Identifying Host-Microbiome Interactions in Genomic Data

A Thesis

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Joshua Lynch

IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Ran Blekhman, Dan Knights

August 2015

Table of Contents

List of Figures

1. Introduction

The microorganisms living on and inside the human body are called the 'human microbiome'. This large and varied collection of organisms has been studied mainly using culturing techniques and focusing on disease-causing organisms. With the development of genome sequencing technology it has become possible to study entire microbial communities. A popular method known as '16S sequencing' focuses on sequencing just one gene, the 16S ribosomal RNA gene, across all organisms in a sample. This allows identification of the taxonomic structure of a community. Comprehensive genetic sequencing of all microbial DNA in a sample, called 'metagonomic shotgun sequencing', can identify the functional potential of a microbial community.

The human microbiome has been compared to an organ (Baquero and Nombela, 2012; Evans et al. 2013). Until recently it has been a 'forgotten' or 'neglected' organ, but 16S and metagenomic shotgun sequencing have brought a new appreciation for the significance of the human microbiome to understanding human biology and treating human disease.

The microbiota of the human body consist of thousands of taxonomically distinct organisms including members of the Archea, Prokaryota, and single-celled Eukaryota plus their attendant viruses. Variation of microbiome between individuals is very high but there are discernable similarities between family members (Spor et al., 2011). Conservative estimates put the number of microbial cells living on and inside a human host at three times the number of the host's cells and having a combined weight of several pounds. While microbial genomes are relatively small compared with the human genome, collectively the human microbiome is believed to harbor more genes than its host (Turnbaugh et al., 2007).

These microbes and their genes provide a wide range of beneficial functions not otherwise available from the human genome. There are many studies in animal models that have found complex interactions between the microbiome and the host related to development and homeostasis of the immune system. Studies in Drosophila and

zebrafish have found that the immune system operates not only to attack pathogens but also to develop and maintain the population of commensal gut microorganisms by inhibiting inflammatory responses to desirable species. Studies in mice have described similar mechanisms and in addition found evidence that the host immune system influences the spatial distribution of gut microbiota. In humans it has been shown that a particular commensal bacteria, *Bacteroides fragilis*, plays a role in immune system development as a promoter of $CD4^+$ T cells (Chu and Mazmanian, 2013).

In addition to its complex interactions with the immune system, the human gut microbiome provides metabolic functions not present in the human genome. These include genes enabling fermentation of complex carbohydrates to produce short-chain fatty acids that can be absorbed by the host, and for synthesizing and metabolizing vitamins (Yatsunenko et al., 2012). In addition, the by-products of microbial metabolism in the gut signal the host intestine, liver, muscle and adipose tissues (Tremaroli and Backhed, 2012), and perhaps the brain as well (Collins et al., 2012).

Increased understanding of the widespread commensal relationships between human hosts and their microbiome has made it clear that the genetic basis of human health and disease is not strictly encapsulated by the human genome. In this sense the microbiome is a new frontier for understanding and treating disease. Several studies have identified potential biomarkers between colorectal cancer and microbiome. In one study stool samples from subjects with colorectal cancer were found to have increased abundance of *Fusobacterium* and *Porphyromonas* and reduced numbers of species considered beneficial as compared with healthy subjects (Zackular et al., 2014). A study of microbiome from biopsied tissue found increased microbial diversity and increased abundance of predicted virulence genes at the tumor site as compared with subject-matched normal tissue (Burns et al., 2015). Significant disturbance of the host microbiome, or dysbiosis, is associated with several disease states including inflammatory bowel disease (Baumgart and Carding, 2007), *C. dificile* infection (Kachrimanidou and Malisiovas, 2011), and obesity (Ley, 2010). In some cases directly treating the microbiome has led to improved outcomes (Khoruts et al., 2010).

The evidence pointing to significant relationships between human health and disease with the state of the human microbiome is compelling, and naturally leads to questions of how the human genome and human microbiome interact. This question has been mainly addressed in mouse studies and human twins studies.

A 2011 review by Spor et al. reports on several studies of human monozygotic (MZ) and dizygotic (DZ) twins with conflicting results. A nucleic acid fingerprinting-based study of not more than 20 pairs of MZ and DZ twins reported an effect of host genetics on gut microbiome (Stewart et al., 2005). A later study using 16S and shotgun metagenomic sequences failed to find any heritable taxa in stool samples from 31 MZ and 23 DZ twin pairs (Turnbaugh et al., 2009). Both studies were probably hampered by small sample sizes.

A quantitative trait loci (QTL) study (Benson et al., 2010) of 645 mice defined a 'core measureable microbiota' of 64 microbial taxa found in all or nearly all mice. Maternal and litter effects were found to account for 26% of the variation in these 64 taxa. The core measurable microbiota were analyzed for quantitative trait association with 530 SNP markers. Of the 64 core measureable taxa, 26 were significantly related to 13 genomic regions and 6 suggestive quantitative trait loci.

A recent human twins study (Goodrich et al., 2014]) analyzed stool samples from 416 twin pairs from the TwinsUK cohort. This study found that the microbiomes of MZ twins were more highly correlated than those of DZ twins, and identified a network of heritable microorganisms. One particular family, Christensenellaceae, was found to be the most heritable taxon. Furthermore, the interaction of host genetics with microbiome was tested directly with a member of the Christensenellaceae family in a mouse model. Germ-free mice were inoculated with a human gut microbiome associated with obesity to which had been added a cultured species, *Christensenella minuta*. This modified microbiome reduced weight gain in recipient mice relative to mice that received the same human gut microbiome without added *C. minuta*. In addition, over time the two treatments resulted in significant differences in gut microbiome. The authors interpret this result as showing interaction between microbiome and host genotype.

With recognition that the human microbiome is an additional component of genetic variation that interacts in complex ways with its host, it becomes clear that understanding individual susceptibility to disease and response to treatment will require more than just decoding the human genome in isolation. The recent study by Goodrich et al., 2014 finds strong evidence of host genetic variation interacting with the human microbiome to influence obesity, and many more discoveries of disease-related interactions are anticipated. Mapping the full range of host-microbiome interactions will require studies of large cohorts from diverse populations and will ultimately address questions of basic human evolution and biology in addition to informing clinical practices.

We are interested in finding these interactions between the human microbiome and host genetic variations. We have developed a method using LASSO regression to screen for microbial taxon abundances, or other related measurements, with a linear relationship to SNPs in the host genome. Linear relationships are of particular interest because they are likely to represent large effects resulting from interactions among a group of genes. We have applied our method to data from 16S and metagenomic shotgun sequencing published by the Human Microbiome Project.

2. Methods

Our goal is to find linear associations between microbial abundance and single nucleotide polymorphisms in the host genomes of 93 human subjects. The microbial abundance data derived from 16S sequencing is reported for 15 different body sites per subject. For each body site we work with a taxon abundance table of 100 to 400 relative microbial abundances per human subject. In addition KEGG module abundance data derived from metagenomic shotgun sequencing is reported for 250 modules at 5 body sites per subject. Our host genetic variation data is a set of 32,698 SNPs in exons, although we do not have data on every SNP for every subject. We encode subject genotype for a given SNP as the count of alternative nucleotides: 0, 1, or 2.

Our data falls in the category '$p \gg n$', which means the number of predictors exceeds the number of samples by a large margin. In our case we typically have more taxon abundances, or KEGG module abundances, than human subjects by a factor of 2 or more. Additionally, in cases of association between a SNP and microbial abundances we expect a 'sparse' model, meaning only a small number of the taxon abundances are involved. We use LASSO regression (Least Absolute Shrinkage and Selection Operator), which is designed to find sparse linear models and in addition is computationally efficient (Tibshirani, 1996; Efron et al., 2004).

The first step in screening for association between a SNP and microbial abundance data is to fit a model using LASSO regression for each SNP and for abundance data at each body site. Next we estimate the false positive rate of the LASSO regression on each data set to determine if we have found meaningful results. We then use randomized LASSO to choose relevant taxa, and finally pathway and network analysis to place selected genes in context. Figure 2.1 shows the analysis pipeline.

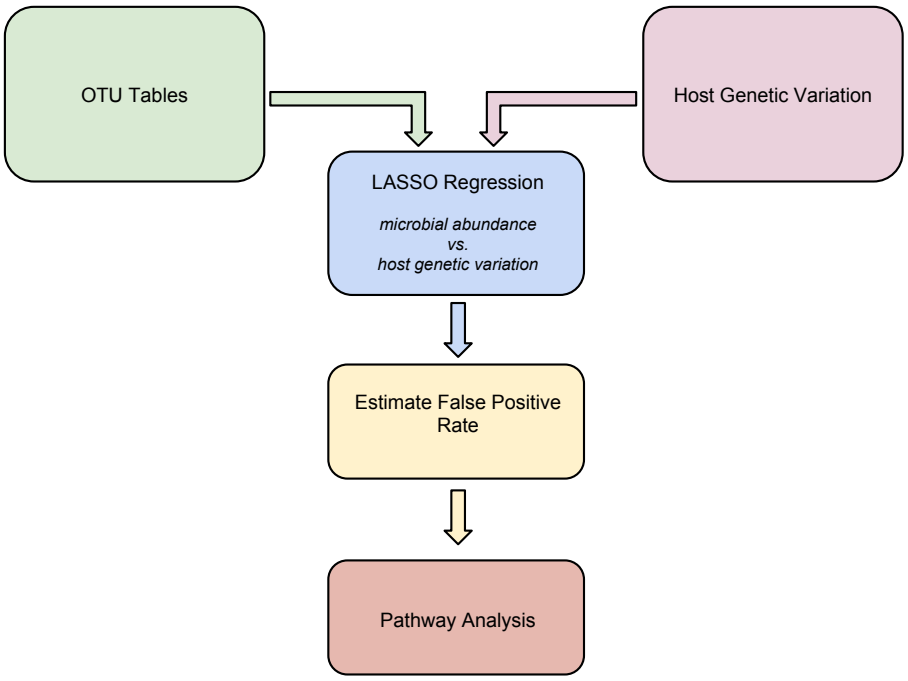

Figure 2.1 Analysis pipeline

LASSO

LASSO regression fits the usual linear model composed of n observations of p features $X_{i,j}$ $(1 \leq i \leq n, 1 \leq j \leq p)$ and n responses y_i $(1 \leq i \leq n)$:

$$X\beta = y$$

in a way that minimizes the loss function

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{p} |\beta_j|$$

In our model of association between a SNP and microbial abundances the matrix X corresponds to a table of abundances (at a particular body site) and the vector y corresponds to the alternate allele count (0, 1, or 2) for each subject for the given SNP. More specifically, each element $X_{i,j}$ is the abundance value of taxon j measured for human subject i. Each element y_i is the observed alternate allele count for human subject i, and each element $\hat{y}_i$ is the predicted alternate allele count for the same subject. Each element β_j is the jth regression coefficient ($1 \leq j \leq p$ where p is the number of abundances at the body site), and α is a scalar tuning parameter.

The α parameter is typically determined by cross-validation for each dataset as part of the model fitting process. The LASSO loss function is convex and so it is solvable by general-purpose convex optimization algorithms. However, there are special-purpose algorithms for LASSO optimization that are simpler and more efficient, such as least-angle regression (LARS) (Efron et al., 2004).

LASSO regression is similar to usual multivariate linear regression. If the tuning parameter α is set to 0 then the LASSO solution will be identical to the multivariate linear regression solution. As α increases fewer features are retained in the LASSO solution, until at some high value no features are retained. The usual method of choosing α is to pick the value that minimizes the MSE of the solution. Figure 2.2 shows an example of the effect of increasing α on the LASSO solution.

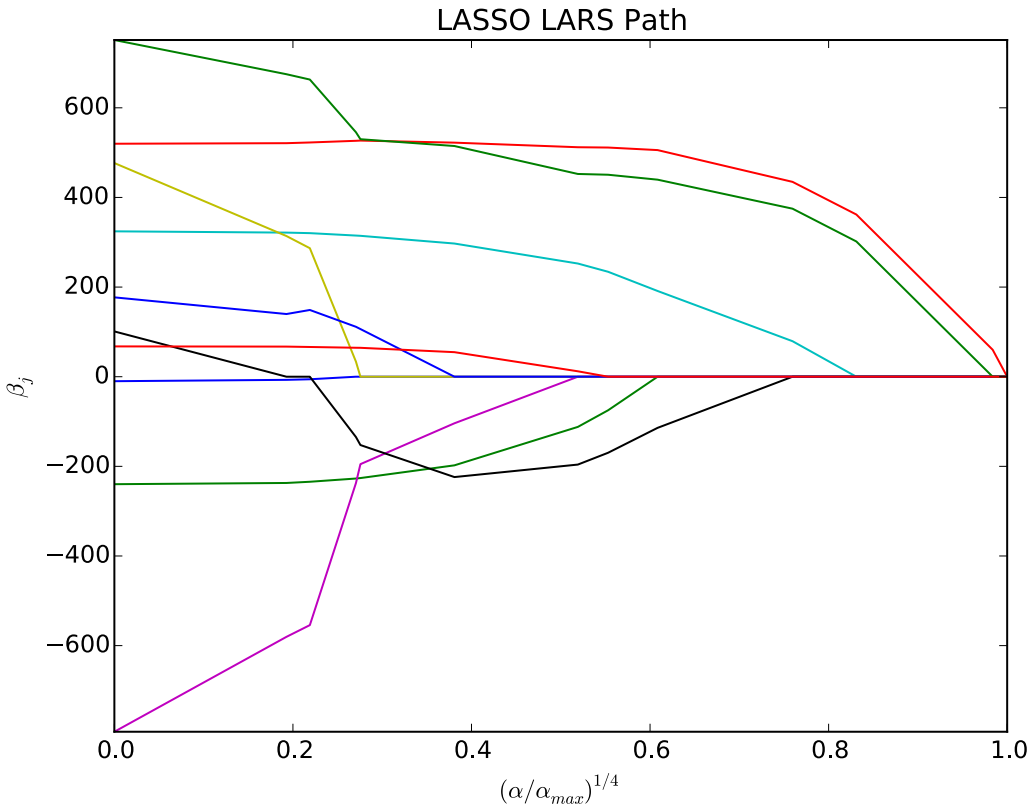

Figure 2.2 LASSO path for a 10-variable example dataset as calculated by the LARS algorithm. Each line corresponds to the regression coefficient of a feature. The LASSO parameter is shown on the horizontal axis increasing from left to right with scaling to separate small values that would otherwise be very close together. For the low values of $\boldsymbol{\alpha}$ *all features have been selected, as indicated by all coefficients having non-zero values. As* $\boldsymbol{\alpha}$ *increases some coefficients are reduced to zero, removing them from the LASSO solution. At the maximum* $\boldsymbol{\alpha}$ *all feature coefficients have been reduced to zero.*

We developed a program using the Python machine-learning library scikit-learn (Pedregosa et al., 2011) to execute LASSO regression with microbiome data as predictors and SNP minor allele count as responses. We use the R^2 statistic from the LASSO model as the indicator of correlation between microbiome predictors and host genetic responses. Our program uses the scikit-learn LARS implementation with default parameters and 5-fold cross-validation to tune the LASSO parameter α.

We observed high variance in R^2 when running LASSO directly on our data. We found the usual 5-fold and 10-fold cross-validation did not reduce the variance sufficiently to give reproducible results. We managed this variance by randomly shuffling the data and repeating the LASSO fitting (with internal 5-fold cross-validation) 100 times splitting 80% of the data for training and 20% for testing. Splits for both the internal 5-fold cross-validation and the external 100-times resampling were constructed to maintain approximately the same genotype distribution as the overall data set. We assigned the median of the 100 R^2 values to be the nominal R^2 for the regression and calculated a 95% percentile bootstrap confidence interval of the median using 10,000 bootstrap samples. We have realized benefits from this approach including reduced effort in significance testing as compared with permutation tests and assessment of data quality.

Estimating False Positive Rate

We estimate the false positive rate of our method as a function of R^2 on each data set by shuffling the allele counts for each SNP among the subjects and repeating the LASSO regression procedure. We can then determine if we have found an enrichment of SNPs with high R^2 relative to this false positive rate. We retain the enriched proportion of those SNPs having a high ratio r of R^2 to width of the 95% confidence interval:

$$r = \frac{R^2}{h_{ci} - l_{ci}}$$

Here h_{ci} is the upper bound and l_{ci} is the lower bound of the 95% confidence interval. This incorporates an estimate of uncertainty in the data into choosing SNPs to analyze further.

Selecting Relevant Taxa

We analyze the retained SNPs to determine the taxa most contributing to correlation with host genetic variation. LASSO regression is known to be sensitive to small variations of the predictors and for this reason it is common to use a resampling method to choose relevant predictors. We have chosen to use stability selection with randomized LASSO (Meinshausen and Bühlmann, 2010). Stability selection is not a LASSO-specific method

but works with any regularized model-fitting algorithm. The stability selection algorithm for a general model is as follows:

1. Choose a set of model tuning parameter values $A = \{\alpha_1, \alpha_2, \cdots, \alpha_n\}$.
2. For each $\alpha_i \in A$, repeat N times:
    a. Randomly choose a subset of the data.
    b. Fit the model to the data subset using α_i.
    c. Record the selected predictors.
3. Calculate the frequency f_p with which each predictor was selected over all trials.
4. Choose a threshold F $(0 < F < 1)$ and select the predictors with $f_p > F$.

Randomized LASSO is a slight variation of the usual LASSO that incorporates random scaling factors in the penalty term:

$$L = \sum_i (y_i - \hat{y}_i)^2 + \alpha \sum_j \frac{|\beta_j|}{W_j}$$

The W_j are drawn randomly from the interval $[\gamma, 1]$ for a fixed 'weakness' $\gamma \in (0,1]$. This variation of the LASSO model is only useful in a setting such as stability selection where the model is repeatedly fit and the results are combined. One advantage of stability selection is that it is less sensitive to the choice of the LASSO parameter than other variable selection methods. Incorporating random scaling in the penalty term improves the consistency of the results.

We use the implementation of stability selection with randomized LASSO from the scikit-learn library, and choose N=1000 trials. We randomly select 80% of the data for each trial keeping the distribution of genotypes the same as in the entire data set. It is necessary to provide a range of values of the LASSO tuning parameter so we repeat the process of model fitting with LARS and 100-times random shuffling to get a series of tuning parameters. We keep the largest α and construct a list of ten evenly spaced values ranging from 0.3α to α. This list contains relatively large tuning parameters to favor including fewer predictors in each model. Figure 2.3 shows stability paths for a simple example dataset.

We do not aggressively tune the threshold F. In the case of synthetic data we will look at a few values, but for real data we will simply choose F such that the number of features selected stays small for most SNPs.

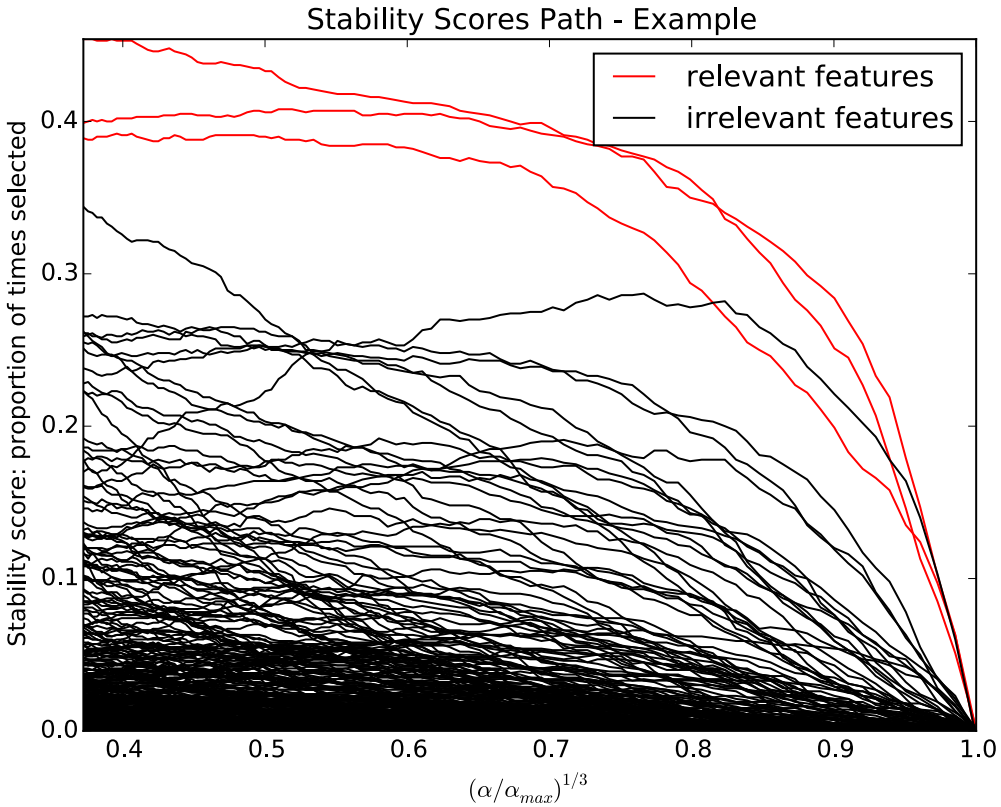

Figure 2.3: Stability paths for an example dataset with 500 features, three of which are correlated to the response. Each line represents the probability that a feature is selected as relevant by the stability selection procedure as a function of the LASSO tuning parameter α. The three paths at the top of the figure are relevant features, the remaining paths are irrelevant features. As the LASSO tuning parameter α decreases from right to left, the LASSO procedure behaves more like regular regression and more variables are selected. The three relevant features are clearly selected most often out of all features.

3. Synthetic Data Results

Synthetic Data

We generated three synthetic datasets to assess the performance of LASSO regression for discovering correlation between SNP minor allele count and microbial abundances. Each synthetic dataset was constructed by varying a different parameter to create several subsets. In synthetic dataset 1 we varied minor allele frequency. In synthetic dataset 2 we varied the correlated taxon count. In synthetic dataset 3 we varied total taxon count. Additionally, all three datasets were constructed with varying degrees of SNP noise, which was generated by randomly shuffling the minor allele counts of a small number of samples.

Each synthetic data subset contained 100 samples (corresponding to human subjects). For each sample we generated a table of synthetic taxon abundances and 500 SNPs with a fixed minor allele frequency and correlated with taxon abundances. Additionally, in synthetic datasets 1 and 3 we generated and tested subsets of 10,000 SNPs with no correlation to the taxon abundances. Synthetic dataset 2 included no uncorrelated SNPs because the parameter under variation was the number of *correlated* taxa per SNP. We generated 200 taxon abundances for each sample in datasets 1 and 2. In dataset 3 we used three different total taxon abundance counts: 100, 300, and 500.

Taxon Abundance

Each synthetic taxon abundance was determined by exponential transform of a random value drawn from a normal distribution with mean 0.0 and standard deviation 1.0. The randomly determined abundances were then normalized to give relative abundances.

SNP Correlation

We constructed SNP alleles with correlation to n synthetic abundances by applying a SNP-specific linear function $f_i: \mathbb{R}^n \to \mathbb{R}$ with randomly selected coefficients $c = (c_1, c_2, \cdots, c_n)$ to a set of randomly selected abundances $a = (a_1, a_2, \cdots, a_n)$:

$$f_i(c, a) = c \cdot a = c_1 a_1 + c_2 a_2 + \cdots + c_n a_n$$

The coefficients c_i were drawn from a normal distribution with mean 0.0 and standard deviation 1.0. The distribution of f_i over the selected taxon abundances was partitioned into three intervals proportional to SNP allele frequencies as determined by the established minor allele frequency. We assigned the SNP allele for each sample by determining to which partition the sample's f_i value belonged. SNP alleles were coded 0, 1, or 2 to indicate the number of alternative nucleotides per SNP.

