Advanced Computational Algorithms for Microbial Community Analysis Using Massive 16S rRNA Sequence Data*

Yijun Sun[§¶], Yunpeng Cai[¶], Volker Mai[†], William Farmerie[§]
Fahong Yu[§], Jian Li[¶], Steve Goodison[‡]

[§]Interdisciplinary Center for Biotechnology Research
[¶]Department of Electrical and Computer Engineering
[†]Department of Microbiology and Cell Science
University of Florida, Gainesville, FL 32610-3622
[‡]Cancer Research Institute
M. D. Anderson Cancer Center, Orlando, FL 32827

Abstract

With the aid of next-generation sequencing technology, researchers can now obtain millions of microbial signature sequences for diverse applications ranging from human epidemiological studies to global ocean surveys. The development of advanced computational strategies to maximally extract pertinent information from massive nucleotide data has become a major focus of the bioinformatics community. Here, we describe a novel analytical strategy including discriminant and topology analyses that enables researchers to deeply investigate the hidden world of microbial communities, far beyond basic microbial diversity estimation. We demonstrate the utility of our approach through a computational study performed on a previously published massive human gut 16S rRNA dataset. The application of discriminant and topology analyses enabled us to derive quantitative disease-associated microbial signatures and describe microbial community structure in far more detail than previously achievable. Our approach provides rigorous statistical tools for sequence based studies aimed at elucidating associations between known or unknown organisms and a variety of physiological or environmental conditions.

Nucleic Acids Research - Methods Paper
submitted in March 2010; accepted in September 2010

*Please address all correspondence to: Dr. Yijun Sun, Interdisciplinary Center for Biotechnology Research, University of Florida, P. O. Box 103622, Gainesville, FL 32610-3622, USA. E-mail: sunyijun@biotech.ufl.edu. Y. Sun and Y. Cai contributed to the paper equally.

Keywords: pyrosequencing, 16S rRNA, massive data, machine learning, data mining, predictive model, microbial community analysis, human microbiome.
Running title: Algorithms for Microbial Community Analysis
Abbreviation: rRNA, ribosomal RNA; OTU, operational taxonomic unit; LOOCV, leave-one-out cross validation; ROC, receiver operating characteristic curve; AUC, area under ROC curve; CI, confidence interval; o, obese; l, lean.

1 Introduction

The biosphere contains an estimated $10^{30} \sim 10^{31}$ microbial cells, at least $2 \sim 3$ orders of magnitude larger than the number of plant and animal cells combined [1]. These microbes play an essential role in processes as diverse as maintenance of human health and biogeochemical activities critical to all life. However, the diversity and the community structure of complex microbial communities are still poorly understood, historically due to our inability to culture most microorganisms using standard microbiological techniques. While there are likely millions of bacterial species, only a few thousand have been formally described to date [2]. Accordingly, researchers lack basic information to compare microbial communities under different physical-chemical conditions, and to model dynamic microbe-microbe and environment-microbe interactions.

The recent development of massively parallel pyrosequencing technology allows researchers to study genetic materials recovered directly from environmental samples, by eliminating the need of laboratory isolation and cultivation of individual species, and thus opens a new window to probe the hidden world of microbial communities [2, 3, 4]. In recognition of the role of marine microbes in biogeochemical processes, the International Census of Marine Microbes (ICoMM) consortium has launched an international effort to catalog the diversity of microbial populations in the oceanic, coastal, and benthic waters. Microbes associated with human health are intensely studied through two large-scale initiatives: the Human Microbiome Project (HMP) sponsored by NIH and MetaHIT sponsored by the EU, which seek to establish a correlation between the composition of the human microbiome and various diseases [5]. These studies leverage the power of deep sequencing that allows for the rapid and cost-effective surveying of complex microbial communities to reveal the presence of known and currently unknown species alike. However, as emphasized specifically by the NIH HMP working group, computational methods for analyzing massive sequence data generated by these initiatives are still in their infancy, and consequently new

computational algorithms and strategies are urgently needed to maximize research yields in these efforts [5].

This paper presents a novel computational strategy specifically designed to address the challenges of analyzing large collections of 16S rRNA pyrosequencing data for various biological and ecological inquires. The key idea is to use taxonomy independent analysis to transform the information encoded in the nucleotide domain into the numerical domain, and then use various advanced machine learning and statistical methods to quantify and visualize the associations between altered microbial community composition with physiological or environmental conditions of interest. We demonstrate the viability of the proposed analytical strategy on a previously published massive human gut 16S rRNA dataset generated by Turnbaugh et al. [9] to investigate correlations between the human gut microbiota and obesity. The work by Turnbaugh et al. and other papers mostly by the same group have reported that an obese phenotype is associated with broad, phylum-level changes in the gut community structure [6, 7]. More specifically, obese individuals appear to have a lower proportion of *Bacteroidetes* and a higher proportion of *Firmicutes* compared to lean individuals. This pattern was initially reported only in a small cohort of 12 subjects (~ 350 sequences at each sampling point), likely too small to develop a good indicator for the overall population [8]. A subsequent study involving a much larger number of samples suggested that it was the ratio between *Bacteroidetes* and *Actinobacteria*, not *Firmicutes*, that differed in the obese group compared to the lean group [9]. It is well established that *Firmicutes* and *Bacteroidetes* are the two largest phyla in the human gut flora, consisting of over 250 and 125 genera, respectively [10]. It is possible that the compositions of only a few genera within these phyla are altered in obesity. Hence, it would be valuable to examine differences in microbial composition at more resolved phylogenetic levels. To this end, we performed a series of data analyses that correlated community structures in the gut with respect to physiological state. Our study showed that while several genus-level OTUs classified as belonging to *Bacteroidetes* were all negatively correlated with obesity, there exist both negatively and positively correlated OTUs within *Firmicutes*, which in part explained some conflicting results observed in previous studies. Through discriminant and topology analyses, we further showed that despite individual diverse gut microbial compositions, common microbial signatures exist that can be used to accurately stratify obese and lean individuals. Our study brought new light onto this human microbiome question that previous methods have been unable to resolve. Our approach is broadly applicable to

other sequence based microbial studies.

2 Methods

Fig. 1 presents the schematic diagram of the proposed analytical strategy. We detail each module in the following subsections.

2.1 Taxonomy Independent Analysis

Providing a detailed description of microbial populations, including high, medium and low abundance components, is frequently the first step to perform in microbial community analysis [10, 11]. PCR-based techniques for selectively generating 16S rRNA amplicons followed by DNA sequencing are currently the most commonly used approach to characterizing microbial communities, and have been successfully used in numerous applications (see for example [12, 13] for excellent review). Existing algorithms for microbial classification using 16S rRNA sequences can be generally categorized into taxonomy dependent or independent analyses. In the former methods, query sequences are first compared against a database and then assigned to the organism of the best-matched reference sequences (e.g., BLAST). Since most microbes have not been formally described yet, these methods are inherently limited by the lack of completeness of reference databases [12]. Taxonomy dependent analysis is performed generally for the purpose of sequence annotation. In this paper, we primarily focus on taxonomy independent analyses, where sequences are compared against each other to form a distance matrix, based on which hierarchical clustering is performed to group sequences into operational taxonomic units (OTUs) of specified sequence variations. Typically, sequences with 1-3% dissimilarity are assigned to the same species, while those with less than 5% dissimilarity are assigned to the same genus [14, 9, 15], although these distinctions are controversial. Various ecological metrics can then be estimated from the clustering information to characterize a microbial community. The analysis does not rely on any reference database, and hence is able to enumerate characterized organisms as well as novel pathogenic and uncultured microbes.

We recently developed a new algorithm, referred to as ESPRIT, for large-scale taxonomy independent analysis [16]. The algorithm consists of four modules: (1) filtering out low-quality sequence reads on the basis of multiple criteria, (2) computing pairwise dis-

tances between input sequences, (3) performing hierarchical clustering to group sequences into OTUs at different distance levels, and (4) performing statistical inferences to estimate various ecological metrics. In contrast to many existing 16S rRNA based studies, ESPRIT uses the Needleman-Wunsch algorithm [17], instead of multiple sequence alignment, to optimally align each pair of 16S rRNA sequences, and the quickdist algorithm [3] to compute pairwise distances. More specifically, each pairwise distance equals mismatches, including indels, divided by a sequence length. To avoid overestimating distances between sequences from rapidly diverging variable regions, end gaps are ignored and gaps of any length are treated as a single evolutionary event or mismatch. Through a benchmark study, we demonstrated that global pairwise alignment provided a much more accurate estimate of microbial richness than multiple sequence alignment. Interested reader may refer to the supplement data for a detailed discussion. Within the ESPRIT framework, we also developed a new clustering algorithm, referred to as Hcluster, to handle large-scale hierarchical clustering analysis. Unlike conventional methods that load a distance matrix directly into memory, Hcluster groups sequences into OTUs on-the-fly while keeping track of linkage information, which overcomes memory limitations. The complete-linkage method was used to ensure that the maximum pairwise genetic distance of the sequences grouped into the same cluster is smaller than the specified distance level defining an OTU. ESPRIT has been used extensively by the research community. Two versions of ESPRIT, one for personal computers and one for computer clusters, are freely available at `http://plaza.ufl.edu/sunyijun/ESPRIT.htm`.

2.2 Constructing Profile Data Matrix

One of the major obstacles of using sequence data to query a biological/ecological hypothesis is that most statistical approaches reported in the literature were designed solely for analyzing numerical-valued data. To overcome this difficulty, we applied taxonomy independent analysis to transform the information encoded in the nucleotide domain (i.e., A, T, C, and G) into the numerical domain. More specifically, we used ESPRIT to hierarchically group sequences into OTU at various distance levels to form a tree-like structure. Using a barcode labeling system for each sample, the origin of each sequence was retrieved and the number of sequences from each sample within each OTU was counted and recorded in a data matrix. Each column of the data matrix represents a sample, and each row

represents an OTU. The data matrix was then normalized along the row direction so that each column vector represents a percentage profile of OTUs in each sample. Analogous to microarray technology that enables researchers to *simultaneously* monitor the expression levels of all genes in a cell or tissue [18, 19], the so-obtained profile data matrix provides microbiologists with a *global* view of how microbial compositions change across individuals or between groups with different physiological states at various phylogenetic levels. Alternatively, a profile data matrix can be generated using taxonomy dependent analysis. However, a massive amount of query sequences would be grouped into the unknown or uncultured category regardless their origins, and the uncertainties in sequence annotation would propagate to the entire downstream data analyses. Once we obtain a profile data matrix, various advanced computational methods can be applied to analyze massive, high-dimensional data. In this paper, we mainly focused on discriminant and topology analyses. In Section 4, we presented a brief discussion of how to use nucleotide sequence data to infer microbial interaction networks.

2.3 Discriminant Analysis

The main purpose of discriminant analysis is to identify a list of OTUs containing the most discriminant information that can be used to characterize microbial communities under different conditions. From clinical perspectives, identifying the pathogenic phylotypes stratifying diseased patients from healthy individuals could be used for disease diagnosis and to help physicians make informed decisions to prescribe personalized antibiotics, rather than broad-spectrum antibiotics, to maximize the treatment efficacy [20]. Note that the primary goal of the recently launched HMP Project is to determine whether there are associations between changes in the microbiome and various diseases and thus to pave the way for future large-scale human epidemiological studies [5]. Discriminant analysis is probably one of the most rigorous analyses one can perform to quantify such associations.

One major characteristic of a profile data matrix is that the number of OTUs is several orders of magnitude larger than the number of samples. For instance, in the case study we present in Section 3, at the 0.05 distance level, the number of observed OTUs is 40,765 while there are only 101 samples. In the statistical literature, this is called a "small N and large P" problem [21, 22], where N is the number of samples and P is the number of OTUs. In this situation, special care must be taken to avoid overfitting problems. A commonly used

practice is to select a small feature subset so that the performance of a learning algorithm is optimized [21, 22, 23]. For the purpose of this paper, we used ℓ_1 regularized logistical regression to perform feature selection and classification simultaneously [23]. Since the objective function optimized by the algorithm is not differentiable, fast implementation of ℓ_1 regularized learning has long been considered a challenging problem in the machine-learning community. We recently developed a new gradient descent based algorithm for large-scale ℓ_1 regularized learning [23] (`http://plaza.ufl.edu/sunyijun/DGM.htm`). The new algorithm makes large-scale studies (e.g., permutation tests) computationally tractable. Due to the small sample size, the leave-one-out cross validation (LOOCV) method was adopted to estimate the prediction performance. In each iteration, one sample was held out for test, and the remaining samples were used for training. The regularization parameter of a logistical regression model was estimated through ten-fold cross validation using the training data, and then a predictive model was trained using the estimated parameter and *blindly* applied to the held-out sample. The experiment was repeated until each sample had been tested. Test samples were not involved in any stage of training process (see Fig. 2 for details). A receiver operating characteristic (ROC) curve obtained by varying a decision threshold was then used to visualize how a prediction model performed at different sensitivity and specificity levels. The area under receiver operating characteristic curves (AUC) provides a quantitative assessment of the predictive value of constructed classifiers (AUC = 1: perfect ability to discriminate and AUC = 0.5: random guess) [24].

