TITLE

Exploring thematic structure in 16S rRNA marker gene surveys

AUTHORS

Stephen Woloszynek (sw424@drexel.edu) [1] [corresponding author]

Joshua Chang Mell (joshua.mell@drexelmed.edu) [2]

Gideon Simpson (simpson@math.drexel.edu) [3]

Gail L Rosen (gailr@coe.drexel.edu) [1]

AFFILIATIONS

[1] Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, United States of America.

[2] Department of Microbiology and Immunology, Drexel University College of Medicine, Philadelphia, PA, United States of America.

[3] Department of Mathematics, Drexel University, Philadelphia, PA, United States of America

KEYWORDS

American Gut, Bayesian, Crohn's Disease, Diet, Inflammatory Bowel Disease, KEGG, Metagenomics, Microbiome, PICRUSt, Topic Model

ABSTRACT

Background: Analysis of microbiome data involves identifying co-occurring taxa associated with a specific set of sample attributes (e.g., disease presence) but is often hindered by the data being compositional, high dimensional, and sparse. Also, the configuration of co-occurring taxa may represent overlapping subcommunities that contribute to host status. Preserving the configuration of co-occurring microbes is superior to detecting indicator species since this approach is more likely to represent underlying microbiome mechanisms and thus facilitate more biologically meaningful interpretations. Moreover, analysis which simultaneously utilizes taxonomic and functional abundances typically requires independent characterization of taxonomic and functional profiles before linking them to sample information. However, this limits investigators from identifying which specific functional components associate with which subsets of co-occurring taxa.

Results: We provide a pipeline to explore co-occurring taxa using topics generated via a topic model approach and then link these topics to specific sample classes. Also, rather than inferring predicted functional content independently from taxonomic information, we instead focus on within-topic functional content, which we parse via estimating pathway-topic interactions through a multilevel fully Bayesian regression model. We apply our methods to two large 16S amplicon sequencing datasets: an inflammatory bowel disease (IBD) dataset from Gevers et al. and data from the American Gut (AG) project. When applied to the Gevers et al. IBD study, we determine that a topic highly associated with Crohn's disease (CD) diagnosis is (1) dominated by a cluster of bacteria known to be linked with CD and (2) uniquely enriched for a subset of lipopolysaccharide (LPS) synthesis genes. In the AG data, our approach found that individuals with plant-based diets were enriched with Lachnospiraceae, *Roseburia*, *Blautia*, and *Ruminococcaceae*, as well as fluorobenzoate degradation pathways, whereas pathways involved in LPS biosynthesis were depleted.

Conclusions: We therefore introduce an approach for uncovering latent thematic structure in the context of host state for 16S rRNA surveys. Using our topic-model approach, investigators can (1) capture sets of co-occurring taxa, (2) uncover their functional potential, and (3) identify gene sets that may help guide future inquiry. These methods have been implemented in a freely available R package https://github.com/EESI/themetagenomics.

LIST OF ABBREVIATIONS

AG, American gut

CD, Crohn's disease

CV, cross validation

IBD, inflammatory bowel disease

LDA, latent Dirichlet allocation

LFC, log-fold change

LPS, lipopolysaccharide

OTU, operational taxonomic unit

PPD, posterior predictive distribution

RF, random forest

STM, structural topic model

BACKGROUND

With the decreasing cost of high-throughput sequencing, large datasets are becoming increasingly available, particularly microbiome datasets rich in sample data. These data consist of categorical and numeric information associated with each sample, which in turn are linked to a set of taxonomic abundances that are derived from clustering sequencing reads. Such clusters are based on taxonomic marker genes – typically a portion of the 16S rRNA gene that meet a fixed degree of sequence similarity, termed Operational Taxonomic Units (OTUs). Analysis of these data often involves identifying co-occurring groups of taxa associated with specific sample features via unsupervised exploratory methods such as principal component analysis, correspondence analysis, multidimensional scaling, and hierarchical clustering, in addition to statistical inference strategies aimed at identifying differentially abundant taxa and differences in alpha and beta diversity. Nevertheless, model building is hindered by the complexity inherent to these data, which have a disproportionate number of samples relative to features (Knights et al., 2011), a substantial degree of sparsity, and are typically strictly positive and constrained to sum to 1, i.e., compositional (Gilbert et al., 2016; Li, 2015).

From an ecological perspective, the configuration of these co-occurring microbiota may represent related, overlapping sets of subcommunities consisting of taxa that correlate with, for example, host phenotype. Identifying subcommunities that contribute to host status as opposed to single indicator species facilitates a more biologically meaningful interpretation by preserving the natural configuration of co-occurring bacteria when making inferences with respect to host phenotype (Shafiei et al., 2015). Recent work has attempted to explore such relationships (Jiang et al., 2012; Ning & Beiko, 2015; Ren et al., 2016; Shafiei et al., 2015).

Still, suitable approaches to uncover these relationships in the context of functional information is deficient – that is, few methods successfully integrate subcommunity-host phenotype with functional profiles specific to these subcommunities. In both metagenomic and 16S rRNA surveys, analyses utilizing taxonomic and functional abundance information typically involve independently characterizing the taxonomic and functional profiles of the samples and subsequently associating these profiles with host information. In the former, this is done directly by analyzing taxonomic and functional sequence information, whereas the latter requires predicting functional profiles from 16S rRNA survey information using methods such as PICRUSt, Tax4fun, and Piphillin (Aßhauer et al., 2015; Iwai et al., 2016; Langille et al., 2013). In either approach, despite obtaining information regarding which taxa co-occur and whether specific taxa or functional categories are associated with sample data, the investigator remains limited from identifying which specific functional components associate with which subsets of co-occurring taxa.

In the context of 16S rRNA gene sequencing data, our objective is therefore two-fold: to implement a modeling framework that can (1) capture sets of co-occurring taxa associated with specific sample data and (2) uncover the functional potential that further characterizes the configuration of these subcommunities.

For our first objective, we will employ a topic model approach. Topic models have had considerable use in natural language processing, but have also been explored as a method for exploring genomic count data. Knights et al. (Knights et al., 2013) utilized latent Dirichlet allocation (LDA) to infer the relative contributions of an unknown number of source environments to a set of indoor samples. Shafiei et al. (Shafiei et al., 2015), alternatively, took a supervised approach where they first trained their model on sets of co-occurring OTUs to learn how they correlate with sample classes of interest. They were then able to predict the class of new samples given the trained model.

Our approach utilizes a structural topic model (STM) (Roberts et al., 2014), which generalizes previously described topic models such as LDA, the correlated topic model (Blei & Lafferty, 2007), and Dirichlet-Multinomial regression topic model (Mimno & McCallum, 2012). Like the Dirichlet-Multinomial regression topic model, the STM permits the influence of sample data on the distribution of samples-over-topics; LDA, on the other hand, can only incorporate sample information if done so in a two-stage process – first performing topic extraction, and then identifying linear relationships between the topic assignments and sample information (Blei et al., 2003; Roberts et al., 2014). A two-stage approach limits the breadth of sample information one can use, typically forcing the user to use only a single vector of covariate information, and moreover prevents propagating uncertainty throughout the model. Similar to the correlated topic model, the STM's logistic normal distribution defines the prevalence of topics across samples and permits correlation between topics.

With the STM, we will uncover a thematic representation of 16S rRNA survey abundance data and jointly measure its relationship with sample information (figure 1.1). Two latent distributions will be estimated: a samples-over-topics and a topics-over-OTU distribution, which represent the probability of a topic occurring in a sample and the probability of an OTU occurring in a topic, respectively (figure 1.2). By utilizing sample information (figure 1.3), we will then be able to determine whether particular sample covariates increase or decrease the probability of a given topic occurring in a set of samples (figure 1.4).

Our second objective will exploit the estimated topics-over-OTUs distribution. These posterior probabilities dictate the taxonomic composition of the topics and therefore should capture meaningful co-occurrences. Moreover, these probabilities resemble relative abundances of samples across taxa. We can therefore infer the functional potential of these topics using tools such as PICRUSt, allowing us to predict topic-specific gene composition, using a database of reference genomes (figure 1.5). Then, by identifying topics of interest based on their relationship to sample covariates, we can subsequently link this predicted within-topic functional profile to both within-topic taxonomic abundances, as well as the specific samples that have high probability of containing these topics.

It should be noted how this approach differs from the naïve approach where taxonomic and functional profiles are independently estimated and then jointly interpreted. A naïve approach will successfully identify taxonomic abundances that associate with covariate information, and the same for (predicted) functional abundances, but the result lacks the ability to infer which sets of functions are directly linked to specific sets of taxa. The ability to uncover such information provides context as to why specific co-occurrences are present.