We evaluated the degree of correlation between each SNP and the associated abundances by linear regression to guarantee a minimum degree of correlation. We retained the synthetic SNP alleles if $R^2 > 0.35$. Otherwise we repeated the procedure with a new f_i until the threshold was met.

SNP Noise

We added random, smooth noise at levels $P = 0.00, 0.05, 0.10$ to the correlated SNPs while maintaining the chosen minor allele frequency. To add noise to the distribution of allele counts for a correlated SNP, we first randomly selected $\frac{PN}{2}$ samples. The SNP value (alternative allele count 0, 1, or 2) of each selected sample was swapped with that of another randomly selected sample. The noise was 'smooth' in the sense that swapped samples were constrained to differ in allele count by 1. This smoothness emulates adding noise to the taxon abundances, which results in swapping allele count 0 with allele count 2 only at very high noise levels. The advantage of swapping a fixed number of allele counts over adding noise directly to the taxon abundances is that we can directly control the number of swaps so all SNPs have a consistent level of noise. This procedure results in PN miscorrelated SNPs while maintaining the original distribution of SNP values. To generate SNP noise at level $P = 0.05$ we generated noise for half the SNPs at $P = 0.04$ and for the other half at $P = 0.06$.

Synthetic Dataset 1

We constructed this dataset to investigate the effect on LASSO regression of variation in minor allele frequency and SNP noise. This dataset consisted of 12 subsets of 500 correlated SNPs formed by all pairings of minor allele frequencies 0.25, 0.30, 0.35, and 0.40 with SNP noise levels of 0.00, 0.05, and 0.10. In all correlated subsets three taxon

abundances were randomly selected to generate correlated SNP alleles. Four subsets of 10,000 uncorrelated SNPs were also generated, one for each minor allele frequency.

In Figure 3.1 sensitivity (ratio of true positives to the sum of true positive and false negatives) is shown as a function of SNP noise and minor allele frequency for our LASSO procedure on synthetic dataset 1. There is small fluctuation in sensitivity but no dependence on minor allele frequency or SNP noise level.

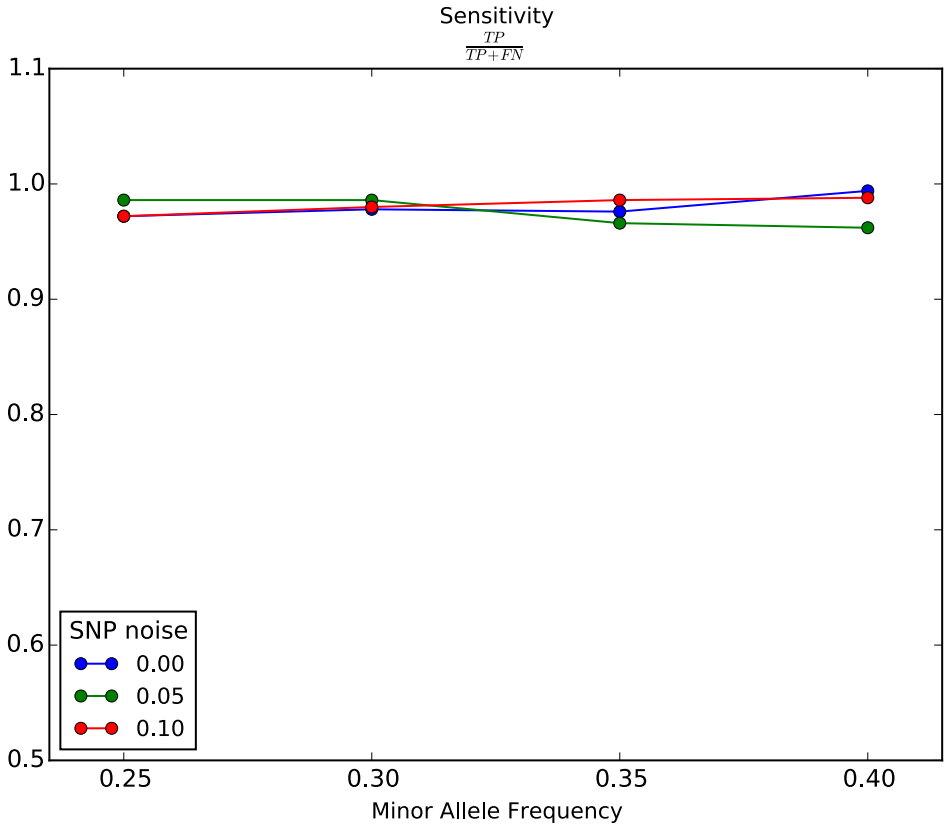

Figure 3.1: Sensitivity as a function of minor allele frequency in synthetic dataset 1.

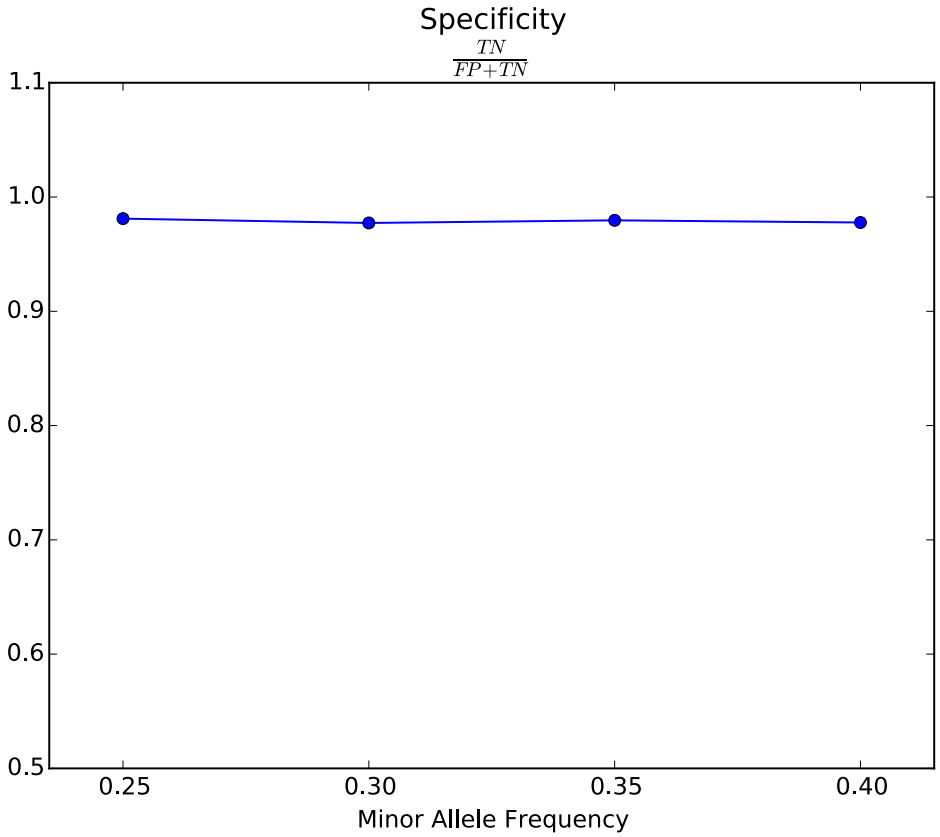

Figure 3.2: Specificity as a function of minor allele frequency in synthetic dataset 1.

Figure 3.2 shows the specificity (ratio of true negatives to the sum of true negatives and false positives) as a function of minor allele frequency for our LASSO procedure on 10,000 uncorrelated SNPs. It can be seen that specificity of the method is high and has no strong dependence on minor allele frequency. The small number of SNPs mis-identified as correlated to some taxa may be correlated by chance since the data set is large.

Minor allele frequency and SNP noise have a greater effect on feature selection. Figure 3.3 shows the performance of correlated taxon identification by stability selection for each level of SNP noise and each minor allele frequency. Overall feature selection performs well on this data. At all three levels of SNP noise and all four minor allele

frequencies, 2/3 or better of correlated taxa are correctly selected on average (mean precision > 0.667) and of the selected taxa 2/3 or better are correlated (recall > 0.667).

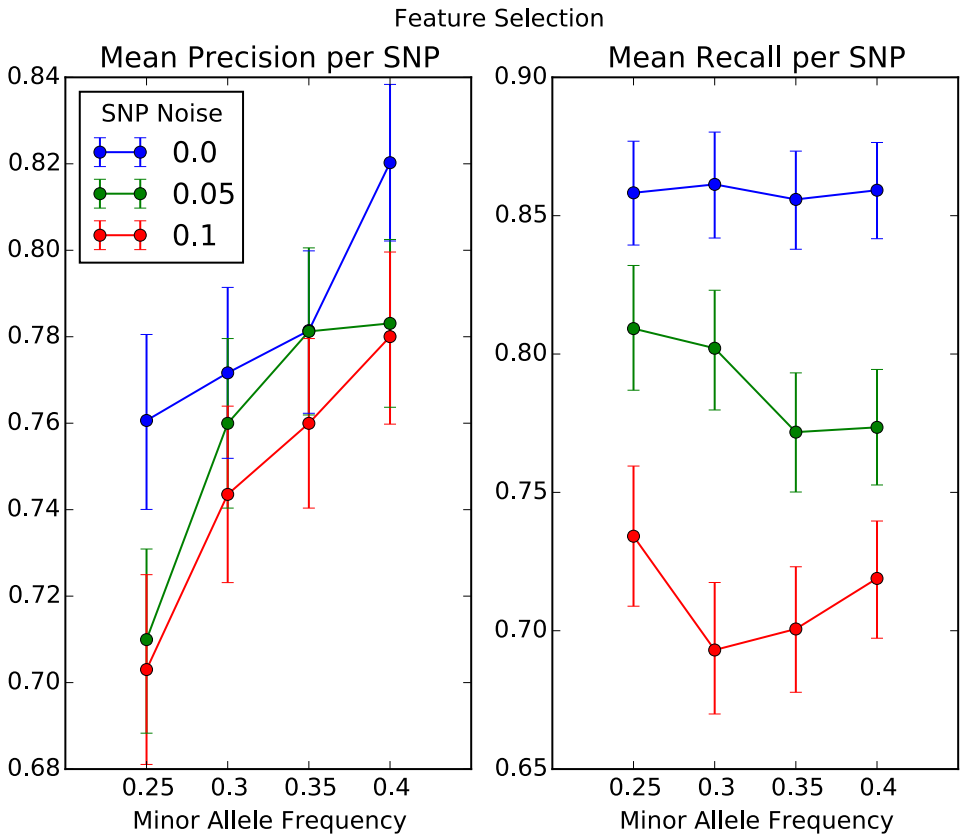

Figure 3.3: Mean precision and recall per SNP of feature selection with 95% confidence intervals in synthetic dataset 1.

Synthetic Dataset 2

We constructed this dataset to observe the effects of varying correlated taxon count and SNP noise on LASSO regression results. This dataset consisted of 9 subsets of 500 correlated SNPs constructed with all pairings of correlated taxon counts 5, 10, and 20 with SNP noise levels of 0.00, 0.05, and 0.10. The SNP minor allele frequency in all subsets was fixed at 0.30 and total taxon count was 200.

Synthetic dataset 2 does not include uncorrelated SNPs since the same parameters are represented by the uncorrelated SNPs in dataset 1 with minor allele frequency 0.3.

Figure 3.4 shows sensitivity as a function of SNP noise and correlated taxon count for our LASSO procedure on synthetic dataset 2. Increasing SNP noise and increasing correlated taxon count cause decrease in sensitivity. This is most likely due to compounding noise with increasing number of noisy SNPs. This suggests our method will be sensitive to noise especially when SNPs are correlated with higher numbers of taxa. With higher numbers of correlated taxa the overall strength of correlation falls because our synthetic taxa are random and so will not tend to reinforce each other.

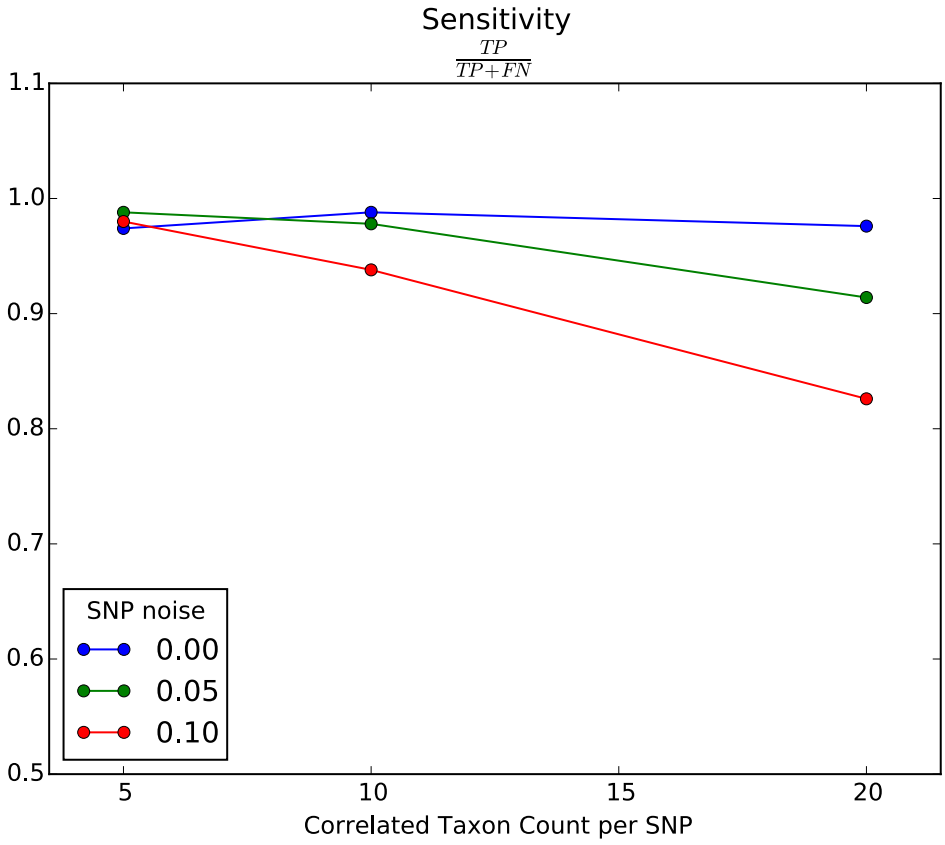

Figure 3.4: Sensitivity as a function of correlated taxon count per SNP in synthetic dataset 2.

Feature selection precision and recall are also strongly influenced by the correlated taxon count per SNP (Figure 3.5). Precision falls to about $1/2$ at 10 correlated taxa and about $1/3$ at 20 correlated taxa. SNP noise has a consistent but small effect on

precision. Recall also falls with increasing numbers of correlated taxa, but stays above 0.5 even at high SNP noise.

Synthetic dataset 2 is very challenging for our method. One reason may be that the random abundances tend to cancel each other's effects and so as more abundances are combined to correlate with a SNP the more difficult it becomes to select the relevant taxa. This behavior may not be representative of actual correlations between microbiome data and host genetic variation.

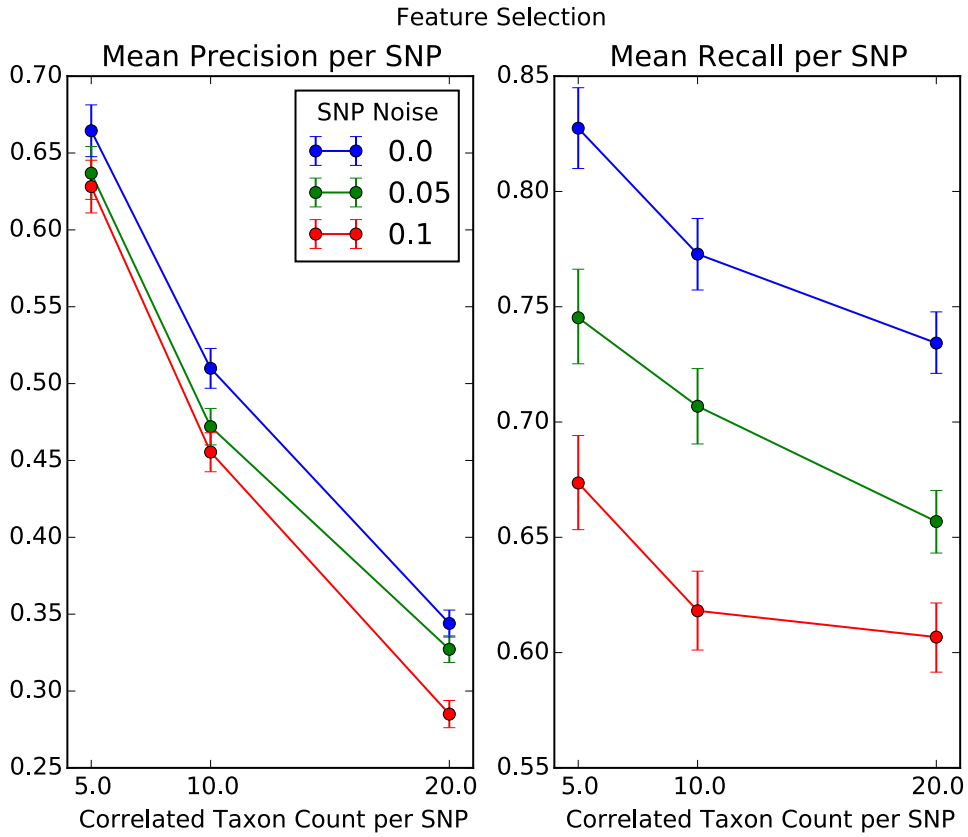

Figure 3.5: Mean feature selection precision and recall per SNP with 95% confidence intervals for synthetic dataset 2.

Synthetic Dataset 3

We constructed this dataset to investigate LASSO regression with varying total taxon count and SNP noise. This dataset consisted of 9 subsets of 500 correlated SNPs

generated from all pairings of total taxon counts 100, 300, and 500 and SNP noise levels of 0.00, 0.05, and 0.10. Three subsets of 10,000 uncorrelated SNPs were generated, one for each total taxon count. The SNP minor allele frequency in all subsets was fixed at 0.30.

Figure 3.6 and Figure 3.7 show sensitivity and specificity as functions of total taxon count. There is some fluctuation in sensitivity but no clear dependence on total taxon count or SNP noise. Specificity is consistent across the different taxon counts.

This means our method is good at identifying SNPs correlated with taxon abundances even at larger taxon counts and in the presence of noise.

Figure 3.8 shows mean precision and mean recall per SNP for feature precision on synthetic dataset 3. The precision of feature selection is largely unaffected by the total taxon count and SNP noise has a consistent but small effect. Feature selection recall is sensitive to increasing taxon table size and to increasing SNP noise.

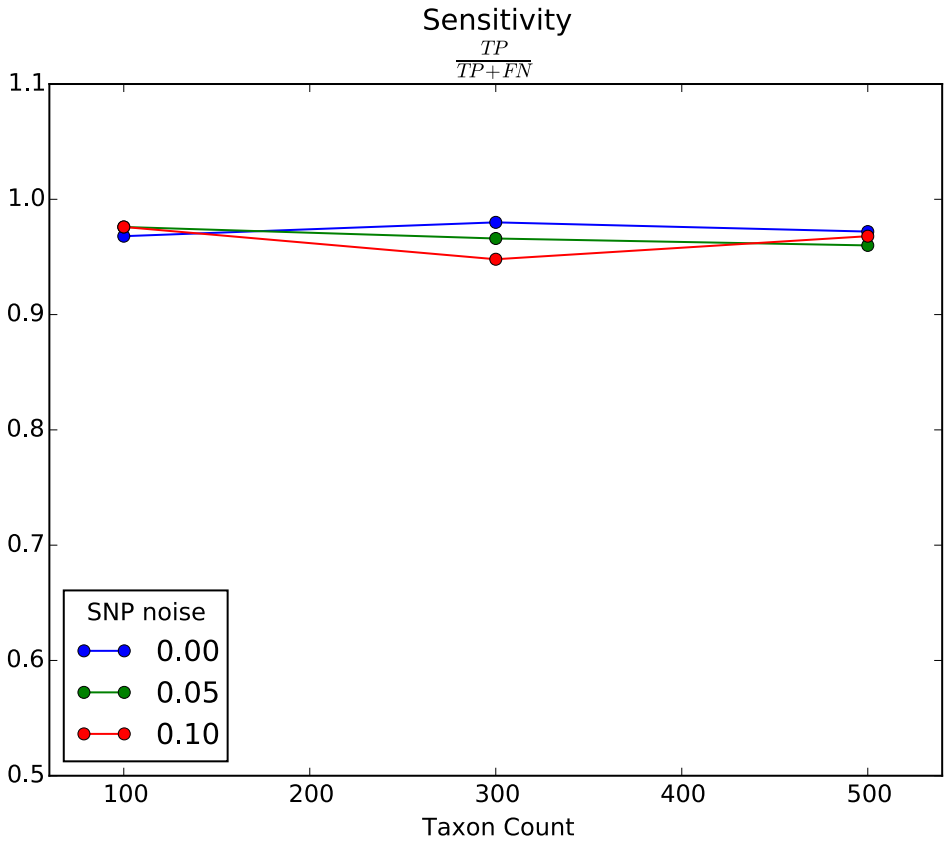

Figure 3.6: Sensitivity as a function of total taxon count in synthetic dataset 3.

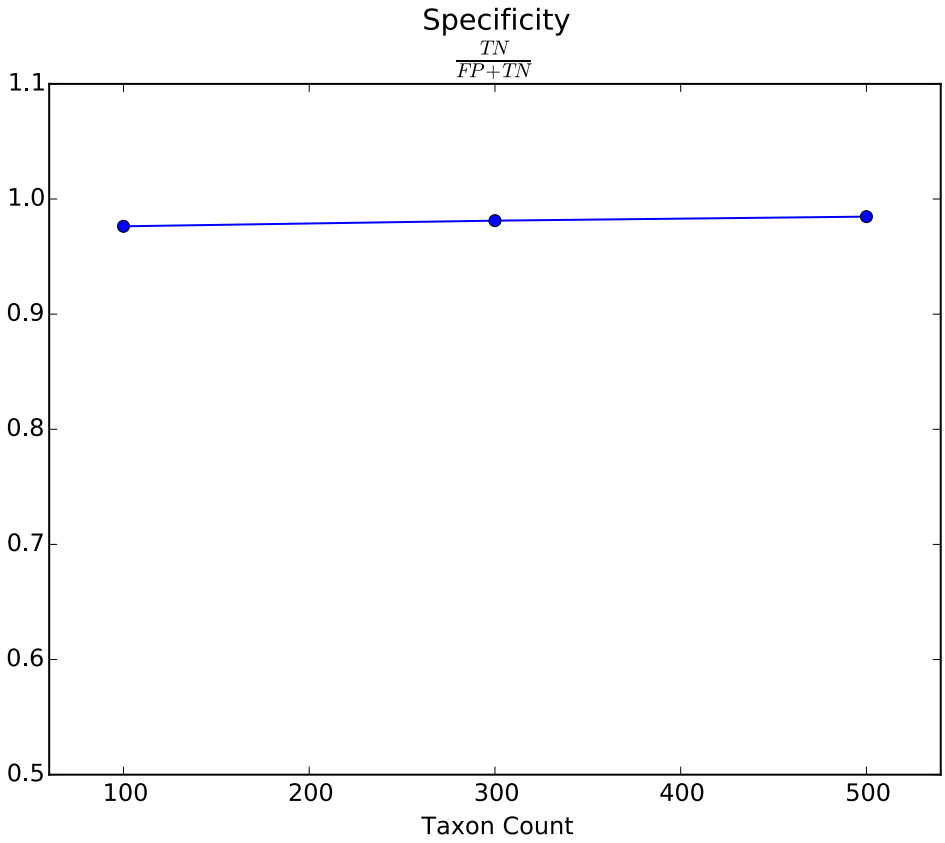

Figure 3.7: Specificity as a function of total taxon count in synthetic dataset 3.