A typical 16S rRNA based microbial study involves only tens or at most hundreds of samples. With a small data size, it is possible that the outcomes of discriminant analysis are due to some random confounding factors of no interest to investigators. We performed a permutation test to estimate the p-value of predictive performance. For computational reasons, in this paper, the permutation test was repeated 1000 times. In each iteration, the class labels were randomly shuffled, the above-described experimental protocol was executed, and the area under the resulting ROC curve was recorded. The p-value was computed as the occurrence frequencies of the iterations where the resulting AUCs outperformed that obtained using the original class labels.

2.4 Topology Analysis

Topology analysis was preformed that enables microbiologists to visualize and study the global topology structure of a complex microbial community. In this analytical strategy, each sequence is regarded as a data point in a high-dimensional nucleotide space, with each coordinate corresponding to a nucleotide base taking values from set {A, T, C, G}. We used the Isomap algorithm [25] to map sequences into a two-dimensional numerical space that optimally preserves the intrinsic geometry or distribution of the data (i.e., two sequences that have a small genetic distance between each other should stay together in a two-dimensional numerical space). In order to make computation feasible, in this paper, we considered only the clusters generated by ESPRIT at the 0.10 distance level, and removed small clusters containing less than ten sequences. However, we should emphasize that the analysis can be performed at all distance levels. We then randomly selected 100 sequences from each cluster (if a cluster contained less than 100 sequences, all sequences were used.), and computed the pairwise inter-cluster distances as $d_{ij} = \frac{1}{N_i N_j} \sum_{s_n \in C_i} \sum_{s_m \in C_j} d(s_n, s_m)$, where d_{ij} is the distance between clusters C_i and C_j, $d(s_n, s_m)$ is the pairwise distance between two globally aligned sequences s_n and s_m, and N_i and N_j are the numbers of sequences from the two clusters that were used in distance computation. The pairwise inter-cluster distances were then fed into the Isomap algorithm to generate a two-dimensional mapping of massive sequence data. The code is available at `http://waldron.stanford.edu/isomap/`. The only free parameter of the algorithm is the number of the nearest neighbors used to construct a neighborhood graph, which was set to 10.

2.5 Sequence Annotation

We used the RDP classifier [26] to annotate all of the sequences due to its computational efficiency. We also used BLAST search against the RDP-II [27] and greengenes [28] databases to phylogenetically classify the sequences within the top ranked OTUs. A query sequence was assigned to the organism of the best-matched reference sequence if the e-value $\leq 10^{-20}$ and the identity percentage $\geq 95\%$. The analysis was performed on the RAST web application [29]. Both the RDP classifier and RAST do not classify sequences below the genus level.

3 Results

We conducted an intensive computational study on a publicly available human gut microbiota dataset to demonstrate the viability of the proposed computational strategy. The dataset was originally used to study the connection between obesity and altered composition of the human gut flora [9]. It contains 1,119,519 sequences with an average length of 219 nucleotides, covering the V2 hyper-variable region of 16S rRNAs collected from the stool samples of 154 individuals from 54 families. Each sample is labeled as obese, lean, or overweight, based on the corresponding body mass index. This is by far the most comprehensive 16S rRNA based survey of the human gut flora available to date. To reduce random sequencing errors, we applied a trimming procedure similar to those used in [9] to remove reads that (1) contain at least one mismatch in the primers, (2) contain ambiguous bases, or (3) have a length less than 200 bp. We performed a taxonomy independent analysis of the data using the ESPRIT tools described in Section 2.1, and generated profile data matrices at various distance levels from 0.03 to 0.18 using the approach outlined in Section 2.2.

3.1 Microbial Signatures Associated with Obesity

We first applied unsupervised learning techniques to visualize the distributions of the samples. In order to reduce the effect of confounding factors such as antibiotics usage and sampling depth, we removed the samples that (1) were obtained from the individuals who were on antibiotics within 6 months of stool sample collection, (2) have less than 3,000 sequences, and (3) have ambiguous class labels (i.e., overweight). This resulted in a total of 101 samples with 26 in the lean group and 75 in the obese group. We then performed a correlation analysis of OTUs with respect to physiological state. The heat map of the top 50 ranked OTUs defined at the 0.08 distance level plotted in Fig. 3 reveals that obese individuals have a distinguishing pattern of microbial profiles compared to lean individuals. Unsupervised hierarchical clustering clearly partitions the samples into two groups, and this pattern was observed over a wide range of phylogenetic levels (Figs. 2S, 3S and 4S).

For a more rigorous analysis, we then applied supervised machine-learning techniques to quantify how the predictive value of microbial profiles varies at different phylogenetic levels. We used ℓ_1 regularized logistical regression to estimate the posteriori probability of a sample belonging to the obese or lean group (see [23] and Methods section for details).

The AUCs obtained at different distance levels ranging from 0.03 to 0.18 are presented in Fig. 4 (left panel). We observe that the microbial profile-based predictive models perform very well over a wide range of distance levels. For example, at the 0.08 distance level, the AUC equals 0.88 (p-value$<$0.001 obtained by a permutation test. Fig. 5S). At the 80% sensitivity level, the model correctly classified 83 out of 101 samples (82%), including 61 obese and 22 lean individuals (Fig. 4 right panel). The dataset under analysis came from a twin study [9], and it was reported that members within the same family had similar gut microbial community structures. In order to avoid information leakage, we also performed a leave-family-out cross validation where all of the samples from the same family were held out and classified by the predictive model constructed using the samples from *other* families. The classification result had no statistical difference from that obtained using LOOCV (p-value$>$0.30 based on a Student's t-test. Fig. 4 left panel). This experiment demonstrates that despite the fact that each individual has diverse gut microbial compositions [10, 9] and that members within the same family have similar overall gut community structures independent of obesity status, there exists a *common microbial signature* that can be used to accurately distinguish obese from lean individuals. Interestingly, the AUC analysis reveals that the discriminant information is contained over a wide range of phylogenetic levels (Fig. 4 left panel). This finding extends previous studies by quantifying the association between changes in the microbiome and obesity and pinpointing OTUs that may have a connection with obesity at more resolved phylogenetic levels.

It is interesting to note that the AUC vs distance level plot has a bell shape (Fig. 4). This makes intuitive sense. When the distance for defining OTUs is large, sequences are grouped into large clusters where discriminant and non-discriminant information are mixed. On the other hand, when the distance level is small (say 0.03 and 0.05), deep sequencing is required to obtain accurate estimates of microbial composition profiles [3, 30]. For the gut microbiota data we considered, the average number of sequences in each sample was 7799 with one standard deviation of 5953. This level of coverage may not be sufficient to fully catalog the microbial species resident in the gut, and it is likely that more exhaustive surveys can lead to derivation of a more accurate microbial signature at the genus or even species phylogenetic levels.

3.2 Topology Structure of Human Gut Microbiota

We next applied topology analysis to the data to visualize the community structure of the human gut microbial community. We used the Isomap algorithm [25] to map the sequences into a two-dimensional numerical space that optimally preserves the geometry of the data (see Methods section for details). Fig. 5 presents the output of the analysis. Each circle represents an OTU defined at the 0.10 distance level, and the diameter represents the number of sequences within the OTU divided by the total number of sequences. We used the RDP classifier [26] to annotate the sequences in each cluster. The face color of each circle represents the percentage of the sequences within that cluster that can be annotated by RDP at the genus level with a confidence level >80%. This figure reveals the following points: (1) *Bacteroidetes* and *Firmicutes* phyla are the two largest groups within the human gut flora, and a large proportion of sequences (>60%) are unclassifiable at the genus level. These results are consistent with the findings reported in [10]. (2) There are clearly two subgroups within *Firmicutes*, supporting a recent suggestion that this phylum is likely to be redefined [31].

We next used Isomap to analyze the top ranked OTUs correlated with obesity. Among the 7491 OTUs defined at the 0.10 distance level, only 266 (<3.6%) OTUs had a significant correlation with weight status with a p-value <0.05. The results of topology analysis are presented in Fig. 6. Unlike the previous results, the face color of each circle represents the magnitude of the corresponding correlation coefficient with obesity. BLAST search against the RDP-II [27] and greengenes [28] databases was used to phylogenetically classify the 52,227 sequences within the 266 top ranked OTUs (Tables 1S, 2S and 3S). For ease of presentation, each cluster was labeled with the name of the phylum it was assigned to. From the analysis, we observed that: (1) The compositions of most OTUs within *Bacteroidetes* and *Firmicutes* phyla have little or no correlation with the disease states. (2) The OTUs within *Bacteroidetes* tend to have a negative correlation with obesity, which is concordant with previous results suggesting obese individuals have a lower proportion of *Bacteroidetes* in the gut [6, 7, 9]. (3) As we observed in the total gut topology structure analysis in Fig. 5, *Firmicutes* is partitioned into two subgroups. Interestingly, one subgroup contains more OTUs that have a positive correlation with obesity, while the other group contains more negatively correlated OTUs. This, together with the first observation, may explain why previous studies did not find a significant connection between *Firmicutes* and obesity

since analyses were largely restricted to the phylum level and treated *Firmicutes* as a single group [9].

The full annotation results of the sequences within the 266 top ranked obesity-associated OTUs are reported in supplementary Tables 1S, 2S and 3S. Notably, as many as 40,000 (>77%) sequences were classified as unknown at the genus level, suggesting that many potentially important gut microbes have yet to be characterized. As this is one of the deepest interrogations of the gut microbiota to date, it is not surprising that there is no prior information available on the association of many OTUs revealed here with obesity or any other human diseases. However, previous reports of phylum level associations and analysis of models using cultivatable species from representative genera provide pointers to potential roles for phylotypes in obesity.

Our analysis revealed that several OTUs classified as belonging to *Bacteroidetes* were all negatively correlated with obesity (Fig. 6). There have been conflicting results with regard to the relationship of *Bacteroidetes* and obesity in human studies. In a study using FISH probes, Duncan et al. found no relationship between obesity and *Bacteroides* populations in individuals on controlled weight-maintenance diets [32]. Zhang et al. also found no difference between the fraction of *Bacteroidetes* in obese and non-obese individuals in a sequence-based study [14]. Conversely, Nadal et al. demonstrated an increase in *Bacteroides* proportions in adolescents on a weight-loss regimen [33], and studies by Ley et al. proposed that a reciprocal relationship between *Bacteroidetes* and *Firmicutes* was evident in obese individuals [7]. While the total abundance of microbes within this phylum may not be an accurate biomarker of obesity in itself as shown above, analysis at the genus level may reveal significant associations between specific members of *Bacteroidetes* and weight status. The two genera from this phylum that were most associated with weight status were *Bacterioides* and *Rikenella* (p-value<0.0001). It has been proposed that *Bacteroides* populations could contribute to the generation of propionate, which may favor a lean phenotype by inhibiting lipid synthesis from acetate [34].

A novel finding derived from applying our new analytical tools is that while some OTUs classified as *Firmicutes* were correlated positively with obesity, others showed a negatively correlation (Fig. 6). The large majority of OTUs in *Firmicutes* were comprised of the class *Clostridia* and the order *Clostridiales*. Notably, *unclassified Clostridiales*, *Clostridiaceae* and *Lachnospiraceae* were the most prevalent components in *Clostridiales*. The classified genera from this phylum that were most associated with a decrease of abundance in obe-

sity were *Megasphaera*, *Phascolarctobacterium*, and *Erysipelothrix*. *Megaspheara* and *Phascolarctobacterium* are genera of the *Acidaminococcaceae* family, anaerobic Gram-negative diplococci that use amino acids as their sole energy source. These genera are routinely found in the gut of mammals, but no direct link between energy extraction efficiency or host physiology has been reported to date. The genera from *Firmicutes* that were increased in obesity included *Roseburia*, *Sporobacter* and *Faecalibacterium*. *Faecalibacterium* is a major component of the gut flora and members are thought to influence colonic health in a number of ways [35].

4 Conclusions

Advances in next-generation DNA sequencing technology allow researchers to obtain millions of DNA sequences rapidly and economically. Consequently, large-scale DNA sequencing is increasingly used as a primary research tool in environmental and human epidemiological studies. Advanced computational algorithms are crucial to efficiently extract pertinent information from massive nucleotide data collections to maximize research yields. While many 16S rRNA based studies were mainly designed to catalog the diversity of microbial populations [12], we report here a novel analytical strategy that enables researchers to deeply investigate the hidden world of microbes beyond basic microbial diversity estimation. We applied the proposed strategy to derive specific microbial signatures associated with obesity and describe microbial community structures in far more detail than previously achievable. Although we still cannot determine the cause/effect relationship between the human gut microbiota and obesity, we have clearly shown that our approach partially addresses the needs of analyzing the HMP data. Whether the association we identified is direct or indirect is a subject of large-scale population studies, and is outside the scope of this method paper. However, the strategy for analyzing the data from population studies largely remains the same.

We herein mainly focused on taxonomy independent analysis, discriminant analysis and topology analysis. The ultimate goal of a microbial community analysis is to establish a microbial interaction network. Since only a small fraction of microbes can be cultivated in laboratories under current technologies, it would be difficult to use a cultivation-based method to perform such studies. Accordingly, little work has been done in this direction [37]. Profile data matrices generated through taxonomy independent analysis contain suffi-

cient statistical information to study dynamic microbe-microbe and environment-microbe interactions. The results of our ongoing network analyses will be reported elsewhere.