We apply our methods on two large 16S rRNA amplicon sequencing datasets: an inflammatory bowel disease (IBD) dataset from Gevers et al. (Gevers et al., 2014) and data from the American Gut (AG) project. After confirming the generalizability of extracted topics, we identified distinct taxonomic subcommunities that, in the case of the Gevers dataset, were consistent with published results. These subcommunities were in turn composed of distinct predicted functional profiles, and moreover, our approach provided gene-sets specific to topics of interest that may warrant further exploration. In a companion paper, we performed simulations to further validate a topic model approach for 16S survey data and to determine a suitable normalization strategy (Woloszynek et al., 2017). Our simulations suggested that predefined taxonomic subcommunities concentrate with high probability to extracted topics and that no library size normalization is required to maximize power or ability to infer taxonomic structure, thus making a topic model approach a more direct, suitable procedure for inferring the subcommunity configuration. Also, in the context of topic models, while DESeq2 normalization outperforms rarefying, it results in decreased power compared to simply using raw, unnormalized abundances.

These methods have been implemented in a freely available R package themetagenomics: https://github.com/EESI/themetagenomics.

METHODS

Review of the Structural Topic Model

The STM is a Bayesian generative model such that, given a set of M samples, each consisting N OTUs, belonging to a vocabulary of V unique OTU terms, K latent topics (chosen *a priori*) are generated from the data. These topics consist of overlapping sets of co-occurring OTUs, potentially sharing some biological context. The samples-over-topics distribution is given a logistic Normal (LN) prior, which allows for estimation of topic-topic correlations, giving a means to infer co-occurring topics across samples. The topics-over-OTUs prior, on the other hand, estimates the deviation of OTU frequencies from a background distribution that encompasses all samples in the dataset (Eisenstein et al., 2011). Word and topic assignments are both generated via V- and K-multinomial distributions, respectively.

The STM is estimated by a semi-collapsed variational expectation maximization procedure. Convergence is reached when a relative change in the variational objective (i.e., the estimated lower bound) falls below a predetermined tolerance.

Datasets and Preprocessing

16S rRNA sequencing data from two human microbiome studies were downloaded from their corresponding repositories. The Gevers et al. dataset ("Gevers") (PRJNA237362, 03/30/2016) is a multicohort, IBD dataset that includes control, Crohn's disease (CD), and ulcerative colitis samples taken from multiple locations throughout the gastrointestinal tract (Gevers et al., 2014). The AG project ("AG") (ERP012803, 02/21/2017), on the other hand, is a crowd sourced dataset that includes user-submitted microbiome samples from a variety of body sites and associated subject information provided through questionnaires (http://americangut.org/).

Human gut microbiota from an inflammatory bowel disease cohort (Gevers). Paired-end reads were joined and quality filtered (maximum unacceptable Phred quality score = 32; maximum number of consecutive low quality base calls before read truncation = 3; minimum number of consecutive high quality base calls included per read as a fraction of input read length = 0.75) using QIIME version 1.9.1. Closed-reference OTU picking was performed using SortMeRNA against GreenGenes v13.5 at 97% sequence identity. This was followed by copy number normalization via PICRUSt version 1.0.0 (Kembel et al., 2012).

We selected only terminal ileum samples. Those with fewer than 1000 total reads were omitted. We subsequently removed OTUs with fewer than 10 total reads across samples and OTUs that lacked a known classification at the Phylum level.

Human gut microbiota from samples differing in terms of diet (AG). Quality trimming and filtering were performed in the following manner on single-end reads using the fastqFilter command found in the dada2 R package. The first 10 bases were trimmed from each read. Reads were then trimmed to position 135 based on visualizing the quality score of sampled reads as a function of base position. Further truncation occurred at positions with quality scores less than or equal to 2. Any truncated read with total expected errors greater than 2 were removed. A portion of AG samples were affected by bacterial blooming during shipment. These sequences were removed using the protocol provided in the AG documentation (02-filter_sequences_for_blooms.md).

OTU picking and copy number normalization were implemented as above. Samples with fewer than 1000 reads, and OTUs with fewer than 10 total reads across samples and lacking any known classification at the Phylum level were discarded. We filtered samples falling into the "baby" age category (thus the minimum age was 3) and retained only fecal samples. Within the diet category, unknown, vegetarian-with-shellfish, and omnivore-without-red-meat diets types were removed. We then merged vegan and vegetarian-without-shellfish into one class, resulting in a binary set of labels: "O" for omnivores and "V" for vegans and vegetarians.

Structural Topic Model Fitting (figure 1.1)

Each resulting OTU table consists of sets of raw counts normalized by 16S rRNA copy number. No other normalization was conducted based on the simulation results in Woloszynek et al. (2017). A series of topic models with different parameterizations in terms of topic number ($K \in$ 15, 25, 50, 75, 100, 150, 250) and sample covariates (e.g., indicators for presence of disease, diet type, etc.) were fit to the OTU tables.

We evaluated each model fit for presence of overdispersed residuals. We also conducted permutation tests where the covariate of interest is randomly assigned to a sample, prior to STM fitting. To compare parameterizations between models, we evaluated predictive performance using held-out likelihood estimation (Blei et al., 2003).

Assessment of topic generalizability

We performed classification to assess the generalizability of the extracted topics. No sample information was used as covariates. OTU tables were first split into 80/20 training-testing datasets. A topic model was trained to estimate the topics-over-OTUs distribution. We then held this distribution fixed; hence, only the testing set's samples-over-topics distribution was estimated. For both the training and testing sets, simulated posterior samples from the samples-over-topics distribution were averaged. The resulting posterior topic probabilities in the training set were then used as predictors to classify sample labels, similar to using $\bar{Z}$ in supervised LDA (Blei et al., 2008). The generalization error was then assessed by using the optimal parametrization based on cross validation performance (CV) on the test set topic probabilities. Classification was performed using a random forest (RF).

For the RF, parameter tuning to determine the number of variables for each split was accomplished through repeated (10x) 10-fold CV, using up- or down-sampling to overcome class unbalance (for Gevers and AG, respectively). We performed a parameter sweep over the number of randomly selected features, while setting the number of trees fixed at 128. The optimal parameterizations were selected based on maximizing ROC area under the curve (AUC).

Assessing Concentration of OTUs as a function of topic number

Comparison of Shannon entropy across topics was performed via ANOVA and Tukey HSD post-hoc analysis. To quantify the relationship between taxonomic abundance and continuous predictors (e.g., PCDAI), we performed negative binomial regression (log link), using total sample coverage as an offset. The family-wise error rate was adjusted via Bonferroni correction. Critical values for hypothesis testing were set at 0.05 unless stated otherwise.

Comparison of topic taxonomic profile to a network approach

To further validate the clusters of high probability taxa identified in the topics-over-OTUs distribution, we compared our results to those generated from an OTU-OTU association network on the raw (copy number normalized) OTU tables using SPIEC-EASI's neighborhood selection method (Kurtz et al., 2015).

Inferring within-topic functional potential (figure 1.5)

We obtained the topics-over-OTUs distribution for each model fit and mapped the within-topic OTU probabilities to integers ("pseudo-counts") using a constant: $10000 \times \beta$. A large constant was used to prevent low probability OTUs from being set to zero, although their contribution to downstream analysis was likely negligible. Gene prediction was then performed on each topic-OTU pseudo-count table using PICRUSt version 1.0.0 (Langille et al., 2013). Recall that copy number normalization was performed prior to topic model fitting.

Identifying topics of interest (figure 1.3, 1.4)

Topics of interest were identified by regressing the sample-specific topic probabilities against their set of sample covariates. We calculated 95% uncertainty intervals using an approximation that accounts for uncertainty in estimation of both the coefficients and the topic probabilities.

Identifying predicted functions that distinguish topics (figure 1.6)

To determine which predicted gene functions best distinguish topics, we utilized the following multilevel negative binomial regression model:

$$\theta_{k,c} = \exp\left[\mu + \beta_k + \beta_c + \beta_{k,c}\right]$$

$$y_{k,c} \sim \mathrm{NB}(\theta_{k,c}, \lambda)$$

where μ is the intercept, β_k is the per topic weight, β_c is the per level-3 gene category weight, $\beta_{k,c}$ is the weight for a given topic-gene category combination, $y_{k,c}$ is the count for a given topic-gene category combination, and λ is the dispersion parameter. All weights were given normal priors. Convergence was assessed across 4 chains using diagnostic plots to assess mixing and by evaluating the Gelman-Rubin convergence diagnostic (Gelman & Rubin, 1992). To reduce model size, we used genes belonging to only 15 (arbitrary number) level-2 KEGG pathway categories (table S1). For large topic models, we fit only the top 25 topics, ranked in terms of the regression weights that measure the degree of association between sample-over-topic probabilities and our covariate of interest.

