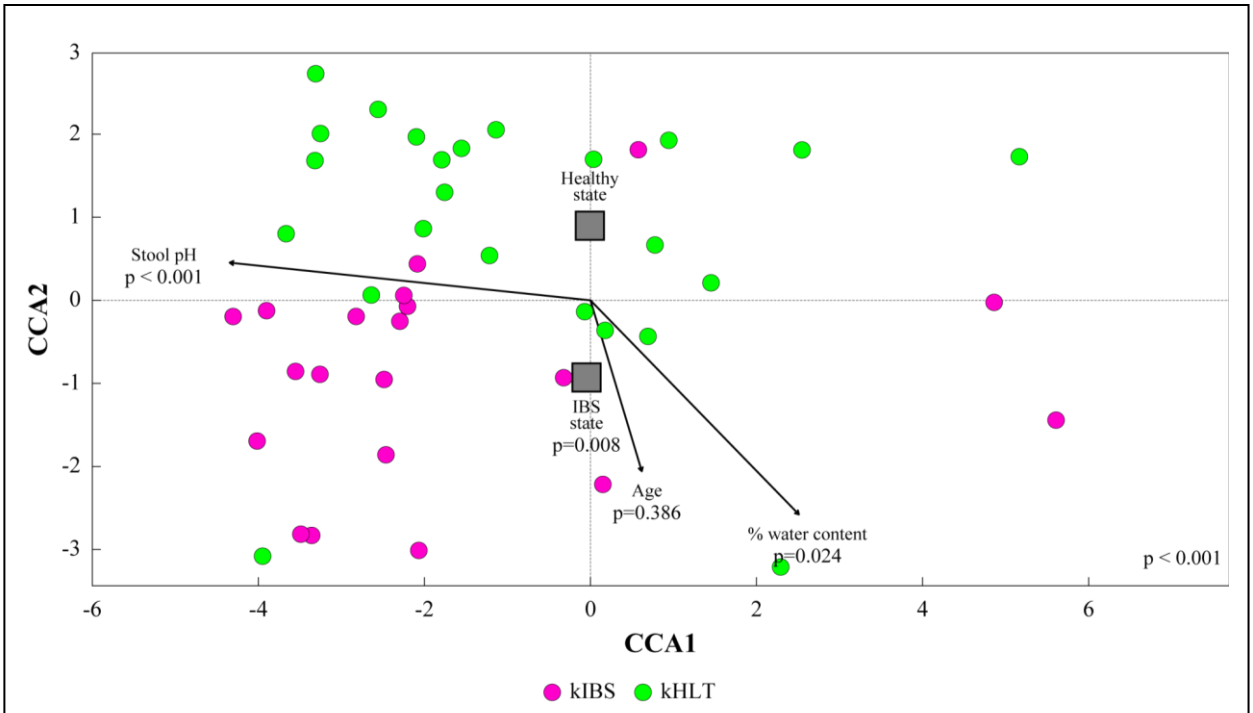

Figure 5. Canonical correspondence analysis on fecal metabolite abundances from healthy and IBS children, constrained by host age, health state, fecal pH and % water content.

The biplot visualization of CCpdA (Figure 5) depicted the ordination of samples based on the constraining variables (fecal pH, % water content, age and health state). The continuous variables (fecal pH, % water content and age) are depicted as arrows, while the categorical variable (health state) is depicted as group centroid. Among the explanatory variables, fecal pH arrow was the longest and aligned mostly with the first canonical axis. The variable with the second longest arrow was the % water content which influenced both the first and the second dimensions. The health state categorical variable mostly separated along the second canonical axis. Cumulatively, constraining variables explained 29.2% of the total inertia within the dataset. The first and the second canonical axes captured 17.0% and 9.4% of the total inertia, respectively. The cumulative pseudo F-statistic for the model was 4.019 with a $p<0.001$. Statistical significance of individual explanatory variables was calculated using iterative

margination of each explanatory variable (71). This permutation test output the following pseudo F-statistic and corresponding p-values for each explanatory variable; Age - 1.239 with a $p=0.386$, Health state - 3.952 with a $p=0.008$, Fecal pH - 7.874 with a $p<0.001$ and Fecal % water content - 3.010 with $p=0.024$.

Discussion:

Among the explanatory variables, fecal pH explained the largest degree of variability, evidenced by the relative length of its arrow compared to the other variables, the value of its pseudo F-statistic and the associated highly significant p-value. It is interesting to note that the axis of variability accounted for by the fecal pH (arrow) was nearly orthogonal to the separation of healthy and IBS sample groups in the canonical ordination space. This implies that fecal pH did not have a large impact on the separation of samples between these groups. This is somewhat of a surprising finding, because changes in luminal pH have been reported in IBS when compared to healthy controls due mostly to altered short-chain fatty acid (SCFA) production (88). It is possible that the differences in methods used for pH measurement in the studies might lead to conflicting results (i.e., fecal pH might not be an accurate representation of luminal pH in the different regions of the large intestine). The explanatory variable that aligns the best with group separation is age (second canonical axis). However, it is important to point out that the inertia captured by age is not only small but also not statistically significant. Therefore, the arrow lengths (which represents the rate at which this variable changes along that direction) alone cannot be used to gauge the importance of an explanatory variable to the fitted constrained ordination. All the known explanatory variables combined only explain 29.2% of the overall

inertia with the dataset. This indicates that a significant portion of the variability has yet to be explained and is due to some unknown environmental gradients.

Sub aim 2b: Determine the variability explained by age, BMI and country in the ordination of fecal microbiota genus abundances profiles from healthy US and Egyptian children.

Analysis method:

Distance-based Redundancy analysis (dbRDA) was performed on genus abundances from the distal gut microbiota of healthy US and Egyptian children with age, BMI and the country of origin as the constraining variables. dbRDA was performed with UniFrac distances in order to take advantage of the phylogenetic relationships of the taxa during sample ordination. Although CCpdA is more suitable for ecological gradients (microbial communities respond unimodally to environmental or independent gradients), dbRDA can still provide comparable performance when the gradients of the measured explanatory variables are short (12). The pseudo F-statistic, and statistical significance testing through comparison to null distribution generated using random permutations described in the previous section was utilized in this sub aim as well, to assess the quality of the constrained ordination. Additionally, variation partitioning of the dbRDA output was used to determine the relative contribution of each explanatory variable to the overall variation explained.

Results:

The outputs of dbRDA and subsequent variation partitioning were depicted as sample-explanatory variable biplot (Figure 6A) and Venn diagram (Figure 6B) respectively. In the

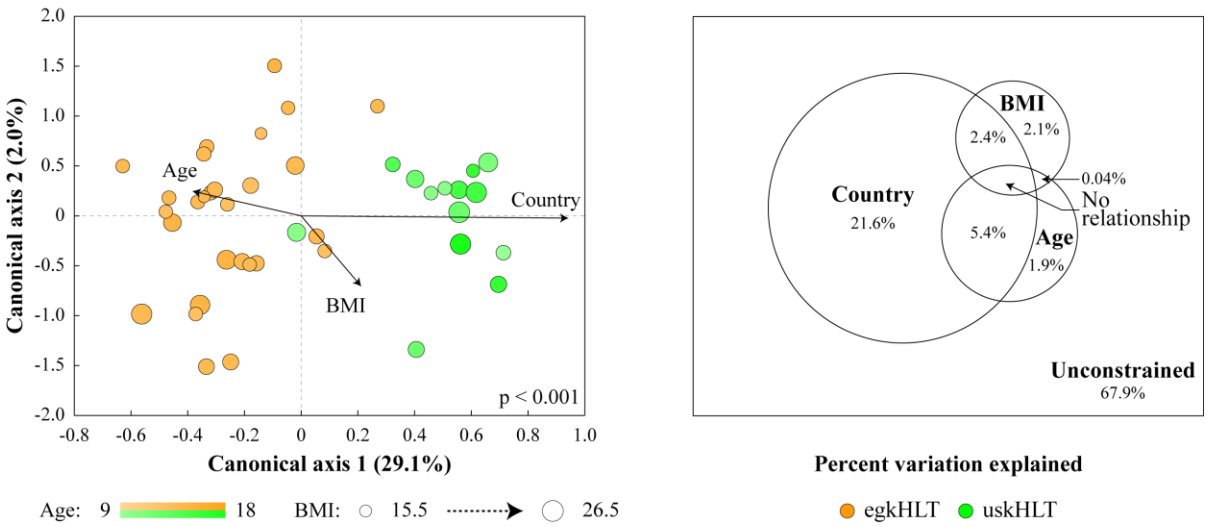

Figure 6. Distance-based Redundancy analysis (A) and subsequent variation partitioning (B) performed on genus abundances from distal gut microbiota of healthy US and Egyptian children, constrained by country, age and BMI.

dbRDA biplot, a clear separation was seen between the sample groups (Egyptian vs US) which coincided with the direction of the "country" arrow and the first canonical axis. The arrow corresponding to the country of origin was the longest among the explanatory variables, followed by age and BMI. The color gradient and size of the sample dots depicted the numerical values of the age and BMI respectively for each sample. The variability captured by the first and the second constrained canonical axes were 29.1% and 2.0%, respectively, and all three explanatory variables cumulatively explained 32.1% of the overall variability within the dataset. In agreement with the arrow lengths in dbRDA biplot, variation partitioning indicated that the country of origin for the samples was the dominant explanatory variable and accounted for 29.3% of the overall variability, while age and BMI contributed considerably less (7.3% and 4.5% respectively). The pseudo F-statistic for the overall model containing all three explanatory variables was 6.403 with a $p<0.001$.

Discussion:

Among the three variables included within the constraining model, country of the sample (Egyptian or US) was the strongest gradient as indicated by the length of the arrow in the dbRDA biplot and the Venn diagram of the variation partitioning. This indicates that the majority of the variations in the genus abundance profiles can be explained by the sample group gradient. Given the large contribution of the variable to the overall inertia, it is likely that there are very distinct and large changes in genus abundance profiles when comparing samples between groups. In contrast, both age and BMI only mildly contribute to the observed overall variability. This is evidenced by the lack of a strong congruency between the variable arrows and the superimposed sample dot information (age and BMI). There is not a clear, distinct pattern (change across) in either the color gradient or the dot size that aligns with the direction of age or BMI arrows.

Sub aim 2c: Elucidate time-dependent changes in the genus abundance profiles of fecal samples collected from patients with *Clostridium difficile* infection before fecal microbiota transplantation therapy (FMT) and subsequent collections after FMT.

Analysis method:

Principal response curves analysis (PRC) was performed on the genus abundances profiles of fecal samples collected from CDI patients before-FMT and after-FMT, and their respective donors to illustrate the changes in their gut microbiota composition with respect to time. Since PRC is built upon the framework of Redundancy analysis (RDA), Euclidean distance was used as the metric to define (dis)similarities between samples. The F-type test statistic and Monte Carlo permutation procedure were used to test RDA under the reduced model framework

(See description of PRC in **Introduction** for details). In order to calculate statistical significance of the PRC model, the data from all three patients and their respective donors were combined into a single dataset. This procedure could only be done for the before-FMT time point, Day 3 and Day 7, because these were the only time points that were shared among all three patients. This step was necessary because permutation-based testing of PRC requires replicates of both conditions (after-FMT and donor microbiota composition). Statistical significance was calculated by comparing the F-type statistic on the overall model and comparing it to its null distribution (generated through calculation of the statistic for the PRC model after random permutation of samples within each time point) (35). In this case, the F-type statistic and the permutation procedure tests the quality of PRC model in explaining the effect of FMT on the change of gut microbiota of CDI patients over time. In order to determine patient-specific gut microbiota changes and to depict long term effects, PRC was also run separately for each patient such that their after-FMT samples (post-FMT time points) were compared to both their respective before-FMT sample and their specific healthy donor sample, resulting in two reference points for each patient curve. Key drivers of CDI and healthy donor state were derived from PRC based on the model weights of the genera. Genera with highly negative scores represented drivers of the pathogenic state and conversely, those with large positive scores depicted members driving toward the healthy donor state.

Results:

PRC analysis run separately for each patient shows a sharp change in community composition moving away from before-FMT profile toward a state that resembles that of each patient's respective donor profile (Figure 7). The change is evident at the earliest time point of

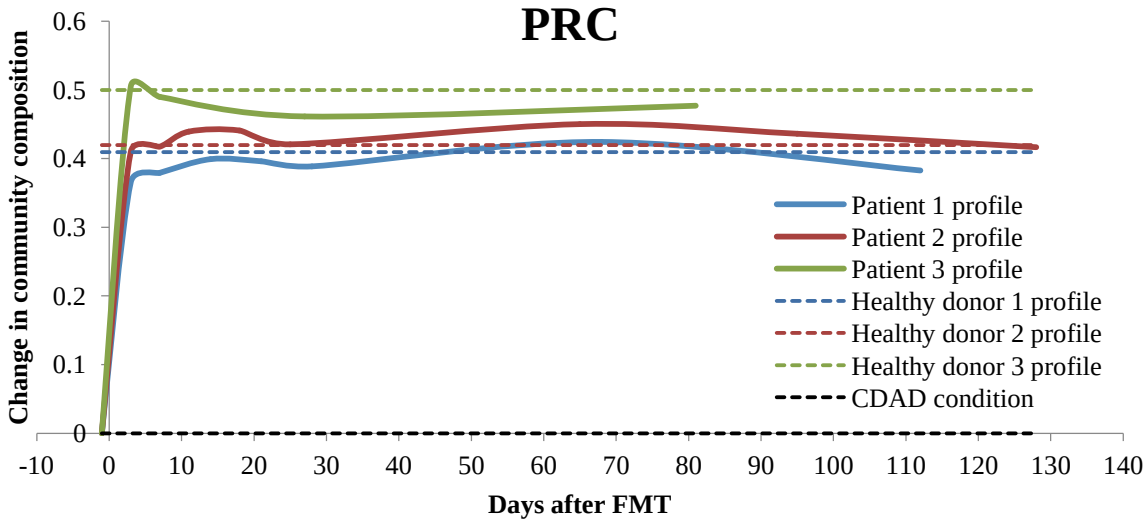

Figure 7. Principal response curve analysis performed on distal gut microbiota profiles from CDI patients before- and after-FMT, and their respective healthy donors.

collection (Day3) after FMT procedure. Also, this effect was seen consistently with all three patients (Figure 7). There was some variability within the donor profiles (green dotted line) where one of the healthy donors had large enough microbiota differences in comparison to CDI condition (black dotted line) to separate from the other two health donor profiles (blue and red dotted line). While there were some fluctuations over time, all three patient curves stably followed their respective donor profiles for time points as far as Day 128. Table 1 shows top 10 genera for each patient based on their negative and positive model weights (species coefficients). While a significant number of both positive and negative drivers were shared among all three patients, there were key patient specific differences (Table 1). The negative drivers comprised mostly of facultative anaerobic genera and conversely, the positive drivers were obligate anaerobes. For the statistical significance testing of model quality, F-type statistic for the overall model (Before-FMT, Day 3 and Day 7) and the associated p-value generated through Monte Carlo permutation were 27.66 with a $p<0.001$. The same analysis was performed with margination of each time point from the overall model to determine time point specific F-type

statistic and associated p-value. This analysis produced the following values; Before-FMT - 1.22 with a p=0.354, Day 3 - 17.75 with a p=0.069 and Day 7 - 17.61 with a p=0.057.