The above bioinformatics analysis can be applied to query multiple research questions. For example, clinical microbiologists may want to derive microbial signatures to characterize microbially caused diseases such as bacterial pneumonia and inflammatory bowel disease; they may also want to perform time series analyses to study how antibiotics usage affects the dynamics of microbial communities over time [4] (in this case, each column of a data profile matrix represents a time point at which a sample is collected.); in our study, we found that *Bacteroidetes* is the second largest phylum present in the human gut. A recent study showed that in elderly individuals, it is *Actinobacteria* that is the second most abundant gut phylum [36]. It would be interesting to study how microbial composition changes over time by collecting gut samples from individuals of different ages. The above are just a few possible applications. We are currently developing a web application that will provide researchers with a complete package of computational tools for microbial community analysis. We hope that the web application will be of high utility for the microbiology community and beyond.

References

[1] Whitman WB, Coleman DC, Wiebe WJ. (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* **95**(12):6578-6583.

[2] Eisen JA. (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol* **5**:e82.

[3] Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci USA* **103**:12115-12120.

[4] Dethlefsen L, Huse S, Sogin ML, Relman DA. (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biology* **6**:e280.

[5] Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, et al. (2009) The NIH Human Microbiome Project. *Genome Res* **19**(12):2317-2323.

[6] Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. (2005) Obesity alters gut microbial ecology. *Proc Natl Acad Sci USA* **102**:11070-11075.

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

[8] Tschöp MH, Hugenholtz P, Karp CL. (2009) Getting to the core of the gut microbiome. *Nat Biotechnol* **27**(4):344-346.

[9] Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, et al. (2009) A core gut microbiome in obese and lean twins. *Nature* **457**(7228):480-484.

[10] Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. (2005) Diversity of the human intestinal microbial flora. *Science* **308**(5728):1635-1638.

[11] Huse SM, Dethlefsen L, Huber JA, Mark Welch D, Relman DA, Sogin ML. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* **4**:e1000255.

[12] Fabrice A, Didier R. (2009) Exploring microbial diversity using 16S rRNA high-throughput methods. *J Comput Sci Syst Biol* **2**(1):074-92.

[13] Hamady M, Knight R. (2009) Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res* **19**(7):1141-1152.

[14] Zhang H, DiBaise JK, Zuccolo A, Kudrna D, Braidotti M, Yu Y, Parameswaran P, Crowell MD, Wing R, Rittmann BE, Krajmalnik-Brown R. (2009) Human gut microbiota in obesity and after gastric bypass. *Proc Natl Acad Sci USA* **106**:2365-2370.

[15] Schloss PD, Handelsman J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol*, **71**:1501-1506.

[16] Sun Y, Cai Y, Liu L, Yu F, Farrell ML, McKendree W, Farmerie W. (2009) ESPRIT: Estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res* **37**(10):e76.

[17] Needleman SB, Wunsch CD. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**(3):443-453.

[18] van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**:530-536.

[19] Sun Y, Goodison S, Li J, Liu L, Farmerie W. (2007) Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics* **23**(1):30-37.

[20] Clarridge JE. 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev* **17**:840-862.

[21] Sun Y. (2007) Iterative RELIEF for feature weighting: algorithms, theories, and applications. *IEEE Trans Pattern Anal Mach Intell* **29**(6):1035-1051.

[22] Sun Y, Todorovic S, Goodison S. (2010) Local learning based feature selection for high dimensional data analysis. *IEEE Trans Pattern Anal Mach Intell* **32**(9):1610-1626.

[23] Cai Y, Sun Y, Cheng Y, Li J, Goodison S. (2010) Fast implementation of ℓ_1 regularized learning algorithms using gradient descent methods. *Proc 10th SIAM International Conference on Data Mining*, 862-871.

[24] Duda RO, Hart PE, Stork DG. (2001) *Pattern Classification*. Wiley, New York.

[25] Tenenbaum JB, de Silva V, Langford JC. (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* **290**(5500):2319-2323.

[26] Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**(16):5261-5267.

[27] Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM, Bandela AM, Cardenas E, Garrity GM, Tiedje JM. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res* **35**:D169-D172.

[28] DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**:5069-5072.

[29] Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, et al. (2008) The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**:75.

[30] Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML. (2007) Microbial population structures in the deep marine biosphere. *Science*, **318**:97-100.

[31] Wolf M, Müller T, Dandekar T, Pollack JD. (2004) Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. *Int J Syst Evol Microbiol* **54**(Pt 3):871-875.

[32] Duncan SH, Lobley GE, Holtrop G, Ince J, Johnstone AM, Louis P, Flint HJ. (2008) Human colonic microbiota associated with diet, obesity and weight loss. *Int J Obes (Lond)* **32**:1720-1724.

[33] Nadal I, Santacruz A, Marcos A, Warnberg J, Garagorri M, Moreno LA, Martin-Matillas M, Campoy C, Marti A, Moleres A, et al. (2009) Shifts in clostridia, bacteroides and immunoglobulin-coating fecal bacteria associated with weight loss in obese adolescents. *Int J Obes (Lond)* **33**(7):758-767.

[34] Wolever TM, Spadafora PJ, Cunnane SC, Pencharz PB. (1995) Propionate inhibits incorporation of colonic [1,2-13C]acetate into plasma lipids in humans. *Am J Clin Nutr* **61**:1241-1247.

[35] Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermdez-Humarn LG, Gratadoux JJ, Blugeon S, Bridonneau C, Furet JP, Corthier G, et al. (2008) Faecalibacterium praus-

nitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc Natl Acad Sci USA* **105**(43):16731-16736.

[36] Andersson AF, Lindberg M, Jakobsson H, Bäckhed F, Nyren P, Engstrand L. (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One* **3**(7):e2836.

[37] Fuhrman JA. (2009) Microbial community structure and its functional implications. *Nature* **459**(7244):193-199.

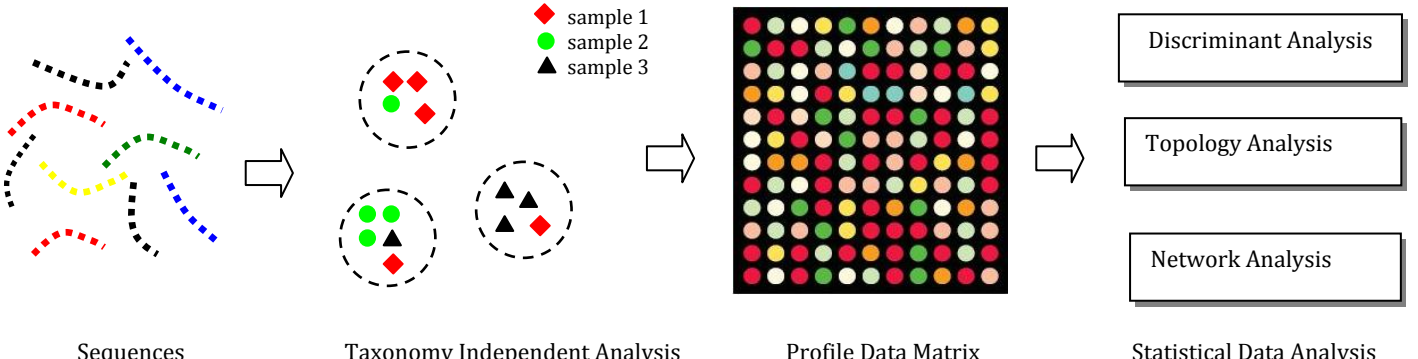

Figure 1: Schematic diagram of the presented analytical strategy.

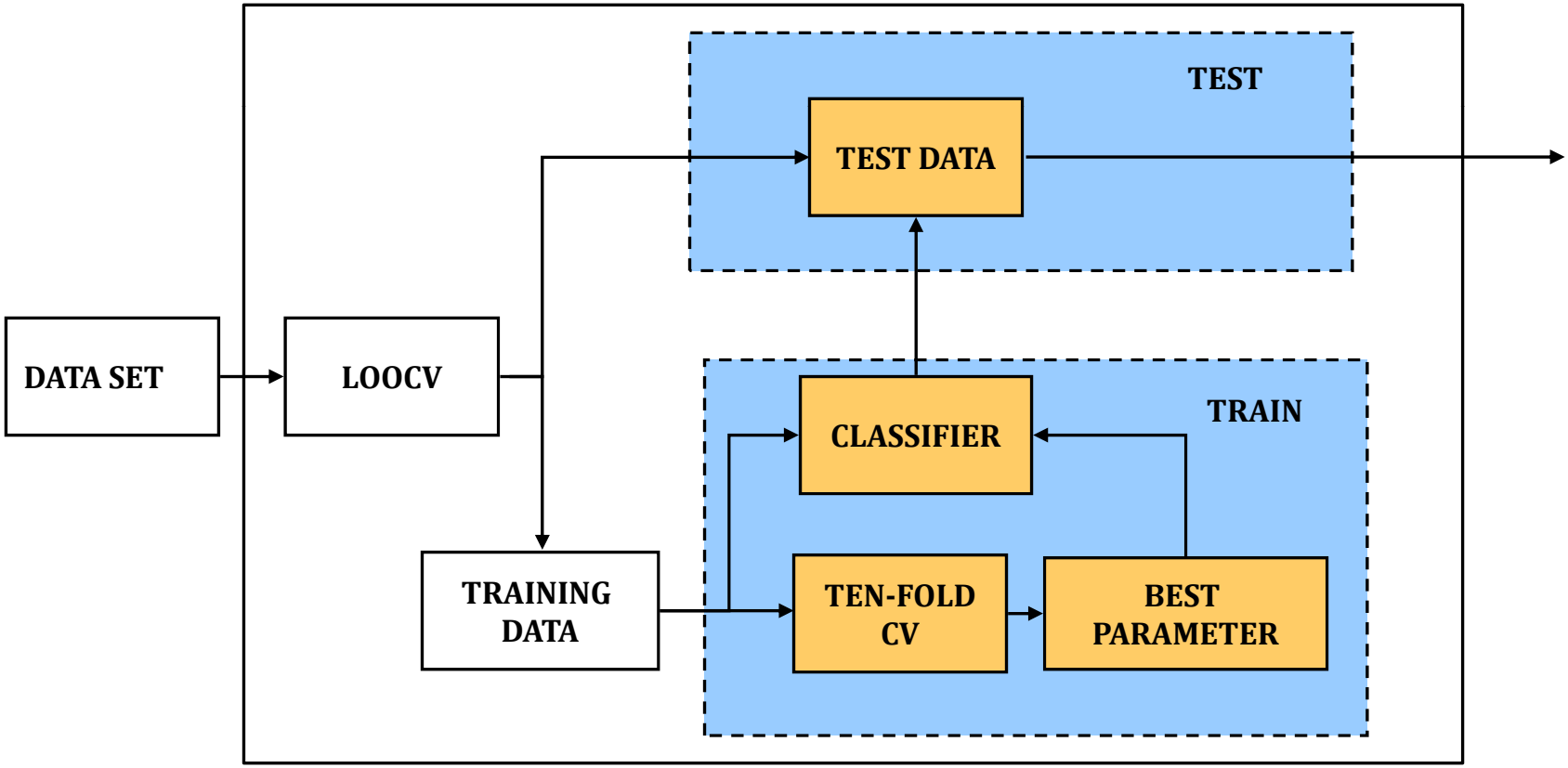

Figure 2: The experimental protocol consisted of an inner and an outer loop. In the inner loop, the regularization parameter of a logistical regression model was estimated through ten-fold cross validation using the training data provided by the outer loop, and in the outer loop a predictive model was trained using the best parameter from the inner loop and held-out samples were blindly classified. The experiment was repeated until each sample had been tested. Test samples were not involved in any stage of training process. LOOCV: leave-one-out cross validation. CV: cross validation.