Comparison of within-topic pathway profile to OTU-table approach

We compared the profile of predicted functions obtained from the hierarchical negative binomial model to a differential abundance approach. We performed (KEGG) functional prediction via PICRUSt on raw OTU abundances that were copy number normalized. The resulting functional abundances were collapsed into level-3 KEGG pathways. Note that we again restricted our genes to the 15 level-2 KEGG pathways used previously to remain consistent. The resulting level-3 pathway abundances underwent DESeq2 differential abundance analysis followed by Bonferroni correction (McMurdie & Holmes, 2014). Adjusted p-values below 0.1 were deemed significant.

Packages utilized

All analysis was done in R version 3.2.3. Topic models, RFs, and NB regression models were fit using stm (Roberts, Margaret E., Stewart & Tingley, 2017), caret (Kuhn, 2008), and rstanarm (Stan Development Team, 2016), respectively. AG filtering was performed using dada2 (Callahan et al., 2015). SPIEC-EASI was fit using the SPIEC-EASI package (Kurtz et al., 2016). DESeq2 differential abundance analysis was conducted with phyloseq (McMurdie & Holmes, 2013).

RESULTS

We will explore the use of a topic model approach on datasets of gut and fecal microbial community profiles, beginning with the IBD data from Gevers, followed by the dietary data from AG. For each dataset, we show that the topics extracted from the STM generalize well to test set data not initially seen by the model, suggesting that co-occurrence profiles identified by the STM are robust to overfitting. Then, we apply our complete pipeline, where we successfully link within-topic predicted functional profiles to taxonomic subcommunity configurations and host features.

Thematic Structure of IBD-Associated Microbiota (Gevers)

Dimensionality reduction using topics facilitates classification of CD diagnosis and generalizes well to test data. We aimed to assess whether (1) topics fit in the absence of sample covariates are associated with positive CD diagnosis (CD+), and (2) they generalize to new data – that is, whether they captured meaningful information inherent to the data while ignoring characteristics associated exclusively with the fitted data.

The 80/20 training/testing splits for terminal ilium samples from Gevers are shown in table S2. We hypothesized that there would be a drop in performance using OTU relative abundances as features compared to topics, since the former has much higher dimensionality and is sparser. These are both relaxed when using topics, since the size of the feature space is decreased through dimensionality reduction. There was little difference between the two approaches during training CV with at least 25 topics (figure S1, table S3). During testing, however, topics outperformed OTU relative abundances, particularly in terms of F1 score, with scores of 0.808 and 0.857 for OTUs and topics (K=25 and K=100), respectively (table S4).

As one example, the largest discrepancy in classification performance between OTUs and topics was in terms of their negative predictive value, with the OTU model being correct only half the time (0.517) when predicting the negative class (CD-), whereas the worst performing topic model (K=15) performed slightly better (0.526), and topic models seemingly improved as the number of topics increased: 0.655 (K=25), 0.559 (50), 0.577 (75), 0.682 (100), and 0.643 (150) (table S4). Such a high proportion of false negatives with the OTU model was likely due to its reliance on few, relatively rare taxa (figure S2).

As another example, OTU 319708 (Clostridiaceae family) was the fourth most important feature for distinguishing classes. It was over twice as common in CD- training samples. Over 10% of correctly classified CD- samples contained this feature. This was also the case for 10% of misclassified CD+ samples, some of which contained this OTU at a greater proportion than other samples in the training set. A similar scenario can be seen for the OTU with the largest importance score, OTU 186723 (Ruminococcaceae family), which associated predominately with disease presence, and hence its absence in CD+ samples resulted in false negatives.

Concentration of high probability OTUs across topics begins to plateau at 75 topics. After assessing the generalizability of extracted topics, we implemented our full pipeline using sample covariates, specifically a binary indicator for IBD diagnosis. After fitting the topic model to the OTU abundance data, we aimed to uncover how OTUs concentrate within topics as a function of topic number. We performed an ANOVA to compare Shannon entropy for individual topics across OTUs for topic models of varying sizes, followed by post-hoc multiple comparisons testing using Tukey HSD (α=0.05) (figure S2). We found a significant difference in the mean Shannon entropy among the models considered. When we tested for differences between pairwise model combinations, we found that the drop in entropy with increasing topic number diminished, such that differences between models with 75, 100, and 150 topics were not significantly different from one another. This suggests that the probability mass of the topics-over-OTUs distribution concentrates on OTU subsets of similar sizes as topic number increases. Analyzing topics in this way may help guide the user in the selection of topic number.

CD diagnosis was associated with unique thematic and hence taxonomic profiles. The configuration of topics K25 and K75 are shown in figures S3 and S4, exemplifying how our pipeline represents 16S rRNA abundance data. For the topics shown, their posterior estimates did not span 0, a result that was also present when we performed permutation tests to confirm (figure S5). The panels are ordered in terms of mean effect estimate using the samples-over-topic distribution against sample diagnosis. We consider topics with larger mean effect estimates as “high-ranking topics.” Both panels show that CD- training samples had a topic distribution that differed from CD+ samples. Moreover, a given topic’s association with disease presence in most influenced by the disease burden of its samples, particularly for K25, where CD+ samples with high probability for T19, T13, and T26 (the topics most associated with CD-) tend to have minimal disease burden. We henceforth focus on the K25 model.

Focusing on these eight key topics, we identified multiple clusters of bacterial species that disproportionately dominated the top topics associated with CD+ (figure 2, top). For example, T2 contained a cluster dominated by *Enterobacteriaceae* taxa, whereas T12’s cluster contained a mixture of *Fusobacteria* and *Enterobacteriaceae*. The T15 cluster contained *Haemophilus* spp., *Neisseria*, *Fusobacteria*, and *Streptococcus*, all of which were noted as having a positive correlation with CD+ subjects in Gevers et al, as well as *Aggregatibacter*, a genus reportedly associated with colorectal cancer (Tjalsma et al., 2012).

Given that T15 contains a cluster of bacteria known for their association with bowel inflammation and this topic occurs disproportionately in subjects with greater disease burden, we asked whether the abundance of these OTUs in CD+ subjects correlated with PCDAI, a clinical measure of CD burden. After performing negative binomial regression (figure 3), we identified significant positive trends as a function of PCDAI for *Aggregatibacter* ($p<0.0001$), *Erwinia* (p=0.0004), *Fusobacterium* (p=0.0001), and *Haemophilus* (p=0.0484).

The topics most associated with CD-, on the other hand, were dominated by taxa belonging to *Lachnospiraceae*, *Roseburia*, *Rubinococcus*, *Blautia*, *Bacteroidetes*, and *Coprococcus*, all of which were noted by Gevers et al. as being negativity associated with CD (figure 2, bottom; figure S6). In addition to these taxa, *Akkermania*, *Dialister*, and *Dorea* contributed to these topics, which is consistent with the findings of Lewis et al. who found a reduction of these taxa in CD+ subjects (Lewis et al., 2015).

Within-topic co-occurrence profiles were confirmed via SPIEC-EASI. We compared the resulting topics to the correlations obtained via a network approach. The SPIEC-EASI network edges for the clusters of high probability OTUs in our most correlated topics are showed in figure S7. For each of these topic clusters, the majority of taxa were connected by a non-zero edge (table S5). Of the 11 taxa in the T15 cluster, 8 had first order connections (direct connections to other taxa within the cluster, OTU_c-$OTU_{c'}$), whereas 9 had second order connections (indirect connections to other taxa within the cluster via an intermediate OTU not present in the cluster, OTU_c-OTU_{nc}-$OTU_{c'}$). Moreover, the two OTUs connected by largest edge weight, *H. parainfluenzae* and *Haemophilus spp.*, had the largest probabilities of the taxa in the topic cluster, 0.320 and 0.245, respectively. Of these 6 topics, none had more than one OTU with zero connections or fewer than 75% of taxa joined by first order connections. Predictably, the taxa that lacked within-cluster connections received low probability from the topic model, with one exception, *Catenibacterium spp.* in T19. Taken together, this reaffirms that the within-topic co-occurrence profiles are consistent with alternative approaches.

Predicted functional potential of notable topics further described their association with CD. We sought to further explore the co-occurrence profiles of these topics, thereby exploiting the posterior estimates of the topic model in a way unique compared to other approaches. To do so, we predicted the topic-specific functional content using PICRUSt and then performed a fully Bayesian multilevel regression analysis on the abundances of each gene function.

Like Gevers et al., we identified an increase in membrane transport associated with CD+, particularly topics T2 and T12; however, through our approach, we were able to pinpoint the specific topics these functional categories associated with. This, in turn, allowed us to link these categories to specific taxa. For example, the two aforementioned topics were dominated by Enterobacteriaceae (figure S8). Topic T15, on the other hand, contained the cluster of *Haemophilus* spp., *Neisseria*, and *Fusobacteria* taxa, and despite being most associated with CD+, had a less substantial shift in membrane transport genes, suggesting that this pathogenic cluster contributed less to the shift of those genes.