Table 1. List of top 10 genera based on PRC weights that either drive toward CDI or healthy state

Top genera driving toward a pathogenic state

Patient 1 Genus	Species scores	Patient 2 Genus	Species scores	Patient 3 Genus	Species scores
Streptococcus	-3.139	***Raoultella***	-3.169	***Lactobacillus***	-2.796
Escherichia/Shigella	-3.090	***Enterobacter***	-3.120	*Enterococcus*	-2.752
Bifidobacterium	-3.041	*Streptococcus*	-3.070	***Enterobacter***	-2.709
Haemophilus	-2.992	***Lactobacillus***	-3.021	***Escherichia/Shigella***	-2.665
Lactobacillus	-2.942	*Klebsiella*	-2.971	***Raoultella***	-2.621
Raoultella	-2.893	***Veillonella***	-2.922	***Veillonella***	-2.578
Enterobacter	-2.844	***Escherichia/Shigella***	-2.872	*Zymophilus*	-2.534
Veillonella	-2.795	*Lactococcus*	-2.823	*Klebsiella*	-2.490
Prevotella	-2.746	*Zymophilus*	-2.773	*Cupriavidus*	-2.447
Ruminococcus	-2.697	*Rothia*	-2.724	*Herbaspirillum*	-2.403

Top genera driving toward a healthy state

Patient 1 Genus	Species scores	Patient 2 Genus	Species scores	Patient 3 Genus	Species scores
Blautia	3.188	***Blautia***	3.219	***Blautia***	2.840
Faecalibacterium	3.139	***Coprococcus***	3.169	***Coprococcus***	2.796
Dorea	3.090	***Faecalibacterium***	3.120	***Faecalibacterium***	2.753
Roseburia	3.041	***Roseburia***	3.070	***Dorea***	2.709
Holdemania	2.992	*Holdemania*	3.021	*Bifidobacterium*	2.665
Subdoligranulum	2.943	***Dorea***	2.971	***Roseburia***	2.622
Bacteroides	2.893	***Papillibacter***	2.922	*Anaerostipes*	2.578
Papillibacter	2.844	*Anaerotruncus*	2.872	*Subdoligranulum*	2.534
Adlercreutzia	2.795	*Bacteroides*	2.823	***Papillibacter***	2.490
Coprococcus	2.746	*Akkermansia*	2.773	*Adlercreutzia*	2.447

Genera that are shared among all three patients are bolded

Discussion:

One of the most striking finding in the analysis was how quickly the healthy donor microbiota established itself with the CDI patients after FMT. Even by day 3, CDI patients' distal gut microbial communities greatly resembled those of the donors'. This rapid shift and

stabilization is clearly depicted in the PRC plot. Another interesting finding is the fact that this 're-colonization' of gut microbiota in the guts of CDI patients is not a transient event and that it is stable even as far as 128 days after FMT. The derivation of variable weights from the PRC model indicate that genera that are responsible for the shift toward a pathogenic CDI state are mostly facultative anaerobes. This finding is indeed consistent with several reports that have shown that the presence of facultative anaerobes in the distal gut is an indication of various disease pathology (89-91). Furthermore, many of these genera are members of the family Enterobacteriacea, to which belong many pathogenic microbes (*Enterobacter, Escherichia, Shigella* and *Raoultella*) (92, 93). Conversely, genera that drive toward a healthy donor state are obligate anaerobes. Genera such as *Blautia*, *Faecalibacterium*, *Dorea* and *Coprococcus* are considered common beneficial members of a healthy gut environment (94). Some of these genera ferment plant polysaccharide to produce short-chain fatty acids (SCFA) which have been shown to have many positive effects of host colonocytes (95). Additionally, *Faecalibacterium* is generally thought to have many positive healthy benefits which include regulation of the mucosal immune responses and intestinal cell differentiation (96). As expected and in agreement with the sharp shift in gut microbiota over as little as 3 days, the statistical testing of the overall model (Before-FMT, Day 3 and Day 7) indeed resulted in a highly significant p value ($p<0.001$). The statistical testing analysis run on individual terms (time points) however did not result in highly significant p-values. This is likely because of the small number of samples per time point. Since the testing procedure relies on generating a null distribution for F-type statistic by randomizing the sample identities, with small number of sample, the randomization likely ran out of combinations. Nevertheless, the inertia captured by the individual terms are a good indication of the shift in the microbial composition. The low value at the Before-FMT time is

based only on inter-personal variability, hence small (F=1.22). In comparison, the values for Day 3 and Day 7 are very large, indicating that the model has captured a large proportion of the variability which is likely due to the FMT therapy (F=17.75 and 17.61, respectively).

Specific aim 3: ***Construct discriminant models, compare performances of classifier techniques, and determine variables that are relevant to separation of samples between groups.***

Rationale

When analyzing multivariate data originating from distinct sample groups, one biologically relevant question to ask is if there are differences in the variables among samples which can be used to build a consistent pattern that explains the separation of samples between groups. Identification of these differences can help explain the potential biological reasons behind the group separation. For example, identification of microbial groups and/or luminal metabolites that differ between healthy controls and patients with colorectal cancer can help us determine the etiology and the subsequent pathology of the condition (97). Alternatively, even if multiple sample groups do not exist in the dataset, but exploratory analyses indicate the presence of distinct clusters in ordination, it might be of interest to determine which variables are responsible for the observed clustering. A class of methods, often referred to as discriminant analyses, which aim to maximize differences between groups specified *a priori*, are best suited to answer such queries. Additionally, because these techniques maximize differences between groups by building sets of patterns using the response variables, these patterns (more popularly referred to as discriminant models) can be used to classify or predict the grouping of new, unknown samples based on their response variables. This feature is highly relevant in a clinical

setting where rapid and accurate identification of sample type can lead to efficient treatment strategies. In this aim, we use discriminant analyses to build and test predictive models, compare the different techniques based on their discrimination performance and identify the discriminatory variables that explain the differences between sample groups in datasets related to human distal gut microbial ecology.

Analysis method:

Three different classifier techniques were used to analyze all three datasets. These techniques are Random forest (RF), Orthogonal projection to latent structures discriminant analysis (OPLS-DA) and Support vector machines (SVM). These three techniques build discriminant models using different strategies (refer to **Introduction** for descriptions of the techniques). Discriminant models were tested for over-fitting and statistical significance using *k*-fold cross-validation, with *k* varying based on the dataset. For OPLS-DA, Q^2 (predictive power of the model), R^2X (variation in response variables not pertaining to class) and R^2Y (variation in response variables pertaining to class) metrics were used to assess the performance of the models. Statistical significance for OPLS-DA was generated using the comparison of Q^2 to the permuted Q^2 threshold (sample identity swapping and recalculation of Q^2). For the statistical significance of RF models, Davies-Bouldin index comparison to null distribution calculated on the multi-dimensional scaling (MDS) of the random forest proximity matrix was used. In order to identify discriminatory variables using each classifier technique, the following strategies were used: For RF, the mean decrease in model accuracy with random permutation of variable values (importance score) was used, for OPLS-DA, the absolute values of weights (coefficients of variables in the discriminant function) was used, and for SVM, the decrease in area under the

Receiver operating characteristics (ROC) curve with random permutation of variable values was used. Classifier performances were assessed by comparing cumulative accuracies of class assignment (model accuracy) and through construction of ROC curves from class assignment probabilities for all three techniques.

Sub aim 3a: Build discriminant models to represent differences in fecal microbiota genus abundance profiles between healthy and IBS patients and identify genera that contribute the most to this discrimination.

Prior to application of the discriminant techniques, the data were preprocessed to remove unwanted artifacts and to satisfy the assumptions of the techniques used. The genus abundance data obtained from Microbiota Array were mean-centered (subtract corresponding means of variables from variable values throughout the dataset which results in each variable mean centered on zero) and scaled (divide each variable array by its standard deviation resulting an even spread for all variables). To test for the over-fitting of each model, a 22-fold cross-validation (CV) procedure was used with each technique to calculate classification accuracy. 22-fold CV was used because this dataset contained 22 samples per group. A 22-fold CV ensures that the folds (fraction) are uniformly divided (2 samples per fold) while still maintaining a large number of CV tests (each fold acts as a test set) for a stable classification (42, 43).

Results:

All three techniques produced highly discriminant models (Table 2). RF produced a statistically significant model indicated by the Davies-Bouldin index and its associated p-value derived from the multidimensional scaling of the proximity matrix (DB index – 3.042 with a

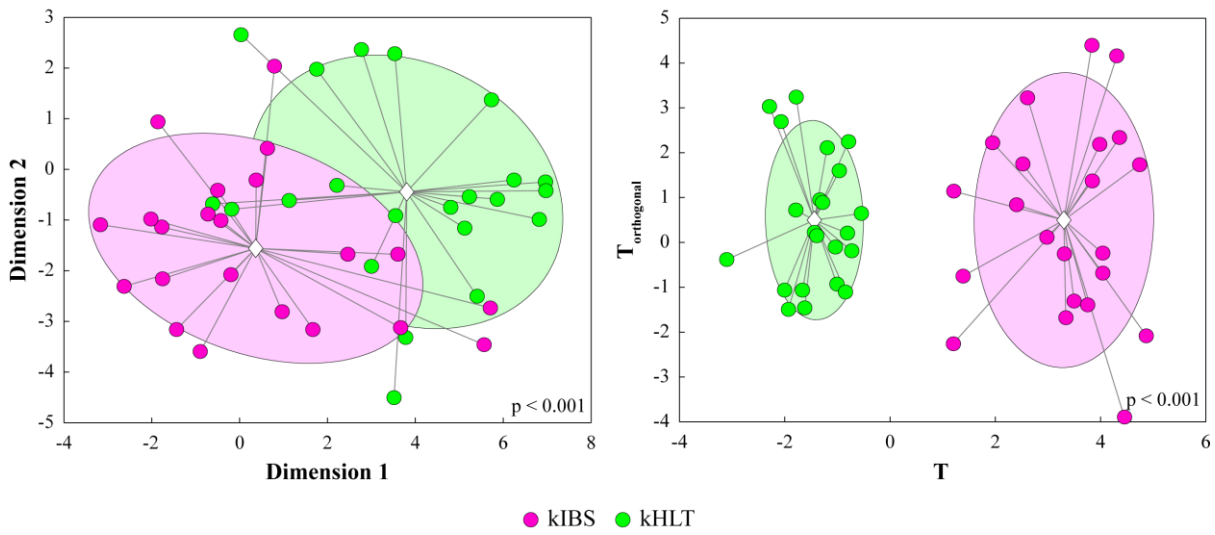

Figure 8. Discriminant analysis using Random forest (A) and Orthogonal projection to latent structures (B) performed on genus abundance data from the distal gut of healthy and IBS children.

p<0.001). The OPLS-DA model was also highly statistically significant with a cumulative Q^2 = 0.427 and a p-value<0.001. DB index and the associated p-value for the T-vs-$T_{orthogonal}$ plot was also statistically significant (DB index – 1.858 with a p<0.001). As expected the visualization of both RF (Figure 8A) and OPLS-DA (Figure 8B) showed distinct clustering of samples based on their sample groups; however OPLS-DA showed much clearer separation of sample groups compared to RF's MDS plot (also indicated by a smaller DB index value compared to that of RF). Other quality parameters for the OPLS-DA model include R^2X = 0.15 and R^2Y = 0.88 with an associated p=0.002. The comparison of accuracy for the three models after 22-fold cross-validation indicated that RF performed the best, followed by OPLS-DA and SVM performed the worst (Table 2). Similarly, comparison of the area under the ROC curves for the three models confirmed the trend seen with the model accuracy (Figure 9). Comparison of the top 10 discriminatory genera from the three models showed that there is a high degree of congruency between the models. The genera *Parasporobacterium*, *Papillibacter*, *Gemella*, *Oxalobacter*,

Solobacterium and *Actinomyces* (bolded in Table 2) were consistently found as top discriminatory genera with all three models.