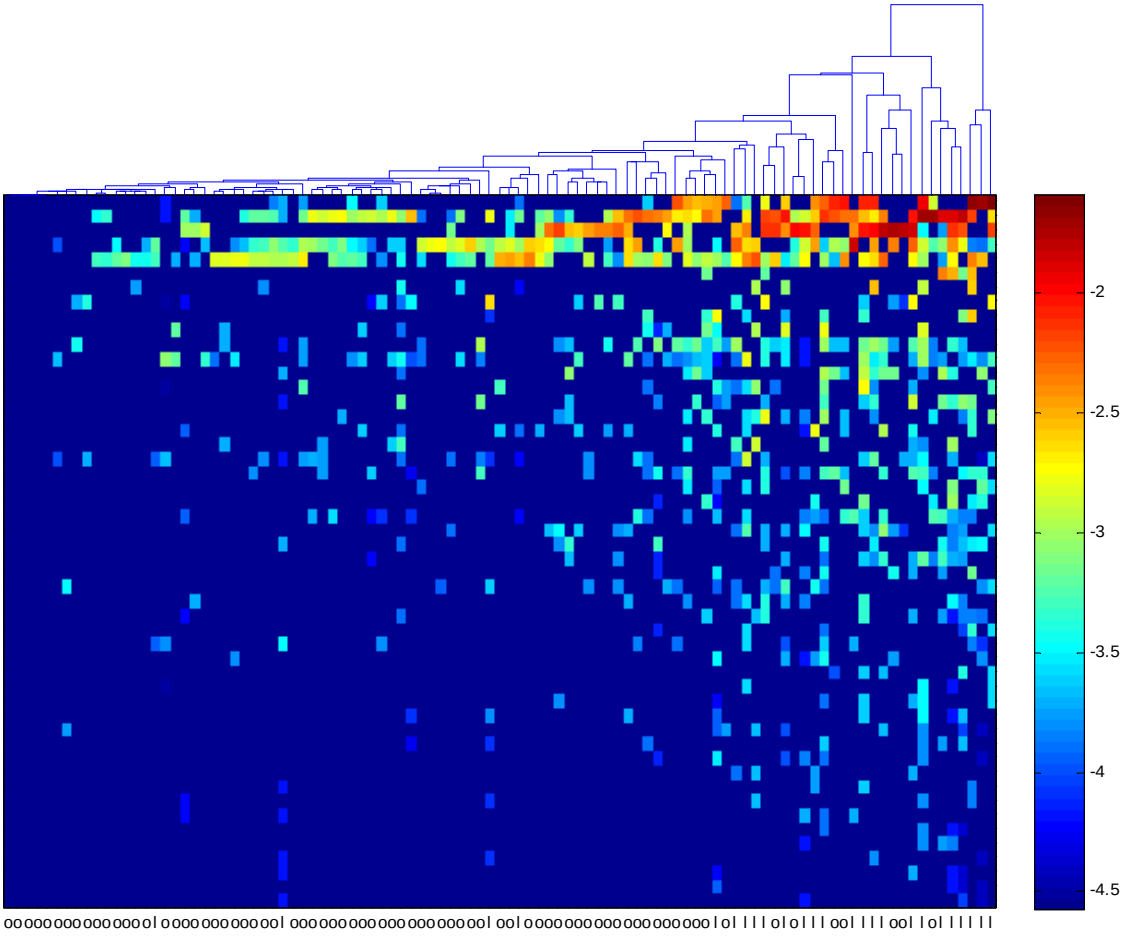

Figure 3: Heatmap of the top 50 ranked gut microbiota phylotypes (rows) defined at the 0.08 distance level. Lean individuals (l) have a distinguishing pattern of microbial composition profiles compared to obese individuals (o). The phylotypes were ranked based on their corresponding correlation coefficients with respect to physiological status.

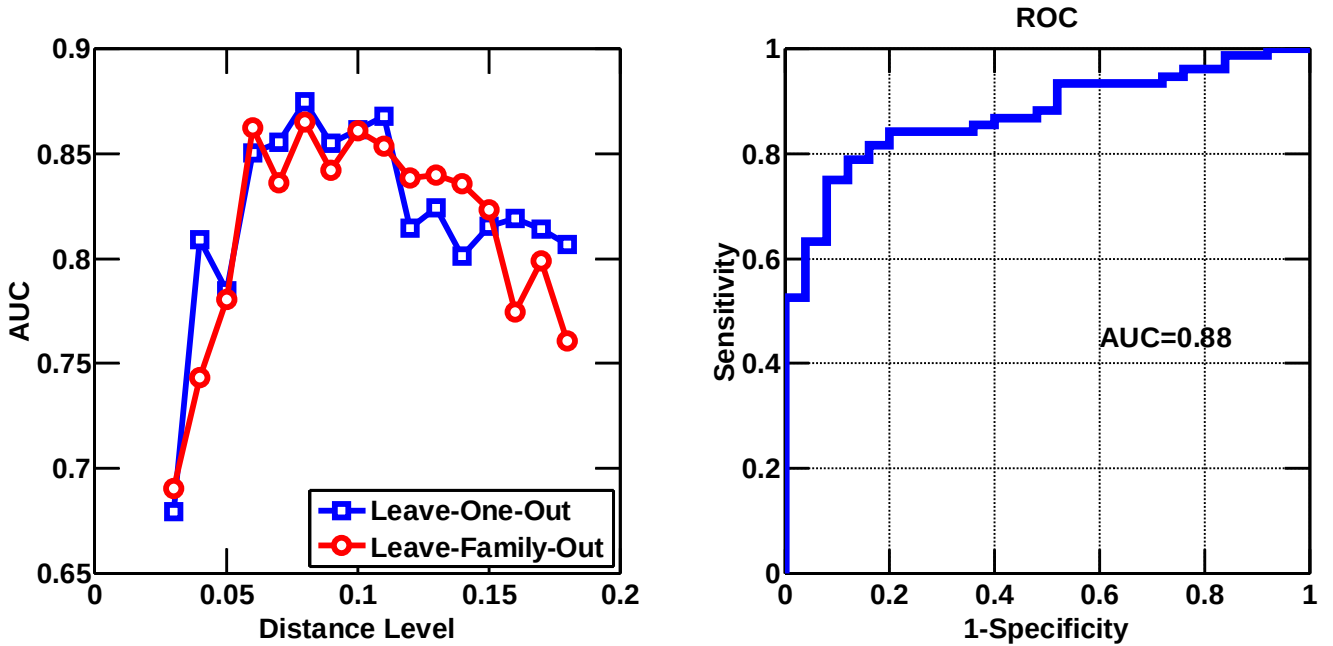

Figure 4: Results of discriminant analysis. (left panel) The area under receiver operating characteristic (AUC) curves obtained at various distance levels ranging from 0.03 to 0.18; (right panel) The receiver operating characteristic (ROC) curve obtained at the 0.08 distance level. The sensitivity and specificity are defined as the rate of correctly predicting obese and lean individuals, respectively.

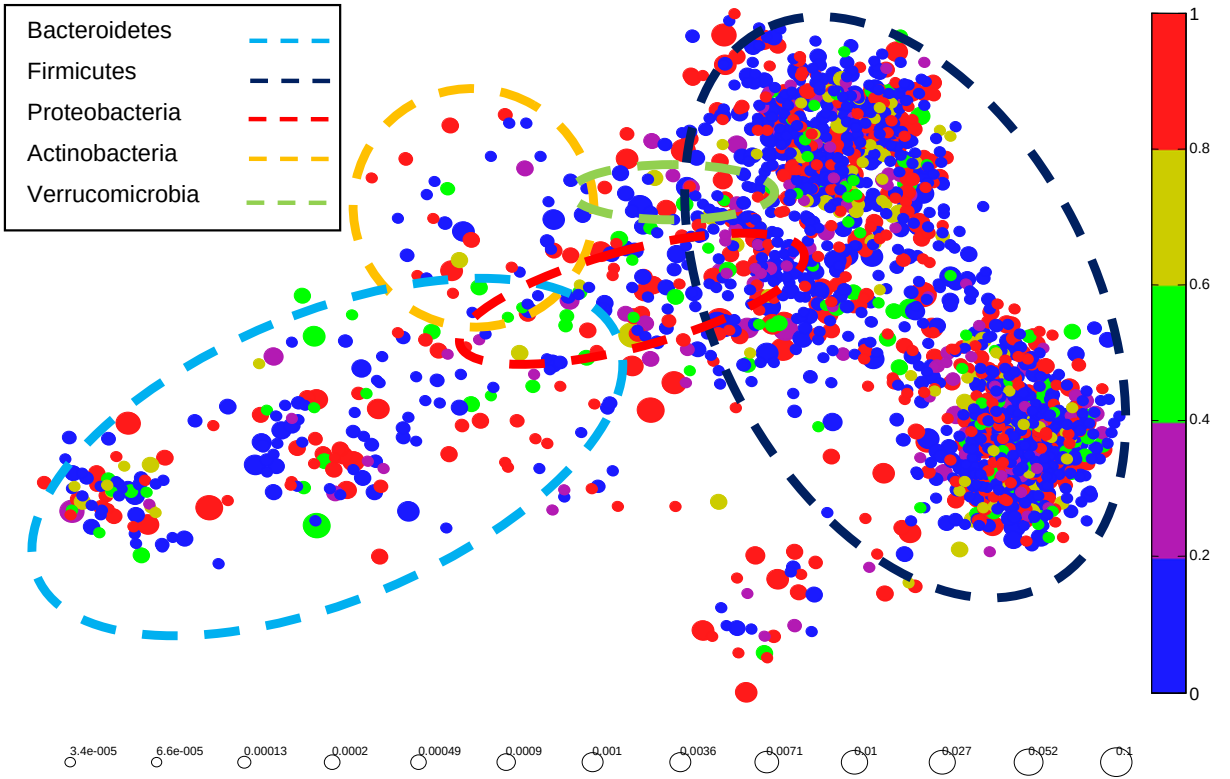

Figure 5: Topology analysis performed on the human gut flora. See the main text for detailed descriptions.

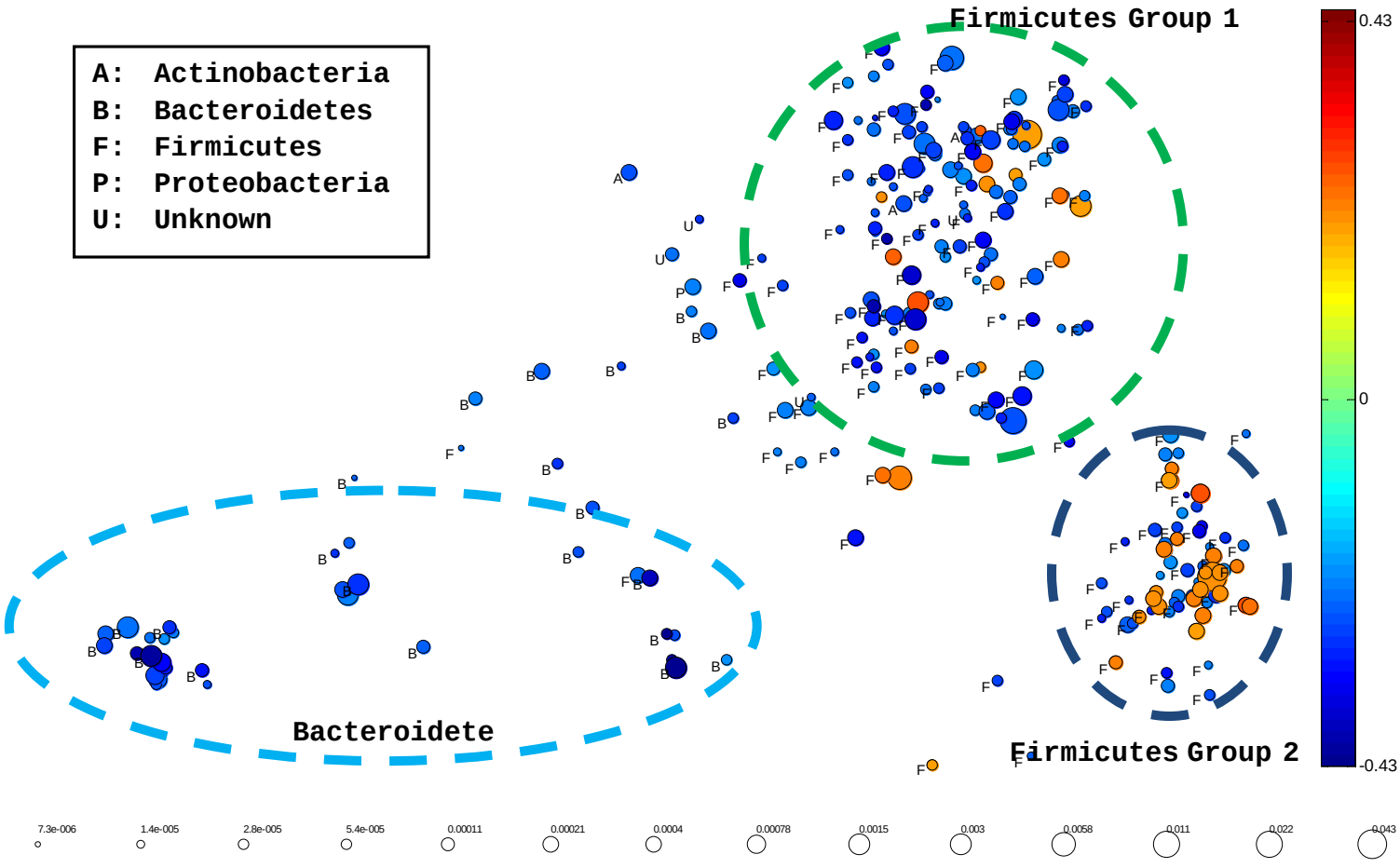

Figure 6: Topology analysis performed on the 266 top ranked phylotypes (p-value<0.05) defined at the 0.10 distance level.