A considerable degree of cell motility genes was found in T19 relative to all other topics, which is consistent with this topic being dominated by mobile bacteria that belong to Lachnospiraceae, Roseburia, and Clostridiales. More specifically, this topic was enriched for genes belonging to the following KEGG categories: bacterial motility proteins, bacterial chemotaxis, and flagellar assembly (figure 4). The aforementioned Enterobacteriaceae-enriched topics were also enriched for siderophore and secretion system related genes. Enrichment of two lipopolysaccharide (LPS) synthesis categories were associated with CD+ topics; however, one of these categories was specific for T15 (table S7).

Considerably more pathways were deemed significant via a DESeq2 approach on the OTU abundance table, hindering interpretability. We compared our within-topic functional profiles to the profiles obtained by performing PICRUSt on the copy-number normalized OTU abundance table and then performing a DESeq2 differential abundance analysis. Of the 160 level-3 KEGG categories, 87 were found significant ($\alpha < .1$) (figure S9) in the DESeq2 approach. Pathways with the largest log-fold change (LFC) associated with CD+ samples included degradation pathways (caprolactam, LFC=0.542; fluorobenzoate, 0.532; geraniol, 0.371; and toluene degradation, 0.371), alphalinolenic acid metabolism (0.641), and electron transfer carriers (0.635). Interestingly, these degradation pathways also demonstrated strong effects between topics; however, they associated with T1, a topic unrelated to disease status. Electron transfer carriers was identified in both approaches, but the topic model approach isolated T12, placing high probability on bacteria also enriched for functions linked to secretion systems, LPS biosynthesis, and motility. The DESeq2 approach also found fewer categories associated with CD- that had large LFC. For example, only 1 category had an LFC less than -0.04, whereas there were 8 greater than 0.04. The categories with the largest LFCs relative to CD- included germination (LFC=-0.450) and sporulation (-0.346). The topic model identified 10 topics with functional profiles significantly enriched or depleted in sporulation genes, three of which were associated with CD- samples. Moreover, multiple topics demonstrated an inverse relationship between sporulation and LPS genes, such that topics that contained taxa enriched in one were depleted in the other.

Thematic Structure in Terms of Diet (AG)

Despite consisting of far more samples, the AG dataset, split into O and V diet groups from self-reported dietary information, offered a new challenge for our approach, given that there were far more data and features (taxa), as well as severe imbalance between classes. Of the 4864 samples that fit into our diet classes, 4527 and only 337 were O and V samples, respectively. This renders comparisons between group means a worse estimate of treatments effects (Gelman & Hill, 2006).

Accounting for potential confounding. Before applying our pipeline, we aimed to eliminate any potential sources of confounding. Male and female samples were distributed similarly with respect to diet (table S9; figure S10). There was no significant difference in mean age between diet groups (t=-0.03, df=373.93, p=0.98). Sample body mass index was not normally distributed (Shapiro-Wilk: W=0.86, p<0.001) and was plagued with many mislabeled heights and weights (figure S11). After attempting to remove samples we deemed unreliable, we found a significant mean difference in body mass index between diet groups via a Mann Whitney U test (p<0.001).

Classification using topics is less conservative and, for low dimensional models, less generalizable. Unlike Gevers, models with fewer topics (K < 75) generalized poorly compared to using OTUs as features, which may be due to AG having nearly 3-times as many unique OTUs, causing too few topics to dampen any meaningful signal (table S11). Interestingly, all parameterizations outperformed the raw data in terms of sensitivity but not specificity (table S11), suggesting that classification using OTU features is more conservative.

Diet was associated with specific taxonomic and predicted functional profiles. We will henceforth report the results from a 100 topic model fit with dietary prior information. As before, we identified our topics of interest by regressing the samples-over-topics distribution against diet and further validated these results via permutation tests, resulting in 9 topics, 5 of which were associated with the O group, and 4 with the V group (figure S13).

Across the 9 topics, members of the family Lachnospiraceae were well represented, which is not surprising given that it typically accounts for over half of bacteria in healthy human fecal samples (Flint, 2012). Within topics, we identified roughly 11 clusters of interest that contained high probability taxa, one of which belonged to T61, the topic most associated with the V group (figure 5). This cluster was dominated by taxa belonging to Lachnospiraceae (11/23), but T61 still placed high probability on *Roseburia*, *Blautia*, and *Ruminococcaceae*. Given T61's associated the V diet, this result is consistent with literature associating *Roseburia* and *Ruminococcaceae* with starch and plant polysaccharide metabolism (Flint et al., 2012) and *Roseburia* and *Blautia* with whole grains (Flint et al., 2015; Martínez et al., 2013). Also, consistent with this topic being dominated by Gram positive bacteria, we identified a significant depletion in predicted LPS biosynthesis genes. (figure 6)

T12 contained a small yet diverse cluster of bacteria within *Acinetobacter*, a genus often associated with fermented foods and beverages, having high topic probability (Tamang et al., 2016). Quinn et al. (2016), investigating the effect home-fermented foods had on human microbiota, identified enrichment of predicted fluorobenzoate degradation pathways (Quinn et al., 2016). This same pathway, T12, had the largest shift of any predicted pathways within a given topic (figure 6). To further investigate relationship between fluorobenzoate degradation pathways and diet group, we performed a logistic regression (logit link) on all samples aged at least 21y. Diet type (V=1) and the z-scored probability of containing T61 were independent variables with alcohol consumption (n_{no}=837, n_{yes}=3692) as the binary outcome (yes=1). Both T61 (β_{T61}=1.10, z=3.64, p<0.001) and diet (β_{diet}=-0.89, z=-6.78, p<0.001) were significant, suggesting a potential relationship with fermented foods (specifically alcohol), *Acinetobacter*, and fluorobenzoate degradation.

Finally, while T76 contained bacteria typically associated with a western lifestyle such as *Clostridiales* (Gorvitovskaia et al., 2016), it also placed the most probability mass on the *Faecalibacterium prausnitzii* (figure S14), as well as predicted butyrate production. This is significant because butyrate has not only critical in the fermentation of plant matter (Gill et al., 2006), but reduction of fecal butyrate has been implicated in obesity and a shift toward a less carbohydrate-rich diet (Duncan et al., 2007). Moreover, the remaining bacteria present in this T76 cluster, *Ruminiococus* and *Roseburia*, have been shown to be elevated after fiber consumption (Flint et al., 2015).

The topics associated with the O group, on the other hand, had predicted enrichment for LPS and secretion system pathways. A noteworthy cluster in T77 was surprisingly quite similar to the aforementioned cluster in T61. Lachnospiraceae composed the majority of each cluster: 47.8% (11/23) of taxa for T61 compared to 20.6% (13/63) for T61. Moreover, the profiles of predicted functional content were analogous for all pathways except carotenoid biosynthesis and porphyrin and chlorophyll metabolism. A notable distinguishing characteristic is the lack of any *Roseburia* in the T77 cluster compared to T61.

T20 also was enriched in predicted carotenoid biosynthesis, but the specific genes differed between the two topic clusters (table S12). T77 contained a disproportionate amount of the gene that codes for the enzyme in the final step of the synthesis of bacterial antioxidant staphyloxanthin (Clauditz et al., 2006) (figure S16). T20 was also abundant in genes belonging associated with secretion system function and LPS biosynthesis, and with respect to T20, a relative shift away from a subset of LPS genes key in one specific branch of the LPS pathway. High probability mass was placed on two taxa (order RF32) belonging to the class alphaproteobacteria, which has been identified in a cluster associated with animal based diets (David et al., 2014).

DISCUSSION

We have introduced our approach for uncovering latent thematic structure in the context of host state for 16S rRNA surveys. We contend that using a topic model to explore taxonomic and predicted functional structure improves interpretability in its natural ability to uncover the relationship between collections of co-occurring taxa (topics) and samples, topics and individual taxa, as well as topics and host covariates. Also, rather than inferring predicted functional content independently from taxonomic information, we shifted our focus to predicting within-topic functional content, which we parse by estimating pathway-topic interactions using a multilevel fully Bayesian regression model. The result not only provides a means to further explore our topics, it also allows us to link functions to specific clusters which can in turn be linked to sample covariates. This has notable implications in that we are drastically reducing the dimensionality of three sources of information, thus achieving a novel means to interpret these data. Moreover, we can identify gene sets of interest from noteworthy topics. For example, when the pipeline was applied to Gevers, we determined that T15 is (1) associated with CD+ samples; (2) dominated by a cluster of bacteria known to be associated with CD; and (3) uniquely enriched for a subset of LPS synthesis genes. Being able to explore this topic's gene profile demonstrates the utility of this topic model approach. Using this information, one could focus on gene subsets associated with topic specific bacterial clusters that are known disease biomarkers, which in turn may facilitate targeted approaches for manipulating the microbiome.