Table 2. Comparison of discriminant analyses using kIBS-kHLT genus abundance dataset (shared genera are bolded)

	RF	OPLS-DA	SVM
Accuracy	86.4%	75.0%	65.9%
AUC	0.898	0.754	0.732
Top 10 discriminatory genera	**Mean decrease in accuracy**	**|Weights|**	**Mean decrease in AUC**
	Parasporobacterium	***Papillibacter***	***Parasporobacterium***
	Oxalobacter	***Parasporobacterium***	***Papillibacter***
	Bryantella	***Gemella***	*Bryantella*
	Papillibacter	***Oxalobacter***	***Gemella***
	Eubacterium	***Solobacterium***	***Oxalobacter***
	Gemella	*Dorea*	***Solobacterium***
	Enterobacter	***Actinomyces***	*Ruminococcus*
	Raoultella	*Roseburia*	*Mogibacterium*
	Solobacterium	*Mitsuokella*	***Actinomyces***
	Actinomyces	*Coprobacillus*	*Roseburia*

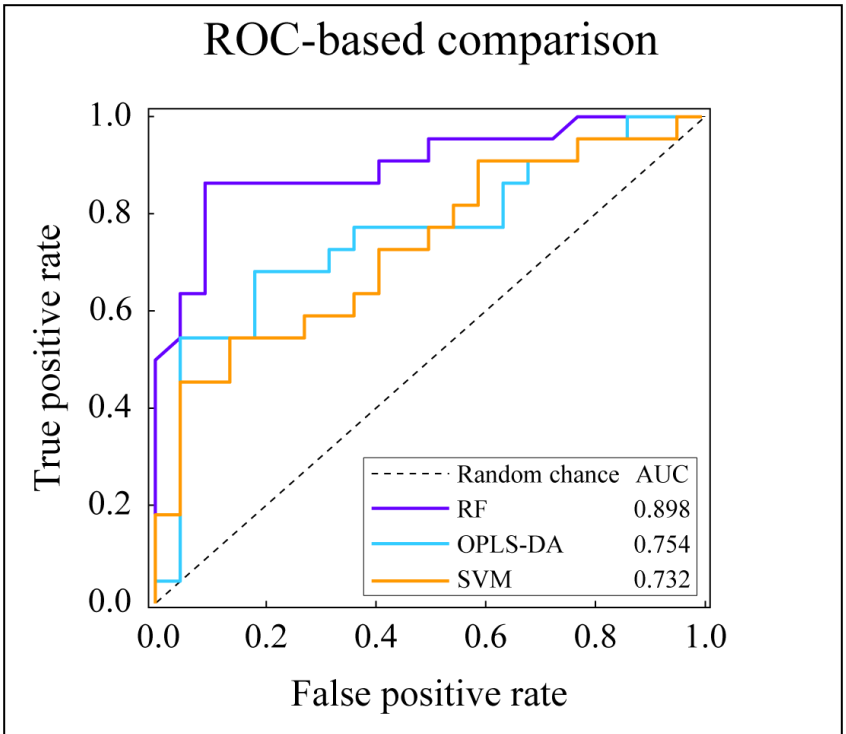

Figure 9. Comparison of kIBS vs kHLT genus abundance model performance using ROC analysis of the class assignment probabilities obtained from each model.

Discussion:

The fact that all three techniques produce good quality discriminant models implies that there are consistent and substantial differences in the distal gut genus abundances profiles when comparing healthy children with those suffering from IBS. This is confirmed by the observation that all three techniques labelled nearly the same list of genera (listed above and in Table 2) as highly discriminatory. The differences in the list of top discriminatory genera are likely because of different biases (or the lack of) introduced by the assumptions and approaches of these techniques (44, 45). Although, all three techniques performed well in their ability to discriminate between the sample groups, RF was shown to perform the best, based on model accuracy and AUC from ROC analysis after cross-validation. This high performance of RF is likely attributable to its decision tree-based approach. Because a very large number of decision trees are used for the voting procedure, the technique is highly stable with regard to outlier trees, model variances and biases. Additionally, this characteristic high performance of RF has recently been reported in a study by Knights and colleagues, where the authors compared the performance of several classifiers, including RF and SVM, using datasets from human-associated microbiota and showed that RF performed the best, followed by SVM (44). Interestingly, despite such a high classification performance, the proximity matrix based visualization of RF output shows only a partial separation compared to the clear separation seen with OPLS-DA's T vs $T_{orthogonal}$ plot (Figure 8). This phenomenon might have resulted from the differences in how the visualizations are generated. With OPLS-DA, the discriminant axis is directly plotted as the T axis (linear combinations of weighted variable scores that explain group separation). Whereas, with RF, the multidimensional scaling of the proximity matrix, an indirect approach, is utilized for visualization. Because MDS rotates and transforms the proximity matrix

to derive and rank axes based on the amount of captured variability, the first two axes (used here for visualization), may not directly be related to the differences between groups.

With regard to the genera that were identified as highly discriminatory between healthy and IBS state, very little information is currently available on the functional capabilities of these members and the potential role that they might play in IBS pathology. A more comprehensive approach, combining newer, more sophisticated techniques such as metagenomics and metatranscriptomics might be able to elucidate the link between these microbiota members and IBS. Also, while it is tempting to use these observations in the context of clinical diagnosis as markers for IBS, it should be noted that these findings might not fit a more general model of IBS simply due to the modest sample size used in this study. Nevertheless, these finding provide an excellent foundation for future studies that could explore the relationship between these genera and IBS.

Sub aim 3b: Model the differences in fecal metabolite profiles of healthy and IBS children and identify the metabolites that contribute the most to model separation.

Similar to **sub aim 3a**, metabolite abundance data obtained from H^1 NMR spectra were mean centered by subtracting the means from variable values and normalized by dividing by the variable standard deviation. Because the same set of samples from **sub aim 3a** were used for NMR-based fecal metabolomic analysis, the models were tested for over-fitting with a 22-fold cross-validation for the same reasons stated in **sub aim 3a** (uniform division of the dataset with a large number of CV tests).

Results:

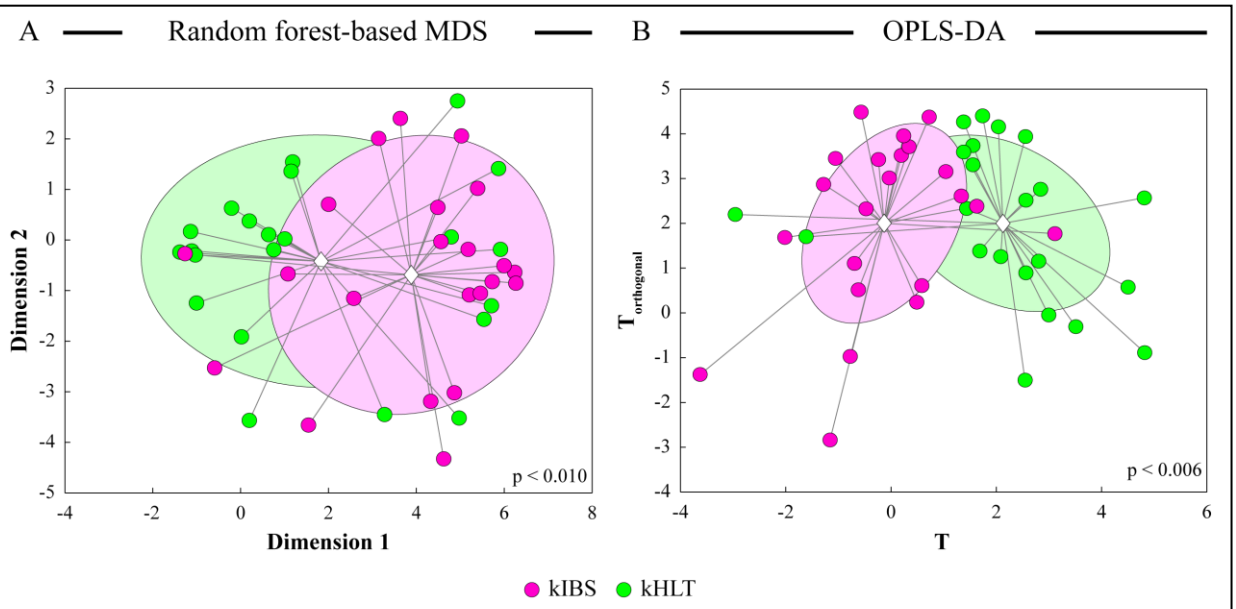

Figure 10. Discriminant analysis using Random forest (A) and Orthogonal projection to latent structures (B) performed on metabolite abundance data from the distal gut of healthy and IBS children.

All three techniques produced a fairly accurate model (Table 3). Although RF had a moderately good model accuracy of 77.3%, the MDS plot of the RF proximity matrix shows only a partial separation of the clusters. Nevertheless, this separation was statistically significant based on a DB index and associated p-value (5.19 with a p=0.010, Figure 10A). The OPLS-DA model was also statistically significant with a cumulative Q^2 value of 0.16 and a p=0.010. Contrary to RF proximity MDS, the OPLS-DA's T-vs-$T_{orthogonal}$ plot showed distinct clustering of samples based on their respective groups. As expected, the DB index calculated on the T-vs-$T_{orthogonal}$ plot was better and more statistically significant compared to RF MDS plot (4.16 with a p=0.006, Figure 10B). The OPLS-DA model captured substantial variability within the dataset that pertained to the separation of sample groups, as indicated by the cumulative R^2Y parameter and its associated p-value (0.38 with a p=0.02). However, R^2X (variability captured not pertaining to sample group separation) was also quite large for this OPLS-DA model (0.53). Comparison of model

Table 3. Comparison of discriminant analyses using kIBS-kHLT met abundance dataset (shared metabolites are bolded)

	RF	OPLS-DA	SVM
Accuracy	77.3%	79.5%	70.5%
AUC	0.734	0.823	0.698
Top 6 discriminatory genera	**Mean decrease in accuracy**	**|Weights|**	**Mean decrease in AUC**
	Tyrosine	**Formate**	**Tyrosine**
	Formate	Pyruvate	**Formate**
	Pyruvate	**Glucose**	Lysine
	Lactate	Lysine	**Glucose**
	Leucine	**Tyrosine**	Leucine
	Glucose	Methylamine	Lactate

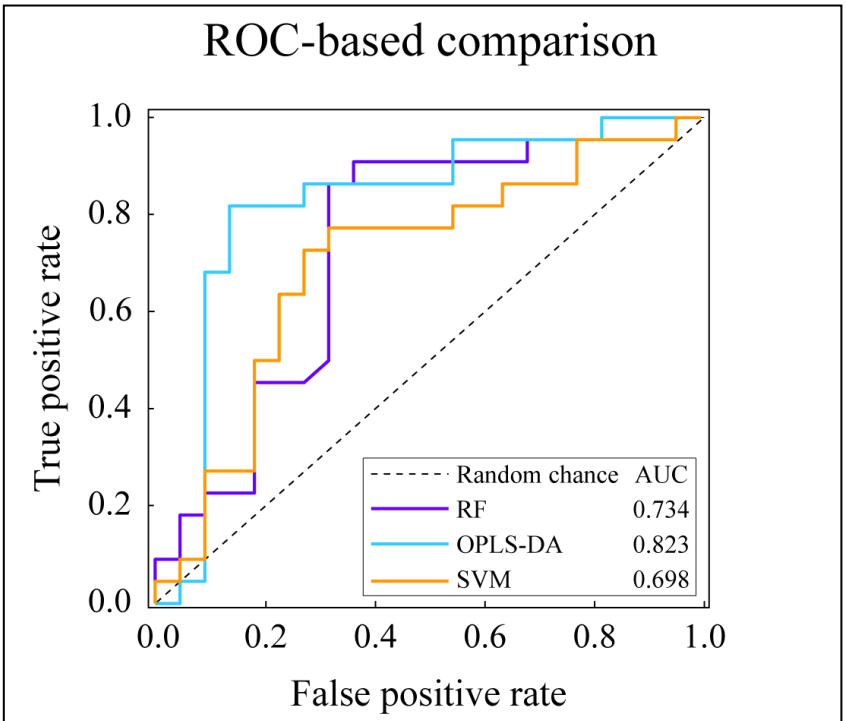

Figure 11. Comparison of kIBS vs kHLT metabolite abundance model performance using ROC analysis of the class assignment probabilities obtained from each model.

accuracies from cross-validation indicate that OPLS-DA model performed the best, followed by RF and SVM (Table 3). This trend was consistent when comparing the models' respective AUCs

from their ROC plots (Figure 11 and Table 3). Comparison of the top discriminatory metabolites indicated that there was a fair degree of congruency among the three models. Specifically, all three models indicated that formate, tyrosine and glucose were key discriminators of the sample groups (bolded in Table 3).

Discussion:

Similar to **sub aim 3a**, discriminant models constructed on metabolite abundances obtained from fecal samples collected from healthy controls and IBS children were statistically significant. This observation indicates that, similar to the distal gut microbiota, the metabolite profiles in the distal gut environment also are different between these two cohorts. There is also a fair degree of consistency between the models, which is indicated by the similarities in comparison of the top discriminatory metabolites (formate, tyrosine and glucose are shared among all three techniques and pyruvate, leucine and lysine are shared among two of the three models). While all three techniques produced statistically significant models, OPLS-DA performed the best among these three techniques. This is somewhat surprising because when using the same cohort with distal gut microbiota data in **sub aim 3a**, RF outperformed the other two techniques. This observation indicates that the structure of the metabolite data better suites the assumptions and requirements of OPLS-DA. Indeed, the number of variables between the distal gut microbiota and metabolite datasets was quite different (more than 50 genera compared to 19 metabolites), and this difference might play a role in the performance of the classifiers on these datasets. In agreement with RF's moderate performance, the MDS plot of the proximity matrix only showed a modest separation. It is possible that the variability pertaining to the

sample group separation might not be among the largest gradients within the dataset and therefore are not being depicted in the first few axes of MDS.

Metabolites identified by the models as key discriminators of healthy-IBS metabolite profiles are indeed consistent with the pathological features of IBS. Elevated levels of tyrosine and several other amino acids (see supplementary material for (58)) in the IBS cohort are in agreement with several reports that claim that there is increased proteolysis in this syndrome (98). The elevated levels of glucose in the IBS cohort might be an indication of incomplete metabolic pathways as a result of reduced metabolite cross-feeding among distal gut community members. This phenomenon has in fact, been reported as a characteristic of IBS in previous studies (59, 74, 98). Finally, elevated levels of formate in the distal gut environment of diarrhea-predominant IBS (IBS-D) children might imply that the microbiota functional pathway for the utilization of formate to produce hydrogen as an end product is missing. This is relevant to the current context because hydrogen gas has been shown to increase the gut transit time and has been linked to the incidence of constipation-predominant IBS (IBS-C) (99). Therefore, the surplus of formate in IBS-D is indeed consistent with the diarrhea-based pathology in our cohort.

Sub aim 3c: Use discriminant models to identify genera that contribute the most to differences in the fecal microbial profiles between healthy US and Egyptian children.