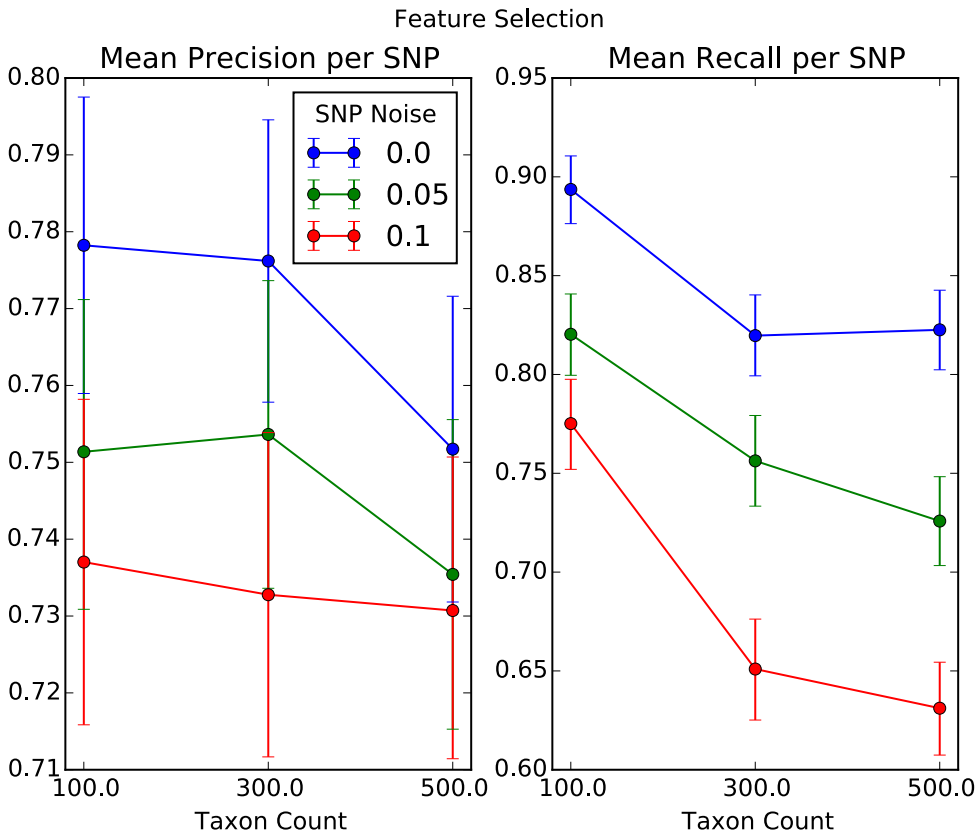

Figure 3.8: Mean feature selection precision and recall per SNP with 95% confidence intervals for synthetic dataset 3.

Discussion

The synthetic datasets are challenging for several reasons. The synthetic abundances span a wide range of values. The relationship between abundances and alternate allele counts created by the f_i functions is indirect and the coefficients of the f_i are allowed to be positive or negative and a few will have very small magnitude because they are drawn from a normal distribution. Nevertheless, our procedure performs well at identifying correlated SNPs even under moderate SNP noise. Identifying the correlated taxa is a more difficult problem but this method is almost always able to identify correctly a third or more of the correlated taxa in the synthetic datasets.

We will use the results of the synthetic data studies to inform our work on real data from the Human Microbiome Project. For example, we will preprocess our real data to reduce

the number of taxa as much as possible. We will also filter out SNPs with low minor allele frequency.

4. HMP 16S Results

The Human Microbiome Project (Human Microbiome Project Consortium, 2012) has collected and published microbiome data from several hundred healthy individuals derived from 16S sequencing for 15 body sites and from shotgun metagenomic sequencing for 5 body sites. In addition, the genomes of 93 of these subjects have been sequenced with 10x coverage from human genetic material present in the same published samples providing data on 3 million SNPs. We used LASSO regression to search for correlations between 32,698 SNPs in exons from these host genomes and microbial abundances.

We acquired OTU tables based on 16S sequencing from the Human Microbiome Project that identified microbial taxa at the genus level. We processed these tables with the QIIME (Caporaso, 2010) script 'summarize_taxa.py' to produce taxon tables of relative abundance for all taxonomic levels from genus to phylum. We also removed taxa with very low abundance (< 0.00005%). The remaining taxon abundances were clustered at using K-medoids clustering with a clustering threshold of 0.95 correlation. We retained the medoid taxon of each cluster for the final taxon table. Taxon abundances were transformed with the square root function followed by the inverse sin function. Figure S1 shows the final taxon count for each body site.

Starting with 32,698 exonic SNPs, we filtered the data to remove SNPs with five or fewer homozygous samples, fewer than 50 samples overall, and minor allele frequency less than 0.2. At each body site approximately 14,000 SNPs met the criteria for testing. There is some variation across body sites in the number of tested SNPs because a small number of subjects do not have abundance data at all 15 body sites. Figure S2 shows the number of SNPs that met the criteria for testing by body site.

In addition we controlled for subject sex by including it as one of the predictor variables. Adding sex to the model rather than regressing it out of the predictors allows the LASSO regression process to determine if sex is a relevant variable on a SNP-by-SNP basis.

The first step is to test for correlation using LASSO regression. This step produces a median R^2 and 95% bootstrap confidence interval for each SNP, body site pair. Of the

tested SNPs, we found 250-500 SNPs per body site having 95% confidence interval of the median R^2 not overlapping 0. We retained these SNPs for further analysis. Figure S3 shows counts of retained SNPs by body site.

The second step is to estimate the false positive rate for the SNP and abundance data. If we find more SNPs correlated with taxon abundances than the false positive rate predicts then we can be confident we have found some meaningful correlations.

We estimate the false positive rate for each body site by twice repeating the LASSO regression procedure on each SNP with the genotypes randomly permuted within the male and female groups. Permuting genotypes across the sexes would result in an underestimate of the false positive rate because sex is a significant predictor for a number of SNPs and we consider those to be false positive identifications. SNPs with a high correlation to sex (stability score > 0.8) were removed from the final list of SNPs correlated with taxon abundances.

We selected $R^2 = 0.1$ as a reasonable cutoff for considering a SNP to be well-correlated with the HMP 16S microbiome data. We then compared the number of well-correlated SNPs in the real data with the number of well-correlated SNPs in the permuted data.

Figure 4.1 shows, for the stool body site, the superimposed distributions of median R^2 values for the real data and the average of the two distributions of permuted data. It is clear that this body site has an enrichment of real correlated SNPs relative to the permuted correlated SNPs. Supplemental figures S4-S17 show the superimposed distributions for the remaining HMP 16S body sites.

The body sites most enriched for real well-correlated SNPs are attached keratinized ginigiva, buccal mucosa, right antecubital fossa, saliva, stool, subgingival plaque, and throat. The hard palate, palatine tonsils, and tongue dorsum are only slightly enriched. The anterior nares, left antecubital fossa, left retroauricular crease, right retroauricular crease, supragingival plaque, and supragingival plaque are not enriched for well-correlated SNPs.

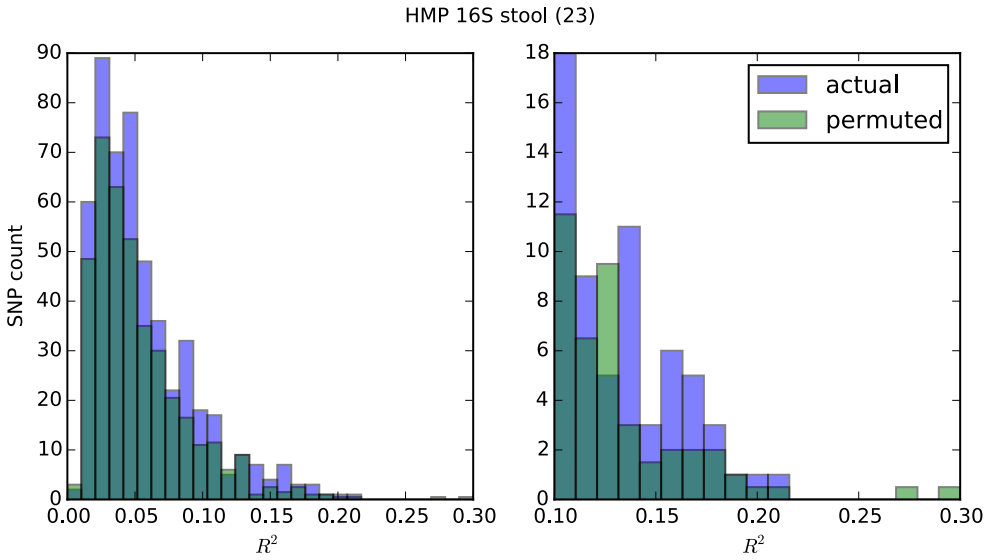

Figure 4.1: Superimposed distributions of R^2 for actual and permuted data from the stool body site. The figure on the left shows the distribution of SNPs with $R^2 > 0$. The figure on the right shows the distribution of SNPs with $R^2 > 0.1$. It is clear in both figures that this body site is enriched for actual correlated SNPs.

Finally, individual well-correlated SNPs were identified as having significant correlation to the HMP 16S microbiome data using the ratio r of median R^2 to the width of the corresponding 95% confidence interval. If a body site had N well-correlated SNPs over the number predicted by the false positive rate then the top N SNPs ordered by r were identified as well-correlated. Identified, well-correlated SNPs and corresponding genes are listed in file 'hmp_16S_snps_genes.xlsx'. SNPs with high correlation (stability score > 0.8) to subject sex were removed. The SNPs found to have a high correlation to subject sex were in genes ALMS1, FAM20A, GLYR1, LIPC, and PRAME. In all, 14 SNPs were removed due to high correlation to subject sex leaving 76 SNPs in 72 genes. Figure 4.2 shows the final count of identified SNPs.

Microbial taxa associated with well-correlated genes were identified by randomized LASSO with stability score threshold 0.5. This threshold is a conservative choice that identifies at least one microbial taxon per well-correlated SNP.

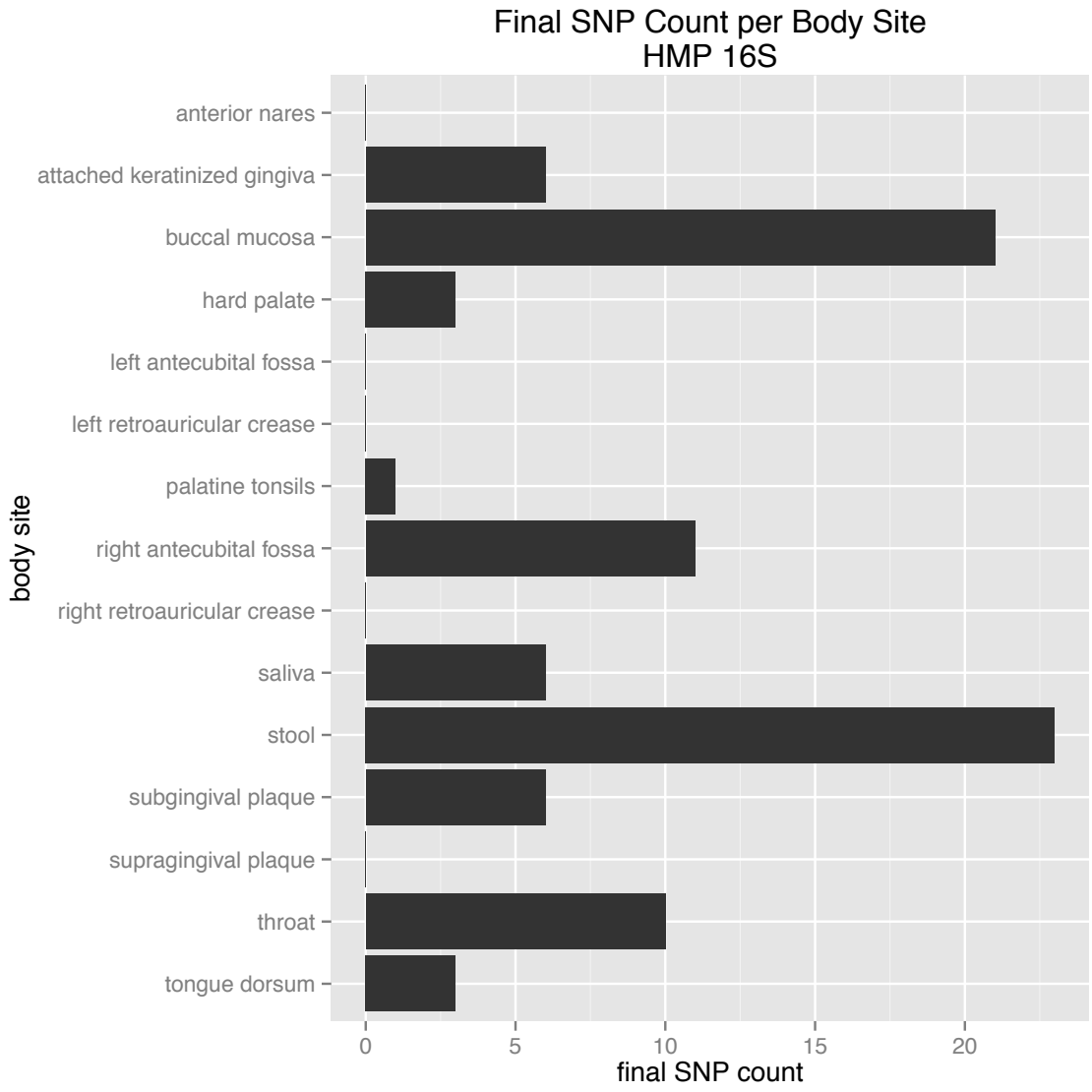

Figure 4.2: Count of SNPs per body site identified as well-correlated with HMP 16S microbial abundance.

Pathway and Network Analysis

We used QUIGEN's Ingenuity Pathway Analysis (IPA, QUIGEN Redwood City, www.quiagen.com/ingenuity) to perform pathway and network analysis on the list of genes containing identified, well-correlated SNPs (the complete list of SNPs and genes is found in supplementary file hmp_16S_snps_genes.xlsx). Figure S18 shows the results of the pathway analysis.

The list was found to be significantly enriched for five pathways:

1. Phospholipases
    a. PLA2G12A from the throat body site
    b. PLD2 from the saliva body site
2. Tryptophan Degradation to 2-amino-3-carboxymuconate Semialdehyde
    a. IDO2 from the buccal mucosa body site
3. Actin Cytoskeleton Signaling
    a. Gene APC from the subgingival plaque body site
    b. Gene MYH7B from the throat body site
    c. Gene PAK7 from the right antecubital fossa body site
4. NAD biosynthesis II (from tryptophan)
    a. Gene IDO2 from the buccal mucosa body site
5. Chondroitin Sulfate Degradation (Metazoa)
    a. Gene ARSB from the buccal mucosa body site
6. Choline Biosynthesis III
    a. Gene ARSB from the buccal mucosa body site
7. Antioxidant action of vitamin C
    a. Gene PLA2G12A from the throat body site
    b. Gene PLD2 from the saliva body site

However, each pathway included only one gene per body site from the list of selected genes. We cannot conclude that our list of selected genes is enriched for any pathways. We can instead consider each SNP as potentially perturbing pathways to which it belongs.

The list of genes was found to overlap two gene networks by more than one gene:

1. Cell-to-Cell Signaling and Interaction, Hematological System Development and Function, Immune Cell Trafficking (15 genes)
2. Gene Expression, Connective Tissue Development and Function, Tissue Development (9 genes)

Figures S19 and S20 show these networks.

Example Genes

IDO2 (synonym: INDOL1)

The non-synonymous SNP rs10109853 in gene IDO2 was found to be associated with microbial abundance of Firmicutes;Clostridia;Clostridiales;Veillonellaceae;Dialister and Firmicutes;Bacilli;Lactobacillales in the buccal mucosa body site. This gene encodes an enzyme that contributes to tryptophan metabolism (Ball et al., 2007). This enzyme participates in three host metabolic pathways: Tryptophan Degradation to 2-amino-3-carbosymuconate Semialdehyde, NAD biosynthesis II (from tryptophan), and Tryptophan Degradation III (Eukaryotic). Figures 4.3, 4.4, and 4.5 show associated microbial abundances and median abundances as functions of genotype.

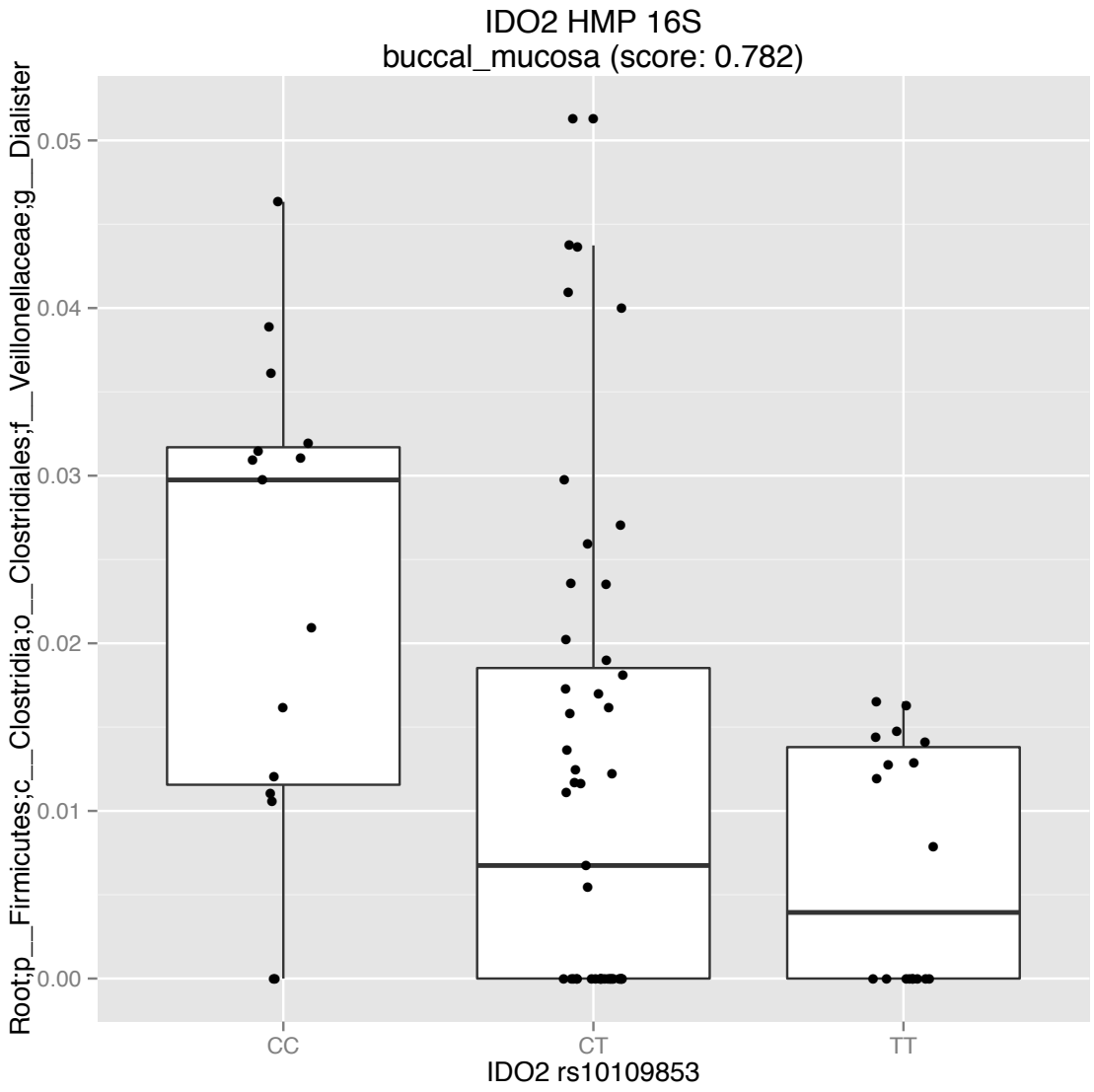

Figure 4.3: Transformed microbial abundance for V. Dialister as a function of genotype. This taxon is associated with gene IDO2.

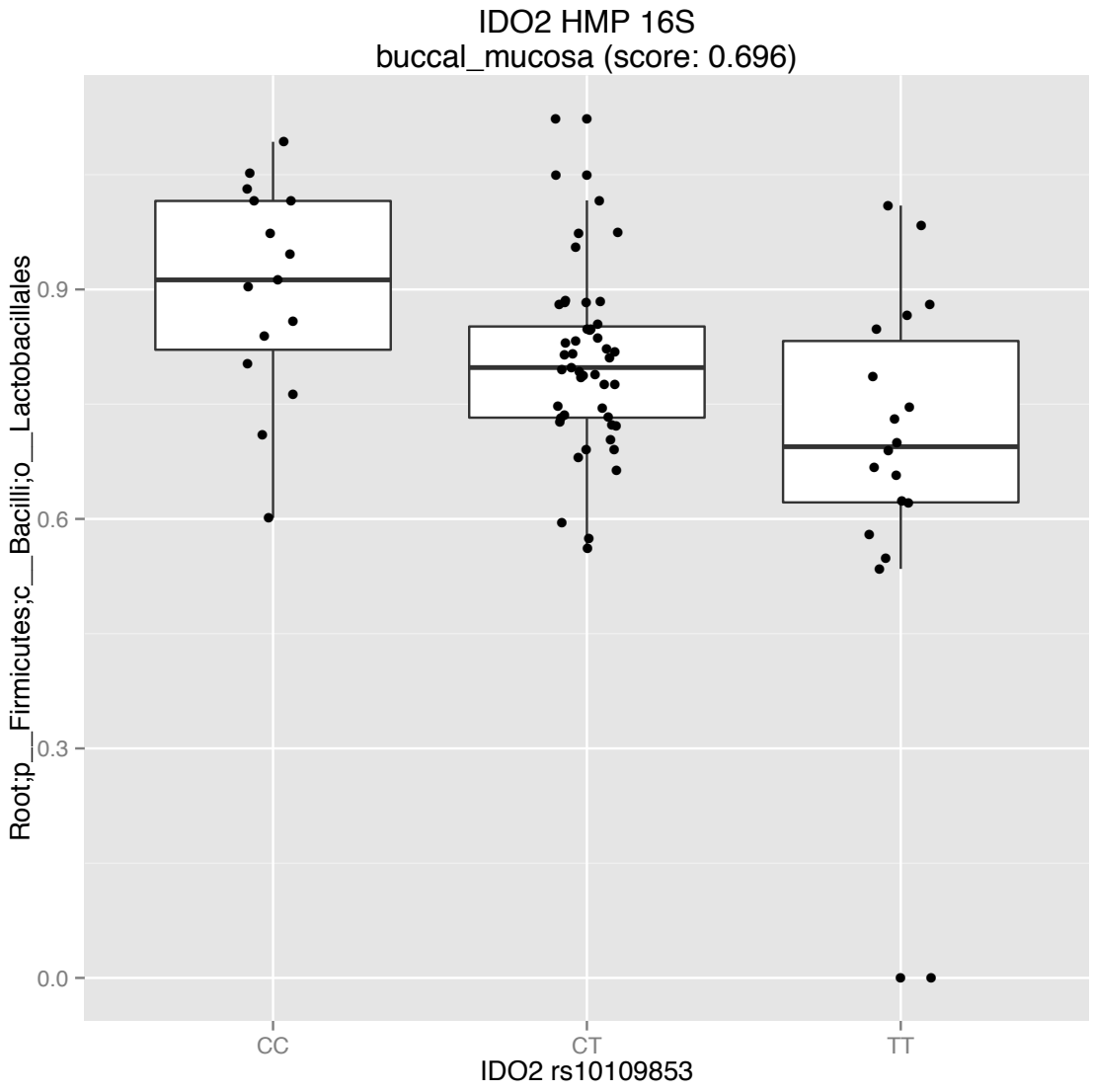

Figure 4.4: Transformed microbial abundance for order Lactobacillales as a function of genotype. This taxon is associated with gene IDO2.