We present our approach at a time when novel means to analyze complex microbiome abundance data is called for. Current methods often link the abundance of a single OTU across samples to some particular sample outcome. These methods routinely identify important subsets of taxa, but ignore OTU co-occurrence. Network methods overcome this concern, but instead fail to do so in the context of sample data and hence are incapable of linking sections of the network with sample subsets of interest. Constrained ordination methods, such as canonical correspondence analysis, do in fact couple inter-community distance with sample information, but the user is limited to specific distance metrics (e.g., Chi-squared) and must follow key assumptions (e.g., the distributions of taxa along environmental gradients are unimodal) (Legendre & Legendre, 1998). Moreover, interpretation of biplots becomes increasingly difficult as more covariates are included, and, unlike our approach, linking key subsets of taxa with corresponding subsets of gene functions is not easily achievable.

The ability to make meaningful inferences is further compounded by the fact that microbiome data is often inadequately sampled (justifying some type of normalization procedure), compositional (due to normalization), sparse, and overdispersed. Compositional data restricts the appropriateness of many statistical methods due to the sum constraint placed across samples. SPIEC-EASI provides a robust network approach for overcoming compositional artifacts in an attempt to infer community level interactions. We hence compared our within topic taxonomic clusters to the first and second order interactions identified by SPIEC-EASI, to which we found coherence between the two approaches, suggesting a topic model approach for compositional data is in fact appropriate.

Others have explored the use of Dirichlet-Multinomial models, which are well equipped at managing overdispersed count data (Brien & Record, 2016; Holmes et al., 2012; De Valpine & Harmon-Threatt, 2013). The fact that Dirichlet-Multinomial conjugacy is exploited for the topics-over-OTUs component of the topics models described above reflects their suitability for abundance data. We selected the recently developed STM for our workflow because of its ability to not only utilize sample data prior information in the flavor of the Dirichlet-Multinomial topic model, but also its ability to capture topic correlation structure and apply partial pooling over samples or regularization across regression weights.

Normalization is also a chief concern when analyzing sequencing abundance data (McMurdie & Holmes, 2014); hence, we found it imperative to determine a suitable approach. In the original LDA paper, the generative process assumed a fixed document length N, but N was considered a simplification and could easily be removed because it is independent of all other components of the model. This allows for the possibility of more realistic document size distributions (Blei et al., 2003). Given this fact, coupled with the ability of the Dirichlet-Multinomial distribution in handling overdispersion, and the results of our simulations, we concluded that raw abundance data could be adequately modeled in our approach (Woloszynek et al., 2017). The variance stabilization through DESeq2, while potentially ideal for large sample sizes with adequate signal, seemed to dampen the ability to identify topic-sample associations. Despite performing well at mapping SCs to topics, the rarefied approach suffered from reduced power when identifying topics with large covariate effects.

Finally, there are limitations to our approach. First, the workflow from OTU abundance table through pathway-topic inference scales poorly in terms of computation time for large numbers of topic, which may be more necessary as datasets continue to grow in size. Regularization and sparsity inducing priors help limit the number of important topics; hence, exploring only a subset of topics during the final regression step can offer substantial speed improvements at little cost, but utilizing the complete set of topic information would be ideal. Also, we utilize Hamiltonian MC via Stan. Other posterior inference procedures such as variational inference using software packages such as Edward may provide additional speed enhancements (Brevdo et al., 2017). Second, we are capable of separately estimating the uncertainty in our topic model, the hierarchical regression model, and the functional predictions from PICRUSt, but we currently do not propagate the uncertainty throughout the workflow. Doing so would improve downstream interpretation with better estimation of the topic-sample covariates and pathway-topic effects, which in turn would greatly improve one's confidence with utilizing within-topic gene sets. Third, we do not incorporate phylogenetic branch length information, which could lead to more meaningful topics.

ACKNOWLEDGMENTS

We would like to thank Zhengqiao Zhao and Mike O'Connor for their feedback on the manuscript and Zhengqiao in particular for his help in developing the simulation. We would also like to thank the Casey Greene Lab for their information on continuous integration and reproducibility, which was invaluable when developing Themetagenomics.

Aßhauer, K.P., Wemheuer, B., Daniel, R., & Meinicke, P. (2015). Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data: Fig. 1. *Bioinformatics* **31**, 2882–2884.

Blei, D.M., & Lafferty, J.D. (2007). A correlated topic model of Science. *Ann. Appl. Stat.* **1**, 17–35.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. **3**, 993–1022.

Blei, D.M., McAuliffe, J.D., & Blei, D.M. (2008). Supervised Topic Models. *Adv. Neural Inf. Process. Syst. 20* **21**, 1–22.

Brevdo, E., Hoffman, M.D., Murphy, K., Blei, D.M., Tran, D., Saurous, R.A., Hoffman, M.D., Brevdo, E., Murphy, K., & Blei, D.M. (2017). Deep Probabilistic Programming.

Brien, J.D.O., & Record, N. (2016). The power and pitfalls of Dirichlet-multinomial mixture models for ecological count data.

Callahan, B.J., Mcmurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J., & Holmes, S.P. (2015). DADA2 : High resolution sample inference from amplicon data. *bioRxiv* **13**, 0–14.

Clauditz, A., Resch, A., Wieland, K.P., Peschel, A., & Götz, F. (2006). Staphyloxanthin plays a role in the fitness of Staphylococcus aureus and its ability to cope with oxidative stress. *Infect. Immun.* **74**, 4950–4953.

David, L. a, Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A. V, Devlin, a S., Varma, Y., Fischbach, M. a, Biddinger, S.B., Dutton, R.J., & Turnbaugh, P.J. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563.

Duncan, S.H., Belenguer, a., Holtrop, G., Johnstone, a. M., Flint, H.J., & Lobley, G.E. (2007). Reduced Dietary Intake of Carbohydrates by Obese Subjects Results in Decreased Concentrations of Butyrate and Butyrate-Producing Bacteria in Feces. *Appl. Environ. Microbiol.* **73**, 1073–1078.

Eisenstein, J., Ahmed, A., & Xing, E.P.E. (2011). Sparse additive generative models of text. *Proc. 28th Int. Conf. Mach. Learn.* , doi: 10.1.1.206.5167.

Flint, H.J. (2012). The impact of nutrition on the human microbiome. *Nutr. Rev.* **70**, S10–S13.

Flint, H.J., Scott, K.P., Duncan, S.H., Louis, P., & Forano, E. (2012). Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* **3**, 289–306.

Flint, H.J., Graf, D., Cagno, R. Di, Fa, F., Nyman, M., Saarela, M., & Watzl, B. (2015). Contribution of diet to the composition of the human gut microbiota ˚. *Microb. Ecol.* **1**, 1–11.

Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models (New York, NY: Cambridge University Press).

Gelman, A., & Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Stat. Sci.* **7**, 457–511.

Gevers, D., Kugathasan, S., Denson, L.A., et al. (2014). The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host Microbe* **15**, 382–392.

Gilbert, J.A., Quinn, R.A., Debelius, J., Xu, Z.Z., Morton, J., Garg, N., Jansson, J.K., Dorrestein, P.C., & Knight, R. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* **535**, 94–103.

Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D. a, Fraser-Liggett, C.M., & Nelson, K.E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355–1359.

Gorvitovskaia, A., Holmes, S.P., & Huse, S.M. (2016). Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. *Microbiome* **4**, 15.

Holmes, I., Harris, K., & Quince, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS One* **7**.

Iwai, S., Weinmaier, T., Schmidt, B.L., Albertson, D.G., Poloso, N.J., Dabbagh, K., & DeSantis, T.Z. (2016). Piphillin: Improved prediction of metagenomic content by direct inference from human microbiomes. *PLoS One* **11**, 1–18.

Jiang, X., Dushoff, J., Chen, X., & Hu, X. (2012). Identifying enterotype in human microbiome by decomposing probabilistic topics into components. *2012 IEEE Int. Conf. Bioinforma. Biomed.* , doi: 10.1109/BIBM.2012.6392720.

Kembel, S.W., Wu, M., Eisen, J.A., & Green, J.L. (2012). Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLoS Comput. Biol.* **8**, 16–18.

Knights, D., Costello, E., & Knight, R. (2011). Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35**, 343–359.

Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R., & Kelley, S.T. (2013). Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8**, 761–763.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **28**, 1–26.

Kurtz, Z., Mueller, C., Miraldi, E., & Bonneau, R. (2016). SpiecEasi: Sparse InversE Covariance estimation for Ecological Association and Statistical Inference.