The genus abundance data from the distal gut microbiota of healthy US and Egyptian children acquired through the use of high-throughput next-generation sequencing was chord-transformed before being used for discriminant analysis. As stated in **sub aim 1c**, chord-transformation of abundance data is especially suited for dealing with datasets that contain a large number of zeroes (this is especially the case when many rare taxa are present in the dataset)

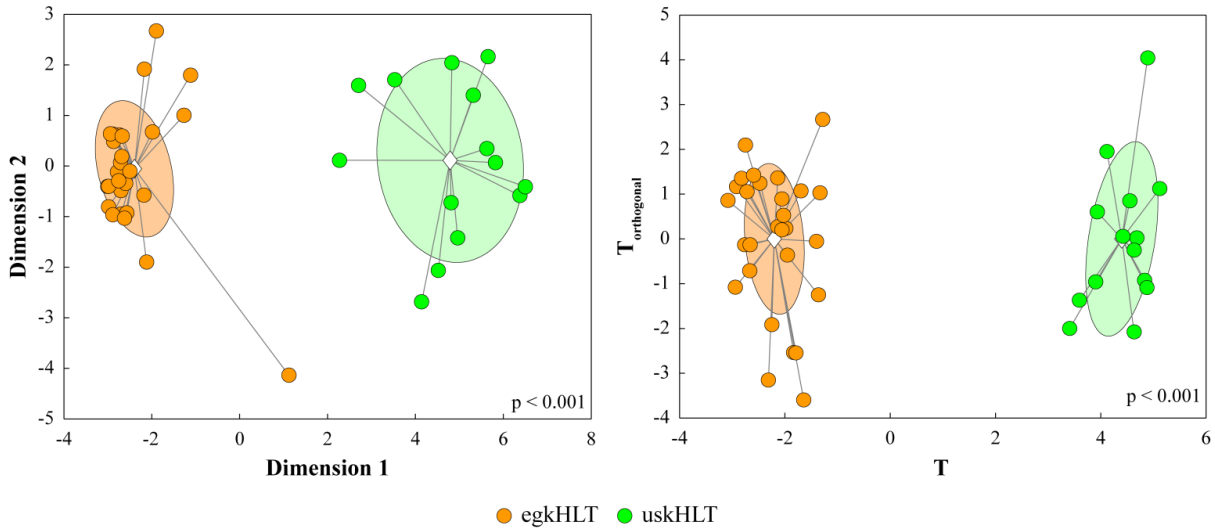

Figure 12. Discriminant analysis using Random forest (A) and Orthogonal projection to latent structures (B) performed on the chord-transformed genus abundance data from the distal gut of healthy US and Egyptian children.

(87). Additionally, the dataset was mean centered and normalized (scaling variable spread) before being analyzed with RF, OPLS-DA and SVM. Because the dataset contains 14 US samples and 28 Egyptian samples, we used a 14-fold cross-validation approach. As previously stated, this ensures that the total number of samples are divided evenly among the folds (3 samples per group). 14 folds, instead of 21 folds were used because the sample groups have unequal number of samples. With 14 folds, each fold can now contain 3 samples and will have a higher probability of representing the distribution of the full dataset (2:1 ratio).

Results:

All three techniques produced models with very high accuracy (Table 4). In agreement with the high accuracy of RF, the MDS plot of the proximity matrix shows very clear separation of sample groups (Figure 12A) and a highly statistically significant DB index (0.88 with a

Table 4. Comparison of discriminant analyses using egkHLT-uskHLT genus abundance dataset (shared genera are bolded)

	RF	OPLS-DA	SVM
Accuracy	100.0%	97.6%	95.2%
AUC	1.000	0.999	0.996
Top 10 discriminatory genera	**Mean decrease in accuracy**	**|Weights|**	**Mean decrease in AUC**
	Bacteroides	***Bacteroides***	***Bacteroides***
	Blautia	***Blautia***	***Catenibacterium***
	Catenibacterium	***Ruminococcus***	***Blautia***
	Coprococcus	***Anaerostipes***	***Coprococcus***
	Prevotella	*Adlercreutzia*	***Prevotella***
	Ruminococcus	***Coprococcus***	*Eubacterium*
	Eubacterium	*Faecalibacterium*	***Ruminococcus***
	Anaerostipes	***Prevotella***	*Mitsuokella*
	Adlercreutzia	***Catenibacterium***	***Anaerostipes***
	Mitsuokella	*Oscillospira*	*Faecalibacterium*

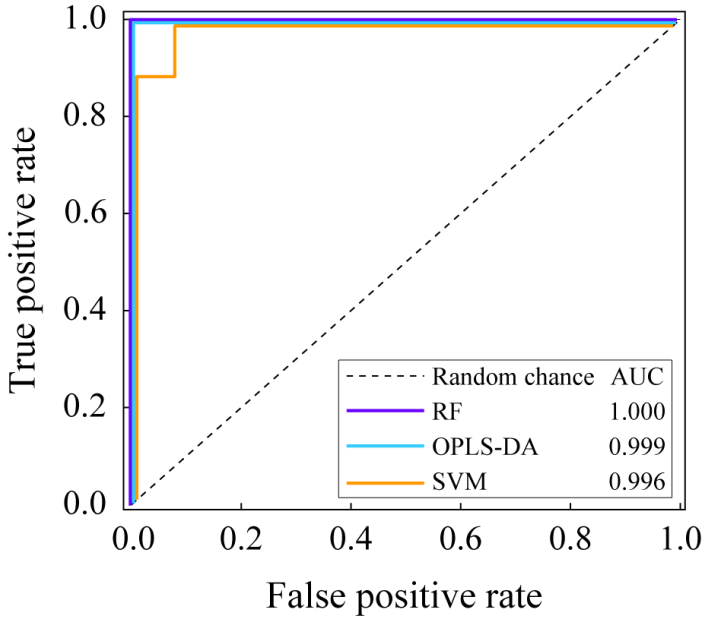

Figure 13. Comparison of egkHLTvs uskHLT genus abundance model performance using ROC analysis of the class assignment probabilities obtained from each model.

$p<0.0001$). Consistent with RF, OPLS-DA model was also highly discriminatory between the sample groups, indicated by a statistically significant cumulative Q^2 (0.82 with a $p<0.001$) and a large and statistically significant cumulative R^2Y (0.98 with a $p<0.001$). The cumulative R^2X (0.13) was relatively small compared to R^2Y, indicating that most of the variability captured by the OPLS-DA model corresponds to separation between sample groups. As expected, the visualization of OPLS-DA's T-vs-$T_{orthogonal}$ plot resulted in a clear separation between the sample groups and generated a highly statistically significant DB index (0.83 with a $p<0.0001$, Figure 12B). Comparison of model performances using accuracy and AUC from ROC analysis indicated that although all three models showed very high discriminatory performance, RF performed the best, followed OPLS-DA and SVM (Figure 13 and Table 4). All three models consistently identified *Prevotella, Bacteroides, Blautia, Catenibacterium, Coprococcus, Ruminococcus* and *Anaerostipes* as the top discriminatory genera.

Discussion:

The RF model showed 100% accuracy for the model classification after the *k*-fold cross-validation. Similarly, OPLS-DA and SVM were closely behind RF in their classification performance. These results are somewhat surprising, given the modest sample sizes and unequal groups sizes. This consistent, very high performance of the discriminant models is likely due to large fundamental differences within the overall distal gut genus abundance profiles between the two population groups (Egyptian and US children). This observation is further supported by the high level of consistency in which genera are identified by the three techniques as highly discriminatory. Similar to the observation in **sub aim 3a** and in contrast to that of **sub aim 3b**, RF outperformed OPLS-DA and SVM with this dataset, albeit only by a small margin. This

further supports the possibility that RF performs better with datasets that contain a large number of variables compared to OPLS-DA (this dataset contained 129 genera). Contrary to the lack of a clear separation between sample groups in the MDS plots of the RF proximity matrix from **sub aim 3a** and **sub aim 3b**, the MDS plot from RF on this dataset shows a very clear separation. This implies that the gradients that correspond to the differences between Egyptian and US children's gut microbiota are among the largest contributors to the overall variability. Indeed, this is clearly evident even in the indirect gradient analysis performed on the same dataset in **sub aim 1c**. Such large difference in the microbiota composition have, in fact, been reported by several studies that have interrogated the distal gut microbial communities in geographically distinct human populations (28, 46, 77).

With regard to the genera that were found to be highly discriminatory between these two sample groups, higher abundance of *Bacteroides* in the US population and the reciprocally higher abundance of *Prevotella* in the Egyptian population (*Bacteroides*: 11.0% vs 2.7%, respectively; *Prevotella*: 7.3% vs 18.0%, respectively) are thought to be due to the substantial differences in the dietary composition of these two host populations. Many members of the genus *Bacteroides* are highly adapted to be able to degrade dietary proteins, a common, highly abundant component of the Western diet (100, 101). Likewise, the genus *Prevotella* comprises several known indigestible polysaccharide degrading members and these members have likely adapted to take advantage of higher relative composition of plant-based fibrous foods in the Egyptian/Mediterranean diet compared to that of the Western diet (102). Similarly, higher abundances of *Ruminococcus*, *Coporoccus* and *Blautia* in the US population are also likely due to diet differences. It is possible that these starch-degrading genera are being selected for by the higher composition of starch in the Western diet (103). The reason for the higher abundance of

Catenibacterium in the Egyptian population is yet to be fully understood; however, some studies have reported higher abundances of this genera in the distal gut environments of other non-western human populations and in co-occurrence with *Prevotella* (104, 105).

Sub aim 3d: Build discriminant models using a cumulative dataset from multiple studies that have interrogated distal gut microbiota from industrialized and non-industrialized countries, and identify the main drivers of the separation between these two groups.

The available datasets from studies that have assessed the distal gut microbiota profiles from human populations of geographically distinct countries were combined and analyzed together with our dataset from **sub aim 3c**. Because the cumulative dataset contained a very diverse set of variables and many rare genera, it resulted in a large number of zeros within the dataset. We transformed the genus abundances from this dataset using the chord transformation to correct for the number of zeros. Additionally, we also mean centered and scaled the dataset to improve the performance of some of the discriminant techniques. For cross-validation testing of the model, we used 37-fold CV, because there were 370 samples within this dataset. This ensured an even splitting of the full dataset.

Results:

Owing to the large number of variables (genera), all three discriminant analyses produced highly accurate models. RF produced a statistically significant model indicated by a DB index of 0.719 with a $p<0.001$ for the MDS plot of the RF proximity matrix (Figure 14A). Visualization of the MDS plot indicates that the first dimension separates samples based on the discriminant axis (industrialized vs non-industrialized). The second MDS axis mostly separates the egkHLT-

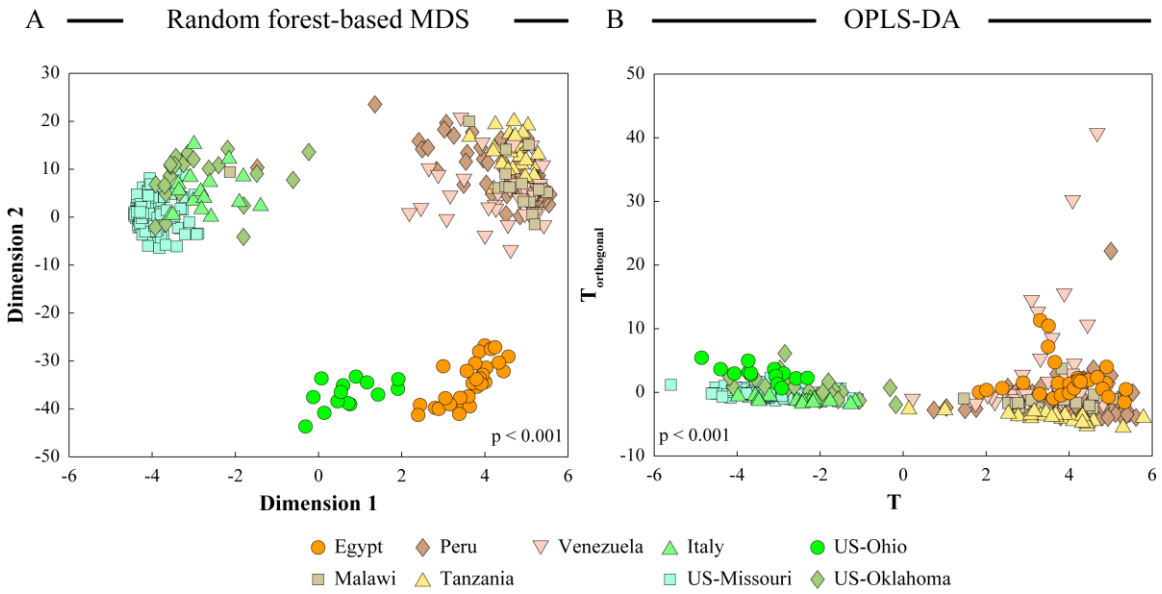

Figure 14. Discriminant analysis using Random forest (A) and Orthogonal projection to latent structures (B) performed on the chord-transformed genus abundance data of fecal samples from industrialized and non-industrialized countries.

uskHLT dataset (**sub aim 3c**) from the rest of the cumulative dataset. The OPLS-DA model, in agreement with RF, was also highly discriminatory between distal gut genus abundance datasets from industrialized and non-industrialized countries, indicated by a statistically significant cumulative Q^2 of 0.787 with a $p<0.001$, a large and statistically significant R2Y (0.928 with $p<0.001$) and a relatively small R2X (0.092). Likewise, a clear separation was seen between the sample groups in the visualization of OPLS-DA using T-vs-$T_{orthogonal}$ plot (Figure 14B). This separation was statistically significant based on a DB index of 1.381 with a $p<0.001$. Comparison of model performance using the CV model accuracies indicate that while all three models were very highly accurate, RF had the best performance, albeit only by a very small margin. Both OPLS-DA and SVM have nearly identical accuracies (97.6%). Similarly, comparison using AUC from ROC analysis indicated that RF outperformed the other two techniques (Figure 15 and Table 5). Based on AUC, SVM was slightly better in terms of