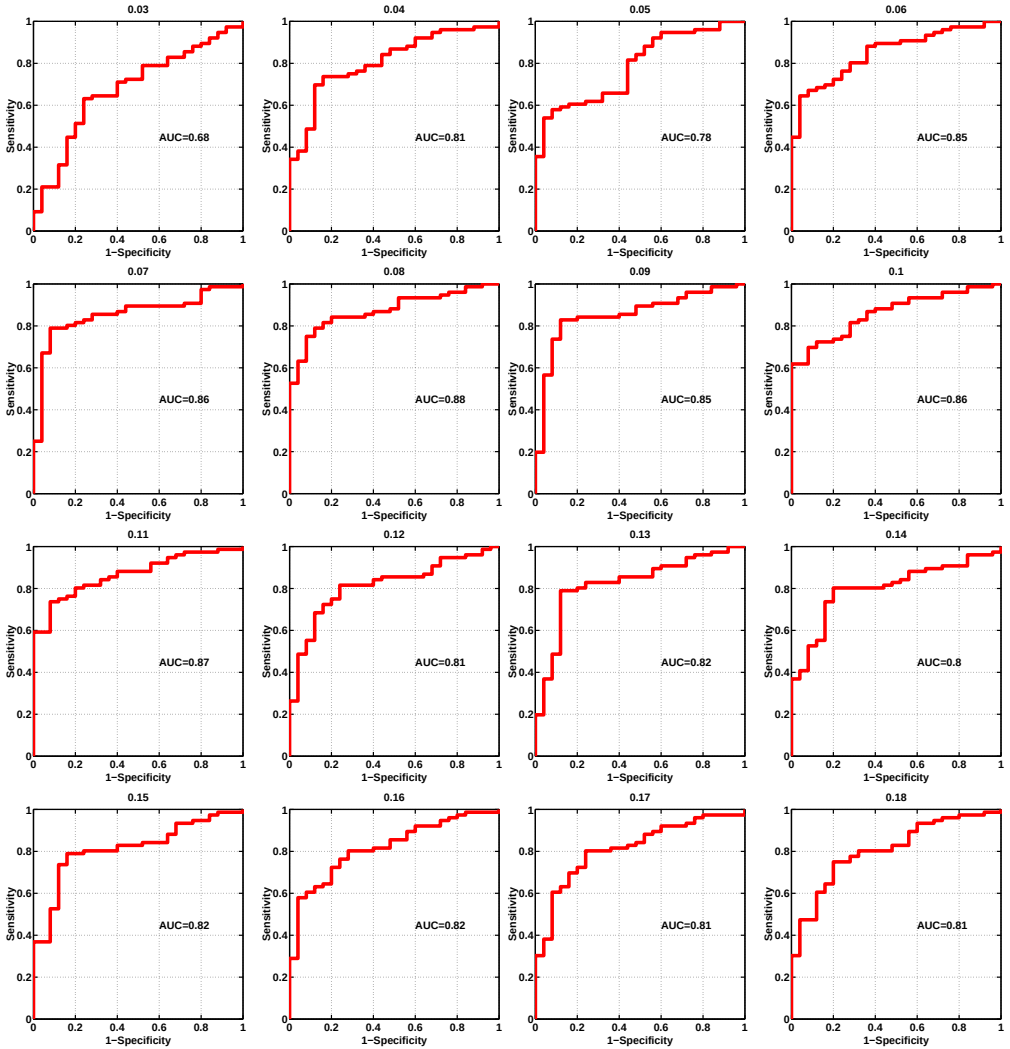

Figure 1S: The receiver operating characteristic curves obtained at various distance levels ranging from 0.03 to 0.18.

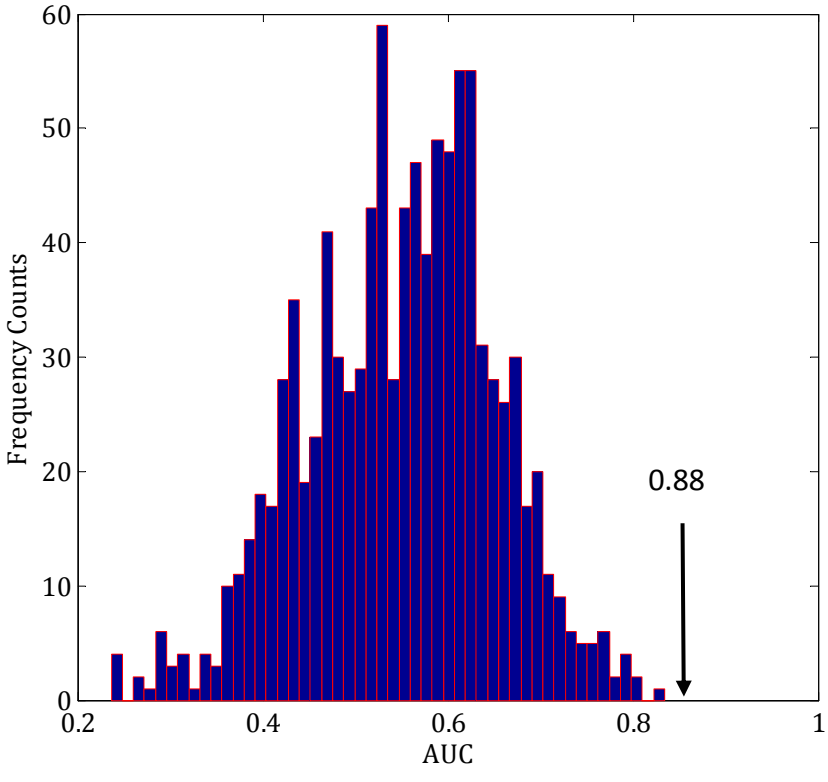

Figure 2S: Result of a permutation test. For computational reasons, the permutation test was repeated 1000 time. In each iteration, the class labels were randomly shuffled, the previously described experimental protocol was executed, and the area under the resulting ROC curve was recorded. The p-value was computed as the occurrence frequencies of the iterations where the resulting AUCs outperformed that obtained using the original class labels.

TABLE I: Breakdown of the 52,227 sequences within the 266 top ranked obesity-associated phylotypes at various phylogenetic levels. The annotation was performed at the RAST web application using BLAST search against RDP-II database. A query sequence was assigned to the organism of the best-matched reference sequence if e-value $< 10^{-20}$ and identity percentage $\geq 95\%$.

Kingdom	Phylum	Class	Order	Family	Genus	# of Reads
Bacteria	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Eggerthella	167
Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Bacteroidaceae	Bacteroides	3267
Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae	unclassified Porphyromonadaceae	996
Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Rikenellaceae	Rikenella	1306
Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Rikenellaceae	unclassified Rikenellaceae	203
Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	unclassified Bacteroidales	unclassified Bacteroidales	1779
Bacteria	Firmicutes	Clostridia	Clostridiales	Acidaminococcaceae	Acidaminococcus	52
Bacteria	Firmicutes	Clostridia	Clostridiales	Acidaminococcaceae	Megasphaera	1064
Bacteria	Firmicutes	Clostridia	Clostridiales	Acidaminococcaceae	Phascolarctobacterium	1330
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Acetivibrio	55
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Clostridium	95
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Dorea	209
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Sporobacter	3025
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	unclassified Clostridiaceae	229
Bacteria	Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Eubacterium	1222
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Anaerofilum	744
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia	853
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Ruminococcus	414
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	unclassified Lachnospiraceae	5509
Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Sedimentibacter	15
Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified Clostridiales	unclassified Clostridiales	22551
Bacteria	Firmicutes	Clostridia	unclassified Clostridia	unclassified Clostridia	unclassified Clostridia	173
Bacteria	Firmicutes	Mollicutes	Incertae sedis	Erysipelotrichaceae	Erysipelothrix	3099
Bacteria	Firmicutes	Mollicutes	unclassified Mollicutes	unclassified Mollicutes	unclassified Mollicutes	160
Bacteria	Firmicutes	unclassified Firmicutes	unclassified Firmicutes	unclassified Firmicutes	unclassified Firmicutes	3893
Bacteria	Proteobacteria	Alphaproteobacteria	unclassified Alphaproteobacteria	unclassified Alphaproteobacteria	unclassified Alphaproteobacteria	106
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	Sutterella	91
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Bilophila	226
Bacteria	unclassified Bacteria	unclassified Bacteria	unclassified Bacteria	unclassified Bacteria	unclassified Bacteria	5096

TABLE II: Breakdown of the 52,227 sequences within the 266 top ranked obesity-associated phylotypes at various phylogenetic levels. The annotation was performed at the RAST web application using BLAST search against greengenes database. A query sequence was assigned to the organism of the best-matched reference sequence if e-value $< 10^{-20}$ and identity percentage $\geq 95\%$.

Kingdom	Phylum	Class	Order	Family	Genus	# of Reads
Bacteria	Actinobacteria	Actinobacteria	Coriobacteridae	Coriobacteriales	Eggerthella	167
Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Bacteroidaceae	Bacteroides	4284
Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae	Tannerella	181
Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae	unclassified Porphyromonadaceae	860
Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Rikenellaceae	Alistipes	1683
Bacteria	Firmicutes	Clostridia	Clostridiales	Acidaminococcaceae	Acidaminococcus	52
Bacteria	Firmicutes	Clostridia	Clostridiales	Acidaminococcaceae	Megasphaera	1064
Bacteria	Firmicutes	Clostridia	Clostridiales	Acidaminococcaceae	Papillibacter	533
Bacteria	Firmicutes	Clostridia	Clostridiales	Acidaminococcaceae	Phascolarctobacterium	1330
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Acetivibrio	261
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Anaerotruncus	356
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Clostridium	338
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Coprobacillus	5094
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Dorea	134
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Faecalibacterium	11058
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Sporobacter	191
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Subdoligranulum	571
Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	unclassified Clostridiaceae	3629
Bacteria	Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Eubacterium	1329
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Anaerostipes	103
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus	21
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospira	25
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia	6521
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Ruminococcus	937
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	unclassified Lachnospiraceae	6654
Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified Clostridiales	unclassified Clostridiales	8553
Bacteria	Firmicutes	Clostridia	unclassified Clostridia	unclassified Clostridia	unclassified Clostridia	777
Bacteria	Firmicutes	unclassified Firmicutes	unclassified Firmicutes	unclassified Firmicutes	unclassified Firmicutes	99
Bacteria	Proteobacteria	Alphaproteobacteria	unclassified Alphaproteobacteria	unclassified Alphaproteobacteria	unclassified Alphaproteobacteria	106
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	Sutterella	91
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	unclassified Desulfovibrionaceae	226
Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified	666

TABLE III: Annotation of the 266 top ranked obesity-associated phylotypes (p-value$<$0.05 based on Pearson's correlation analysis) by using the BLAST search against the RDP-II and greengenes databases. The analysis was performed on the RAST web application. See Methods section for details.
# of Reads: Number of reads within a phylotype
Percentage (L): Averaged microbial abundance (in percentage) of a phylotype in the lean group
Percentage (O): Averaged microbial abundance (in percentage) of a phylotype in the obese group
PCC: Pearson's correlation coefficient of the microbial abundance of a phylotype with respect to physiological status
p-value: p-value of Pearson's correlation analysis
RDP: annotation results of the sequences within a phylotype using BLAST search again RDP database. The number after an organism is the percentage of the sequences within a phylotype belonging to that organism.
greengenes: annotation results of the sequences within a phylotype using BLAST search again greengenes database.