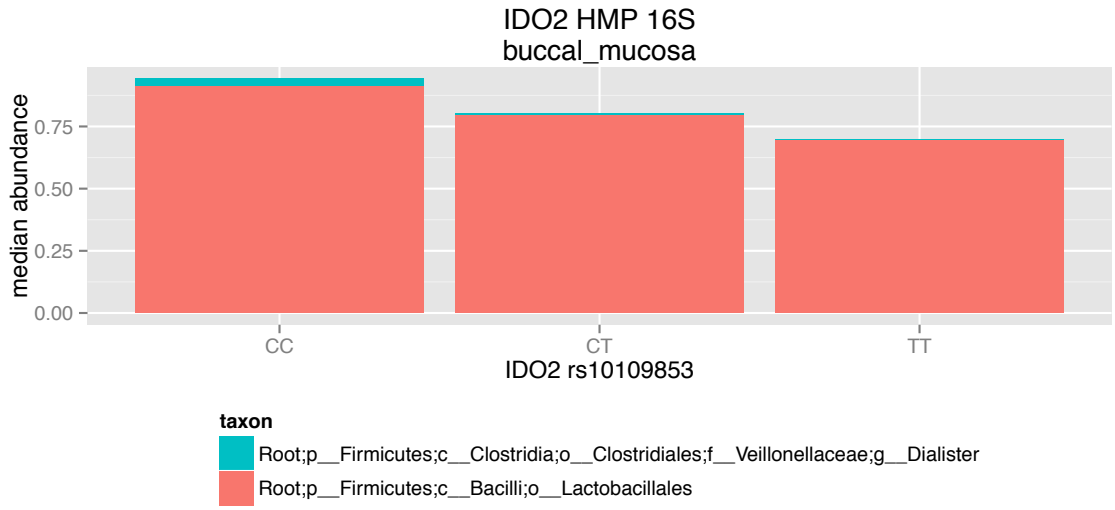

Figure 4.5: Transformed median microbial abundances selected for SNP rs10109853 in gene IDO2 by genotype.

ARSB

The non-synonymous SNP rs1065757 in the gene ARSB was found to be associated with microbial taxa in the buccal mucosa body site:

1. p__Cyanobacteria
2. p__Proteobacteria;c__Gammaproteobacteria,
3. p__Bacteroidetes;c__Flavobacteria;o__Flavobacteriales;f__Flavobacteriaceae;g__Capnocytophaga,
4. p__Tenericutes,
5. p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;Other

This gene participates in the 'Chondroitin Sulfate Degradation (metazoa)' pathway. Figures 4.6, 4.7, 4.8, 4.9, and 4.10 show the associated microbial abundances and median abundances of four of the five microbial taxa found to associate with SNP rs1065757 in ARSB.

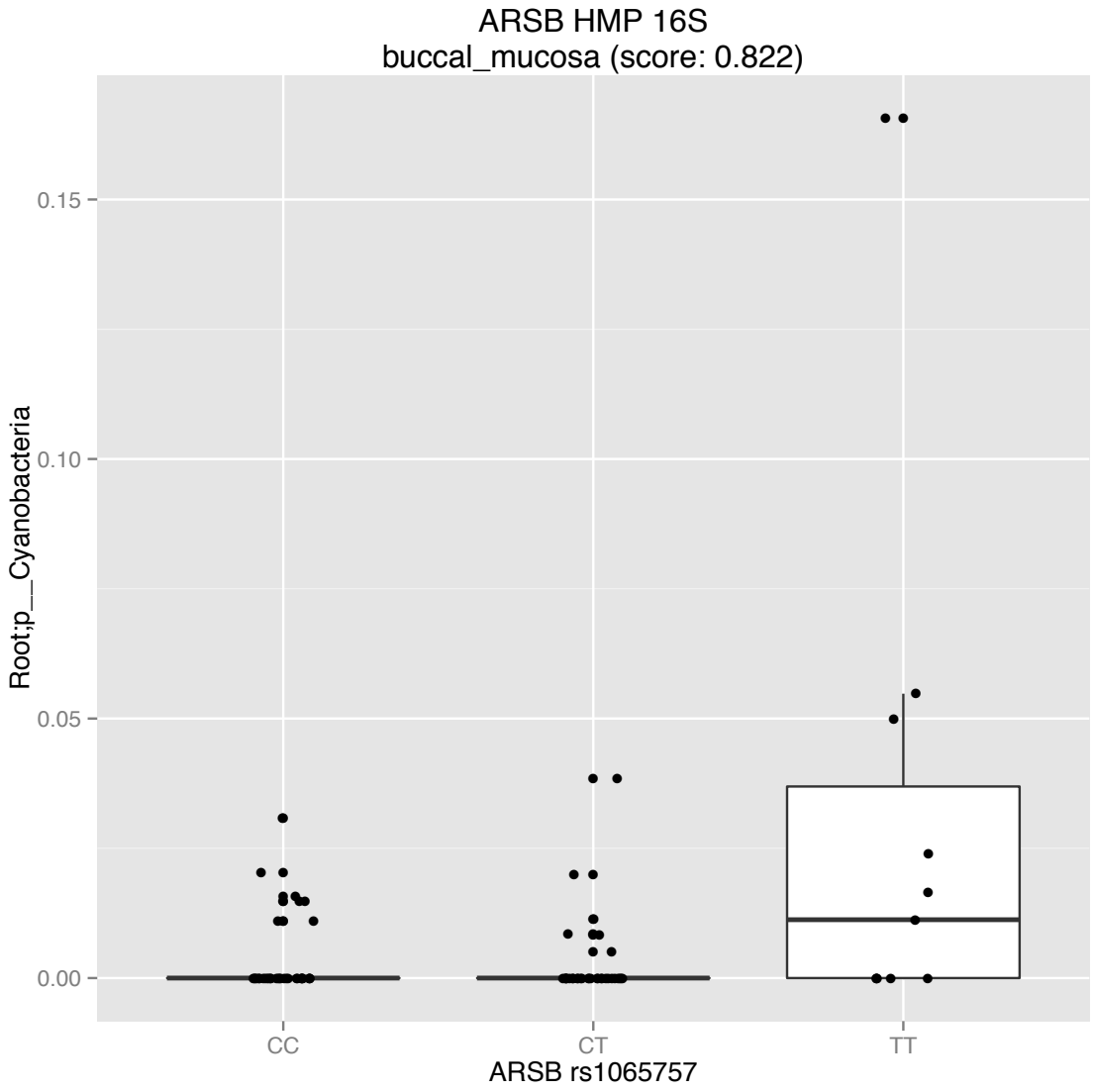

Figure 4.6: Transformed microbial abundance for Cyanobacteria as a function of genotype. This taxon is associated with gene ARSB.

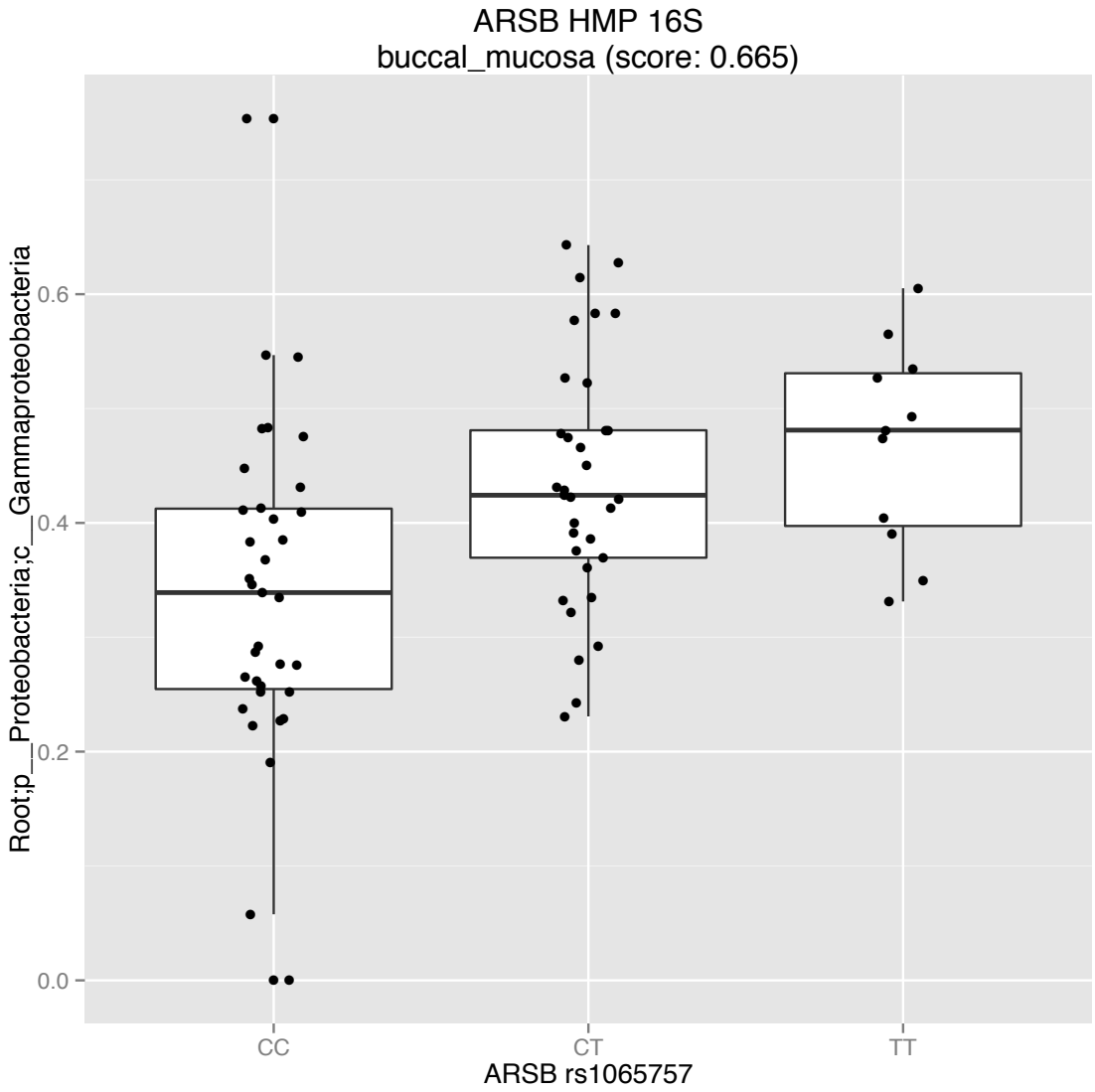

Figure 4.7: Transformed microbial abundance for Gammaproteobacteria as a function of genotype. This taxon is associated with gene ARSB.

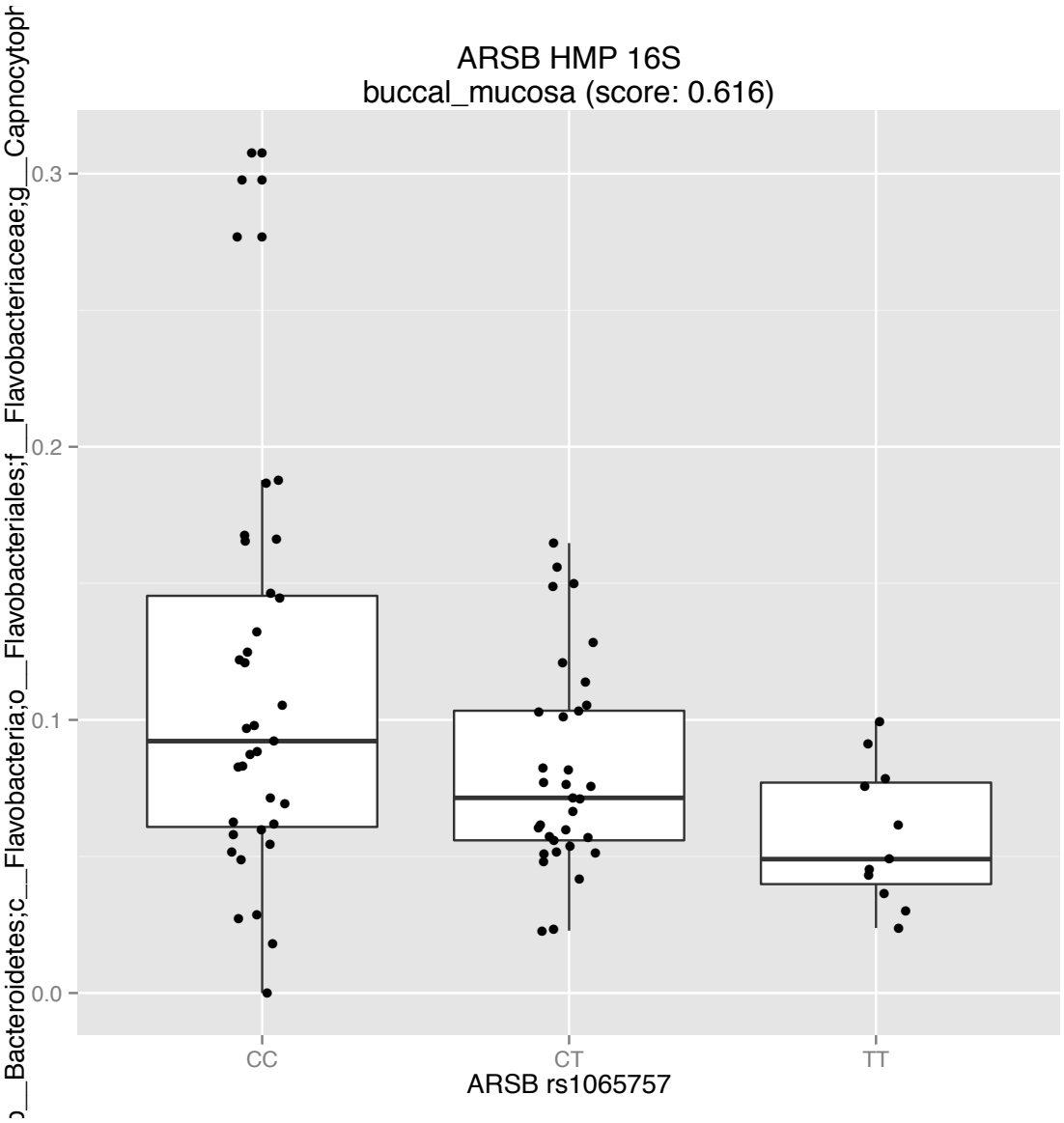

Figure 4.8: Transformed microbial abundance for Capnocytophaga as a function of genotype. This taxon is associated with gene ARSB.

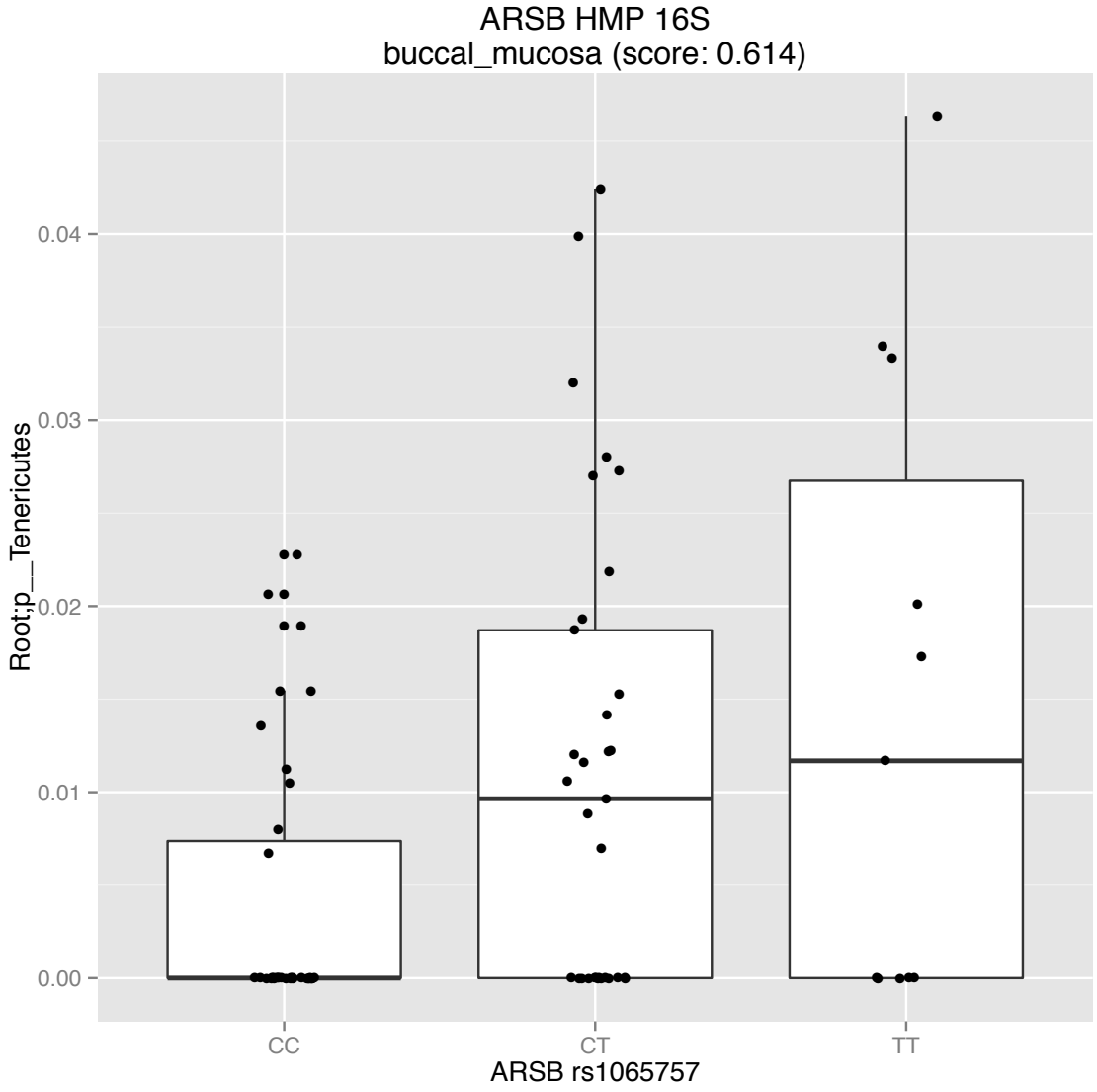

Figure 4.9: Transformed microbial abundance for Tenericutes as a function of genotype. This taxon is associated with gene ARSB.

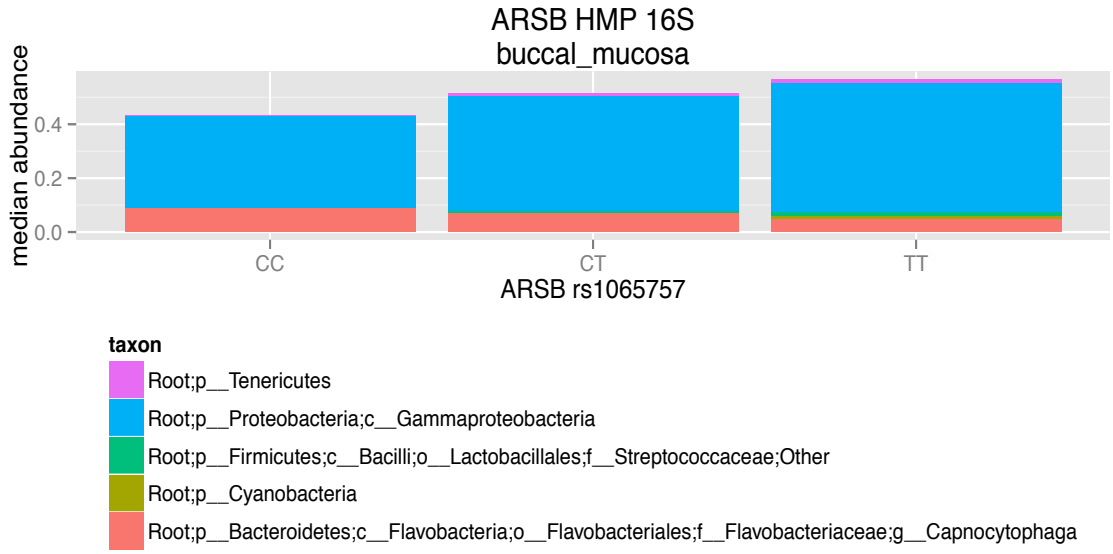

Figure 4.10: Transformed median microbial abundances selected for SNP rs1065757 in gene ARSB by genotype.

Pathway and Network Analysis of Permuted 16S Data

As a check on the validity of our estimated false positive rate, we analyzed the top genes identified in the permuted 16S data for pathway and network enrichment. We compiled a list of the top SNPs by ratio of median R^2 to width of the corresponding 95% confidence interval. We retained the same number of SNPs from each body site as were retained in the real data. Finally we removed SNPs with very high correlation to subject sex. As with the real 16S data, most of the SNPs highly correlated to subject sex in the permuted data were in genes ALMS1 and FAM20A. LUZP2 and C2orf16 were also found to be highly associated with subject sex. The filtered gene list included 64 unique genes. This list is found in the spreadsheet hmp_16S_snps_genes.xlsx.

Pathway analysis of the permuted 16S data reported five significant pathways but none included more than 4 identified genes. The 'Actin Cytoskeleton Signaling' pathway was identified in the permuted 16S data and in the real 16S data. This pathway includes 210 genes.

Network analysis of the genes identified by the permuted 16S data reported two gene networks with more than 1 overlapping gene:

1. Cellular Movement, Cellular Development, Cellular Growth and Proliferation
2. Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry

These networks were not identified by the network analysis of the actual 16S data.

5. HMP MGS KEGG Modules Results

In addition to 16S sequencing, the Human Microbiome Project has released results of metagenomic shotgun sequencing on the samples from five body sites taken from a subset of subjects. The metagenomic data has been processed using HUMANn (Abubucker et al., 2012) to produce tables of KEGG module and pathway abundance.

We acquired KEGG module abundance tables from the HMP DACC. For each body site 250 KEGG module abundances are reported. We did not preprocess the module abundances other than by applying the square-root-arcsin transformation. We again included subject sex as part of the model.

We applied our analysis pipeline to test for correlations between the table of transformed KEGG module abundances and the set of 32,696 exonic SNPs. As before we filtered the data to remove SNPs with five or fewer samples of a single genotype, SNPs with fewer than 50 samples overall, and SNPs with minor allele frequency less than 0.2. Approximately 14,000 SNPs per body site remained after filtering. Figure S21 shows the number of SNPs per body site that met the criteria for testing. After testing we found 300-400 SNPs per body site having 95% confidence interval of the median R^2 not overlapping 0. We retained these SNPs for further analysis. Figure S22 shows the counts of retained SNPs by body site.

We estimated the false positive rate for each body site by running the LASSO regression program twice on each SNP after randomly permuting the SNP allele counts within the male and female groups. Those SNPs with a high correlation to sex were removed from the final list of SNPs correlated to KEGG module abundances. Figure 5.1 shows the distribution of median R^2 values for both the real data and the average of the two distributions of the permuted data. Supplemental figures S23-S26 show the same distributions for the remaining four body sites.