Table 5. Comparison of discriminant analyses using cumulative human distal gut genus abundance dataset from industrialized and non-industrialized countries (shared genera are bolded)

	RF	OPLS-DA	SVM
Accuracy	99.1%	97.6%	97.6%
AUC	0.999	0.995	0.996
Top 10 discriminatory genera	**Mean decrease in accuracy**	**|Weights|**	**Mean decrease in AUC**
	Prevotella	***Prevotella***	*Bacteroidales*
	Bacteroides	***Bacteroides***	***Prevotella***
	Bacteroidales	***Alistipes***	***Succinivibrio***
	Catenibacterium	*Ruminococcus*	***Catenibacterium***
	Alistipes	***Holdemania***	*S24.7*
	Succinivibrio	*X02d06*	***Alistipes***
	S24.7	*CF231*	***Bacteroides***
	YS2	***Succinivibrio***	*YS2*
	Bulleidia	***Bulleidia***	***Holdemania***
	Holdemania	***Catenibacterium***	***Bulleidia***

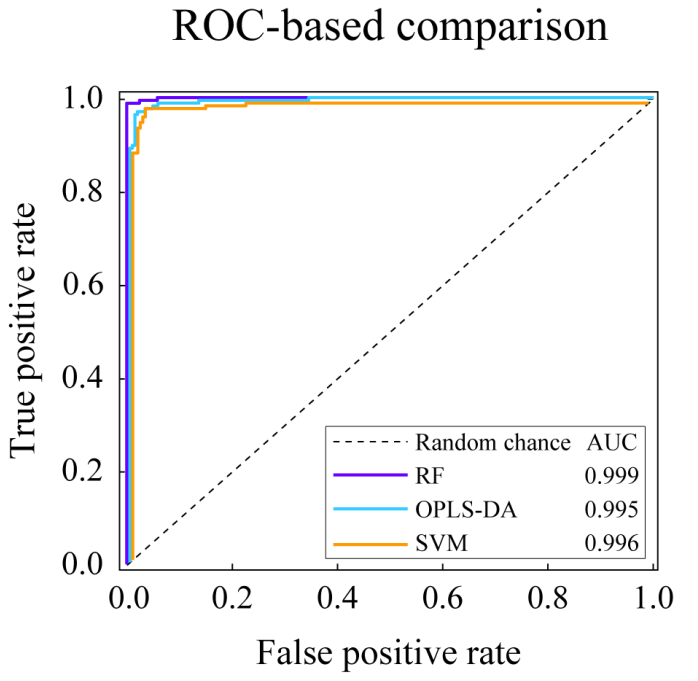

Figure 15. Comparison of cumulative dataset genus abundance model performance using ROC analysis of the class assignment probabilities obtained from each model.

discriminatory performance compared to OPLS-DA (AUC: 0.996 vs 0.995). All three models identified nearly the same set of genera as top discriminatory variables from the dataset, however, there was some variability in the ranks of the genera among the techniques (Table 5). The genera that were consistently found by all three techniques as highly discriminatory between these sample groups are *Prevotella, Bacteroides, Catenibacterium, Alistipes, Succinivibrio, Bulleidia* and *Holdemania.*

Discussion:

The very high accuracy achieved by all three models after CV testing implies that there are large differences in the profiles of distal gut microbiota in these sample groups. The high accuracy is also likely due to the large number of variables, because with a larger pool of variables there would be a higher probability to find combination (or patterns/decisions) of variables that can explain the discrimination between the sample groups (106, 107). Comparison of model performances using both accuracy and AUC after CV indicated the RF had the best performance. This observation, in combination with the results of the other sub aims, further validates the idea that RF has a very high discriminatory performance when the number of variables is large within the dataset. Although, with this particular dataset, both OPLS-DA and SVM were only slightly behind RF in terms of discriminatory performance. This implies that there are other parameters or features (other than number of variables) of the dataset that affect the performance of these techniques. An interesting observation to note with RF is the visualization of the MDS plot from the RF proximity matrix. While the first dimension did correspond to the separation between the sample groups, the second dimension separated our dataset from the rest of the cumulative dataset. This likely indicates a technical problem with the

processing of the sequence data, given the fact that all of the studies (including ours) interrogated the distal gut microbiota communities using the V4 variable region of the 16S rRNA gene. Also interesting to note is the fact that despite this separation of our dataset from the rest in the MDS plot, the DB index for this analysis was better than that of the OPLS-DA's T-vs-$T_{orthogonal}$ plot, which showed slightly better clustering (0.719 vs 1.381). This difference is likely because of the differences in the distance between the centroids in these two plots (DB index is a function of this distance). It is important to indicate that the quality of the clustering between these two plots should not be directly compared since the method used to visualize these two plots are very different, as indicated in the discussion from **sub aim 3a**.

With regard to the biological significance of the top discriminatory genera, *Prevotella* (higher in non-industrialized group) and *Bacteroides* (higher in industrialized group) being identified as the top two genera by RF and OPLS-DA is not surprising when considering their functional role in the processing of dietary components. As mentioned previously in the discussion of **sub aim 3c**, these two genera comprise members that are capable of degrading very different dietary substrates. The differences in the diets consumed by the populations that are part of this cumulative dataset is thought to the likeliest reason for the differences in abundances that we see with these genera. Additionally, a recent publication that attempted to cluster distal gut microbiota communities from throughout the world based on compositional similarities have reported the existence of three major clusters. These clusters, also now popularly referred to as 'enterotypes', are characterized based on key drivers of the entire cluster (microbial members that are consistently found at high abundance within the cluster of samples). The study indicated that the three major enterotypes were driven by the abundances of *Prevotella*, *Bacteroides* and *Ruminococcus* respectively (108). The observations from our analysis, in combination with this

study, indicate that this cumulative dataset likely comprises two different enterotypes, each driven either by *Prevotella* or *Bacteroides.* As indicated in the discussion of **sub aim 3d**, the higher abundance of *Catenibacterium* is linked to the presence of *Prevotella* in the non-industrialized group. This observation might be due to a yet-to-be characterized cross-feeding relationship between the members of these two genera. Finally, it is interesting to note the higher abundance of *Succinivibrio* in the non-industrialized group, since this genus comprises several pathogenic members that have been associated with gastrointestinal diseases (28, 109). This observation is indeed in agreement with a higher incidence of pathogen-related gastrointestinal diseases in developing, non-industrialized countries (110).

Specific aim 4: ***Identify and evaluate associations among response variables across datasets using correlation based network analyses.***

Rationale:

When two or more sets of response variables can be independently measured from the same set of samples, it presents a unique opportunity to find associations or links between these sets of variables. For example, measurement of microbial abundances and metabolite profiles for the same set of samples allows for an integrative analysis approach and lets us extract the microbe-metabolite relationships in the context of host health or disease. While direct gradient analysis can be used to analyze these types of dataset, due to the large number of constraining variables, the output can become difficult to interpret (16). A viable strategy for analysis of such datasets is to construct correlation networks within and between the variable sets. Although correlation does not directly imply causation, this type of analysis can be used as a hypothesis

generation tool to test possible associations using additional future experiments. In this aim, we use correlation analyses to construct bipartite networks with two sets of variables (metabolite and genus abundances) measured for the same set of samples, to identify and evaluate biologically relevant interactions between these variables.

Analysis method:

In order to build associations between the genus abundances and quantified metabolite levels measured from the human distal gut environment Spearman rank correlation analysis was used. Spearman rank correlation was used instead of Pearson because this method is rank-based and non-parametric, and therefore does not assume that the dataset has a specific distribution. The associated p-value for each correlation was corrected for multiple-hypothesis testing using the Benjamini-Hochberg's false discovery rate (FDR) correction (111). FDR defines that the correlation is significant if:

$$p_i \leq \frac{k_i \times \propto}{m}$$

where p_i is the associated p-value for the correlation, k_i is the rank of the p-value, α is the %FDR threshold and m is the number of tested comparisons. The associations that were found to be statistically significant after FDR correction were visualized as bipartite networks where the nodes represented the variables and edges connecting the nodes represented the correlations between the variables. The size of the nodes was used to represent the relative abundance of the variables. The color and the thickness and the transparency of the edges were used to represent the direction and the magnitude of the correlation between the connected nodes.

Sub aim 4a: Determine the differences in putative fecal microbiota-metabolite associations between healthy and IBS children.

The genus abundance data acquired from the distal gut microbiota of children with IBS and healthy controls using the Microbiota Array and the metabolite levels for the same cohort acquired using H^1 NMR were analyzed using Spearman rank correlations. Only the top 46 most abundant genera (based on an abundance cutoff of 0.1%) were used for this analysis. This filtering step ensured that spurious correlations between metabolites and very low abundance genera could be avoided, since these tend to often be erroneous associations due to the high variability in the values of low abundance variables (112). For this dataset, correlation analyses were run separately for the IBS and healthy cohorts, so that differences in the correlation networks could be identified. In addition to the correlation analysis, the correlation matrix for each set were bi-clustered by metabolites and genera, in order to identify local clusters of associations that are similar. For multiple hypothesis testing using FDR, the α threshold was set to 10% (we expect 10% of the statistically significant associations detected by correlation analysis to be false). This threshold has been shown to be appropriate for datasets of this size and nature (113). Associations that were identified as statistically significant were used to construct the bipartite network.

Results:

Heat map visualization of the correlations between metabolite and genera abundances performed separately for healthy controls and IBS cohorts revealed many strong associations in the healthy set (Figure 16A), while only weak associations were identified in the IBS set (Figure

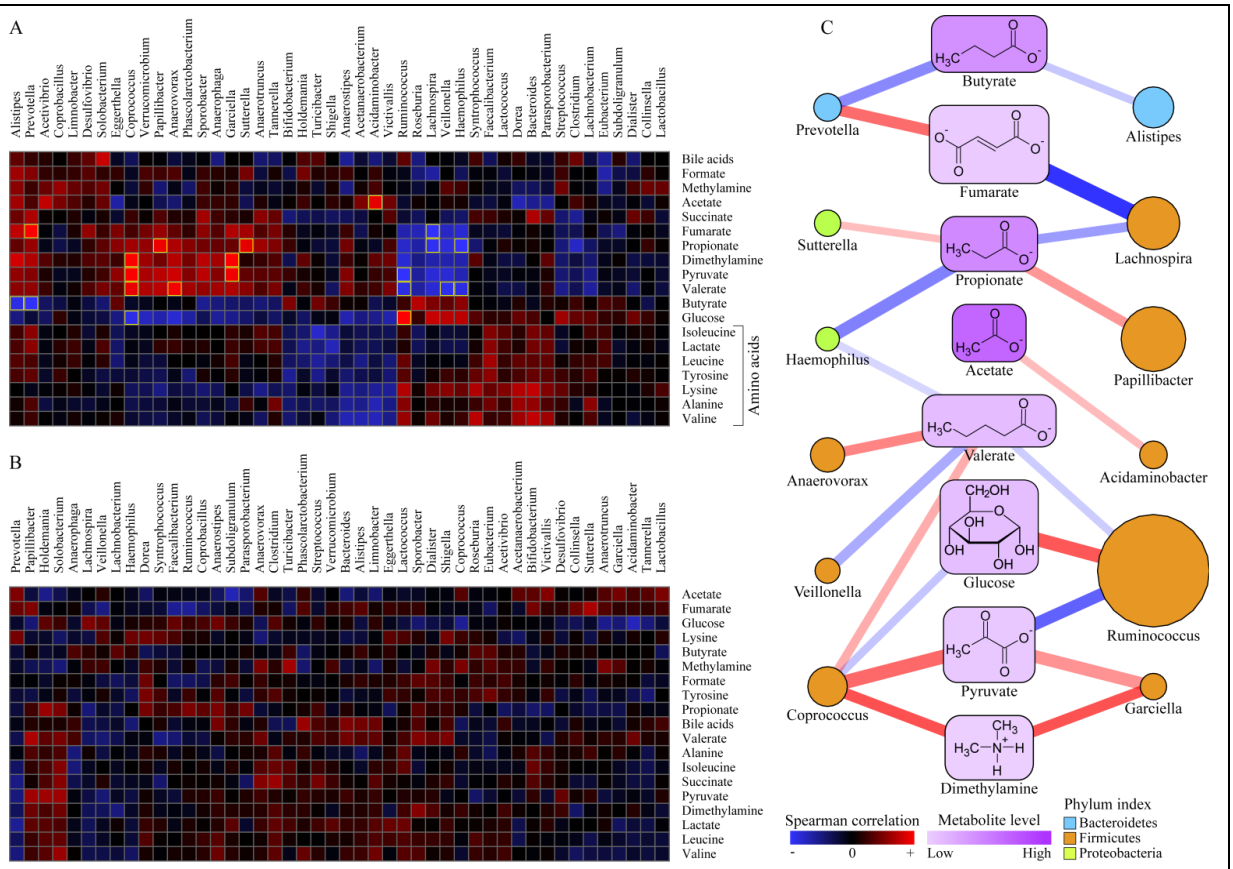

Figure 16. Spearman rank based correlations between the distal gut metabolite and genus abundance profiles displayed as bi-clustered heatmaps for healthy controls (A), and IBS children (B). Statistically significant associations from the healthy set are also displayed as a bipartite network (C).

16B). Bi-clustering of the correlation matrices resulted in the formation of organized local clusters in the healthy set, but not in the IBS set. For example, clear clustering of all amino acids and clustering of carbohydrate metabolism intermediates (SCFAs, fumarate, succinate and pyruvate) were evident in the healthy set. With FDR threshold set at 10%, 21 statistically significant microbe-metabolite associations were found in the healthy set (Figure 16C). No statistically significant correlations were found in the IBS set after FDR correction. The strong, statistically significant associations found in the healthy set included a positive correlation between *Ruminococcus* and glucose, a negative correlation between *Coprococcus* and glucose, and positive correlations between *Acidaminobacter* and acetate, *Coprococcus* and valerate and,

Prevotella and fumarate (Figure 16C). No significant correlations were found between the measured metabolites and the cellulolytic and other polysaccharide-degrading genera such as *Bacteroides* and *Bifidobacterium.*

Discussion:

The significant loss in both the strength and the number of associations in the IBS set, compared to those of the healthy set is an interesting observation. It has been reported previously that the pathology of IBS is accompanied by a significant loss in microbe-microbe associations (59). The observations from this analysis, in combination with the previous reports imply that there is a loss in the microbe-microbe cross-feeding interactions in the distal gut environment of IBS. Additionally, the loss in the associations and a lack of a clear organization of clusters with bi-clustering in the IBS dataset are thought to be indicators of dysbiosis (an imbalance in the intestinal homeostasis of the microbial communities), which is a common symptom in most subtypes of IBS (114, 115).