Phylotype ID	# of Reads	Percentage (L)	Percentage (O)	PCC	p-value	RDP	greengenes
Cluster1	29	1.18e-002	1.08e-003	-0.43	0.00001	Rikenella 97%	Alistipes 97%
Cluster2	2051	5.29e-001	1.35e-001	-0.42	0.00001	Rikenella 100%	Alistipes 100%
Cluster3	1347	4.62e-001	5.33e-002	-0.40	0.00003	Bacteroides 100%	Bacteroides 100%
Cluster4	30	8.20e-003	6.16e-004	-0.39	0.00005	Rikenella 97%	Alistipes 97%
Cluster5	70	2.30e-002	3.04e-003	-0.39	0.00006	unclassified Clostridia 100%	unknown 100%
Cluster6	49	1.56e-002	2.82e-003	-0.38	0.00007	Phascolarctobacterium 63% unknown 37%	Phascolarctobacterium 63% unknown 37%
Cluster7	21	7.79e-003	7.82e-004	-0.38	0.00008	unknown 100%	unknown 95%
Cluster8	76	2.93e-002	2.55e-003	-0.38	0.00009	Bacteroides 100%	Bacteroides 100%
Cluster9	128	4.72e-002	6.21e-003	-0.37	0.00012	Rikenella 96%	Alistipes 96%
Cluster10	76	2.15e-002	3.14e-003	-0.36	0.00018	Bacteroides 99%	Bacteroides 98%
Cluster11	725	1.85e-001	7.65e-002	-0.35	0.00030	unclassified Clostridiales 100%	unclassified Clostridia 100%
Cluster12	2228	5.89e-001	1.84e-001	-0.35	0.00033	Phascolarctobacterium 100%	Phascolarctobacterium 100%
Cluster13	41	2.07e-002	2.19e-003	-0.34	0.00052	Sporobacter 98%	unclassified Clostridiales 98%
Cluster14	15	3.68e-003	1.31e-004	-0.33	0.00064	Eubacterium 13% unclassified Clostridiales 60% unknown 20%	Roseburia 13% unclassified Clostridiales 33% unclassified Lachnospiraceae 27% unknown 20%
Cluster15	88	3.00e-002	7.13e-003	-0.33	0.00081	Phascolarctobacterium 74% unknown 26%	Phascolarctobacterium 74% unknown 26%
Cluster16	111	4.70e-002	9.23e-004	-0.33	0.00082	unknown 98%	unknown 98%
Cluster17	109	2.92e-002	9.04e-003	-0.33	0.00084	unclassified Clostridiales 99%	unclassified Clostridiales 97%
Cluster18	117	3.72e-002	1.09e-002	-0.33	0.00086	unclassified Clostridiales 80% unclassified Firmicutes 12%	Faecalibacterium 99%
Cluster19	85	2.78e-002	5.43e-003	-0.32	0.00099	unclassified Lachnospiraceae 95%	unclassified Clostridiales 95%
Cluster20	123	3.21e-002	1.07e-002	-0.32	0.00111	unknown 100%	unknown 100%
Cluster21	25	8.11e-003	8.85e-004	-0.32	0.00116	unclassified Clostridiales 64% unclassified Lachnospiraceae 12% unknown 20%	unclassified Clostridiales 48% unknown 44%
Cluster22	74	1.64e-002	4.72e-003	-0.32	0.00117	unknown 100%	unknown 97%
Cluster23	42	1.17e-002	1.73e-003	-0.32	0.00131	unclassified Clostridiales 98%	unclassified Clostridiales 98%
Cluster24	179	4.46e-002	8.19e-003	-0.31	0.00141	Erysipelothrix 46% unclassified Firmicutes 54%	Coprobacillus 100%
Cluster25	394	1.90e-001	1.62e-002	-0.31	0.00144	Bacteroides 99%	Bacteroides 99%
Cluster26	256	6.87e-002	6.12e-003	-0.31	0.00148	unclassified Clostridiales 99%	unclassified Clostridiaceae 50% unclassified Lachnospiraceae 48%
Cluster27	241	4.63e-002	1.65e-002	-0.31	0.00156	unclassified Lachnospiraceae 95%	unclassified Clostridiaceae 71% unclassified Clostridiales 25%
Cluster28	66	1.86e-002	1.14e-003	-0.31	0.00185	unknown 100%	unclassified Clostridiales 100%
Cluster29	130	2.99e-002	9.98e-003	-0.30	0.00213	Rikenella 38% unknown 62%	Alistipes 48% unknown 52%
Cluster30	54	2.69e-002	3.86e-003	-0.30	0.00222	Bacteroides 94%	Bacteroides 94%
Cluster31	700	1.46e-001	4.78e-002	-0.30	0.00225	Erysipelothrix 100%	Coprobacillus 100%
Cluster32	37	1.01e-002	1.32e-003	-0.30	0.00229	unknown 100%	unknown 100%
Cluster33	43	1.54e-002	2.13e-003	-0.30	0.00244	unclassified Clostridia 95%	unclassified Clostridia 95%
Cluster34	19	5.66e-003	8.13e-004	-0.30	0.00251	Roseburia 100%	unclassified Clostridiales 100%
Cluster35	47	1.08e-002	2.00e-003	-0.29	0.00278	unknown 98%	unknown 100%
Cluster36	68	2.59e-002	3.83e-003	-0.29	0.00279	unclassified Clostridiales 49% unclassified Firmicutes 40% unknown 12%	unclassified Clostridiales 97%
Cluster37	14	3.41e-003	4.73e-004	-0.29	0.00281	unknown 100%	unknown 100%
Cluster38	18	5.50e-003	5.03e-004	-0.29	0.00288	Eubacterium 22% unclassified Clostridiales 17% unknown 61%	Anaerostipes 33% Eubacterium 17% unknown 50%
Cluster39	25	7.12e-003	1.07e-003	-0.29	0.00290	unknown 96%	unknown 96%
Cluster40	76	1.66e-002	4.43e-003	-0.29	0.00298	Erysipelothrix 99%	Coprobacillus 99%
Cluster41	13	5.10e-003	9.38e-004	-0.29	0.00312	Phascolarctobacterium 85% unknown 15%	Phascolarctobacterium 85% unknown 15%
Cluster42	15	6.87e-003	1.80e-004	-0.29	0.00316	Eubacterium 40% unknown 47%	unclassified Clostridiales 27% unknown 73%
Cluster43	12	3.24e-003	2.86e-004	-0.29	0.00317	unclassified Clostridiales 17% unknown 83%	Faecalibacterium 17% unclassified Clostridiaceae 42% unknown 42%
Cluster44	60	1.84e-002	5.45e-003	-0.29	0.00332	Bacteroides 100%	Bacteroides 100%
Cluster45	17	5.81e-003	3.35e-004	-0.29	0.00335	unclassified Clostridiales 59% unknown 41%	unclassified Clostridiaceae 29% unclassified Lachnospiraceae 41% unknown 29%
Cluster46	35	6.39e-003	1.78e-003	-0.29	0.00342	unclassified Clostridiales 91%	unclassified Clostridiales 91%
Cluster47	119	3.77e-002	1.24e-002	-0.29	0.00344	unclassified Clostridiales 93%	unclassified Clostridiaceae 100%
Cluster48	747	1.72e-001	1.81e-002	-0.29	0.00370	Ruminococcus 100%	unclassified Clostridiales 100%
Cluster49	45	1.74e-002	1.73e-003	-0.29	0.00378	unknown 98%	unclassified Clostridiales 78% unknown 22%
Cluster50	137	7.72e-002	7.14e-004	-0.29	0.00383	Clostridium 99%	Clostridium 99%
Cluster51	15	4.53e-003	5.19e-004	-0.28	0.00395	unclassified Clostridiales 40% unknown 53%	unclassified Clostridiales 80% unknown 13%
Cluster52	15	5.11e-003	3.08e-004	-0.28	0.00396	unclassified Clostridiales 93%	unclassified Clostridiaceae 13% unclassified Lachnospiraceae 80%
Cluster53	13	3.64e-003	4.23e-004	-0.28	0.00409	Eubacterium 23% unknown 77%	Eubacterium 23% unknown 77%
Cluster54	12	5.73e-003	2.43e-004	-0.28	0.00428	unclassified Clostridiales 100%	Ruminococcus 100%
Cluster55	114	2.19e-002	7.67e-003	-0.28	0.00434	Bacteroides 97%	Bacteroides 96%
Cluster56	15	4.06e-003	7.12e-004	-0.28	0.00440	unclassified Bacteroidales 53% unknown 40%	Bacteroides 67% unknown 33%
Cluster57	1680	3.64e-001	1.09e-001	-0.28	0.00440	unclassified Bacteroidales 12% unclassified Porphyromonadaceae 88%	unclassified Porphyromonadaceae 92%
Cluster58	260	5.75e-002	2.29e-002	-0.28	0.00445	Anaerofilum 97%	Subdoligranulum 98%
Cluster59	24	6.02e-003	1.09e-003	-0.28	0.00445	unclassified Clostridiales 96%	unclassified Clostridiales 96%

Continued on next page

TABLE III – continued from previous page

Phylotype ID	# of Reads	Percentage (L)	Percentage (O)	PCC	p-value	RDP	greengenes
Cluster60	2456	5.59e-001	2.72e-001	-0.28	0.00446	unclassified Clostridiales 87% unclassified Firmicutes 13%	unclassified Clostridiaceae 100%
Cluster61	13	5.38e-003	9.02e-004	-0.28	0.00456	Bacteroides 85% unknown 15%	Bacteroides 83% unknown 17%
Cluster62	485	1.34e-001	2.64e-002	-0.28	0.00459	Roseburia 27% unclassified Lachnospiraceae 72%	unclassified Clostridiales 95%
Cluster63	22	1.01e-002	7.42e-004	-0.28	0.00463	Sporobacter 90%	unclassified Clostridiales 91%
Cluster64	40	8.31e-003	2.29e-003	-0.28	0.00487	unclassified Lachnospiraceae 82% unknown 18%	unclassified Clostridiaceae 32% unclassified Clostridiales 50% unknown 18%
Cluster65	18	6.07e-003	1.91e-004	-0.28	0.00508	unclassified Lachnospiraceae 67% unknown 28%	Anaerotruncus 67% unknown 28%
Cluster66	156	3.72e-002	1.38e-002	-0.28	0.00523	unclassified Clostridiales 84%	Faecalibacterium 99%
Cluster67	655	2.34e-001	4.82e-002	-0.28	0.00524	unclassified Clostridiales 70% unclassified Firmicutes 30%	unclassified Clostridiales 100%
Cluster68	59	1.70e-002	5.28e-003	-0.27	0.00581	unclassified Clostridiales 100%	unclassified Clostridiales 100%
Cluster69	23	5.62e-003	1.26e-003	-0.27	0.00588	Roseburia 13% unclassified Clostridiales 30% unclassified Lachnospiraceae 52%	Ruminococcus 87%
Cluster70	14	5.47e-003	5.34e-004	-0.27	0.00649	unknown 100%	unclassified Clostridia 93%
Cluster71	37	8.34e-003	2.36e-003	-0.27	0.00665	Roseburia 35% unclassified Clostridiales 14% unclassified Lachnospiraceae 46%	Ruminococcus 14% unclassified Clostridiaceae 22% unclassified Clostridiales 41% unclassified Lachnospiraceae 14%
Cluster72	17	5.94e-003	8.38e-004	-0.27	0.00675	Rikenella 12% unclassified Clostridiales 29% unclassified Lachnospiraceae 29% unknown 24%	Anaerostipes 24% Ruminococcus 24% unknown 47%
Cluster73	31	7.92e-003	2.04e-003	-0.27	0.00677	unclassified Bacteroidales 48% unclassified Porphyromonadaceae 29% unknown 23%	unclassified Porphyromonadaceae 74% unknown 26%
Cluster74	20	4.58e-003	1.01e-003	-0.27	0.00684	unknown 100%	unknown 100%
Cluster75	45	2.23e-002	8.52e-004	-0.27	0.00703	unclassified Clostridiales 100%	Anaerotruncus 100%
Cluster76	111	3.92e-002	2.68e-003	-0.27	0.00711	unclassified Lachnospiraceae 66% unknown 34%	unclassified Clostridiaceae 67% unknown 33%
Cluster77	144	5.26e-003	2.53e-002	0.27	0.00718	Roseburia 53% unclassified Clostridiales 43%	Roseburia 78% unclassified Lachnospiraceae 17%
Cluster78	107	2.17e-002	9.80e-003	-0.27	0.00731	Roseburia 42% unclassified Clostridiales 13% unknown 33%	Ruminococcus 10% unclassified Clostridiales 23% unclassified Firmicutes 34% unknown 30%
Cluster79	120	2.48e-002	9.92e-003	-0.27	0.00736	Sporobacter 100%	Papillibacter 100%
Cluster80	62	1.50e-002	3.68e-003	-0.26	0.00746	unclassified Clostridiales 100%	unclassified Clostridiales 100%
Cluster81	518	1.14e-001	4.90e-002	-0.26	0.00783	Sporobacter 100%	unclassified Clostridiales 100%
Cluster82	596	4.41e-002	1.07e-001	0.26	0.00795	unclassified Clostridiales 61% unclassified Lachnospiraceae 36%	Roseburia 61% unclassified Lachnospiraceae 38%
Cluster83	17	7.51e-003	1.37e-004	-0.26	0.00809	unknown 100%	unknown 100%
Cluster84	1952	0.00e+000	3.03e-001	0.26	0.00826	Megasphaera 100%	Megasphaera 100%
Cluster85	24	1.41e-002	1.01e-003	-0.26	0.00836	unknown 100%	unknown 96%
Cluster86	172	3.97e-002	1.98e-002	-0.26	0.00838	unclassified Clostridiales 89%	unclassified Clostridiaceae 95%
Cluster87	99	2.75e-002	9.58e-003	-0.26	0.00876	unclassified Clostridiales 98%	unclassified Clostridiales 94%
Cluster88	129	3.52e-002	1.26e-002	-0.26	0.00881	Bacteroides 100%	Bacteroides 100%
Cluster89	303	8.97e-002	2.92e-002	-0.26	0.00888	Sporobacter 100%	unclassified Clostridiales 100%
Cluster90	756	1.62e-001	3.25e-002	-0.26	0.00942	Bacteroides 100%	Bacteroides 100%
Cluster91	16	5.05e-003	1.07e-003	-0.26	0.00971	unclassified Clostridiales 50% unknown 50%	unclassified Clostridiaceae 50% unknown 50%
Cluster92	35	1.55e-002	1.56e-003	-0.26	0.00975	unknown 97%	unclassified Clostridiaceae 14% unknown 83%
Cluster93	13	3.82e-003	5.39e-004	-0.26	0.00976	unclassified Clostridiales 100%	unclassified Clostridiaceae 100%
Cluster94	44	1.17e-002	3.96e-003	-0.26	0.00979	unclassified Clostridiales 16% unclassified Lachnospiraceae 55% unknown 23%	Roseburia 27% unclassified Clostridiales 14% unclassified Lachnospiraceae 20% unknown 20%
Cluster95	28	7.10e-003	1.61e-003	-0.26	0.00984	Clostridium 20% unclassified Clostridiales 80%	Faecalibacterium 57% unknown 43%
Cluster96	73	1.49e-002	4.29e-003	-0.26	0.00992	unclassified Lachnospiraceae 99%	unclassified Lachnospiraceae 99%
Cluster97	250	3.96e-002	1.53e-002	-0.26	0.01005	Eggerthella 100%	Eggerthella 100%
Cluster98	256	7.15e-002	2.04e-002	-0.25	0.01010	Sporobacter 25% unknown 75%	Sporobacter 25% unknown 75%
Cluster99	67	1.42e-002	3.76e-003	-0.25	0.01013	unknown 100%	unknown 100%
Cluster100	23	3.69e-003	4.73e-004	-0.25	0.01018	Sporobacter 87% unknown 13%	unclassified Clostridiaceae 87% unknown 13%
Cluster101	13	3.02e-003	5.00e-004	-0.25	0.01040	unclassified Clostridiales 100%	Faecalibacterium 100%
Cluster102	19	9.50e-003	0.00e+000	-0.25	0.01051	unclassified Lachnospiraceae 100%	unclassified Lachnospiraceae 100%
Cluster103	562	8.72e-002	3.13e-002	-0.25	0.01109	unclassified Bacteroidales 83% unclassified Porphyromonadaceae 15%	Bacteroides 99%
Cluster104	16	3.25e-003	3.45e-004	-0.25	0.01112	Sporobacter 75% unknown 19%	Papillibacter 75% unknown 19%
Cluster105	18	5.63e-003	1.03e-003	-0.25	0.01164	Anaerofilum 22% unknown 78%	Subdoligranulum 22% unknown 78%
Cluster106	338	2.47e-002	5.97e-002	0.25	0.01169	unclassified Clostridiales 90%	Roseburia 59% unclassified Lachnospiraceae 31%
Cluster107	22	7.64e-003	1.10e-003	-0.25	0.01183	Rikenella 100%	Alistipes 100%
Cluster108	11560	1.77e+000	6.29e-001	-0.25	0.01183	Erysipelothrix 66% unclassified Firmicutes 34%	Coprobacillus 100%
Cluster109	63	1.65e-002	6.45e-003	-0.25	0.01188	unclassified Bacteroidales 92%	Alistipes 92%
Cluster110	25	8.07e-003	1.29e-003	-0.25	0.01194	unclassified Clostridiaceae 60% unknown 32%	unclassified Clostridiaceae 24% unknown 64%
Cluster111	192	6.73e-002	4.31e-003	-0.25	0.01195	unknown 93%	unknown 100%
Cluster112	124	0.00e+000	2.15e-002	0.25	0.01199	Megasphaera 98%	Megasphaera 98%
Cluster113	18	5.35e-003	7.26e-004	-0.25	0.01217	unclassified Clostridiales 73% unknown 18%	Faecalibacterium 67% unknown 33%
Cluster114	93	2.18e-002	9.02e-003	-0.25	0.01218	unclassified Clostridiales 83% unclassified Lachnospiraceae 13%	Ruminococcus 94%
Cluster115	16	5.80e-003	2.54e-004	-0.25	0.01218	unknown 100%	unknown 94%
Cluster116	1644	4.50e-001	5.75e-002	-0.25	0.01224	unclassified Lachnospiraceae 99%	unclassified Clostridiaceae 99%
Cluster117	14	3.29e-003	8.12e-004	-0.25	0.01236	unclassified Clostridiales 100%	unclassified Clostridiaceae 100%
Cluster118	1748	4.18e-001	1.47e-001	-0.25	0.01241	Sporobacter 99%	unclassified Clostridiales 99%
Cluster119	11	6.36e-003	0.00e+000	-0.25	0.01292	unknown 100%	unknown 100%
Cluster120	101	2.59e-002	9.83e-003	-0.25	0.01294	Bacteroides 99%	Bacteroides 99%
Cluster121	15	5.29e-003	1.18e-003	-0.25	0.01302	unclassified Lachnospiraceae 100%	Anaerotruncus 100%
Cluster122	35	9.92e-003	2.48e-003	-0.25	0.01316	Rikenella 100%	Alistipes 100%
Cluster123	15	4.59e-003	1.12e-003	-0.25	0.01331	Bacteroides 80% unknown 20%	Bacteroides 87% unknown 13%
Cluster124	35	8.08e-003	2.47e-003	-0.25	0.01348	unclassified Clostridiales 89% unclassified Lachnospiraceae 11%	Ruminococcus 89% unclassified Lachnospiraceae 11%
Cluster125	18	5.65e-003	0.00e+000	-0.24	0.01375	Ruminococcus 100%	unknown 100%
Cluster126	33	6.90e-003	1.73e-003	-0.24	0.01424	unknown 100%	unknown 100%
Cluster127	51	1.13e-002	1.21e-003	-0.24	0.01435	Ruminococcus 84% unknown 16%	unclassified Clostridiales 92%