Microbial taxa associated with well-correlated genes were again identified by randomized LASSO with stability score threshold 0.5.

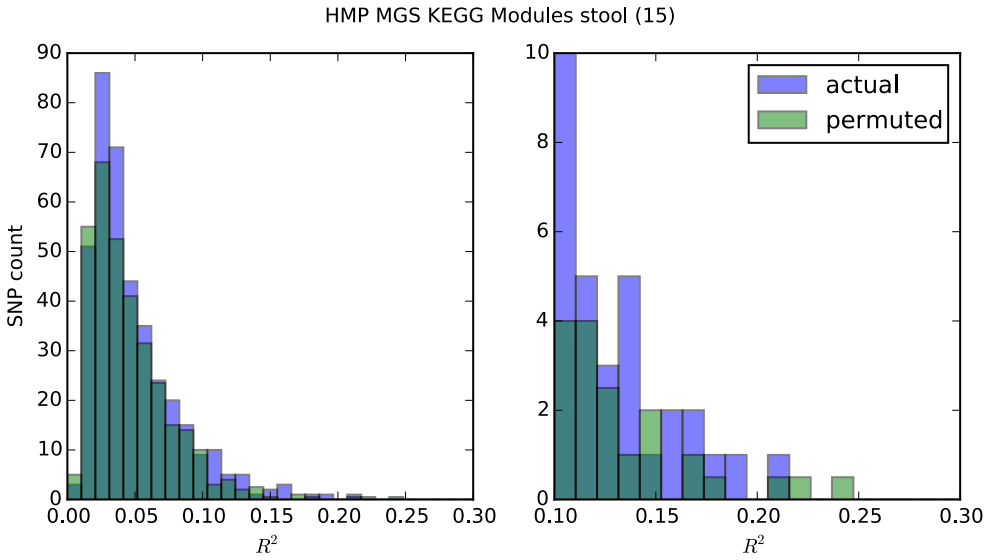

Figure 5.1: Superimposed distributions of R^2 for actual and permuted data from the stool body site. The figure on the left shows the distribution of SNPs with $R^2 > 0$. The figure on the right shows the distribution of SNPs with $R^2 > 0.1$. It is clear in both figures that this body site is enriched for actual correlated SNPs.

We again selected $R^2 = 0.1$ as a reasonable cutoff for considering a SNP to be well-correlated with the HMP MGS KEGG module abundances. The body sites most enriched for actual well-correlated SNPs were stool and tongue dorsum. The supragingival plaque body site was only slightly enriched. The anterior nares and buccal mucosa sites had no enrichment of well-correlated SNPs.

Individual SNPs with high ratio r of median R^2 to the width of the corresponding 95% confidence interval were identified as being well-correlated to the HMP MGS KEGG module abundances. The supplemental spreadsheet 'hmp_mgs_kegg_modules_snps_genes.xlsx' lists the identified SNPs and associated genes. Figure S27 shows the final counts of identified SNPs per body site.

Pathway and Network Analysis

We used QUIGEN's Ingenuity Pathway Analysis (IPA, QUIGEN Redwood City, www.quiagen.com/ingenuity) to perform pathway and network analysis on the list of genes in hmp_mgs_kegg_modules_snps_genes.xlsx. Figure S28 shows the results of pathway analysis.

The list was found to be enriched for six pathways:

1. Pregnenolone Biosynthesis (gene MICAL3)
2. Histidine Degradation VI (gene MICAL3)
3. Ubiquinol-10 Biosynthesis (Eukaryotic) (gene MICAL3)
4. Mitochondrial L-carnitine Shuttle Pathway (gene CPT2)
5. Coagulation System (gene PLG)
6. Neuroprotective Role of THOP1 in Alzheimer's Disease (gene PLG)

However, as with the results from 16S data these pathways include only one gene each from the list of selected genes. As before we cannot conclude that our list of selected genes is enriched for these pathways.

The list of genes was found to overlap seven gene networks:

1. Cancer, Cell Cycle, Cell Morphology (gene PLEKHG2)
2. Cellular Assembly and Organization, DNA Replication, Recombination, and Repair, Cell Cycle (gene RAI14)
3. Cellular Movement, Cell-To-Cell Signaling and Interaction, Cellular Development (gene TJP3)
4. Cellular Growth and Proliferation, Developmental Disorder, Hereditary Disorder – (gene CPT2)
5. DNA Replication, Recombination, and Repair, Infectious Diseases, Hereditary Disorder (gene ASCC3)
6. Cellular Development, Embryonic Development, Hair and Skin Development and Function (gene PSMB1)
7. Cellular Movement, Cell-To-Cell Signaling and Interaction, Cellular Growth and Proliferation (gene PLG)

However, no network overlapped the list of selected genes by more than one gene.

Example Gene

PSMB1 (proteasome subunit, beta type, 1)

The non-synonymous SNP rs12717 in gene PSMB1 was found to be associated with three KEGG modules from the stool body site:

1. M00096: C5 isoprenoid biosynthesis, non-mevalonate pathway
2. M00090: Phosphatidylcholine (PC) biosynthesis, choline => PC
3. M00321: Bicarbonate transport system

This gene codes for a subunit of the proteasome, an enzyme complex that participates in degrading cellular proteins as well as MHC class 1 antigen presentation (Coux et al., 1996). Figures 5.2, 5.3, 5.4, and 5.5 show the KEGG module abundances associated with this SNP.

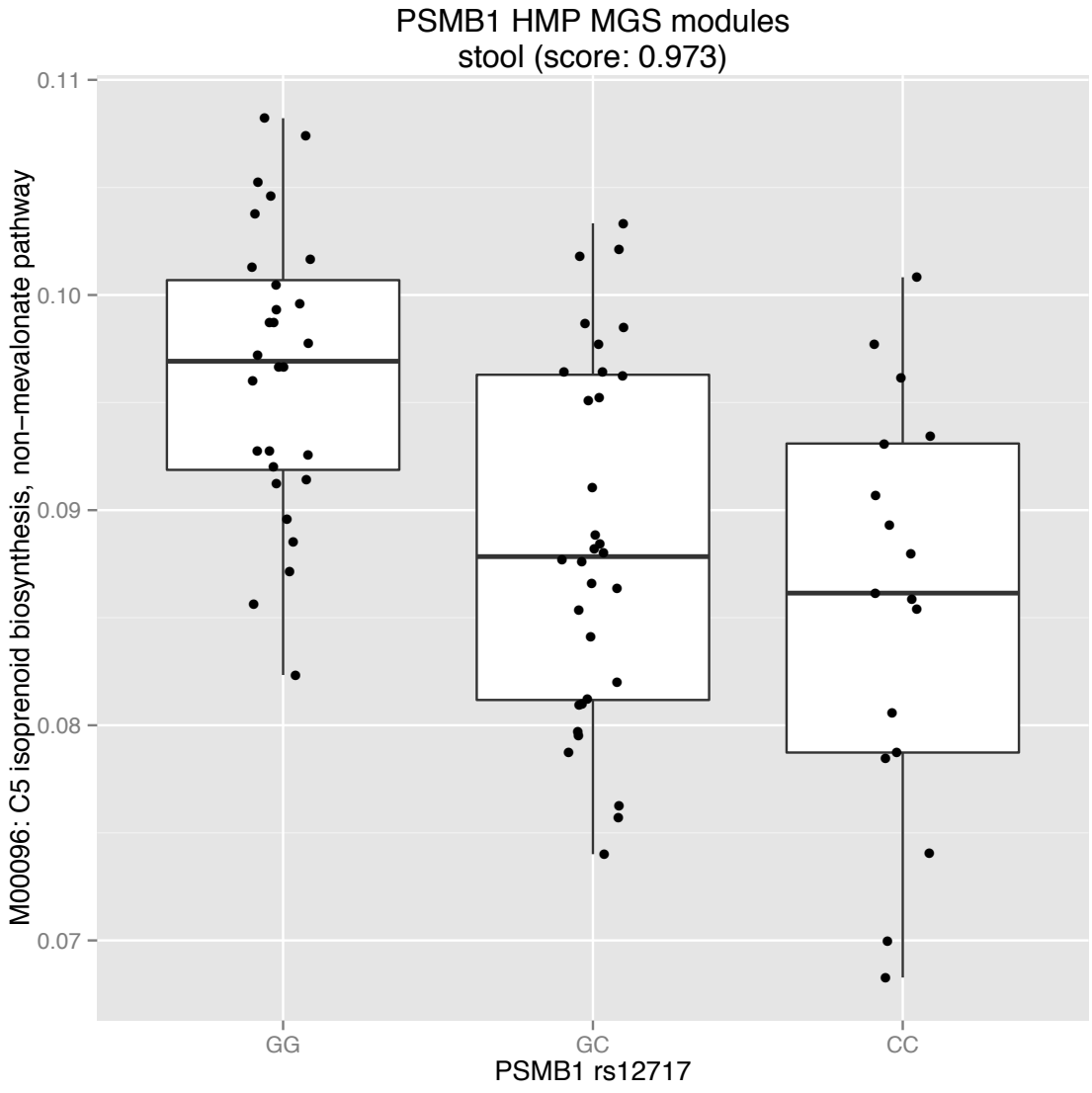

Figure 5.2: Transformed abundances for KEGG Module M00096: C5 isoprenoid biosynthesis, non-mevalonate pathway as a function of genotype. This module is associated with gene PSMB1.

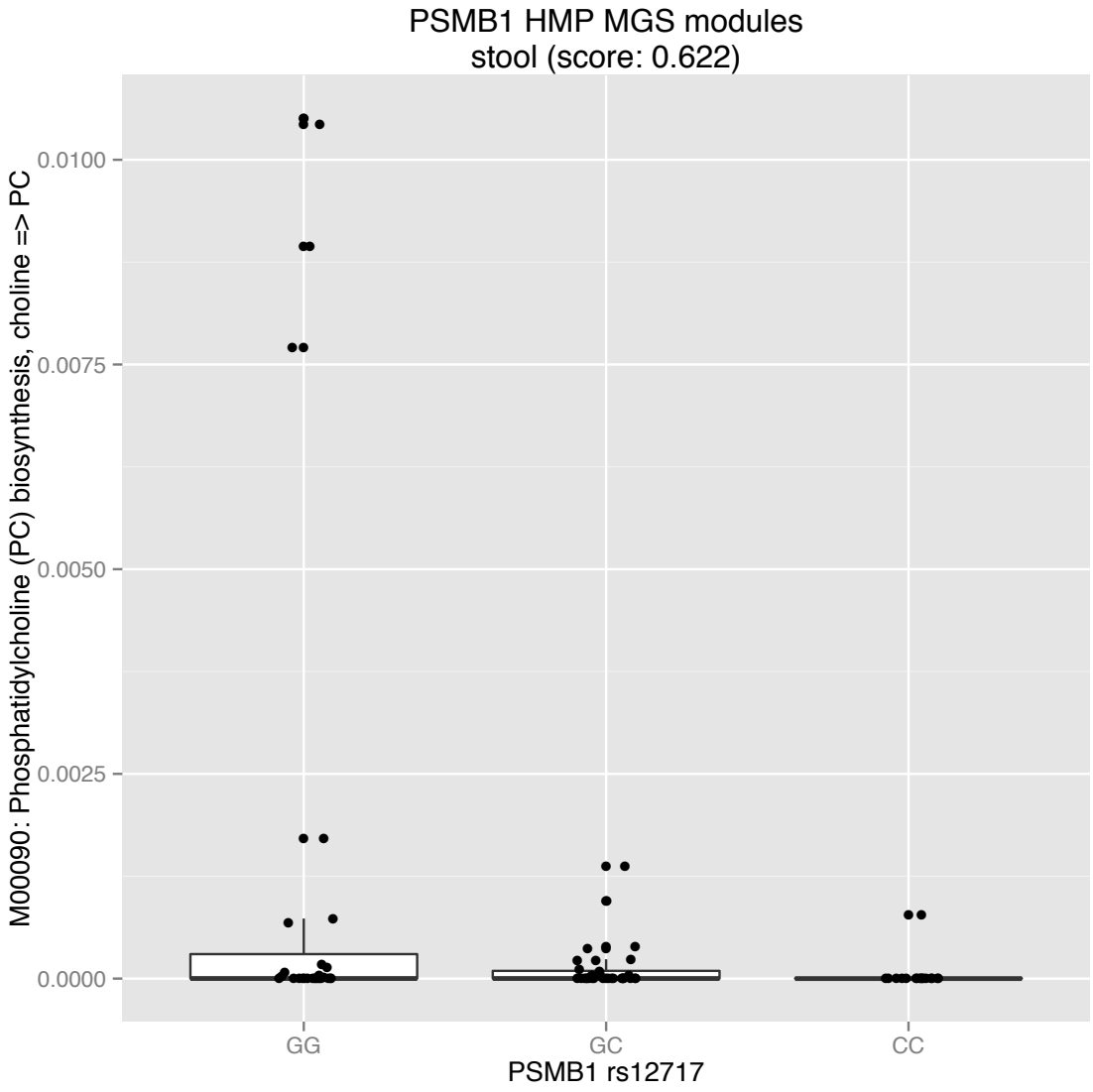

Figure 5.3: Transformed abundances for KEGG Module M00090: Phosphatidylcholine (PC) biosynthesis, choline => PC as a function of genotype. This module is associated with gene PSMB1.

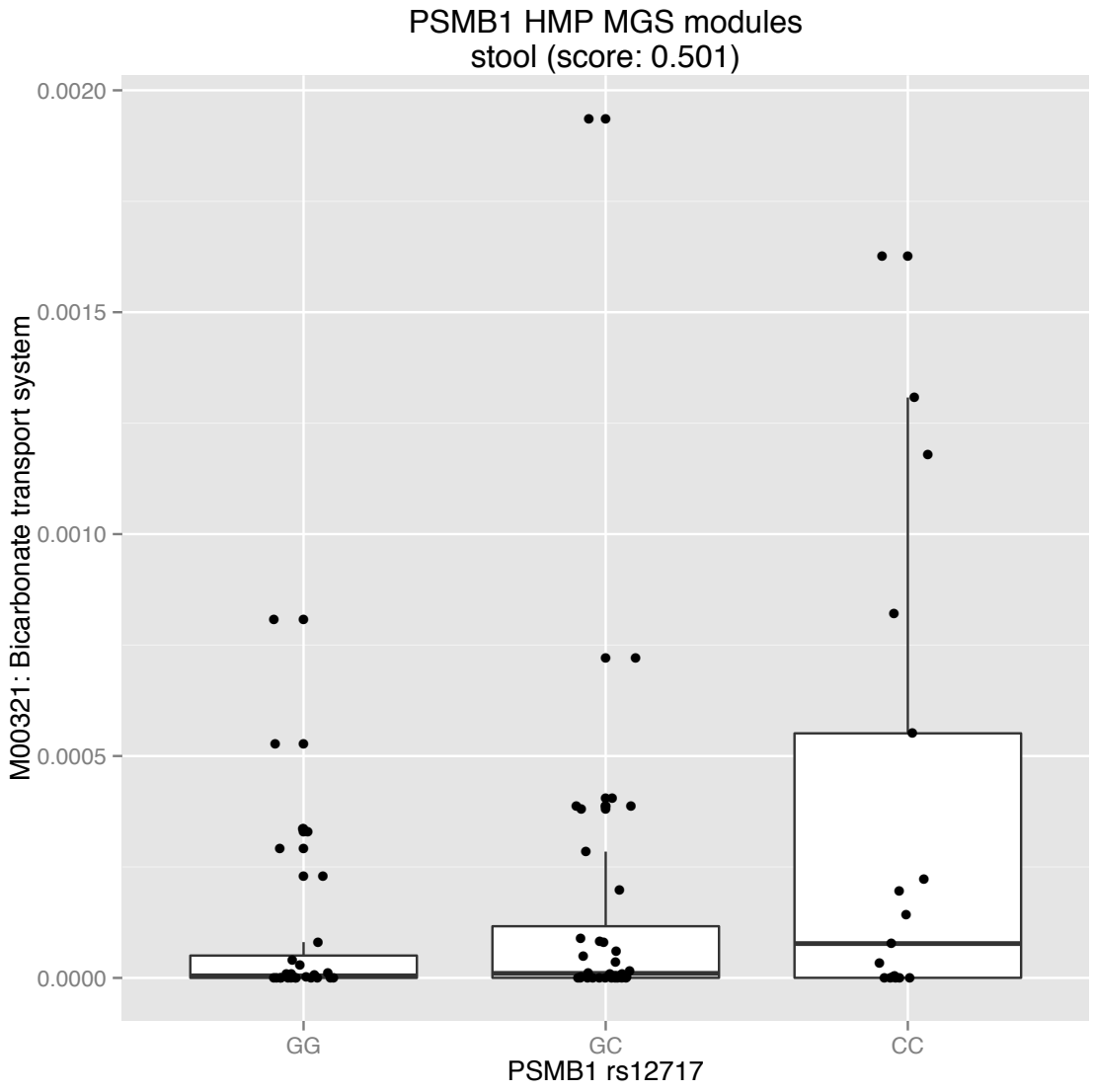

Figure 5.4: Transformed abundances for KEGG Module M00321: Bicarbonate transport system as a function of genotype. This module is associated with gene PSMB1.

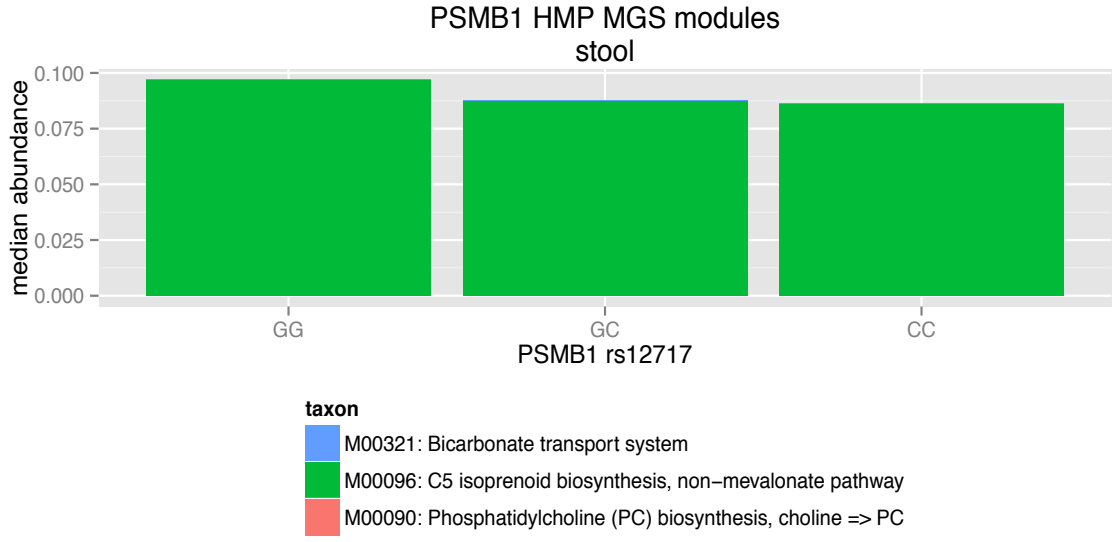

Figure 5.5: Transformed median microbial abundances selected for SNP rs12717 in gene PSMB1 by genotype. Abundances of modules M00096 and M00090 are small relative to the abundances of module M00321.

Pathway and Network Analysis of Permuted MGS KEGG Module Data

As with the permuted 16S data, we analyzed the top genes identified in the permuted KEGG module abundance data for pathway and network enrichment as a check on the validity of our estimated false positive rate. We compiled a list of the top SNPs by ratio of median R^2 to width of the corresponding 95% confidence interval from the permuted data. We retained the same number of SNPs from each body site as were retained in the real KEGG module abundance data. Finally we removed SNPs with very high correlation to subject sex (stability score > 0.8). All of the SNPs highly correlated to subject sex in the permuted data were in genes ALMS1 and FAM20A. The filtered gene list included 23 unique genes. This list is found in the spreadsheet hmp_mgs_kegg_modules_snps_genes.xlsx.

Pathway analysis of the permuted KEGG module abundance data reported five significant pathways but none included more than 2 identified genes. None of these pathways was identified by the analysis of the real KEGG abundance data.

Network analysis of the genes identified by the permuted KEGG module abundance data reported one gene network with more than 1 overlapping gene:

1. Gastrointestinal Disease, Hepatic System Disease, Liver Fibrosis

This network was not identified by the network analysis of the actual KEGG module abundance data.

6. Discussion

We have developed an analysis pipeline based on LASSO regression to screen for linear associations between human microbiome data, including abundance of taxonomic units as determined by 16S sequencing as well as KEGG module abundances derived from metagenomic shotgun sequencing, and host genetic variation in the form of SNPs.

Using synthetic data we established the ability of our pipeline to detect associations we imagine might exist in real data. We also observed limitations of our method with regard to noise in the data. Our method is robust to noise in determining that correlation exists between host genetic variation and microbiome abundance. Noise in the data has a more significant impact on feature selection.

We applied our pipeline to two different types of microbiome data that share a common set of host genetic variation data, all of which came from the Human Microbiome Project. One set of microbiome data was taxon abundance determined by 16S sequencing, the second was KEGG module abundance determined by metagenomic shotgun sequencing. While we do not find that the genes selected from the 16S data are enriched for any pathways, we find two genes each with a non-synonymous SNP that participate in pathways with potential interactions with human microbiota. The genes identified by KEGG module abundances derived from metagenomic sequencing were not significantly enriched for any pathways either. We do find a non-synonymous SNP in one gene that participates in a pathway that may be relevant to the microbiome.

One of the genes identified by the 16S data at the buccal mucosa body site is IDO2, which participates in the pathway 'Tryptophan Degradation to 2-amino-3-carboxymuconate Semialdehyde'. This pathway is present in eukaryotes and prokaryotes ("MetaCyc L-Tryptophan Degradation to 2-Amino-3-Carboxymuconate Semialdehyde." 2015), and has been implicated as an important immune system pathway that contributes to suppressing inflammation (Opitz et al. 2007). Suppressing the inflammatory response is a mechanism for maintaining commensal microbial communities (Chu and Mazmanian, 2013), and we observe a reduction in microbial abundances for the two associated taxa (family Veillonellaceae, genus *Dialister* and

order Lactobacillales) with increasing alternate allele count. The order Lactobacillales includes many organisms consumed by humans, while a species belonging to the Veillonellaceae *Dialister* genus, *D. pneumonsintes*, may be associated with periodontal disease (Doan et al., 2000). Perhaps this host-genetic variation affects the hospitality of the oral cavity toward certain subcommunities.

Another gene identified by the 16S data is ARSB, which participates in the pathway 'Chondroitin Sulfate Degradation (metazoa)'. This pathway influences connective tissue permeability and bodily fluid viscosity. Furthermore it may be promoted by bacteria in bacterial pathogenesis (Girish and Kemparaju, 2007; "Homo Sapiens Chondroitin Sulfate Degradation (metazoa)." 2015). We observe one bacterial taxon that increases in abundance with increasing alternate allele count in SNP rs1065757: genus *Capnocytophaga*. This organism is a commensal member of oral communities but can act as a pathogen causing periodontal infections (Trude et al., 2005). It is possible that variation in ARSB can give this organism a competitive advantage in the oral cavity, or it may give this organism the opportunity to colonize periodontal tissue.