Of the microbe-metabolite associations in the healthy set that were found to be statistically significant, many were novel findings, while some have previously been reported in literature. For example, the positive association between *Ruminococcus* and glucose can be explained by the fact that members of this genus are polysaccharide-degraders that release extra-cellular enzymes to cleave off glucose from complex polysaccharides (116). Likewise, the negative association between glucose and *Coprococcus* can be justified by the reports that have shown that members of this genus utilized glucose under anaerobic conditions (117). Similarly, positive associations between *Acidaminobacter* and acetate, *Coprococcus* and valerate and,

Prevotella and fumarate are likely because members of these genera produce these metabolite end-products as a result of anaerobic fermentation in the intestine (118).

Sub aim 4b: Uncover statistically significant relationships between fecal microbiota and metabolites that are common to both healthy US and Egyptian children.

In order to determine associations between distal gut metabolites and microbes that are shared between the healthy US and Egyptian children, Spearman rank correlation analysis was performed on fecal metabolite abundances acquired through H1 NMR and genus abundances acquired through high-throughput next generation sequencing. The correlation analysis was performed on the combined dataset containing both Egyptian and US samples because there are unequal number of samples within each group (28 and 14). In such situations, running correlation analysis separately for each group will result in incomparable p-values for correlations between groups. For the correlation analysis, only the top 40 most abundant genera were used, for reasons stated in the **sub aim 4a.** Multiple hypothesis testing using FDR was tested at α thresholds of 10%, 5% and 2.5%, but only the results from 2.5% were used for in-depth analysis and biological interpretation. Such a stringent criterion was used because a very large number of statistically significant associations were uncovered at higher percents, making the interpretation of the associations very complex.

Results:

Spearman rank based correlation analysis performed on the combined microbiota-metabolite datasets from both sample groups (Egyptian and US children) resulted in many statistically significant correlations at 10% and 5% FDR thresholds (144 and 70 significant

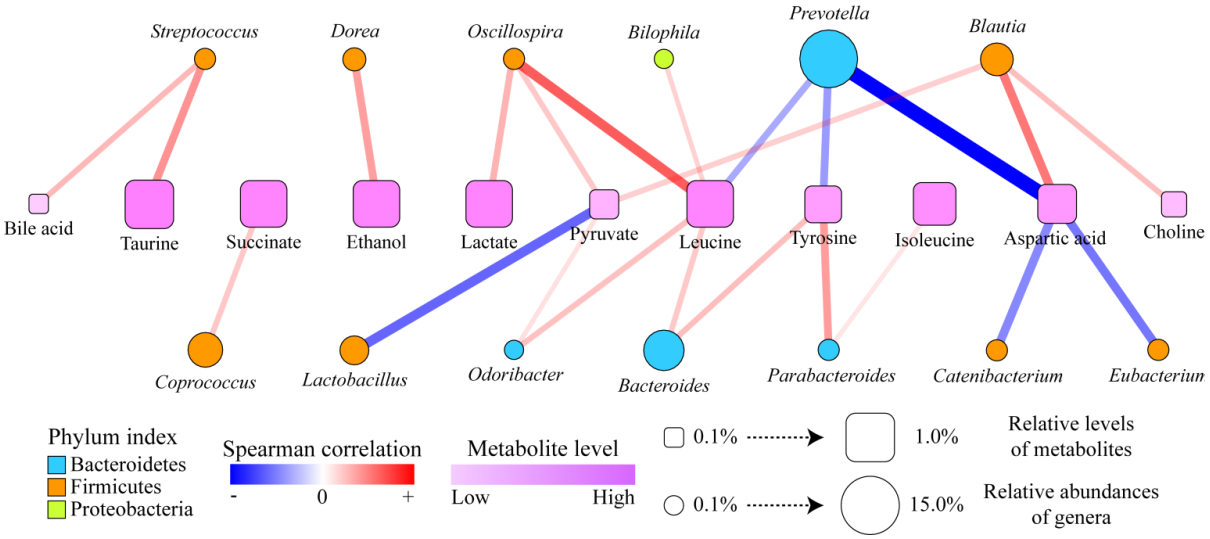

Figure 17. Statistically significant associations based on the Spearman rank based correlations between the distal gut metabolite and genus abundance profiles from healthy US and Egyptian children

correlations respectively). At 2.5% FDR, 23 associations were found to be statistically significant (Figure 17). These associations include positive correlations between bile acid and *Streptococcus*, taurine and *Strepotococcus*, *Dorea* and ethanol, and negative correlations between *Prevotella* and aspartate (aspartic acid), and *Lactobacillus* and pyruvate. All of the statistically significant correlations were between metabolic intermediates and various genera. No significant correlations were detected for simple sugars such as glucose, galactose and sucrose or nucleotide metabolism related metabolites such as uracil, hypoxanthine and cytosine. Many strong correlations were associated with several amino acids such as aspartate, leucine, and tyrosine. Among the genera, the strongest correlations belonged to *Prevotella*, *Lactobacillus*, *Blautia* and *Oscillospira* (Figure 17).

Discussion:

It is interesting to note that despite the observations from **sub aim 1c, sub aim 2b** and **sub aim 3c** which imply that there are large differences in the distal gut environment,

specifically in the microbiota composition, between US and Egyptian children, the Spearman rank correlation analysis uncovered many statistically significant correlations that are shared or common between the two populations. The reason for the large number of strong correlations is likely because of the number of samples used for this analysis. For each metabolite-microbe comparison, a total of 42 samples are used (14 US and 28 Egyptian samples), which is significantly larger than what was used in **sub aim 4a** (22 samples in each analysis set). The increase in the number of samples likely resulted in the enhancement of the p-values (higher confidence resulting in smaller p-values) assigned to the correlations by the Spearman rank analysis.

While many of the uncovered putative associations are novel and have yet to be experimentally proven to be biologically true, some of them have been reported previously in several publications. For example, the relationship between bile acids, *Streptococcus* and taurine has been explored in the context of intestinal metabolite bio-transformation. Members of the genus *Streptococcus* that reside in the gut have been shown to be able to metabolize primary bile acids such as glycocholate and taurocholate to release secondary bile acids, and taurine and choline (119). Similarly, members of *Dorea* have been reported to ferment glucose to produce ethanol, so a positive association between *Dorea* and ethanol from the correlation analysis could a result of this interaction (120). A negative correlation between a genus and a metabolite can imply that the metabolite is being consumed by the members of the genus. A negative correlation between *Prevotella* and aspartic acid could be justified by experimental evidences from reports that claimed that anaerobic growth of members from this genus were enhanced with the addition of aspartic acid to the growth medium (118). Similarly, a negative correlation between *Lactobacillus* and pyruvate can be supported by experimental evidence from literature that have

shown that many gut residing members of *Lactobacillus* (*L.acidophilus, L. bulgaricus, L. casei, L. delbrueckii , L. lactis and L. plantarum*) consumed pyruvate as part of their fermentation pathway (121).

Specific aim 5: ***Construct a protocol using previous aims for the exploratory and hypothesis-driven analysis of microbiota-related multivariate datasets.***

Rationale:

With the advent of high-throughput molecular techniques such as microarrays and next generation sequencing, it is now possible to interrogate many samples and many variables simultaneously in the field of microbial ecology. While the rate and the amount of data generated has increased exponentially, the analysis of such multivariate data has yet to match them (17, 122). New, sophisticated techniques and analyses methods for multivariate data have been developed, but their application in the field of microbial ecology is severely limited. Many studies still rely on the most generic and oldest of ordination techniques such as PCA for analysis of microbial multivariate datasets (17). And often, such practices lead to the misuse of these techniques, primarily due to the unfamiliarity with the statistical frameworks of the techniques (assumptions, distance used, etc.,) (17, 18). In order to facilitate the appropriate use of multivariate analysis techniques in the field of microbial ecology and to increase the ease of the use of these techniques by biologists, we attempt to build a generalized protocol or a set of guidelines for the analysis of microbiota-related high-dimensional dataset in this aim. The general procedure is depicted in Figure 18.

Figure 18. Workflow for the analysis of multivariate datasets in the field of microbial ecology

Protocol:

I) Select a technique based on the scientific question or the goal for the analysis.

Exploratory analysis:

It is often recommended to start the analysis procedure of a new multivariate dataset with exploratory techniques, because observations from these initial analyses can lead to testable hypotheses (both statistically and experimentally). If the goal of the analysis is to condense the complexity within the dataset, and visualize large patterns in low-dimensional space, it is generally a viable strategy to use indirect gradient analysis. Popular techniques within this category include PCA, PCoA and (D)CA (Figure 18). PCA is preferred if testing for linearity has shown that the variables change in a linear fashion with respect to some external gradient (usually environmental) or if the length of the gradient is small (for example, pH in the gastrointestinal tract) (123). Alternatively, if the expected response model (change of variable values with respect to external gradient or effect) is unimodal, CA is preferred. The detrending procedure can be used if artifacts such as the arch effect are visible in the visualization of CA (30). PCA uses Euclidean distance to define the relationships between samples based on their variable values. Similarly, CA (and DCA) uses the chi-square distance for the purpose of ordination. If an alternative way to define this distance is preferred, PCoA is the technique of choice. For example, when the distance between samples can be defined by taxa abundances as well as the phylogenetic relationship between the taxa, it might be appropriate to incorporate this information in the ordination of taxa abundances. A popular distance for this approach is the UniFrac beta diversity metric (27).

Hypothesis-driven analysis:

Once the presence of large gradients has been identified using exploratory approaches, it might of interest to characterize and quantify them. If the goal of the analysis is to associate or explain gradients of variability within the response dataset with independently measured environmental variables, it is recommended to use direct gradient analysis. Popular techniques in this category include (db)RDA, PRC and CCpdA (Figure 18). RDA is the preferred technique when a linear response model is expected with respect to the selected explanatory variables or if the measurement of the explanatory variables is over a small scale (16, 18, 123). RDA is a constrained version of PCA (constrained by the explanatory variables) and therefore uses Euclidean distance. However, if a specialized distance such as UniFrac is available to define inter-sample (dis)similarity, it is recommended to take advantage of this information when performing RDA (this is referred to as dbRDA or distance-based RDA). When one of the environmental variable is time, and the research goal is to characterize effect of time on changes in the response variable, PRC is the preferred technique (35). Finally, CCpdA is the preferred technique if a unimodal response model is expected as a result of changes in the measured explanatory variables or if the model contains categorical explanatory variables (group designation, for example).

If the goal of the analysis is to determine if the largest variability gradients correspond to differences between sample groups and to identify these differences that contribute the most to group separation, discriminant analyses can be used. Currently popular discriminant analyses in the field of microbial ecology include RF, OPLS-DA and SVM (Figure 18). While all three techniques accomplish the same end goal, there might be differences in their performance based

on the features of the input dataset. For example, based on the findings from ***Specific aim 3***, RF had consistently good performance in almost all of the datasets. However, OPLS-DA outperformed RF when used on a dataset with low number of variables (19 metabolites). Although it needs further testing with more datasets, OPLS-DA might be more suited for datasets with low number of variables, whereas RF is the preferred choice in the rest of the situations. Since SVM performed consistently worse than the other two techniques, it is not the preferred choice in the context of the tested datasets. Although, the relatively poor performance of SVM might be because of the use of a linear kernel function. If the groups within the datasets are harder to separate using a linear function, more complex functions need to be formulated. It is important however to note that increasing the complexity of the kernel function can result in over-fitting and a loss of generality.

Finally, if the goal of the analysis is to determine the pairwise associations or relationships between two different sets of response variables measured from the same set of samples, correlation-based bipartite analysis is an ideal choice. One of the more popular non-parametric methods for assigning correlation coefficients between variables is Spearman rank correlation (57). Because this type of analysis involves multiple simultaneous comparisons, the statistically significance testing needs to incorporate multiple testing correction. A good example of such a procedure is the Benjamini-Hochberg's False discovery rate correction (111, 113). Statistically significant correlations can be visualized as bipartite networks to identify associations that can be experimentally tested.

II) Pre-process the dataset if necessary, based on the assumptions and the requirements of the chosen technique.

Various multivariate analysis algorithms are built based on some underlying assumptions or requirements with regard to the characteristics of the input datasets. These assumptions or requirements usually refer to two different aspects of datasets; i) the type of the data, and ii) the underlying structure and relationship within variables. When the data types, structures and relationships do not conform to the defined assumptions and requirements of the selected multivariate statistical technique, the performance and reliability of the results of the technique become questionable. In such cases, it is best to transform and standardize the dataset so that its characteristics better fits the chosen statistical analytical technique. For example, if RDA is chosen as the technique, but it is expected that the response variables change unimodally, it is recommended that the dataset be transformed to linearity before performing analysis (26) or it is better to use CCpdA instead.

Many ordination analyses can result in false patterns as a result of the sparseness of the datasets (many zeros due to rare variables). To correct for such situations, chord transformation or Hellinger transformation of the dataset are recommended (87). If dataset containing relative abundances (compositional) is to be used for ordination analyses, centered log-ratio transformation of the dataset is recommended prior to analysis to reduce the effects of the constant sum constraint (proportion of a constant total). This feature of compositional datasets leads to variable inter-dependence and can cause false patterns in ordination (124, 125). Centered log-ratio transformation of datasets with many zeros is not recommended, since this transformation procedure involves adding a very small value below the detection limit to the entire dataset to facilitate the log-transformation procedure (because zeros cannot be log-transformed). If inter-individual variability is a very large contributor to the overall variability

within the dataset, Mahalanobis scaling of the dataset prior to ordination analysis is recommended (126, 127).

III) Define a metric based on the chosen multivariate technique to represent the alternative hypothesis of the scientific question.