Continued on next page

TABLE III – continued from previous page

Phylotype ID	# of Reads	Percentage (L)	Percentage (O)	PCC	p-value	RDP	greengenes
Cluster128	44	0.00e+000	9.33e-003	0.24	0.01440	unclassified Clostridiales 65% unclassified Firmicutes 17%	Faecalibacterium 95%
Cluster129	64	1.50e-002	4.15e-003	-0.24	0.01456	Bacteroides 100%	Bacteroides 100%
Cluster130	649	1.02e-001	3.55e-002	-0.24	0.01489	Erysipelothrix 57% unclassified Firmicutes 38%	Coprobacillus 95%
Cluster131	171	3.70e-002	1.84e-002	-0.24	0.01607	unknown 100%	unknown 100%
Cluster132	18	5.40e-003	1.25e-003	-0.24	0.01639	unclassified Bacteroidales 44% unclassified Porphyromonadaceae 56%	Tannerella 11% unclassified Porphyromonadaceae 83%
Cluster133	21	1.03e-002	1.59e-003	-0.24	0.01646	unclassified Clostridiales 90%	unclassified Clostridiaceae 90%
Cluster134	59	1.91e-002	5.85e-003	-0.24	0.01745	Eubacterium 78% unknown 22%	unclassified Clostridiales 78% unknown 22%
Cluster135	113	0.00e+000	3.13e-002	0.24	0.01754	unclassified Firmicutes 100%	Eubacterium 100%
Cluster136	229	4.58e-002	2.19e-002	-0.24	0.01797	Anaerofilum 98%	Subdoligranulum 99%
Cluster137	663	1.89e-001	8.04e-002	-0.23	0.01813	Bacteroides 97%	Bacteroides 96%
Cluster138	473	1.75e-002	6.36e-002	0.23	0.01844	Roseburia 100%	Roseburia 100%
Cluster139	162	3.23e-002	1.60e-002	-0.23	0.01853	Bacteroides 97%	Bacteroides 85% unknown 15%
Cluster140	238	2.11e-002	4.15e-002	0.23	0.01861	Anaerofilum 11% unclassified Clostridiales 85%	Faecalibacterium 99%
Cluster141	14	4.49e-003	7.57e-004	-0.23	0.01895	unknown 100%	unknown 100%
Cluster142	14	4.06e-003	7.83e-004	-0.23	0.01904	Anaerofilum 50% unknown 50%	Subdoligranulum 50% unknown 50%
Cluster143	127	3.38e-002	1.21e-002	-0.23	0.01939	Bacteroides 98%	Bacteroides 90% unknown 10%
Cluster144	124	5.00e-002	1.10e-004	-0.23	0.01951	unknown 94%	Tannerella 10% unknown 90%
Cluster145	46	1.77e-002	4.13e-003	-0.23	0.01954	Sporobacter 93%	unclassified Clostridiales 93%
Cluster146	22	8.17e-003	3.36e-004	-0.23	0.02037	unclassified Clostridiales 100%	unclassified Clostridiales 100%
Cluster147	12	2.29e-003	4.98e-004	-0.23	0.02058	Bacteroides 100%	Bacteroides 100%
Cluster148	67	1.29e-002	5.30e-003	-0.23	0.02077	unclassified Clostridiales 90%	Ruminococcus 90%
Cluster149	491	7.63e-002	3.61e-002	-0.23	0.02103	unclassified Lachnospiraceae 100%	unclassified Lachnospiraceae 99%
Cluster150	519	3.12e-002	1.08e-001	0.23	0.02104	Sporobacter 99%	unclassified Clostridiales 100%
Cluster151	17	3.69e-003	9.74e-004	-0.23	0.02143	Anaerofilum 100%	Subdoligranulum 100%
Cluster152	11	3.23e-003	5.56e-004	-0.23	0.02144	unknown 91%	unknown 82%
Cluster153	11	2.58e-003	4.99e-004	-0.23	0.02241	Bilophila 45% unknown 55%	unclassified Desulfovibrionaceae 45% unknown 55%
Cluster154	71	1.52e-002	5.23e-003	-0.23	0.02257	unknown 100%	unclassified Clostridia 97%
Cluster155	67	1.97e-002	2.33e-003	-0.23	0.02265	unclassified Clostridiales 12% unknown 81%	unknown 91%
Cluster156	13	5.34e-003	7.37e-004	-0.23	0.02275	unknown 100%	unknown 100%
Cluster157	89	4.64e-003	1.27e-002	0.23	0.02303	Dorea 75% unclassified Lachnospiraceae 20%	Dorea 73% Ruminococcus 24%
Cluster158	319	7.43e-002	3.24e-002	-0.23	0.02329	unknown 100%	unknown 100%
Cluster159	2089	1.24e-003	5.88e-001	0.23	0.02331	unclassified Firmicutes 100%	Eubacterium 100%
Cluster160	133	2.67e-002	1.14e-002	-0.23	0.02356	Sporobacter 99%	unclassified Clostridiales 98%
Cluster161	36	1.10e-002	3.10e-003	-0.23	0.02358	Sporobacter 97%	unclassified Clostridiales 97%
Cluster162	64	2.06e-002	4.59e-003	-0.22	0.02381	Sporobacter 100%	unclassified Clostridiales 100%
Cluster163	1056	3.69e-001	4.65e-002	-0.22	0.02429	Bacteroides 100%	Bacteroides 100%
Cluster164	62	0.00e+000	1.69e-002	0.22	0.02431	unclassified Firmicutes 85% unknown 15%	Eubacterium 90%
Cluster165	68	2.93e-003	1.13e-002	0.22	0.02433	unclassified Clostridiales 96%	unclassified Clostridiales 91%
Cluster166	37	8.03e-003	2.86e-003	-0.22	0.02478	unclassified Clostridiales 70% unknown 27%	Anaerostipes 46% Ruminococcus 24% unknown 27%
Cluster167	27	7.99e-003	1.57e-003	-0.22	0.02490	Bacteroides 100%	Bacteroides 96%
Cluster168	218	6.88e-002	1.47e-002	-0.22	0.02514	unknown 98%	unknown 98%
Cluster169	86	3.03e-002	8.76e-003	-0.22	0.02528	unclassified Clostridiales 97%	unclassified Clostridiales 85%
Cluster170	17	1.00e-002	1.42e-003	-0.22	0.02534	unknown 100%	unclassified Clostridiales 100%
Cluster171	96	2.39e-002	1.08e-002	-0.22	0.02550	Sporobacter 99%	Papillibacter 100%
Cluster172	13	2.97e-003	7.28e-004	-0.22	0.02552	Roseburia 31% unclassified Clostridiales 46% unknown 23%	Roseburia 31% unclassified Clostridiales 38% unknown 23%
Cluster173	44	1.15e-002	3.14e-003	-0.22	0.02558	Roseburia 52% unclassified Clostridiales 16% unclassified Lachnospiraceae 32%	Roseburia 32% unclassified Clostridiales 14% unclassified Lachnospiraceae 45%
Cluster174	17	5.98e-003	4.51e-004	-0.22	0.02659	unknown 100%	unclassified Alphaproteobacteria 12% unknown 88%
Cluster175	13	6.66e-003	0.00e+000	-0.22	0.02659	unclassified Clostridiales 92%	unclassified Clostridiales 46% unknown 54%
Cluster176	12	1.84e-003	3.70e-004	-0.22	0.02661	Anaerofilum 83% unknown 17%	Subdoligranulum 92%
Cluster177	20	4.40e-003	1.28e-004	-0.22	0.02704	Anaerofilum 80% unknown 20%	Subdoligranulum 90% unknown 10%
Cluster178	22	8.52e-003	1.16e-003	-0.22	0.02738	unclassified Clostridiales 100%	unclassified Clostridiales 100%
Cluster179	155	9.78e-003	2.36e-002	0.22	0.02794	unclassified Clostridiales 87%	Faecalibacterium 99%
Cluster180	101	4.07e-003	1.91e-002	0.22	0.02800	unclassified Clostridiales 24% unclassified Lachnospiraceae 70%	unclassified Clostridiales 83%
Cluster181	147	5.97e-003	2.05e-002	0.22	0.02843	unclassified Lachnospiraceae 82%	Roseburia 18% Ruminococcus 44% unclassified Clostridiales 17% unclassified Lachnospiraceae 10%
Cluster182	5996	8.72e-001	4.46e-001	-0.22	0.02854	unclassified Lachnospiraceae 100%	unclassified Clostridiaceae 20% unclassified Clostridiales 80%
Cluster183	93	5.52e-003	1.50e-002	0.22	0.02878	unclassified Clostridiales 12% unknown 87%	Eubacterium 51% unknown 47%
Cluster184	2222	5.84e-001	2.15e-001	-0.22	0.02900	Eubacterium 100%	unclassified Clostridiales 100%
Cluster185	17	5.31e-003	0.00e+000	-0.22	0.02947	unknown 100%	unclassified Clostridiales 100%
Cluster186	17	5.31e-003	0.00e+000	-0.22	0.02947	unclassified Clostridiales 100%	Acetivibrio 100%
Cluster187	299	1.58e-002	3.96e-002	0.22	0.02951	unclassified Clostridiales 87%	Roseburia 25% Ruminococcus 13% unclassified Lachnospiraceae 54%
Cluster188	15	2.83e-003	4.54e-004	-0.22	0.02956	unclassified Clostridiales 100%	Faecalibacterium 100%
Cluster189	57	3.22e-003	1.10e-002	0.22	0.02961	unclassified Clostridiales 98%	Papillibacter 98%
Cluster190	15	6.90e-003	1.19e-003	-0.22	0.03003	unclassified Firmicutes 87% unknown 13%	Granulicatella 47% unknown 47%
Cluster191	77	3.61e-003	1.28e-002	0.22	0.03009	unclassified Clostridiales 69% unknown 23%	Coprococcus 16% Roseburia 27% Ruminococcus 12% unclassified Lachnospiraceae 20% unknown 22%
Cluster192	72	3.38e-002	9.17e-004	-0.22	0.03035	unclassified Lachnospiraceae 88% unknown 12%	unclassified Clostridiaceae 88% unknown 12%
Cluster193	116	7.10e-003	1.95e-002	0.22	0.03061	unclassified Clostridiales 79% unknown 19%	Roseburia 38% unclassified Lachnospiraceae 38% unknown 18%
Cluster194	120	3.60e-002	1.15e-002	-0.22	0.03083	Eubacterium 95%	unclassified Clostridiales 95%
Cluster195	139	2.14e-002	9.99e-003	-0.21	0.03096	unclassified Lachnospiraceae 86% unknown 14%	unclassified Clostridiaceae 35% unclassified Clostridiales 53% unknown 12%
Cluster196	184	5.44e-002	0.00e+000	-0.21	0.03122	unclassified Mollicutes 100%	unknown 92%
Cluster197	38	7.21e-003	1.25e-003	-0.21	0.03139	unclassified Lachnospiraceae 87% unknown 13%	unclassified Clostridiaceae 82% unknown 13%