Kurtz, Z.D., M??ller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., & Bonneau, R.A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Comput. Biol.* **11**, 1–25.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J. a, Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., Beiko, R.G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821.

Legendre, P., & Legendre, L. (1998). Numerical Ecology - Second English Edition.

Lewis, J.D., Chen, E.Z., Baldassano, R.N., et al. (2015). Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. *Cell Host Microbe* **18**, 489–500.

Li, H. (2015). Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. *Annu. Rev. Stat. Its Appl.* **2**, 73–94.

Martínez, I., Lattimer, J.M., Hubach, K.L., Case, J.A., Yang, J., Weber, C.G., Louk, J.A., Rose, D.J., Kyureghian, G., Peterson, D.A., Haub, M.D., & Walter, J. (2013). Gut microbiome composition is linked to whole grain-induced immunological improvements. *ISME J.* **7**, 269–280.

McMurdie, P.J., & Holmes, S. (2013). Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8**.

McMurdie, P.J., & Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Comput. Biol.* **10**.

Mimno, D., & McCallum, A. (2012). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv Prepr. arXiv1206.3278* , doi: 10.1.1.140.6925.

Ning, J., & Beiko, R.G. (2015). Phylogenetic approaches to microbial community classification. *Microbiome* **3**, 47.

Quinn, R.A., Navas-Molina, J.A., Hyde, E.R., et al. (2016). From Sample to Multi-Omics Conclusions in under 48 Hours. *mSystems* **1**, e00038-16.

Ren, B., Bacallado, S., Favaro, S., Holmes, S., & Trippa, L. (2016). Bayesian Nonparametric Ordination for the Analysis of Microbial Communities. *arXiv Prepr. arXiv1601.05156*.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., & Rand, D.G. (2014). Structural topic models for open-ended survey responses. *Am. J. Pol. Sci.* **58**, 1064–1082.

Roberts, Margaret E., Stewart, B.M., & Tingley, D. (2017). stm: R Package for Structural Topic Models.

Shafiei, M., Dunn, K.A., Boon, E., MacDonald, S.M., Walsh, D.A., Gu, H., & Bielawski, J.P. (2015). BioMiCo: a supervised Bayesian model for inference of microbial community structure. *Microbiome* **3**, 8.

Stan Development Team (2016). rstanarm: Bayesian applied regression modeling via Stan.

Tamang, J.P., Watanabe, K., & Holzapfel, W.H. (2016). Review: Diversity of microorganisms in global fermented foods and beverages. *Front. Microbiol.* **7**.

Tjalsma, H., Boleij, A., Marchesi, J.R., & Dutilh, B.E. (2012). A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. *Nat. Rev. Microbiol.* **10**, 575–582.

De Valpine, P., & Harmon-Threatt, A.N. (2013). General models for resource use or other compositional count data using the Dirichlet-multinomial distribution. *Ecology* **94**, 2678–2687.

Woloszynek, S., Zhao, Z., Simpson, G., Mell, J. C., and Rosen, G. Gauging the use of topic models as a means to understand microbiome data structure. *In prep.*

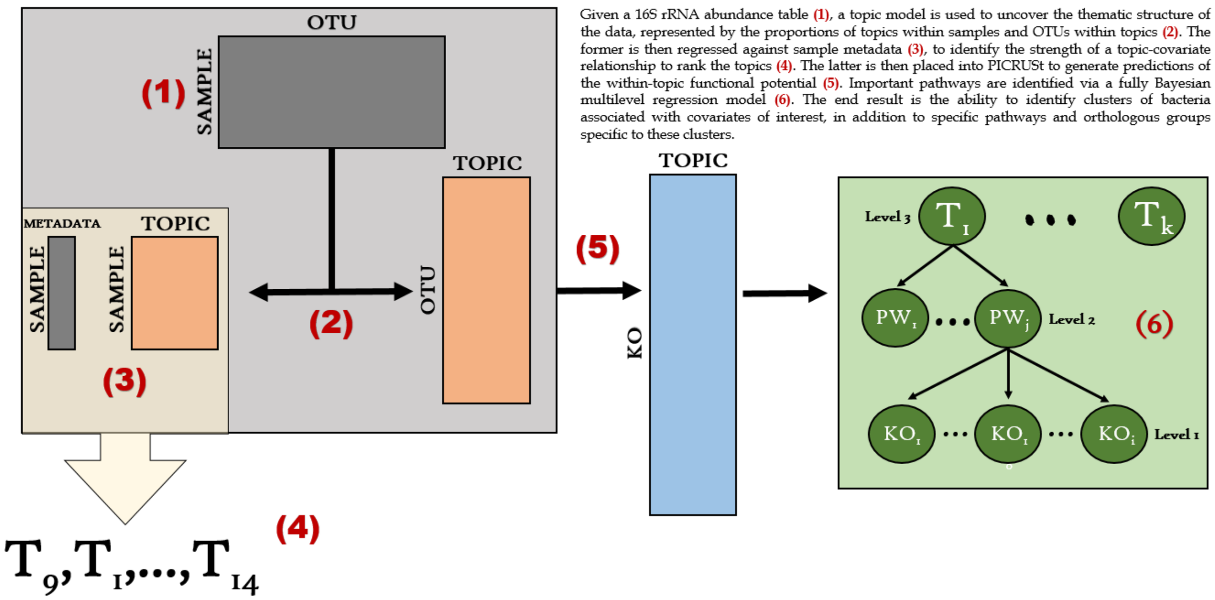
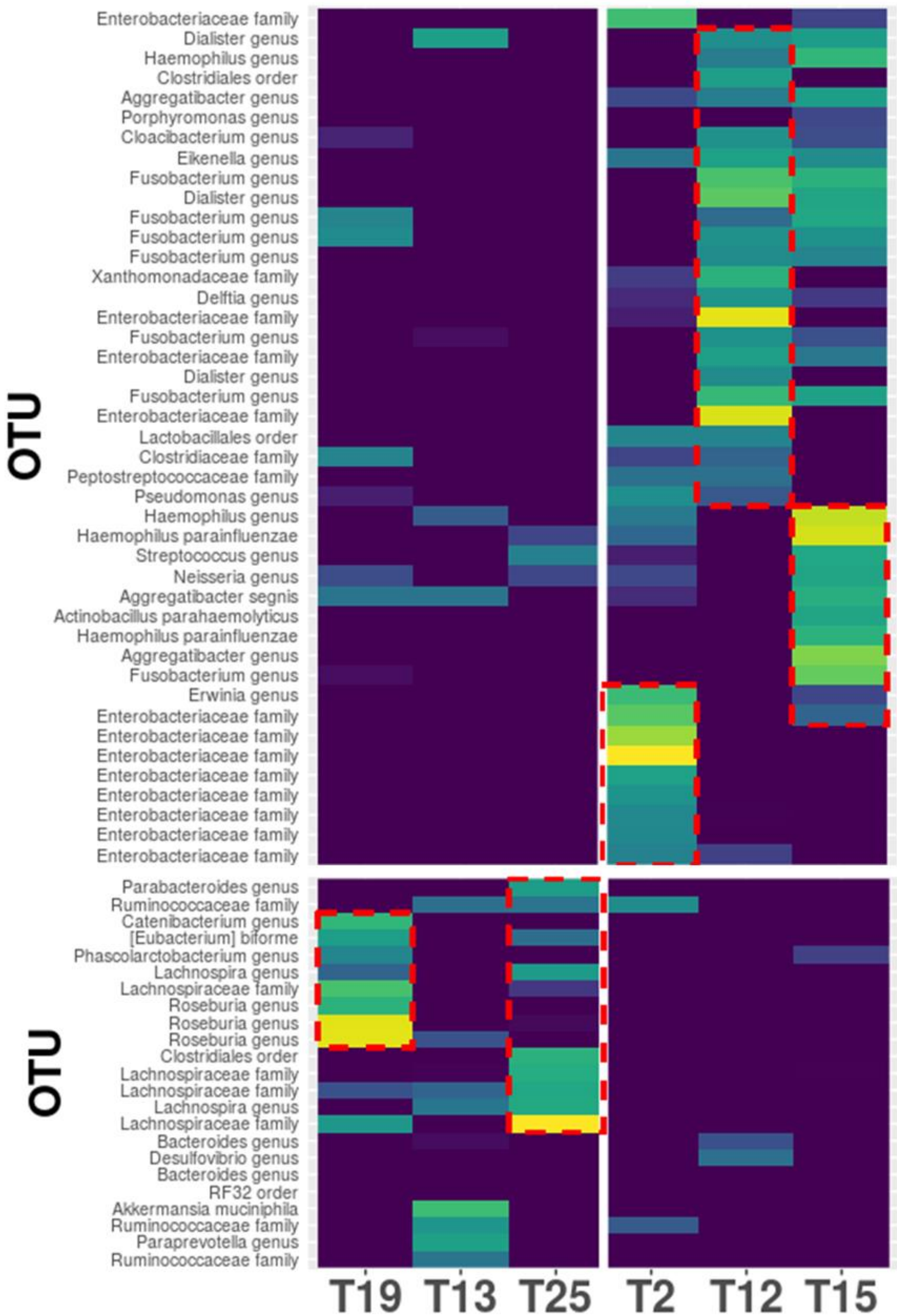


Figure 1. Given a 16S rRNA abundance table (1), a topic model is used to uncover the thematic structure of the data, represented by the proportions of topics within samples and OTUs within topics (2). The former is then regressed against sample data (3), to identify the strength of a topic-covariate relationship to rank the topics (4). The latter is then placed into PICRUSt to generate predictions of the within-topic functional potential (5). Important pathways are identified via a fully Bayesian multilevel regression model (6). The end result is the ability to identify clusters of bacteria associated with covariates of interest, in addition to specific pathways and orthologous groups specific to these clusters.