In the KEGG module abundance data derived from metagenomic shotgun sequencing we found association between a non-synonymous SNP in the gene PSMB1. Of the three KEGG modules associated with this SNP, the strongest effect is seen in M00096: C5 isoprenoid biosynthesis, non-mevalonate pathway. While isoprenoid biosynthesis is a necessity for all cells, the non-mevalonate pathway is used by bacteria and can initiate an immune response (Heuston et al., 2012). Abundance of this module decreases with increasing alternate allele count, which seems to indicate a reduction in pathogenic activity related to this pathway.

In looking at our results from real data we might at first be disappointed in the small number of SNPs we are able to say with confidence are correlated with microbiome data. Considering our host genetic variation data is relatively small in terms of the number of samples per SNP we are optimistic that our results would be improved with larger data. Furthermore the fact that no single gene was identified as well-correlated in both real and permuted data (except for those very highly correlated with subject sex)

gives us confidence that our estimate of the false positive rate is good. Overall we believe we have constructed a reasonable method for finding linear relationships between the human microbiota abundances and host genetic variation.

Despite our powerful genetic sequencing tools, the interpretation of genomic data is still in the early stages. The information within sequenced genomes remains largely unavailable. Methods such as ours are needed to automatically detect unknown relationships between host genetics and microbiome.

Our method is quite specific in screening for linear associations. The advantage of this tightly focused approach is computational efficiency and relatively simple implementation. The disadvantage is that we will overlook relationships that do not follow the pattern our method recognizes. A weakness that could be addressed is that we screen SNPs in isolation. Our method could perhaps be extended to operate on groups of SNPs.

At the other end of the complexity spectrum are deep-learning methods that make direct use of subtle genetic information such as splicing instructions found in introns (Xiong et al., 2015). These methods have not been applied to microbiome data yet, but this is a growing area of research in machine learning and bioinformatics that may be best suited to uncovering unanticipated host-microbiome interactions.

Supplementary Figures

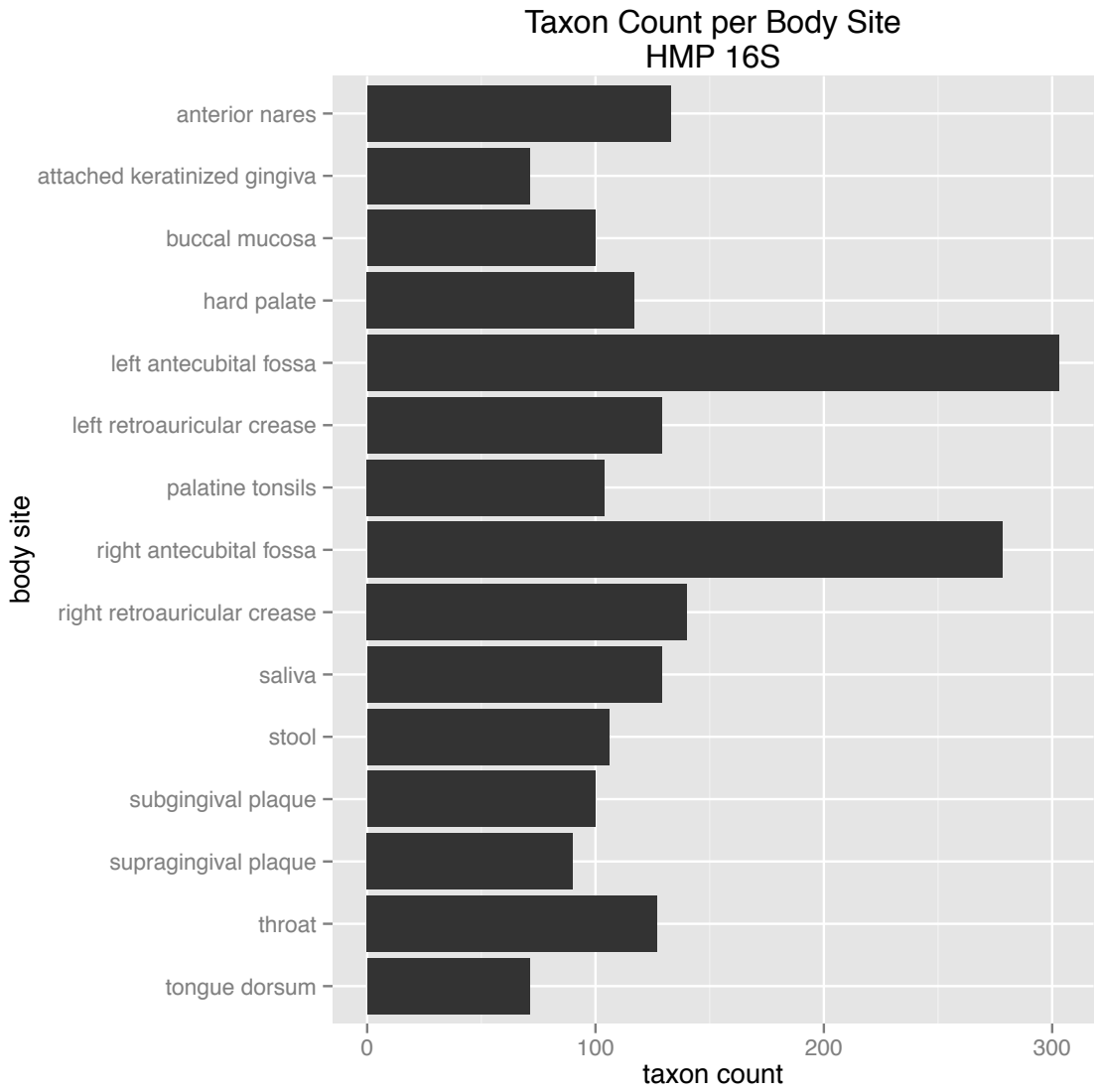

Figure S1: Final count of taxa per body site for HMP 16S data.

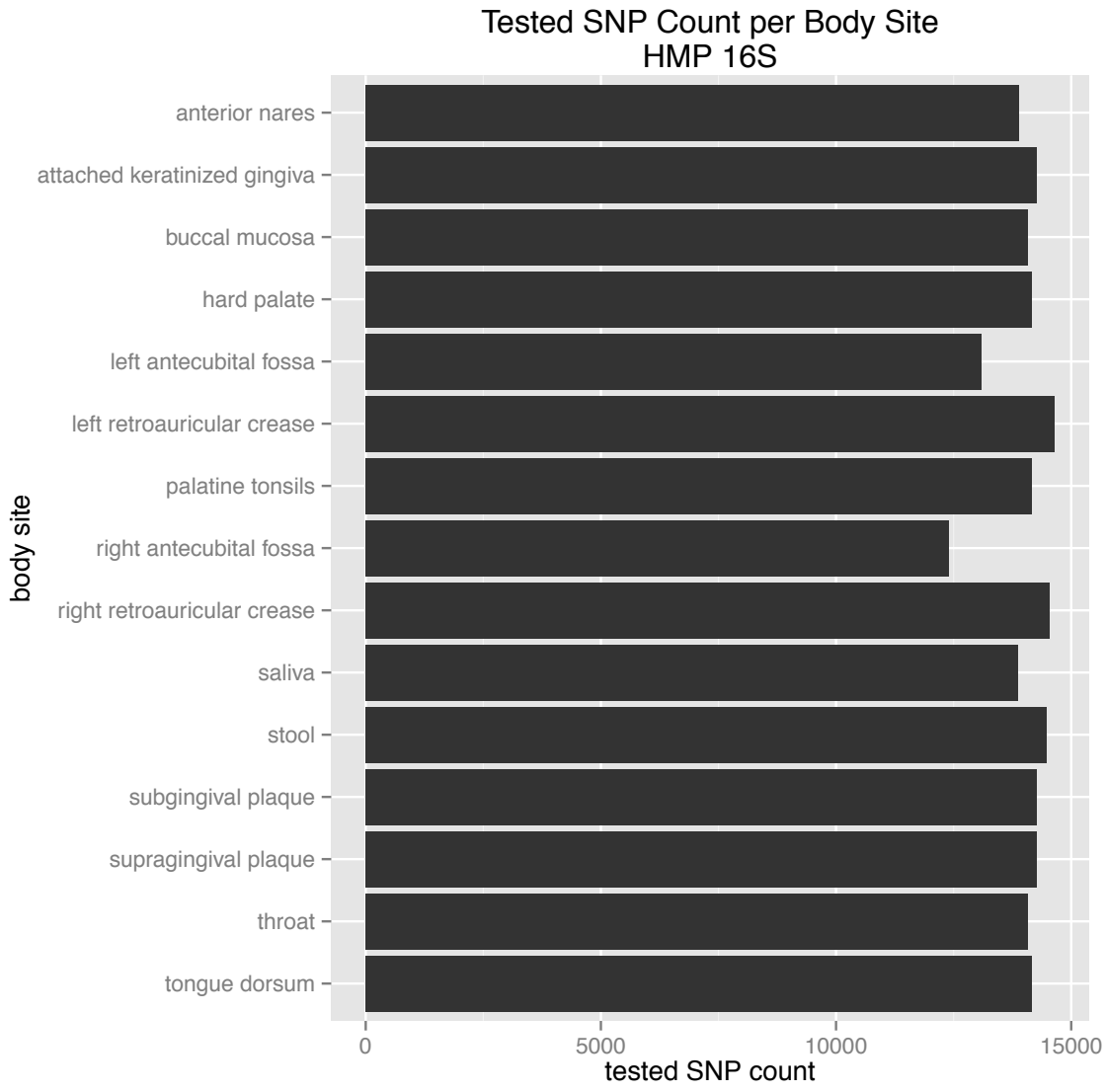

Figure S2: Count of tested SNPs per body site for HMP 16S data.

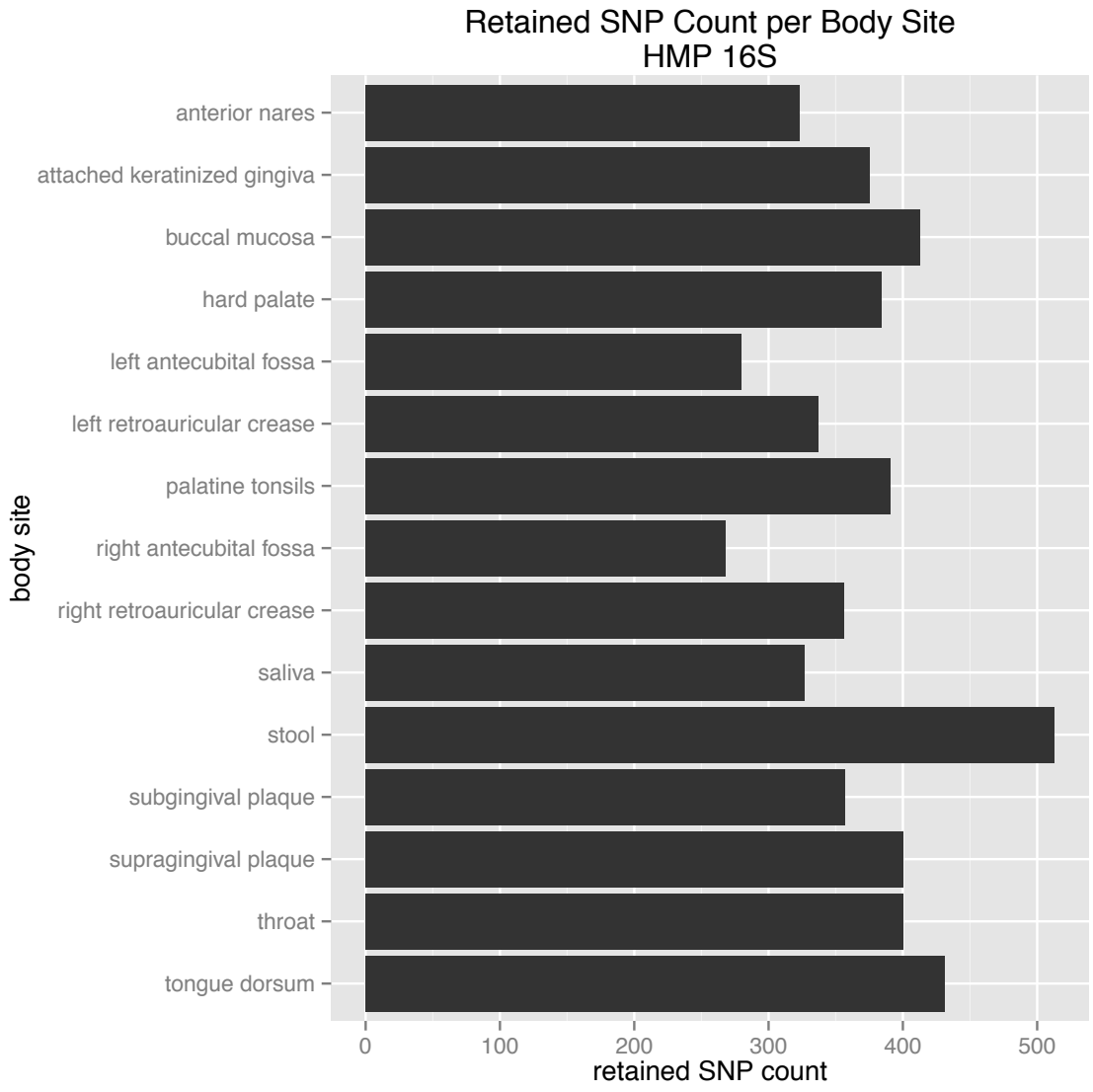

Figure S3: Count of SNPs with 95% confidence interval of median $\boldsymbol{R^2}$ *greater than 0.*

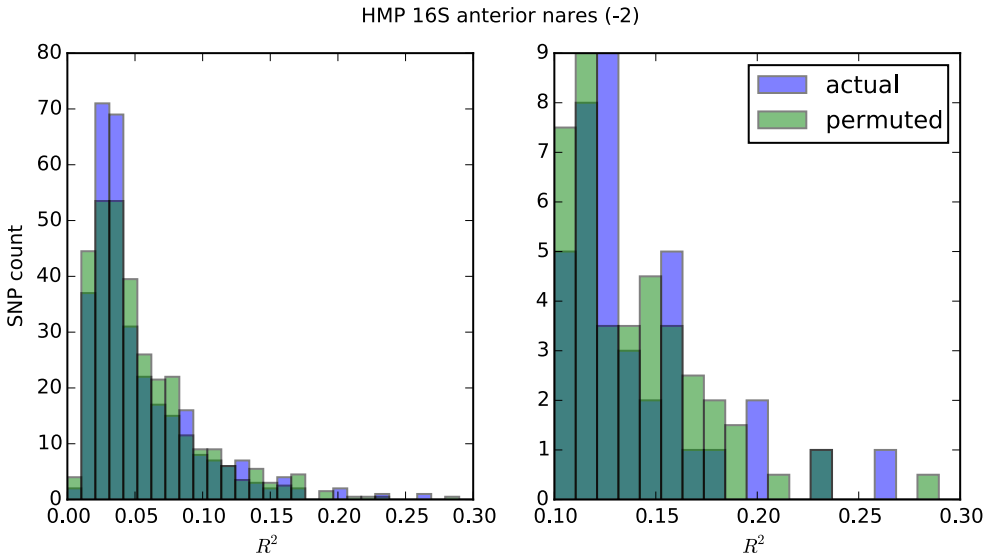

Figure S4: Superimposed distributions of R^2 for actual and permuted data from the anterior nares body site. The figure on the left shows the distribution of SNPs with $R^2 > 0$. The figure on the right shows the distribution of SNPs with $R^2 > 0.1$.

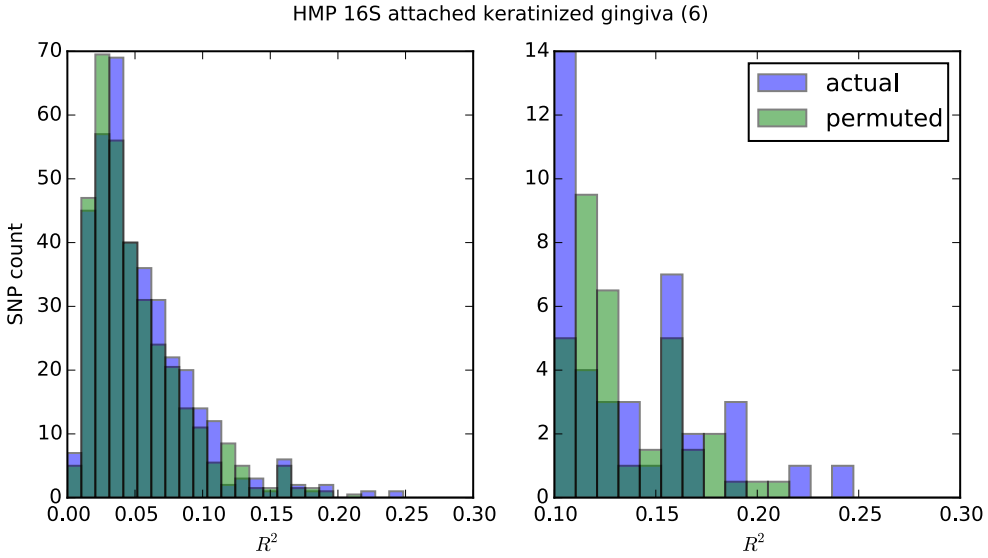

Figure S5: Superimposed distributions of R^2 for actual and permuted data from the attached keratinized gingiva body site.

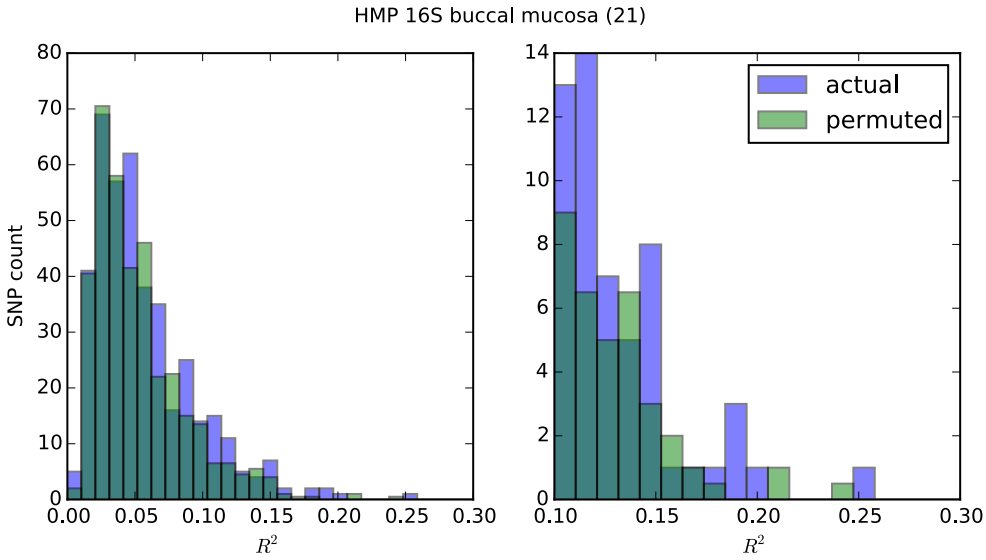

Figure S6: Superimposed distributions of R^2 for actual and permuted data from the buccal mucosa body site.

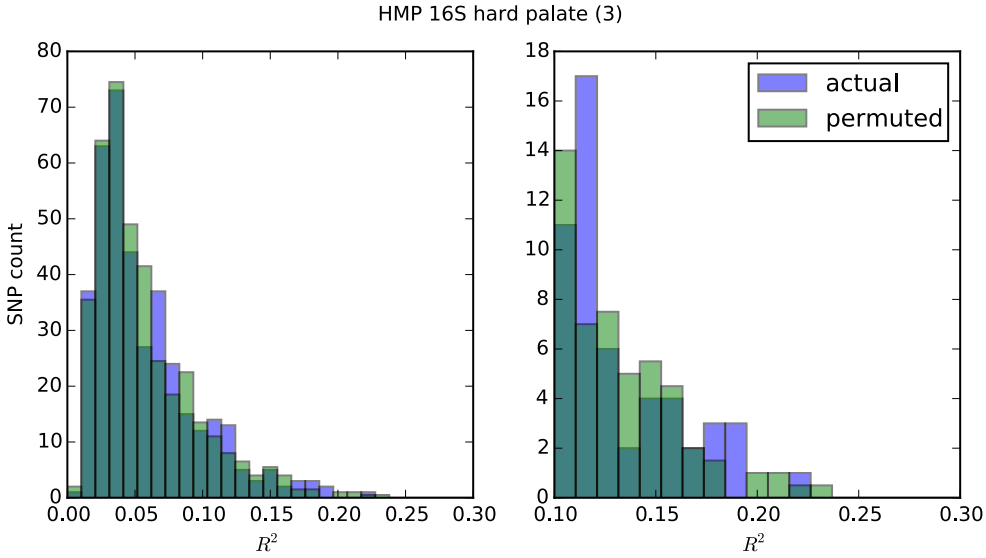

Figure S7: Superimposed distributions of R^2 for actual and permuted data from the hard palate body site.

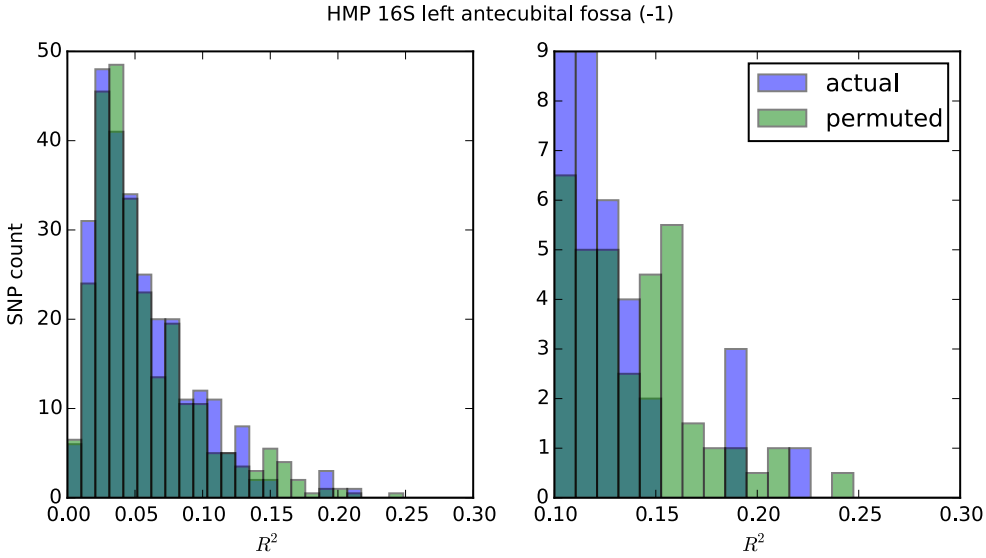

Figure S8: Superimposed distributions of R^2 for actual and permuted data from the left antecubital fossa body site.