Continued on next page

TABLE III – continued from previous page

Phylotype ID	# of Reads	Percentage (L)	Percentage (O)	PCC	p-value	RDP	greengenes
Cluster198	71	1.97e-002	5.77e-003	-0.21	0.03140	Sporobacter 97%	unclassified Clostridiales 100%
Cluster199	12	4.94e-003	1.08e-003	-0.21	0.03152	Desulfonispora 100%	unclassified Peptococcaceae 100%
Cluster200	81	4.90e-002	1.90e-003	-0.21	0.03198	unclassified Lachnospiraceae 100%	Ruminococcus 73% unknown 26%
Cluster201	168	1.35e-002	2.94e-002	0.21	0.03202	unclassified Clostridiales 92%	Faecalibacterium 100%
Cluster202	13	3.27e-003	2.73e-004	-0.21	0.03205	unclassified Lachnospiraceae 85% unknown 15%	unclassified Clostridiaceae 85% unknown 15%
Cluster203	23	5.71e-003	1.91e-003	-0.21	0.03229	unclassified Bacteroidales 43% unclassified Porphyromonadaceae 52%	unclassified Porphyromonadaceae 87%
Cluster204	378	1.18e-001	3.07e-002	-0.21	0.03273	Sporobacter 97%	unclassified Clostridiales 99%
Cluster205	104	4.04e-002	5.54e-003	-0.21	0.03301	unclassified Lachnospiraceae 96%	unclassified Clostridiales 88%
Cluster206	27230	2.10e+000	3.78e+000	0.21	0.03304	unclassified Clostridiales 98%	Roseburia 44% unclassified Lachnospiraceae 55%
Cluster207	160	4.73e-002	4.83e-003	-0.21	0.03325	unknown 99%	unclassified Alphaproteobacteria 96%
Cluster208	15	5.62e-003	2.73e-004	-0.21	0.03366	unknown 100%	unknown 100%
Cluster209	26	7.36e-003	2.21e-003	-0.21	0.03383	Sporobacter 73% unknown 27%	unclassified Clostridiales 77% unknown 23%
Cluster210	289	2.12e-002	5.13e-002	0.21	0.03431	unclassified Clostridiales 99%	Ruminococcus 26% unclassified Clostridiales 22% unclassified Lachnospiraceae 44%
Cluster211	84	1.30e-002	4.03e-003	-0.21	0.03446	Erysipelothrix 57% unclassified Firmicutes 42%	Coprobacillus 99%
Cluster212	91	3.17e-003	1.75e-002	0.21	0.03450	Roseburia 16% unclassified Lachnospiraceae 76%	Roseburia 73% unclassified Clostridiales 20%
Cluster213	106	2.51e-002	8.76e-003	-0.21	0.03488	unclassified Clostridiales 100%	unclassified Clostridiales 100%
Cluster214	188	7.52e-003	2.63e-002	0.21	0.03495	Dorea 80%	Dorea 55% Roseburia 11% unclassified Clostridiales 22%
Cluster215	34	9.07e-003	1.97e-003	-0.21	0.03510	unclassified Clostridiales 100%	unknown 100%
Cluster216	20	4.99e-003	1.54e-003	-0.21	0.03520	Phascolarctobacterium 100%	Phascolarctobacterium 100%
Cluster217	111	5.05e-003	1.78e-002	0.21	0.03523	unknown 89%	unknown 88%
Cluster218	95	2.56e-002	6.54e-003	-0.21	0.03539	unclassified Clostridiales 100%	Anaerotruncus 100%
Cluster219	21	8.64e-003	1.70e-003	-0.21	0.03691	unclassified Clostridiales 14% unknown 86%	Anaerovorax 14% unknown 86%
Cluster220	18	3.93e-003	1.05e-003	-0.21	0.03695	Roseburia 11% unclassified Clostridiales 11% unclassified Lachnospiraceae 61% unknown 17%	Ruminococcus 78% unknown 17%
Cluster221	16	5.89e-003	9.47e-004	-0.21	0.03709	unknown 94%	unknown 94%
Cluster222	26	7.39e-003	2.49e-003	-0.21	0.03716	unknown 96%	unknown 92%
Cluster223	21	4.60e-003	1.09e-003	-0.21	0.03746	unclassified Lachnospiraceae 86%	Ruminococcus 95%
Cluster224	101	5.79e-002	7.23e-003	-0.21	0.03749	unknown 98%	Sutterella 97%
Cluster225	1200	2.03e-001	9.72e-002	-0.21	0.03805	unclassified Bacteroidales 89% unclassified Porphyromonadaceae 10%	Bacteroides 99%
Cluster226	16	5.37e-003	6.61e-004	-0.21	0.03825	unclassified Clostridiales 100%	unclassified Clostridiales 100%
Cluster227	98	1.90e-002	8.54e-003	-0.21	0.03829	Anaerofilum 22% unclassified Clostridiales 60% unknown 14%	Faecalibacterium 64% Subdoligranulum 23% unknown 12%
Cluster228	31	9.55e-003	1.47e-003	-0.21	0.03829	unclassified Clostridiales 23% unclassified Lachnospiraceae 74%	Ruminococcus 97%
Cluster229	57	1.96e-002	3.34e-003	-0.21	0.03851	Sporobacter 54% unknown 46%	unclassified Clostridiales 74% unknown 26%
Cluster230	42	0.00e+000	5.93e-003	0.21	0.03921	Megasphaera 79% unknown 21%	Megasphaera 88% unknown 12%
Cluster231	19	8.02e-003	0.00e+000	-0.21	0.03937	unknown 100%	unknown 100%
Cluster232	119	7.00e-003	1.83e-002	0.20	0.03986	unclassified Clostridiales 96%	Roseburia 77% unclassified Lachnospiraceae 21%
Cluster233	120	7.69e-003	2.15e-002	0.20	0.03995	unclassified Clostridiales 45% unclassified Lachnospiraceae 39% unknown 12%	Ruminococcus 86%
Cluster234	27	1.30e-002	7.61e-004	-0.20	0.04068	Bacteroides 100%	Bacteroides 100%
Cluster235	52	1.99e-002	2.51e-003	-0.20	0.04076	unclassified Clostridiales 81% unclassified Lachnospiraceae 15%	unclassified Clostridiales 83%
Cluster236	52	1.40e-002	5.95e-003	-0.20	0.04102	Bacteroides 100%	Bacteroides 100%
Cluster237	73	2.80e-003	9.85e-003	0.20	0.04116	unclassified Clostridiales 77% unclassified Lachnospiraceae 15%	Anaerostipes 96%
Cluster238	335	2.08e-002	4.69e-002	0.20	0.04118	unclassified Clostridiales 48% unclassified Lachnospiraceae 51%	Roseburia 27% unclassified Lachnospiraceae 67%
Cluster239	52	5.04e-004	9.19e-003	0.20	0.04124	Acidaminococcus 88% unknown 12%	Acidaminococcus 90%
Cluster240	72	2.24e-002	7.02e-003	-0.20	0.04135	unclassified Clostridiales 96%	unclassified Clostridiales 96%
Cluster241	297	4.97e-002	2.54e-002	-0.20	0.04135	Bilophila 100%	unclassified Desulfovibrionaceae 100%
Cluster242	19	6.22e-003	1.35e-003	-0.20	0.04153	Lachnospira 47% unclassified Clostridiales 11% unclassified Lachnospiraceae 16% unknown 26%	Lachnospira 68% unknown 26%
Cluster243	186	1.08e-002	3.10e-002	0.20	0.04372	unclassified Clostridiales 56% unclassified Lachnospiraceae 41%	Ruminococcus 91%
Cluster244	35	1.80e-002	1.52e-003	-0.20	0.04386	unknown 97%	unknown 97%
Cluster245	20	5.61e-003	1.36e-003	-0.20	0.04416	unclassified Clostridiaceae 95%	Clostridium 95%
Cluster246	16	4.03e-003	1.26e-003	-0.20	0.04501	unknown 100%	unknown 100%
Cluster247	17	4.61e-003	1.05e-003	-0.20	0.04502	Bacteroides 88% unknown 12%	Bacteroides 83% unknown 17%
Cluster248	29	5.79e-003	1.57e-003	-0.20	0.04554	Rikenella 100%	Alistipes 93%
Cluster249	108	3.11e-002	1.28e-002	-0.20	0.04576	unclassified Clostridiales 98%	unclassified Clostridiales 99%
Cluster250	97	2.03e-002	8.13e-003	-0.20	0.04582	unclassified Clostridiales 88% unknown 12%	Papillibacter 99%
Cluster251	311	8.75e-002	2.39e-003	-0.20	0.04608	unclassified Firmicutes 100%	Acetivibrio 100%
Cluster252	51	2.63e-002	1.49e-003	-0.20	0.04644	unclassified Clostridiales 16% unclassified Lachnospiraceae 71% unknown 12%	unclassified Clostridiales 78% unknown 16%
Cluster253	175	6.87e-002	1.82e-002	-0.20	0.04686	unclassified Clostridiales 98%	unclassified Clostridiaceae 98%
Cluster254	18	3.78e-003	1.34e-003	-0.20	0.04703	Roseburia 17% unclassified Clostridiales 11% unknown 72%	unclassified Firmicutes 20% unknown 80%
Cluster255	47	1.92e-002	2.01e-003	-0.20	0.04759	Bacteroides 98%	unknown 100%
Cluster256	99	3.17e-002	1.84e-003	-0.20	0.04776	unclassified Clostridiales 100%	unclassified Clostridiales 100%
Cluster257	328	1.16e-001	3.71e-002	-0.20	0.04820	unclassified Clostridiales 100%	Clostridium 100%
Cluster258	947	9.01e-002	1.60e-001	0.20	0.04848	unclassified Clostridiales 81% unclassified Firmicutes 13%	Faecalibacterium 99%
Cluster259	96	1.59e-002	8.54e-003	-0.20	0.04860	unknown 97%	Alistipes 12% unknown 88%
Cluster260	74	4.73e-003	1.44e-002	0.20	0.04884	Sporobacter 14% unclassified Clostridiales 65% unknown 22%	unclassified Clostridiales 85% unknown 15%
Cluster261	13	6.45e-003	6.40e-004	-0.20	0.04890	unknown 92%	unknown 100%
Cluster262	15	3.52e-003	1.08e-003	-0.20	0.04919	Sporobacter 79% unclassified Clostridiales 14%	Papillibacter 93%
Cluster263	20	4.91e-003	1.66e-003	-0.20	0.04926	Roseburia 75% unknown 20%	Roseburia 75% unknown 20%
Cluster264	272	2.10e-002	4.37e-002	0.20	0.04954	unclassified Clostridiales 90%	Roseburia 12% unclassified Lachnospiraceae 75%
Cluster265	24	2.52e-004	5.14e-003	0.20	0.04971	Roseburia 25% unclassified Clostridiales 12%	Roseburia 35% unknown 39%

Continued on next page

TABLE III – continued from previous page

Phylotype ID	# of Reads	Percentage (L)	Percentage (O)	PCC	p-value	RDP	greengenes
						unclassified Lachnospiraceae 33% unknown 17%	
Cluster266	22846	2.20e+000	3.44e+000	0.20	0.04991	unclassified Clostridiales 95%	Faecalibacterium 100%