Because the interpretation of the raw results of many ordination techniques can be difficult, it is generally recommended to define a descriptive statistic that has a straightforward and intuitive meaning which can be used to represents the quality of the results of the multivariate technique. For example, when evaluating the quality of the sample group clustering in two or higher dimensional plots of ordination results, the use of the DB index and the Calinski-Harabasz (CH) index are highly recommended (128). These indices are built based on the features of clusters such as the size of the cluster and the distance between clusters, so the interpretation of the values is simple and easy. Additionally, these indices are versatile in that they can be applied to any ordination output that contains sample or variable coordinates. For direct gradient analyses, since the common question asked is, how well the explanatory variables explain the variability within the response set, F-type statistics or the pseudo-F statistics are the recommended choice, because these statistical indices compare variation captured by the explanatory variables to the overall inertia (total variability). For discriminant analyses, two popular metrics are available. Since one of the most important questions with these analyses is robustness of the discriminant model (alternatively, how badly does the model over-fit the data), model accuracy and area under the ROC curves after model cross-validation can be used.

Technique specific parameters such as Q^2 or R^2Y for OPLS-DA are also recommended metrics for assessment of model quality.

IV) Generate a null distribution of the chosen metric.

In order to determine if the results of the multivariate analysis techniques are statistically significant, the observed value for the metric needs to be compared to a null distribution which represents values generated due to random chance. In order for such a comparison, a null distribution based on the metric needs to be generated. A viable strategy for the generation of a null distribution is defined by the Monte Carlo Permutation Procedure which involves the random swapping of samples within the dataset to create a 'random' configuration (81). The metric is calculated and recorded for this random configuration of samples. The process is then repeated thousands of times to generate to generate the null distribution of metric values which can be compared to the observed value. This strategy has been used for DB index in the indirect gradient analyses and discriminant analyses, the pseudo-F statistic and the F-type test statistic in the direct gradient analyses and the model parameters of the OPLS-DA model in the discriminant analyses. One of the advantages of this method is its versatility and its independence of any parameterization of the data (does not rely on specific assumptions of the data structure) (129). This strategy does however suffer from one critical reliance. The number of samples within the dataset must be large enough to generate enough combinations for a reference distribution, the lack of which can results in false or incorrect distribution. The effect of a low number of samples used in the permutation procedure can be seen in **sub aim 2c** where calculation of statistically significance for the model at individual time points generated only modest p-values.

V) Apply the technique, interpret the results and test for statistical significance of the analysis.

Many ordination analyses produce 2 or 3 dimensional plots of the output which can be used to interpret the results and quality of the analyses. Most indirect gradient analyses produce the sample coordinates plot in which the distance between the samples represent the relationships between samples (sample close to one another have similar variable distribution). Some indirect gradient analyses such as CA and DCA also produce variable coordinate plots which can be overlaid with sample points to represent the relationship between samples and variables (variables appear close to sample points that they are present in). If the DB index was used to define specific clusters in the ordination, if the observed value for the index is less than the predefined alpha threshold for the null distribution, then the clustering output of the indirect ordination analysis is considered statistically significant.

Direct gradient analyses produce constrained versions of the ordination plots which are depicted as tri- or bi-plots with continuous environmental or explanatory variables represented as arrows and categorical variables represented as centroid points. In bi-plots of explanatory variables with sample plots, the sample points represent the weighted value of the explanatory variable in that sample and changes along the direction of arrow. Likewise, in a bi-plot of response variables with explanatory variables, the response variable point along the arrow represents weighted mean (or another metric depending on the technique) of the response variable value along the explanatory gradient arrow. A tri-plot contains all three elements in a single plot (sample points, response variables and explanatory variables) (30). Variation

partitioning or variable margination can be used to determine the relative contribution of each explanatory variable to the overall variability for ranking purposes (18). If the observed metric value (pseudo-F statistic or F-type test statistic) is larger than the predefined alpha threshold for the null distribution, then the observed value is considered statistically significant.

Discriminant analyses applied for classification purposes usually produce model accuracy and class classification probabilities after cross-validation as output. The class classification probabilities can be used to build ROC curves. Some techniques inherently produce ordination plots (like OPLS-DA), while other do not (like SVM). Additionally, strategies described in the **Analysis methods** of ***Specific aim 3*** can be used to identify variables that contribute the most to sample group separation. DB index can be used with ordination output to define separation of sample groups and statistical significance of separation by comparison to null distribution. Similarly, technique specific parameters such as Q^2 or R^2Y for OPLS-DA can be compared to their null distribution to test for model statistical significance.

Correlation-based bipartite analyses produce a matrix of correlation coefficient between the two sets of variables and the associated p-values for each comparison pair. The associated p-values is compared to the significant threshold (q-value) of that comparison generated by FDR at a specific alpha value (10%, 5% or 2.5%, etc.,). If the observed p-value is less than the q-value, the correlations coefficient for the comparison pair is considered statistically significant (111, 113).

V. Dissertation summary

Specific aim conclusions:

Specific aim 1: In the tested datasets, indirect gradient analyses successfully showed that the largest gradients of variability corresponded to the separation of samples based on sample groups.

Sub aim 1a: Indirect gradient analysis of genus abundances from fecal microbiota communities distributed samples from healthy and IBS patients into distinct clusters in ordination space.

Sub aim 1b: Unconstrained ordination of quantitative fecal metabolite levels from healthy and IBS patients separated samples between sample groups in ordination space.

Sub aim 1c: Indirect gradient analysis of genus abundance profiles from fecal microbiota communities distributed samples from healthy US and Egyptian children into distinct clusters based on the country of origin (US or Egypt).

Sub aim 1d: Differences in the microbial phylotype abundances from fecal samples collected from patients with *Clostridium difficile* infection before and after fecal transplantation therapy resulted in separate clusters in ordination space.

Specific aim 2: In the tested datasets, direct gradient analyses was successfully used to explain a significant portion of the overall variability present in the response variables using known independent variables.

Sub aim 2a: Fecal pH, fecal percent water content, host age and health state contributed to a significant portion of the variability in the fecal metabolite profiles acquired from IBS and healthy patients. Fecal pH contributed the most, and age contributed the least.

Sub aim 2b: A large proportion of the variability within fecal microbiota genus abundances profiles from healthy US and Egyptian children can be explained by host age, BMI and country of origin. Country of origin explained the most and BMI explained the least of the overall variability.

Sub aim 2c: Time-dependent changes in the distal gut microbiota communities in CDI patients coincided with their fecal microbiota transplantation state. Significant change was observed for time-points after-transplantation, but not for the time-point before transplantation. Transplanted communities maintained their composition in the recipients for up to 4 months with very little fluctuations.

Specific aim 3: Highly accurate discriminant models were constructed using multiple discriminant analysis techniques, their performances were compared and the top discriminatory variables were identified for each dataset.

Sub aim 3a: Discriminant models to represent differences in fecal microbiota genus abundance profiles between healthy and IBS patients were successfully constructed and compared using different classifier techniques. RF performed the best and SVM performed the worst. The genera *Parasporobacterium*, *Papillibacter*, *Gemella*, *Oxalobacter*, *Solobacterium* and *Actinomyces* were highly discriminatory between the sample groups.

Sub aim 3b: Differences in fecal metabolite profiles of healthy and IBS children were successfully modelled and identified using different discriminant analyses. OPLS-DA performed the best and SVM performed the worst. The metabolites identified as

important for the separation of IBS samples from healthy controls were formate, tyrosine and glucose.

Sub aim 3c: Discriminant models were successfully constructed to identify genera that contributed the most to differences in the fecal microbial profiles between healthy US and Egyptian children using multiple techniques. RF performed the best and SVM performed the worst. Although, all three tested techniques resulted in very high accuracy. Discriminant analyses identified *Prevotella, Bacteroides, Blautia, Catenibacterium, Coprococcus, Ruminococcus* and *Anaerostipes* as the top discriminatory genera.

Sub aim 3d: Discriminant models using a cumulative distal gut microbiota dataset from industrialized and non-industrialized countries were successfully constructed to identify the main drivers of the separation between these two groups using multiple techniques. RF performed the best, while SVM performed the worst, albeit, all tested techniques showed very high performance. Discriminant analyses identified *Prevotella, Bacteroides, Catenibacterium, Alistipes, Succinivibrio, Bulleidia* and *Holdemania* as highly discriminatory between industrialized and non-industrialized populations.

Specific aim 4: Correlation-based bipartite analysis was successfully used to identify and statistically test pair-wise associations between two different sets of response variables measured for the same set of samples.

Sub aim 4a: Spearman rank correlation analysis was successfully used to identify statistically significant putative associations between microbiota and metabolites from the distal environment of healthy and IBS children as well as determine the differences in the

associations between these two sample groups. There was a severe loss in the number of statistically significant microbiota-metabolite associations in the IBS group.

Sub aim 4b: Spearman rank correlations analysis successfully identified statistically significant putative associations between distal gut microbiota and metabolites that are common to US and Egyptian children.

Specific aim 5: We were able to integrate the approaches and insights obtained from the various aims into a viable protocol for the analysis of multivariate datasets from field of microbial ecology.

VI. References:

1. Gao Z, Tseng CH, Pei ZH, & Blaser MJ (2007) Molecular analysis of human forearm superficial skin bacterial biota. *P Natl Acad Sci USA* 104(8):2927-2932.

2. Grice EA & Segre JA (2011) The skin microbiome. *Nat Rev Microbiol* 9(4):244-253.

3. Kent AD & Triplett EW (2002) Microbial communities and their interactions in soil and rhizosphere ecosystems. *Annu Rev Microbiol* 56:211-236.

4. McLellan SL, Huse SM, Mueller-Spitz SR, Andreishcheva EN, & Sogin ML (2010) Diversity and population structure of sewage-derived microorganisms in wastewater treatment plant influent. *Environ Microbiol* 12(2):378-392.

5. Olson JB & Kellogg CA (2010) Microbial ecology of corals, sponges, and algae in mesophotic coral environments. *Fems Microbiol Ecol* 73(1):17-30.

6. Belenguer A, *et al.* (2006) Two routes of metabolic cross-feeding between Bifidobacterium adolescentis and butyrate-producing anaerobes from the human gut. *Appl Environ Microb* 72(5):3593-3599.

7. Duncan SH, Louis P, & Flint HJ (2004) Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. *Appl Environ Microb* 70(10):5810-5817.

8. Flint HJ, Bayer EA, Rincon MT, Lamed R, & White BA (2008) Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nat Rev Microbiol* 6(2):121-131.

9. Ley RE, Turnbaugh PJ, Klein S, & Gordon JI (2006) Microbial ecology - Human gut microbes associated with obesity. *Nature* 444(7122):1022-1023.

10. Panke-Buisse K, Poole AC, Goodrich JK, Ley RE, & Kao-Kniffin J (2015) Selection on soil microbiomes reveals reproducible impacts on plant function. *Isme J* 9(4):980-989.

11. Ruff SE, *et al.* (2015) Global dispersion and local diversification of the methane seep microbiome. *P Natl Acad Sci USA* 112(13):4015-4020.

12. Paliy O & Shankar V (2016) Application of multivariate statistical techniques in microbial ecology. *Mol Ecol* 25(5):1032-1057.

13. Pielou EC (1984) *The interpretation of ecological data: a primer on classification and ordination* (John Wiley & Sons).

14. Gauch HG (1982) *Multivariate analysis in community ecology* (Cambridge University Press).

15. Johnson RA & Wichern DW (1992) *Applied multivariate statistical analysis* (Prentice hall Englewood Cliffs, NJ).

16. Palmer M (2006) Ordination methods—an overview.

17. Ramette A (2007) Multivariate analyses in microbial ecology. *Fems Microbiol Ecol* 62(2):142-160.

18. Buttigieg PL & Ramette A (2014) A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *Fems Microbiol Ecol* 90(3):543-550.

19. Ushio M, *et al.* (2015) Microbial communities on flower surfaces act as signatures of pollinator visitation. *Sci Rep-Uk* 5.

20. Koren O, *et al.* (2013) A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. *Plos Comput Biol* 9(1).

21. Jones SE, Shade AL, McMahon KD, & Kent AD (2007) Comparison of primer sets for use in automated ribosomal intergenic spacer analysis of aquatic bacterial communities: an ecological perspective. *Appl Environ Microb* 73(2):659-662.

22. Hong PY, *et al.* (2010) Comparative Analysis of Fecal Microbiota in Infants with and without Eczema. *Plos One* 5(3).

23. Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Philos Mag* 2(7-12):559-572.

24. Bryant FB & Yarnold PR (1995) Principal-components analysis and exploratory and confirmatory factor analysis.

25. Legendre P & Legendre L (2012) *Numerical Ecology* (Elsevier, Amsterdam) 3rd edition Ed.

26. Austin M (2007) Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecol Model* 200(1-2):1-19.

27. Lozupone C & Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microb* 71(12):8228-8235.

28. Schnorr SL, *et al.* (2014) Gut microbiome of the Hadza hunter-gatherers. *Nat Commun* 5.

29. Fierer N, *et al.* (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *P Natl Acad Sci USA* 109(52):21390-21395.

30. ter braak CJF (1986) Canonical Correspondence-Analysis - a New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology* 67(5):1167-1179.

31. Wang TT, *et al.* (2012) Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *Isme J* 6(2):320-329.

32. Rojo D, *et al.* (2015) Clostridium difficile heterogeneously impacts intestinal community architecture but drives stable metabolome responses. *Isme J* 9(10):2206-2220.

33. Ward NL, Pieretti A, Dowd SE, Cox SB, & Goldstein AM (2012) Intestinal aganglionosis is associated with early and sustained disruption of the colonic microbiome. *Neurogastroent Motil* 24(9):874-+.

34. Wang M, *et al.* (2015) Fecal Microbiota Composition of Breast-Fed Infants Is Correlated With Human Milk Oligosaccharides Consumed. *J Pediatr Gastr Nutr* 60(6):825-833.

35. Van den Brink PJ & Ter Braak CJ (1999) Principal response curves: Analysis of time-dependent multivariate responses of biological community to stress. *Environmental Toxicology and Chemistry* 18(2):138-148.