713



OTU
Enterobacteriaceae family
Dialister genus
Haemophilus genus
Clostridiales order
Aggregatibacter genus
Porphyromonas genus
Cloacibacterium genus
Eikenella genus
Fusobacterium genus
Dialister genus
Fusobacterium genus
Fusobacterium genus
Fusobacterium genus
Xanthomonadaceae family
Delftia genus
Enterobacteriaceae family
Fusobacterium genus
Enterobacteriaceae family
Dialister genus
Fusobacterium genus
Enterobacteriaceae family
Lactobacillales order
Clostridiaceae family
Peptostreptococcaceae family
Pseudomonas genus
Haemophilus genus
Haemophilus parainfluenzae
Streptococcus genus
Neisseria genus
Aggregatibacter segnis
Actinobacillus parahaemolyticus
Haemophilus parainfluenzae
Aggregatibacter genus
Fusobacterium genus
Erwinia genus
Enterobacteriaceae family
Enterobacteriaceae family
Enterobacteriaceae family
Enterobacteriaceae family
Enterobacteriaceae family
Enterobacteriaceae family
Enterobacteriaceae family
Enterobacteriaceae family
OTU
Parabacteroides genus
Ruminococcaceae family
Catenibacterium genus
[Eubacterium] biforme
Phascolarctobacterium genus
Lachnospira genus
Lachnospiraceae family
Roseburia genus
Roseburia genus
Roseburia genus
Clostridiales order
Lachnospiraceae family
Lachnospiraceae family
Lachnospira genus
Lachnospiraceae family
Bacteroides genus
Desulfovibrio genus
Bacteroides genus
RF32 order
Akkermansia muciniphila
Ruminococcaceae family
Paraprevotella genus
Ruminococcaceae family
T19
T13
T25
T2
T12
T15

714

Figure 2. Subsections of the heatmap for the Gevers data of the topics over OTU distribution in log space generated by the K25 topic model with covariate prior information. Shown are the top 3 topics associated with CD- and CD+, ordered by mean regression estimate (left to right, respectively, separated by the white line). Clusters of interest are marked with red dotted lines. Clustering was performed via Ward's method on Bray-Curtis distances. Low probabilities ($p < 1x10^{-5}$) are set to 0 to minimize the range of the color gradient to ease visualization. Yellow=high probability, Blue=low probability.

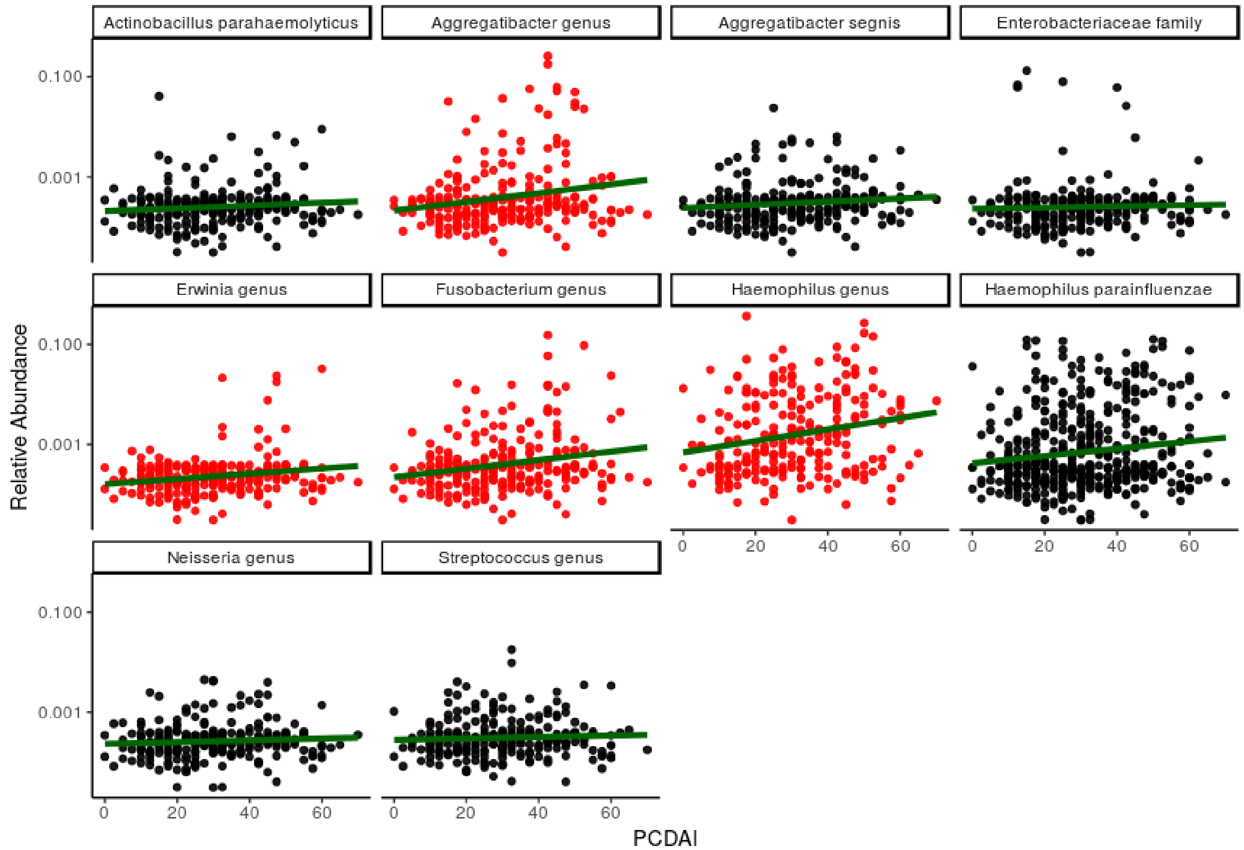


Figure 3. Scatterplots of Gevers data for the relative abundance of taxa that compose a high probability cluster in T15 versus PCDAI, a clinical measure of CD disease burden. Red points reflect significance (alpha=.05) for negative binomial regression (log linked, sample coverage offset) with Bonferroni correction.

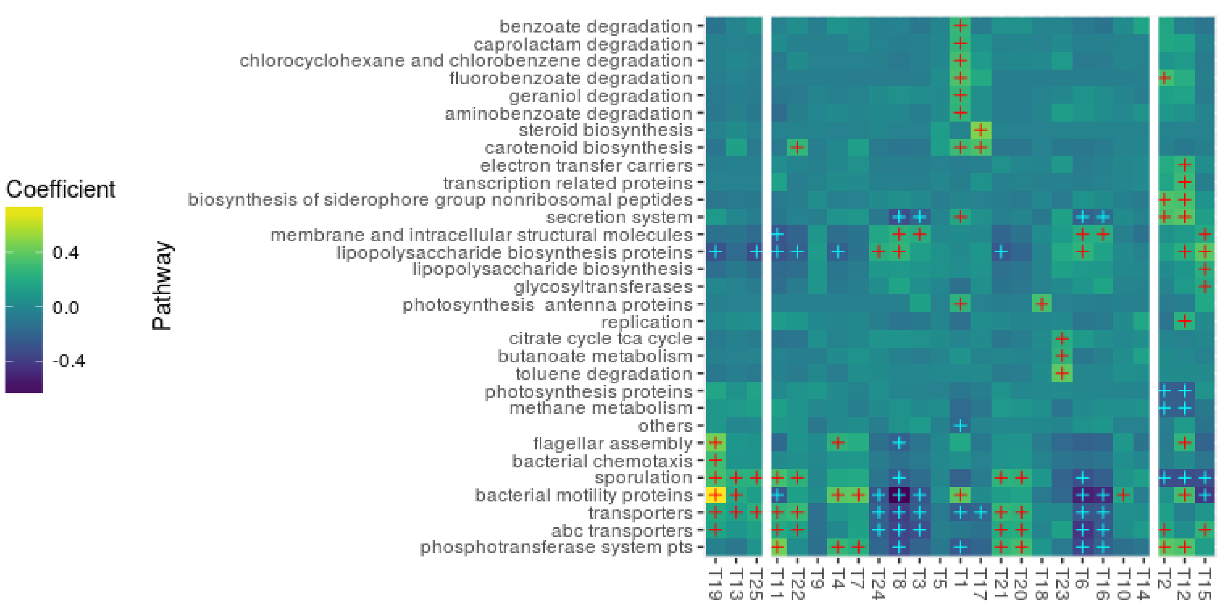


Figure 4. Heatmap for Gevers data of the level-3 pathway category-topic interaction regression coefficients from the multiple level negative binomial model. KEGG information was predicted via PICRUSt on the topics over OTU distribution from the K25 topic