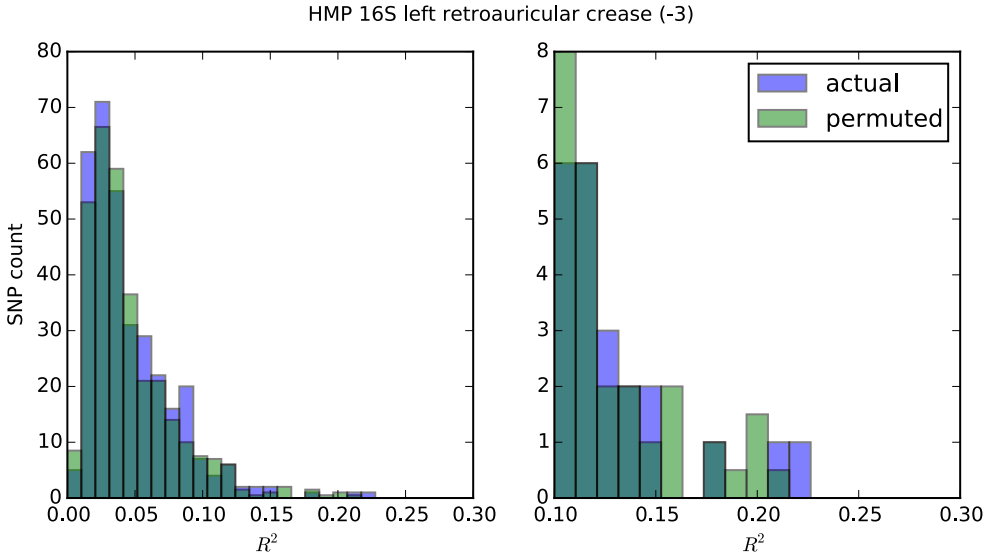

Figure S9: Superimposed distributions of R^2 for actual and permuted data from the left retroauricular crease body site.

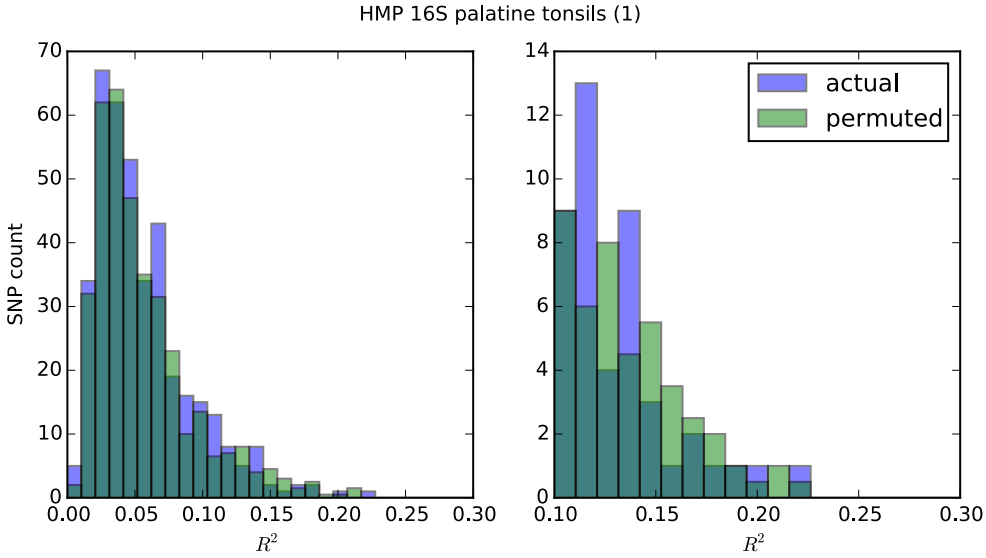

Figure S10: Superimposed distributions of R^2 for actual and permuted data from the palatine tonsils body site.

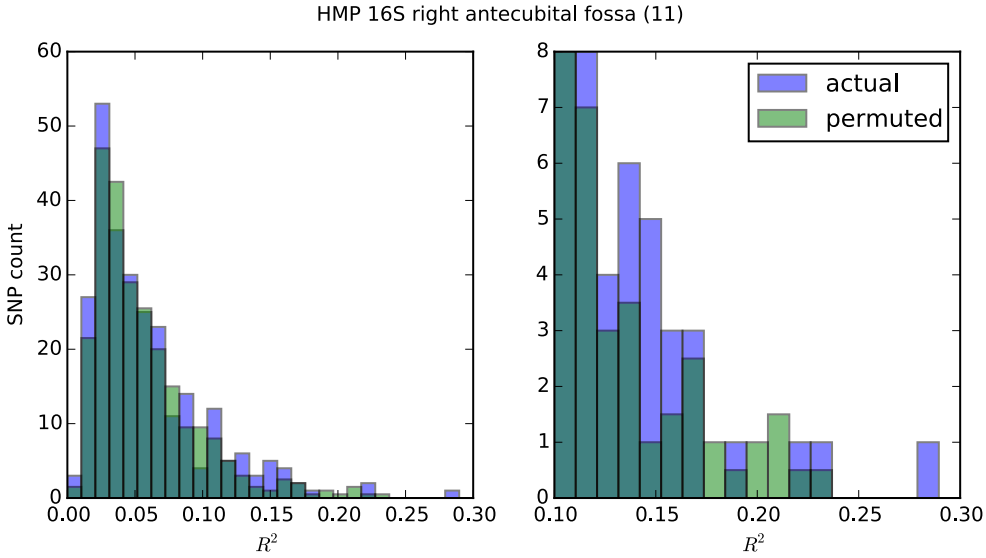

Figure S11: Superimposed distributions of R^2 for actual and permuted data from the right antecubital fossa body site.

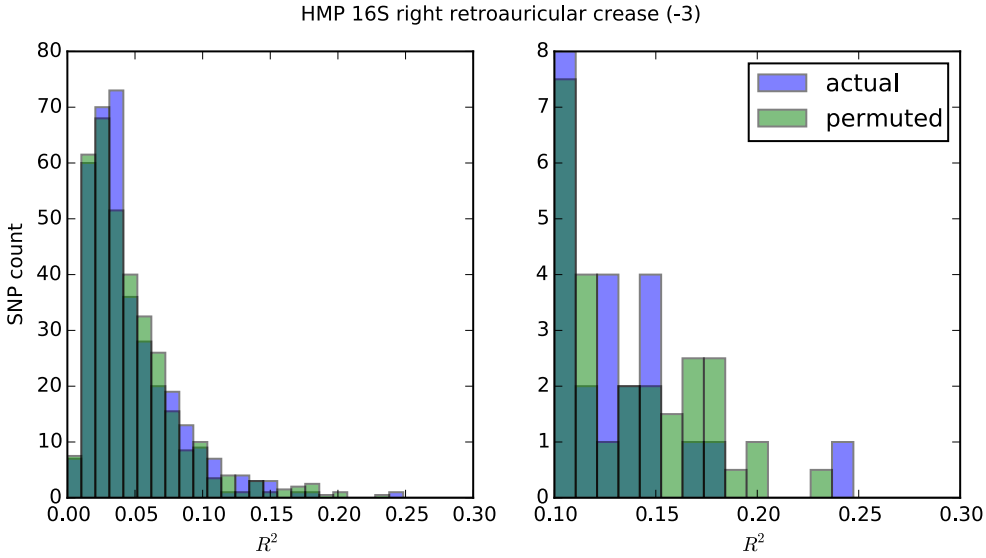

Figure S 12: Superimposed distributions of R^2 for actual and permuted data from the right retroauricular crease body site.

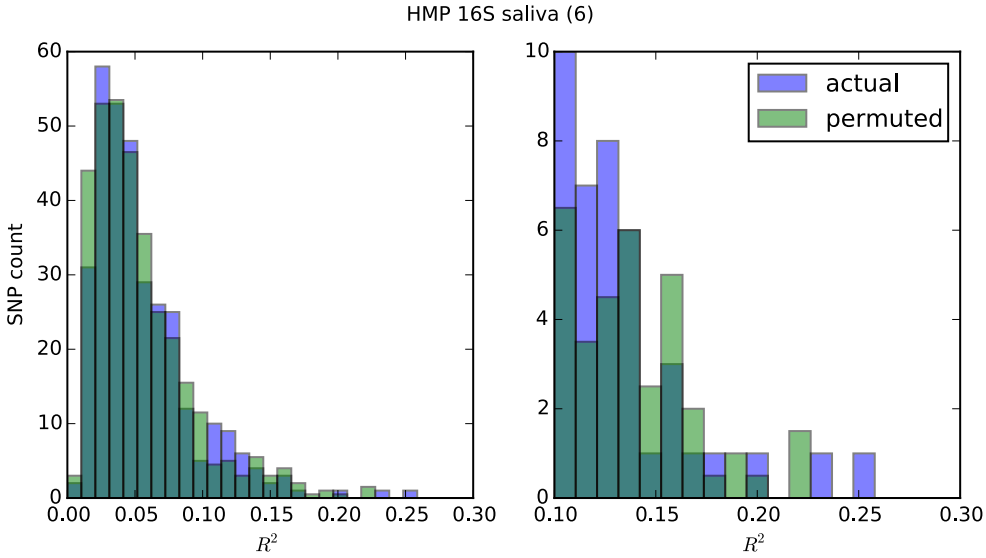

Figure S13: Superimposed distributions of R^2 for actual and permuted data from the saliva body site.

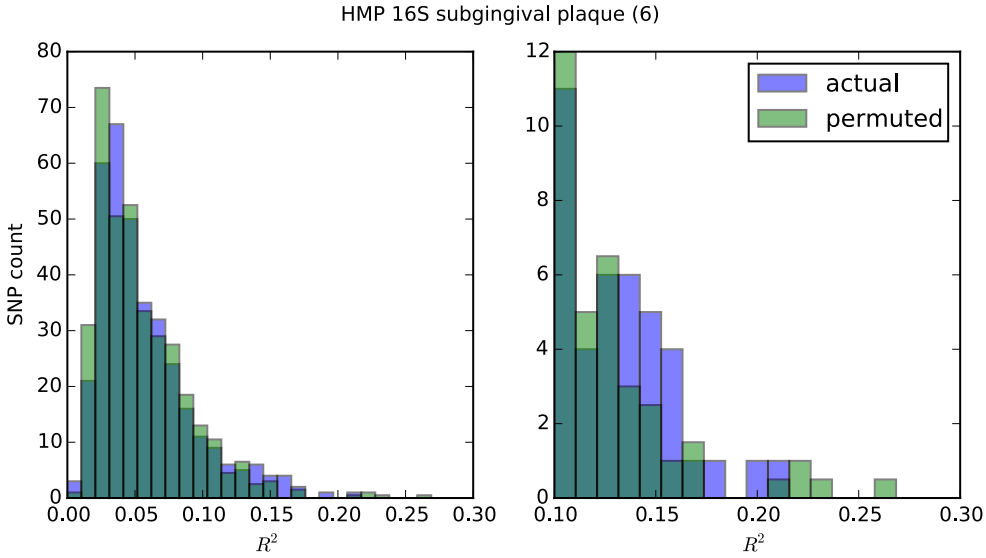

Figure S14: Superimposed distributions of R^2 for actual and permuted data from the subgingival plaque body site.

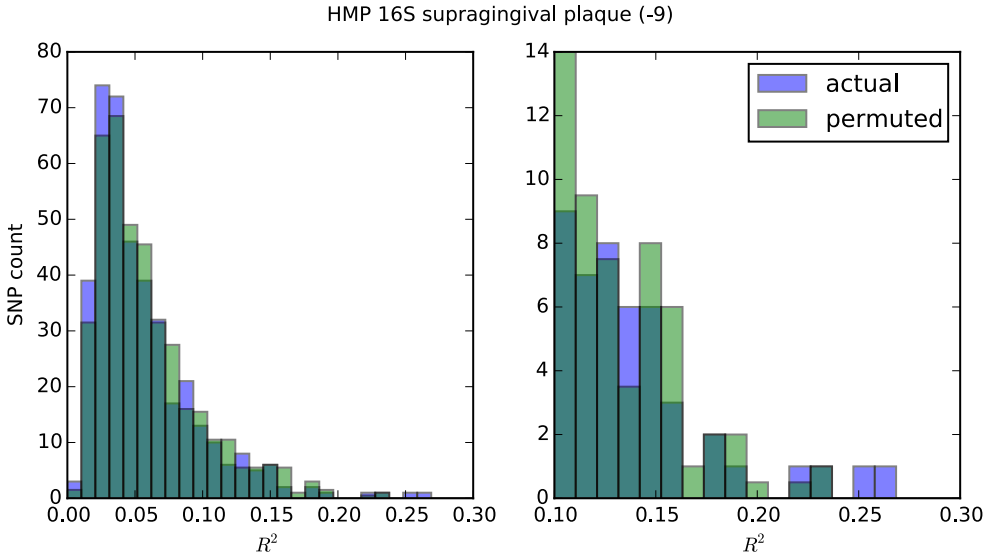

Figure S15: Superimposed distributions of R^2 for actual and permuted data from the supragingival plaque body site.

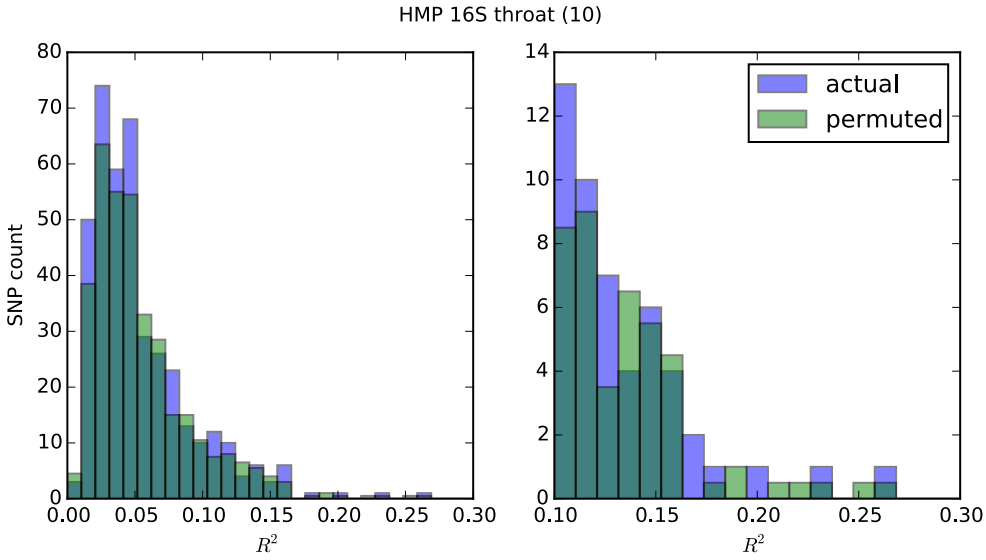

Figure S16: Superimposed distributions of R^2 for actual and permuted data from the throat body site.

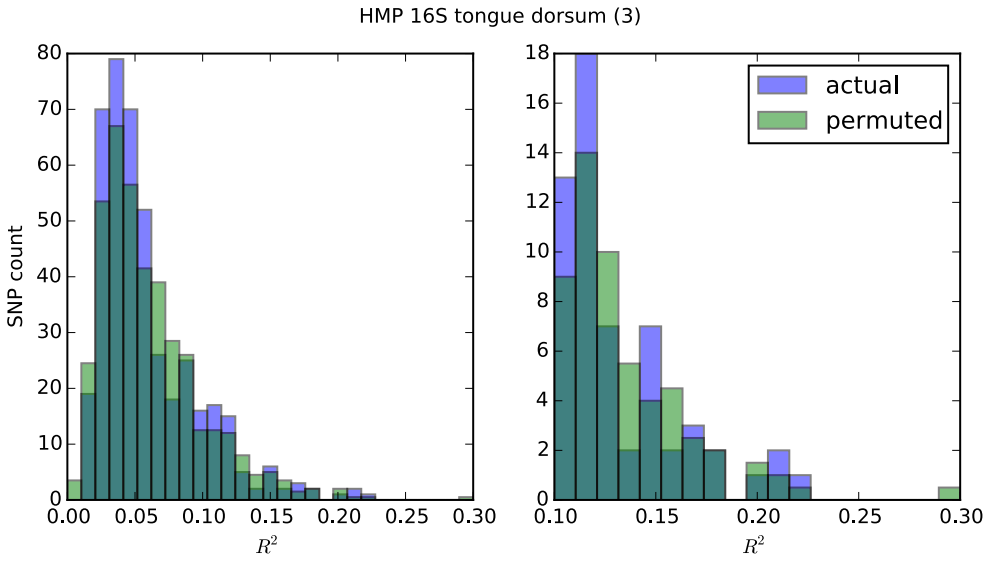

Figure S17: Superimposed distributions of R^2 for actual and permuted data from the tongue dorsum body site.

Analysis: 16S_gene_list_v08 - 2015-07-17 12:44 AM

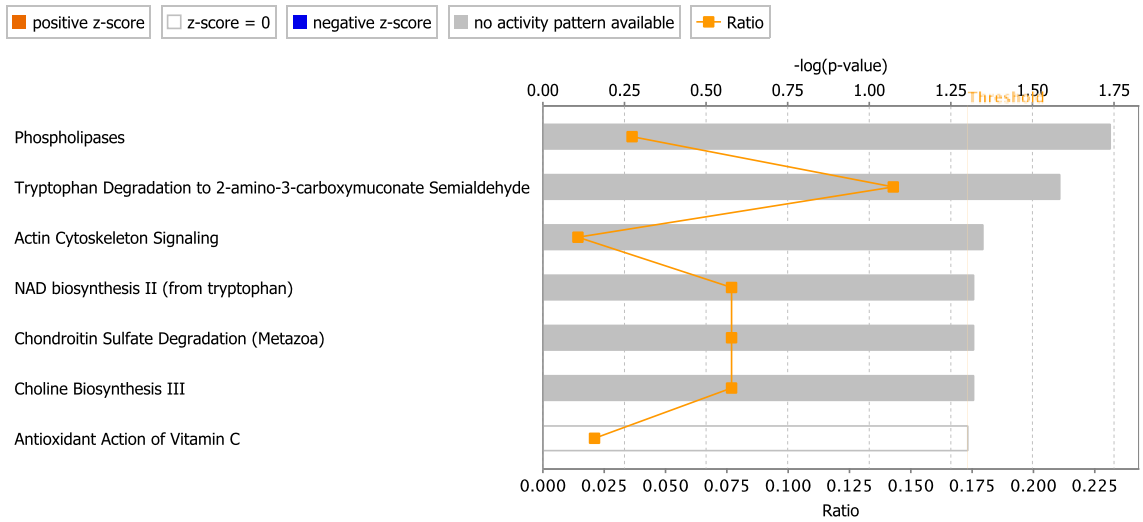

Figure S18: Pathway analysis report from Ingenuity IPA for genes identified with HMP 16S data.

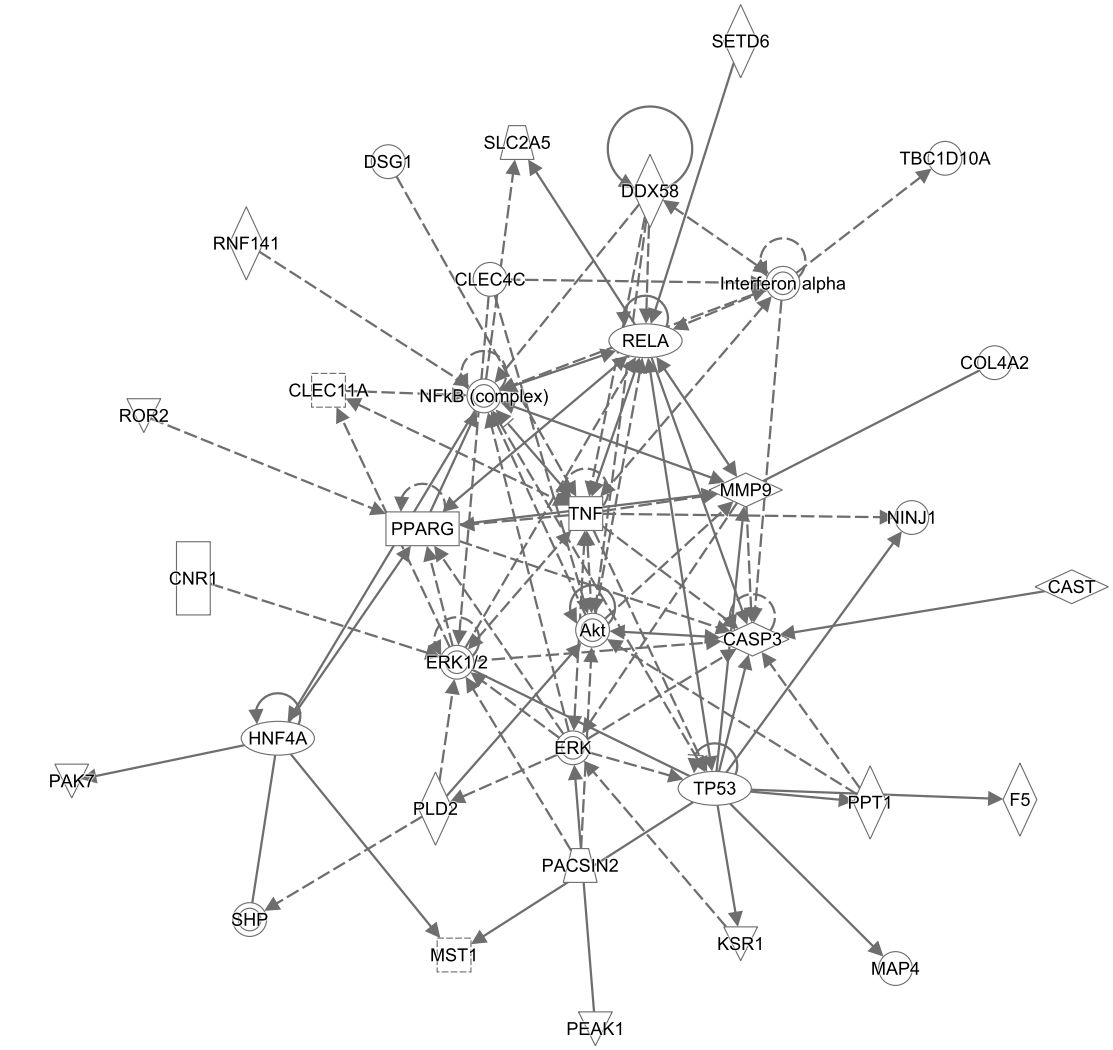

Figure S19: Ingenuity IPA gene network 1 associated with HMP 16S gene list. Functions of the network are Cell-to-Cell Signaling and Interaction, Hematological System Development and Function, and Immune Cell Trafficking.