36. Fuentes S, *et al.* (2014) Reset of a critically disturbed microbial ecosystem: faecal transplant in recurrent Clostridium difficile infection. *Isme J* 8(8):1621-1633.

37. Amato KR & Righini N (2015) The howler monkey as a model for exploring host-gut microbiota interactions in primates. *Howler Monkeys*, (Springer), pp 229-258.

38. James G, Witten D, Hastie T, & Tibshirani R (2013) *An introduction to statistical learning* (Springer).

39. Borcard D, Gillet F, & Legendre P (2011) Numerical Ecology with R. *Use R*:1-300.

40. Shankar V, Reo NV, & Paliy O (2015) Simultaneous fecal microbial and metabolite profiling enables accurate classification of pediatric irritable bowel syndrome. *Microbiome* 3.

41. Westerhuis JA, *et al.* (2008) Assessment of PLSDA cross validation. *Metabolomics* 4(1):81-89.

42. Stone M (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*:111-147.

43. Picard RR & Cook RD (1984) Cross-validation of regression models. *J Am Stat Assoc* 79(387):575-583.

44. Knights D, Costello EK, & Knight R (2011) Supervised classification of human microbiota. *Fems Microbiol Rev* 35(2):343-359.

45. Li C, Wang J, Wang L, Hu L, & Gong P (2014) Comparison of classification algorithms and training sample sizes in urban land classification with Landsat thematic mapper imagery. *Remote Sensing* 6(2):964-983.

46. Yatsunenko T, *et al.* (2012) Human gut microbiome viewed across age and geography. *Nature* 486(7402):222-+.

47. Schloss PD, Iverson KD, Petrosino JF, & Schloss SJ (2014) The dynamics of a family's gut microbiota reveal variations on a theme. *Microbiome* 2.

48. Morton ER, *et al.* (2015) Variation in Rural African Gut Microbiota Is Strongly Correlated with Colonization by Entamoeba and Subsistence. *Plos Genet* 11(11).

49. Abdi H (2010) Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics* 2(1):97-106.

50. Trygg J & Wold S (2002) Orthogonal projections to latent structures (O-PLS). *J Chemometr* 16(3):119-128.

51. Rogers GB, *et al.* (2014) Functional divergence in gastrointestinal microbiota in physically-separated genetically identical mice. *Sci Rep-Uk* 4.

52. Worley B & Powers R (2013) Multivariate Analysis in Metabolomics. *Current Metabolomics* 1(1):92-107.

53. Li M, *et al.* (2008) Symbiotic gut microbes modulate human metabolic phenotypes. *P Natl Acad Sci USA* 105(6):2117-2122.

54. Hofmann M (2006) Support Vector Machines—Kernels and the Kernel Trick.

55. Gokcen I & Peng J (2002) Comparing Linear Discriminant Analysis and Support Vector Machines. *Lect Notes Comput Sc* 2457:104-113.

56. Yang CY, *et al.* (2006) An ecoinformatics tool for microbial community studies: Supervised classification of Amplicon Length Heterogeneity (ALH) profiles of 16S rRNA. *J Microbiol Meth* 65(1):49-62.

57. Weisstein EW (2006) Correlation coefficient.

58. Shankar V, *et al.* (2015) The networks of human gut microbe-metabolite associations are different between health and irritable bowel syndrome. *Isme J* 9(8):1899-1903.

59. Shankar V, Agans R, Holmes B, Raymer M, & Paliy O (2013) Do gut microbial communities differ in pediatric IBS and health? *Gut microbes* 4(4):347-352.

60. McDonald JH (2009) *Handbook of biological statistics*.

61. Chok NS (2010) Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data. (University of Pittsburgh).

62. Martin FPJ, *et al.* (2010) Dietary Modulation of Gut Functional Ecology Studied by Fecal Metabonomics. *J Proteome Res* 9(10):5284-5295.

63. Le Gall G, *et al.* (2011) Metabolomics of fecal extracts detects altered metabolic activity of gut microbiota in ulcerative colitis and irritable bowel syndrome. *Journal of proteome research* 10(9):4208-4218.

64. Jacobs DM, *et al.* (2008) H-1 NMR metabolite profiling of feces as a tool to assess the impact of nutrition on the human microbiome. *Nmr Biomed* 21(6):615-626.

65. Rigsbee L, Agans R, Foy BD, & Paliy O (2011) Optimizing the analysis of human intestinal microbiota with phylogenetic microarray. *Fems Microbiol Ecol* 75(2):332-342.

66. Anderson PE, *et al.* (2011) Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data. *Metabolomics* 7(2):179-190.

67. Paliy O, Kenche H, Abernathy F, & Michail S (2009) High-Throughput Quantitative Analysis of the Human Intestinal Microbiota with a Phylogenetic Microarray. *Appl Environ Microb* 75(11):3572-3579.

68. Caporaso JG, *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335-336.

69. DeSantis TZ, *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microb* 72(7):5069-5072.

70. Hammer Ø, Harper DAT, & Ryan PD (2001) PAST-PAlaeontological STatistics, ver. 1.89.

71. Oksanen J, *et al.* (2013) Package 'vegan'. *Community ecology package, version 2.9*.

72. Micallef L & Rodgers P (2014) euler APE: Drawing area-proportional 3-Venn diagrams using ellipses. *Plos One* 9(7):e101717.

73. Brown KR, *et al.* (2009) NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics* 25(24):3327-3329.

74. Rigsbee L, *et al.* (2012) Quantitative profiling of gut microbiota of children with diarrhea-predominant irritable bowel syndrome. *The American journal of gastroenterology* 107(11):1740-1751.

75. Hamilton MJ, Weingarden AR, Sadowsky MJ, & Khoruts A (2012) Standardized Frozen Preparation for Transplantation of Fecal Microbiota for Recurrent Clostridium difficile Infection. *American Journal of Gastroenterology* 107(5):761-767.

76. Hamilton MJ, Weingarden AR, Unno T, Khoruts A, & Sadowsky MJ (2013) High-throughput DNA sequence analysis reveals stable engraftment of gut microbiota following transplantation of previously frozen fecal bacteria. *Gut microbes* 4(2):125-135.

77. Obregon-Tito AJ, *et al.* (2015) Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat Commun* 6.

78. ter braak CJF & Prentice IC (1988) A Theory of Gradient Analysis. *Adv Ecol Res* 18:271-317.

79. Davies DL & Bouldin DW (1979) A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (2):224-227.

80. Amro (2010) Ellipse around the data in MATLAB.

81. Dwass M (1957) Modified Randomization Tests for Nonparametric Hypotheses. *Ann Math Stat* 28(1):181-187.

82. Salonen A, *et al.* (2014) Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. *Isme J* 8(11):2218-2230.

83. Hildebrand F, *et al.* (2013) Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. *Genome Biol* 14(1).

84. Brussow H (2013) Microbiota and healthy ageing: observational and nutritional intervention studies. *Microb Biotechnol* 6(4):326-334.

85. Spiller R, *et al.* (2007) Guidelines on the irritable bowel syndrome: mechanisms and practical management. *Gut* 56(12):1770-1798.

86. Rajilic-Stojanovic M, *et al.* (2015) Intestinal Microbiota And Diet in IBS: Causes, Consequences, or Epiphenomena? *American Journal of Gastroenterology* 110(2):278-287.

87. Legendre P & Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia* 129(2):271-280.

88. Ringel-Kulka T, *et al.* (2015) Altered Colonic Bacterial Fermentation as a Potential Pathophysiological Factor in Irritable Bowel Syndrome. *American Journal of Gastroenterology* 110(9):1339-1346.

89. Stecher B, Maier L, & Hardt WD (2013) 'Blooming' in the gut: how dysbiosis might contribute to pathogen evolution. *Nat Rev Microbiol* 11(4):277-284.

90. Winter SE & Baumler AJ (2014) Dysbiosis in the inflamed intestine: chance favors the prepared microbe. *Gut microbes* 5(1):71-73.

91. Lilja HE, Wefer H, Nystrom N, Finkel Y, & Engstrand L (2015) Intestinal dysbiosis in children with short bowel syndrome is associated with impaired outcome. *Microbiome* 3.

92. Seekatz AM & Young VB (2014) Clostridium difficile and the microbiota. *J Clin Invest* 124(10):4182-4189.

93. Ling ZX, *et al.* (2014) Impacts of infection with different toxigenic Clostridium difficile strains on faecal microbiota in children. *Sci Rep-Uk* 4.

94. Brahe LK, *et al.* (2015) Specific gut microbiota features and metabolic markers in postmenopausal women with obesity. *Nutr Diabetes* 5.

95. den Besten G, *et al.* (2013) The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *J Lipid Res* 54(9):2325-2340.

96. Wrzosek L, *et al.* (2013) Bacteroides thetaiotaomicron and Faecalibacterium prausnitzii influence the production of mucus glycans and the development of goblet cells in the colonic epithelium of a gnotobiotic model rodent. *Bmc Biol* 11.

97. Zhu QC, Gao RY, Wu W, & Qin HL (2013) The role of gut microbiota in the pathogenesis of colorectal cancer. *Tumor Biol* 34(3):1285-1300.

98. Steck N, Mueller K, Schemann M, & Haller D (2012) Bacterial proteases in IBD and IBS. *Gut* 61(11):1610-1618.

99. Ghoshal UC, Srivastava D, Verma A, & Misra A (2011) Slow Transit Constipation Associated With Excess Methane Production and Its Improvement Following Rifaximin Therapy: A Case Report. *J Neurogastroenterol* 17(2):185-188.

100. Elhenawy W, Debelyy MO, & Feldman MF (2014) Preferential Packing of Acidic Glycosidases and Proteases into Bacteroides Outer Membrane Vesicles. *Mbio* 5(2).

101. Gibson SAW & Macfarlane GT (1988) Studies on the Proteolytic Activity of Bacteroides-Fragilis. *J Gen Microbiol* 134:19-27.

102. Dodd D, Mackie RI, & Cann IKO (2011) Xylan degradation, a metabolic property shared by rumen and human colonic Bacteroidetes. *Mol Microbiol* 79(2):292-304.

103. Lingstrom P, van Houte J, & Kashket S (2000) Food starches and dental caries. *Crit Rev Oral Biol M* 11(3):366-380.

104. Nakayama J, *et al.* (2015) Diversity in gut bacterial community of school-age children in Asia. *Sci Rep-Uk* 5.

105. Wu GD, *et al.* (2011) Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science* 334(6052):105-108.

106. Hastie T, Buja A, & Tibshirani R (1995) Penalized discriminant analysis. *The Annals of Statistics*:73-102.

107. Guyon I & Elisseeff A (2003) An introduction to variable and feature selection. *The Journal of Machine Learning Research* 3:1157-1182.

108. Arumugam M, *et al.* (2011) Enterotypes of the human gut microbiome. *Nature* 473(7346):174-180.

109. De Filippo C, *et al.* (2010) Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *P Natl Acad Sci USA* 107(33):14691-14696.

110. Fletcher SM, McLaws M, & Ellis JT (2013) Prevalence of gastrointestinal pathogens in developed and developing countries: systematic review and meta-analysis. *Journal of public health research* 2(1):9.

111. Benjamini Y & Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* 57(1):289-300.

112. Svetnik V, Liaw A, & Tong C (2004) Variable selection in random forest with application to quantitative structure-activity relationship. *Proceedings of the 7th Course on Ensemble Methods for Learning Machines*.

113. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, & Ploner A (2005) False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* 21(13):3017-3024.

114. Chassard C, *et al.* (2012) Functional dysbiosis within the gut microbiota of patients with constipated-irritable bowel syndrome. *Aliment Pharm Ther* 35(7):828-838.

115. Cremon C, *et al.* (2010) Intestinal dysbiosis in irritable bowel syndrome: etiological factor or epiphenomenon? *Expert Rev Mol Diagn* 10(4):389-393.

116. Iakiviak M, Mackie RI, & Cann IKO (2011) Functional Analyses of Multiple Lichenin-Degrading Enzymes from the Rumen Bacterium Ruminococcus albus 8. *Appl Environ Microb* 77(21):7541-7550.

117. Holdeman LV & Moore WEC (1974) New Genus, Coprococcus, 12 New Species, and Emended Descriptions of 4 Previously Described Species of Bacteria from Human Feces. *Int J Syst Bacteriol* 24(2):260-277.

118. Takahashi N & Yamada T (2000) Pathways for amino acid metabolism by Prevotella intermedia and Prevotella nigrescens. *Oral Microbiol Immun* 15(2):96-102.

119. Hill MJ & Drasar BS (1968) Degradation of Bile Salts by Human Intestinal Bacteria. *Gut* 9(1):22-&.

120. Taras D, Simmering R, Collins MD, Lawson PA, & Blaut M (2002) Reclassification of Eubacterium formicigenerans Holdeman and Moore 1974 as Dorea formicigenerans gen. nov., comb. nov., and description of Dorea longicatena sp nov., isolated from human faeces. *Int J Syst Evol Micr* 52:423-428.

121. Pessione E (2012) Lactic acid bacteria contribution to gut microbiota complexity: lights and shadows. *Front Cell Infect Mi* 2.

122. Muir P, *et al.* (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol* 17:53.

123. Lepš J & Šmilauer P (2003) *Multivariate analysis of ecological data using CANOCO* (Cambridge university press).

124. Faust K, *et al.* (2012) Microbial Co-occurrence Relationships in the Human Microbiome. *Plos Comput Biol* 8(7).

125. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, & Bahler J (2015) Proportionality: A Valid Alternative to Correlation for Relative Data. *Plos Comput Biol* 11(3).

126. Chilingaryan A, Gevorgyan N, Vardanyan A, Jones D, & Szabo A (2002) Multivariate approach for selecting sets of differentially expressed genes. *Math Biosci* 176(1):59-69.

127. Mao KZ & Tang WY (2011) Recursive Mahalanobis Separability Measure for Gene Subset Selection. *Ieee Acm T Comput Bi* 8(1):266-272.

128. Calinski T (1968) A Dendrite Method for Cluster Analysis. *Biometrics* 24(1):207-&.

129. Phipson B & Smyth GK (2010) Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology* 9(1).