model with covariate prior information. Topics are ordered based on their mean regression weight when using topic probabilities as linear predictors for disease presence, where leftmost topics are most associated with CD-, whereas right most topics are most associated with CD+. Clustering was performed via Ward's method on Bray-Curtis distances. Red and blue crosses indicate weights or pathway-topic combinations that do not span 0 with 80% uncertainty and are positive or negative, respectively. Only pathways with at least one such combination are shown.

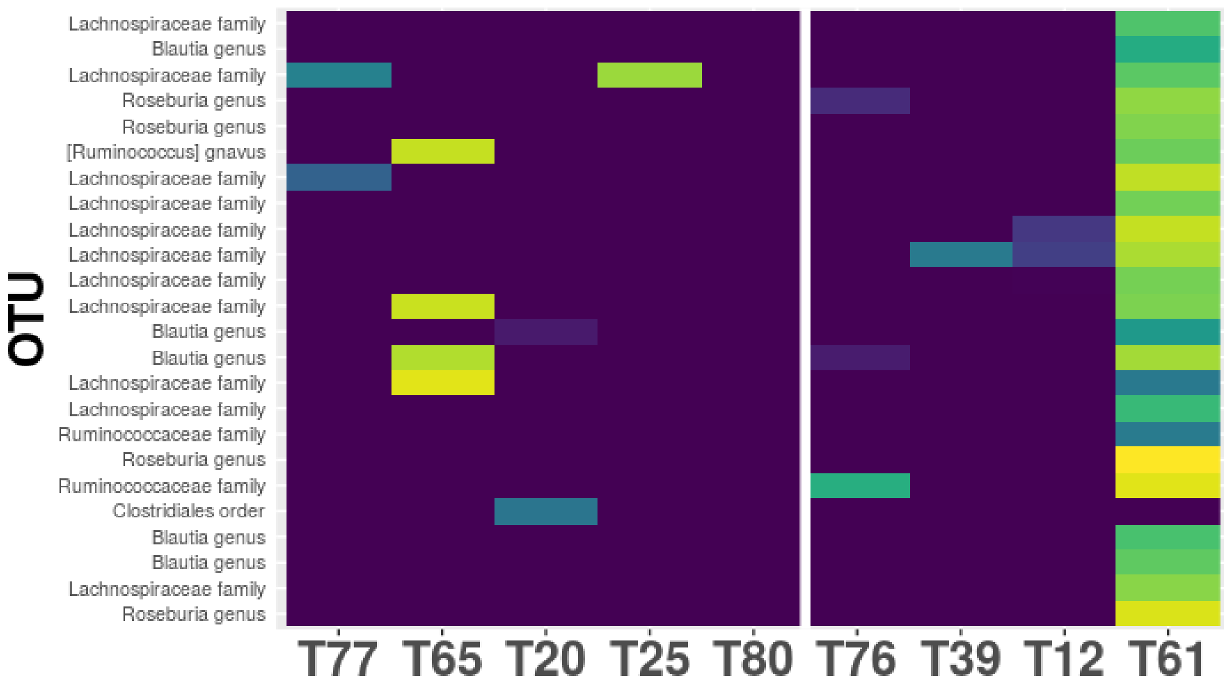


Figure 5. Subsection of the heatmap for AG data for the topics over OTU distribution in log space generated by the K100 topic model with covariate prior information. Shown are the topics with 95% uncertainty intervals that do not enclose 0 when regressed against diet type (O=0, V=1), ordered negative to positive by increasing mean regression estimate (left to right), such that T77 is most associated with O and T61 is most associated with V. The white light signifies a shift from positive to negative means regression

estimates. Clustering was performed via Ward's method on Bray-Curtis distances. Low probabilities ($p < 1x10^{-5}$) are set to 0 to minimize the range of the color gradient to ease visualization. Yellow=high probability, Blue=low probability.

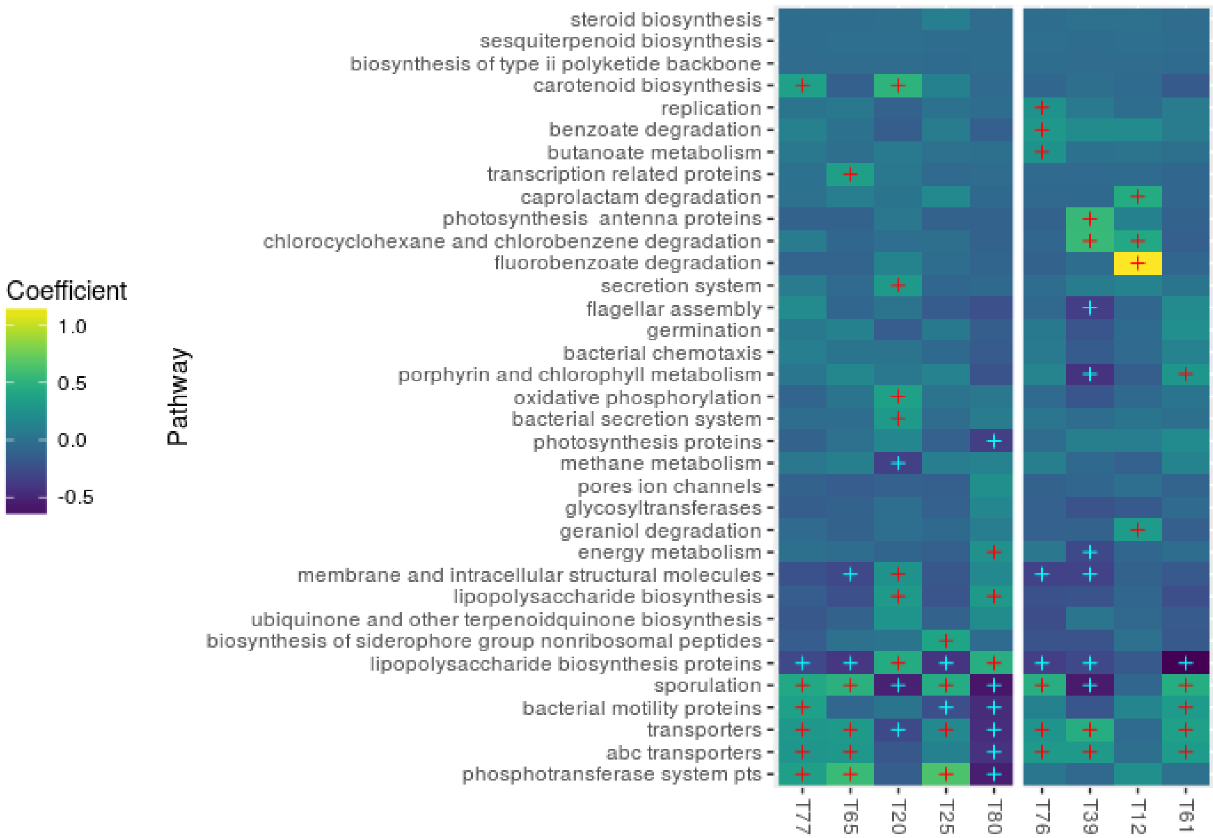


Figure 6. Heatmap for AG data of the level 3 pathway category-topic interaction regression coefficients from the multiple level negative binomial model. KEGG information was predicted via PICRUSt on the topics over OTU distribution from the K100 topic model with covariate prior information. Only the top 25 topics based on mean regression weight (when using topic probabilities as linear predictors for disease presence) were chosen for the negative binomial to alleviate computational concerns. Topics are ordered based on their mean regression weight, where leftmost topics are most associated with O, whereas right most topics are most associated with V, separated by the white line. Clustering was performed via Ward's method on Bray-Curtis distances. Red and blue crosses indicate weights for pathway-topic combinations that do not enclose 0 with 80% uncertainty and are positive or negative, respectively. Only pathways with at least one such combination are shown.