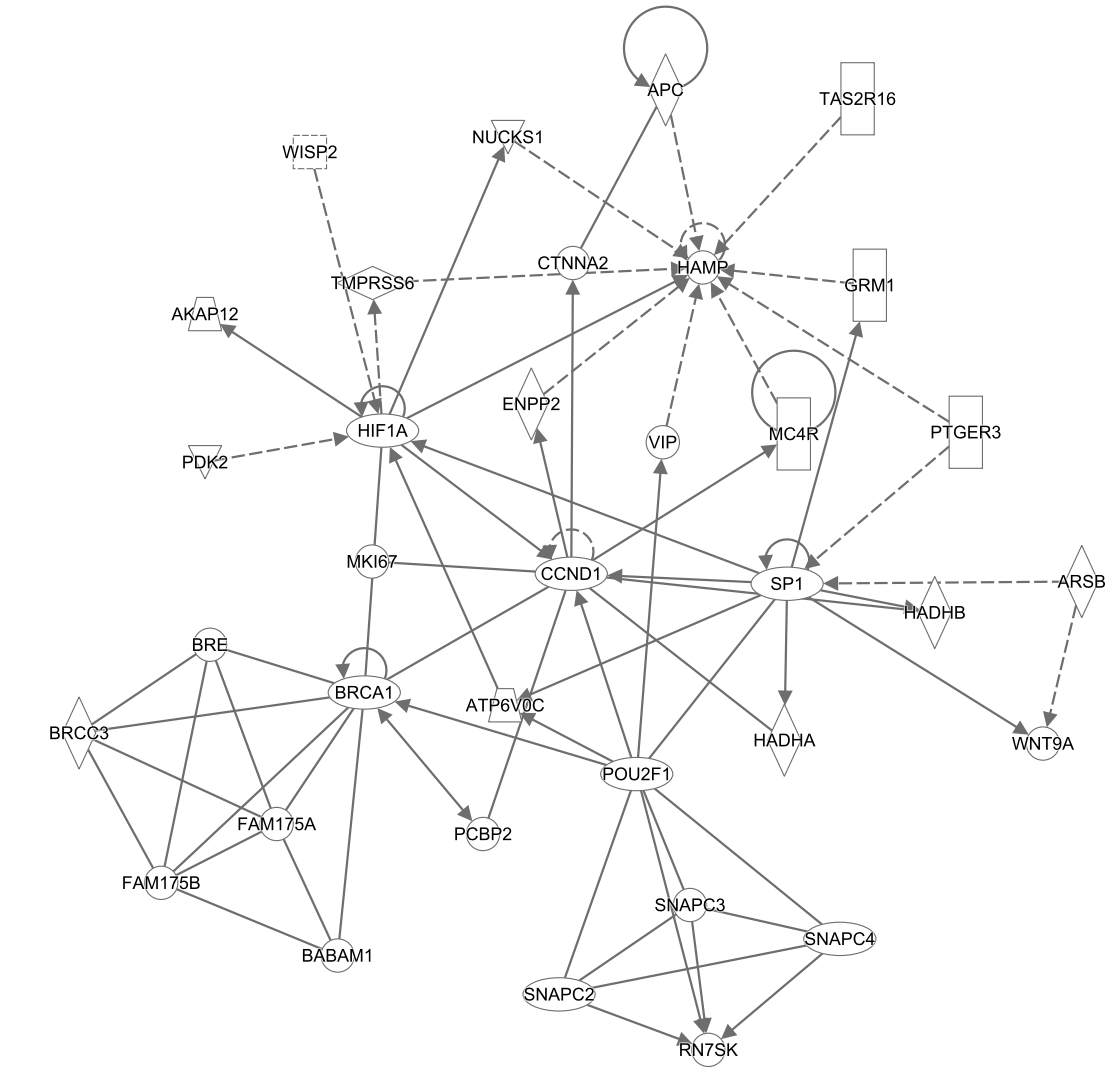

Figure S20: Ingenuity IPA gene network 2 associated with HMP 16S gene list. Functions of the network are Gene Expression, Connective Tissue Development and Function, Tissue Development.

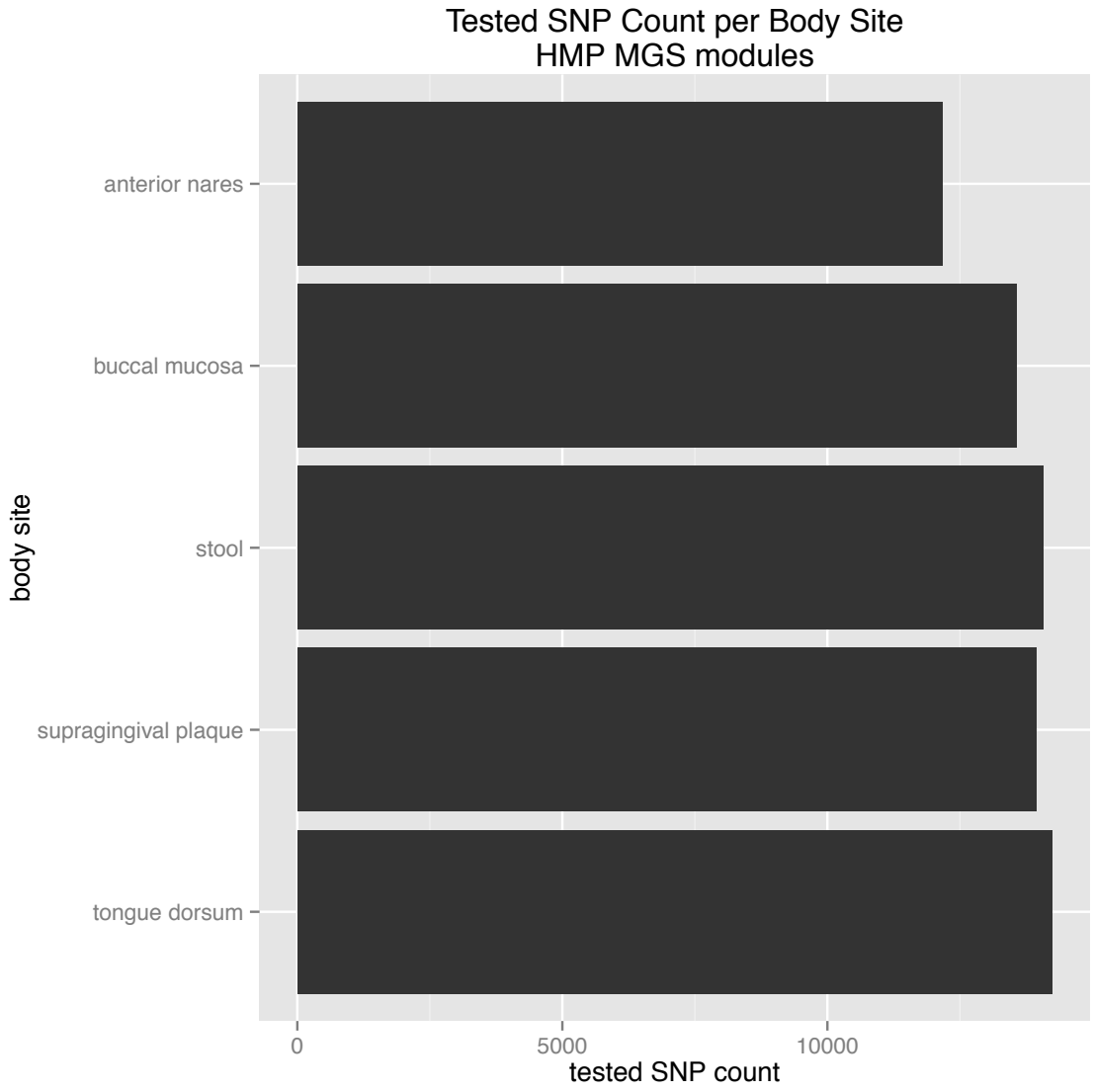

Figure S21: Tested SNP counts per body site for HMP MGS HUMANn KEGG module abundances.

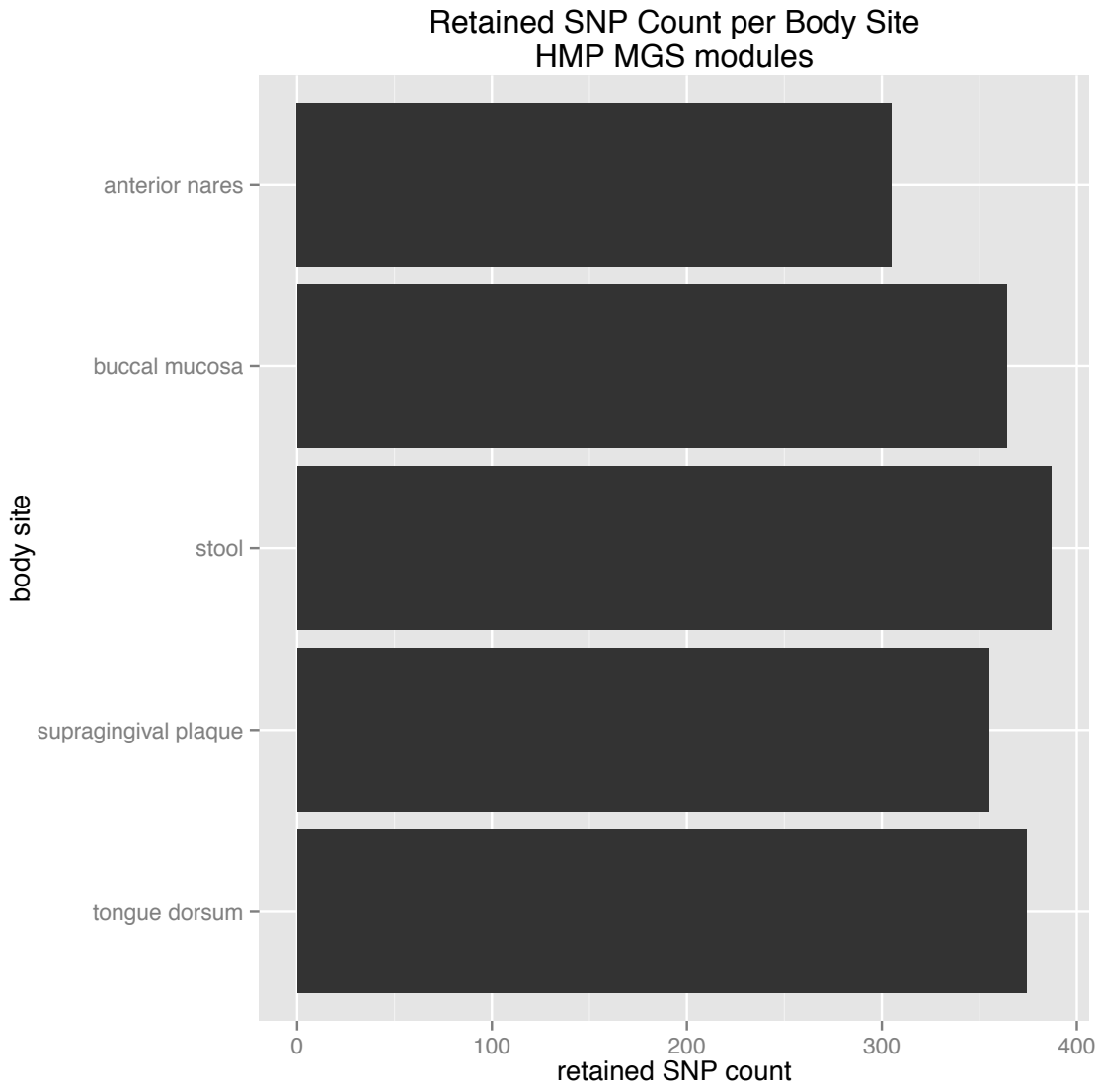

Figure S22: Retained SNP counts per body site for HMP MGS HUMANn KEGG module abundances.

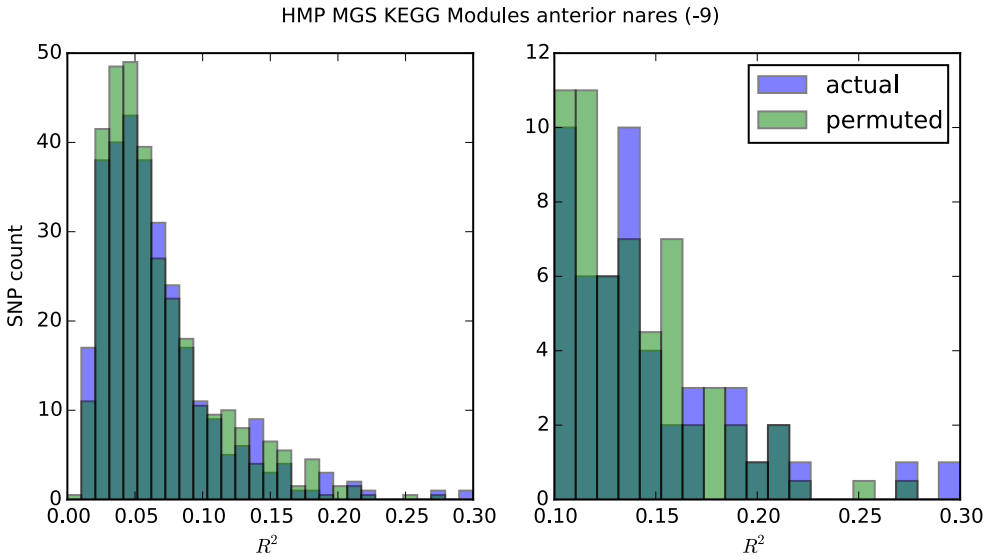

Figure S23: Superimposed distributions of R^2 for actual and permuted data from the anterior nares body site. The figure on the left shows the distribution of SNPs with $R^2 > 0$. The figure on the right shows the distribution of SNPs with $R^2 > 0.1$.

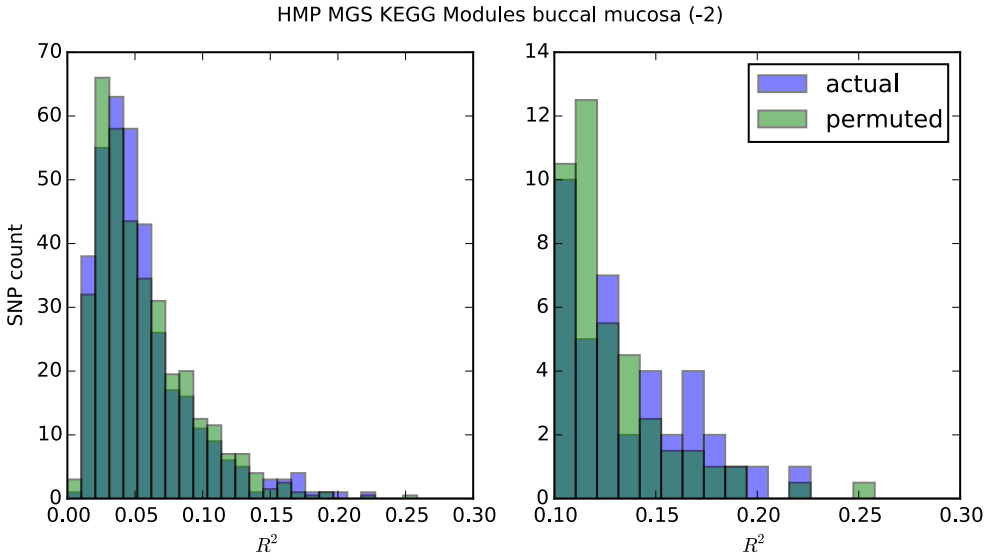

Figure S24: Superimposed distributions of R^2 for actual and permuted data from the buccal mucosa body site. The figure on the left shows the distribution of SNPs with $R^2 > 0$. The figure on the right shows the distribution of SNPs with $R^2 > 0.1$.

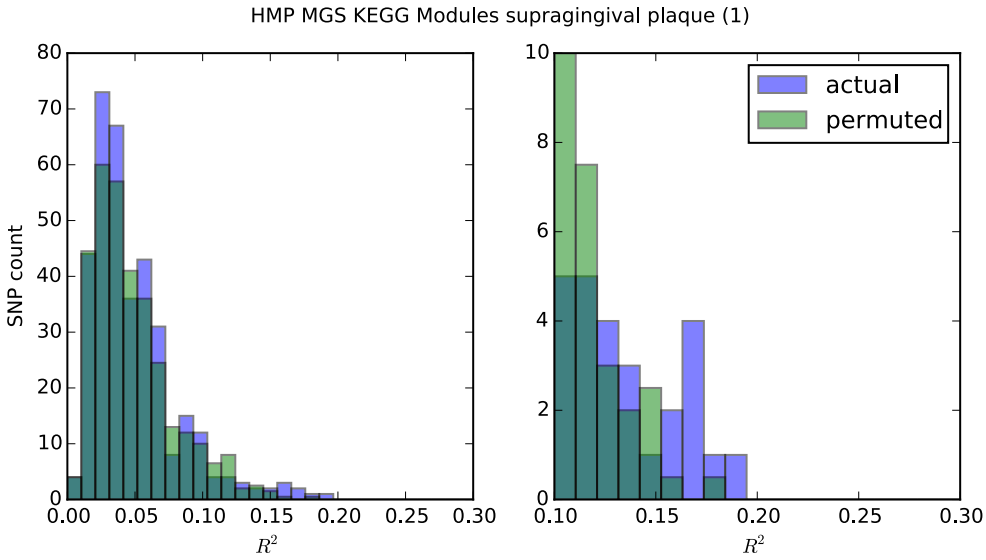

Figure S25: Superimposed distributions of R^2 for actual and permuted data from the supragingival plaque body site. The figure on the left shows the distribution of SNPs with $R^2 > 0$. The figure on the right shows the distribution of SNPs with $R^2 > 0.1$.

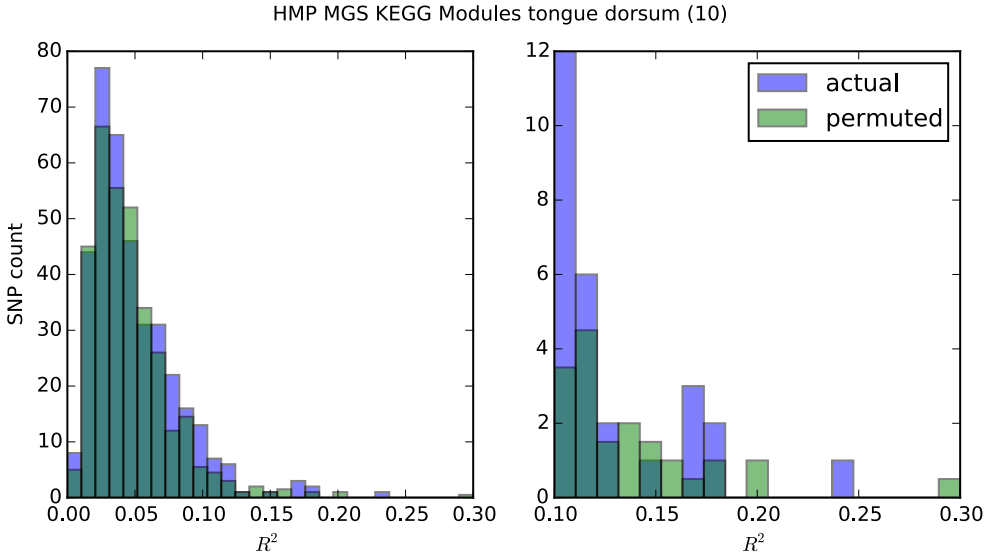

Figure S26: Superimposed distributions of R^2 for actual and permuted data from the tongue dorsum body site. The figure on the left shows the distribution of SNPs with $R^2 > 0$. The figure on the right shows the distribution of SNPs with $R^2 > 0.1$.

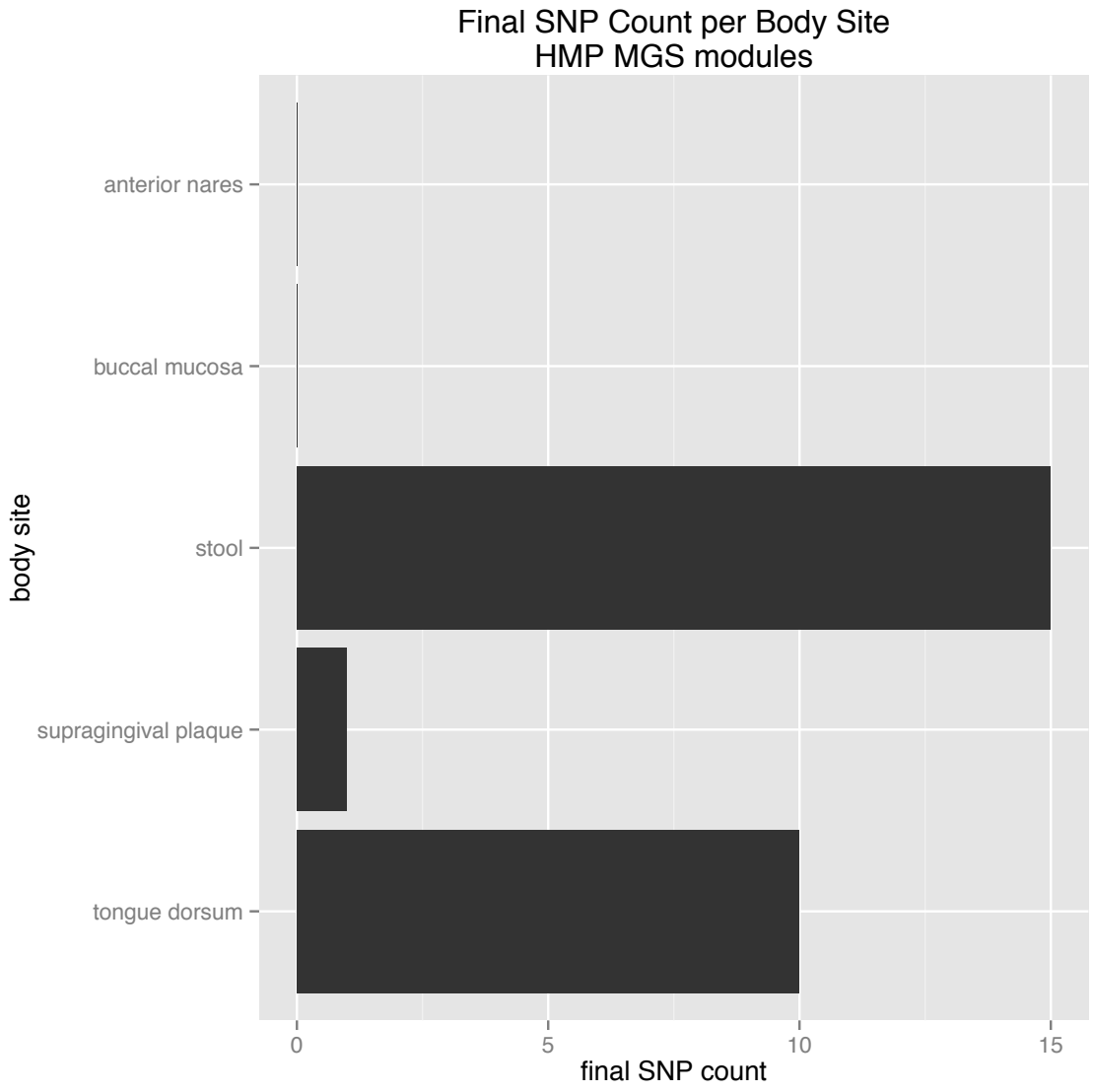

Figure S27: Final SNP counts per body site for HMP MGS HUMANn KEGG module abundances.

Analysis: modules_gene_list_v08 - 2015-07-17 02:04 AM

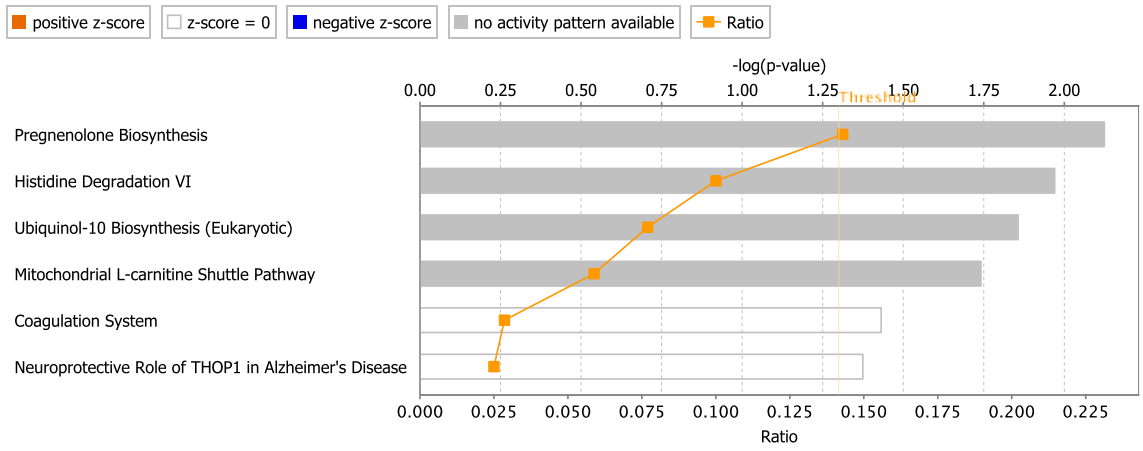

Figure S28: Ingenuity IPA pathway analysis results for genes with SNPs found to be correlated with HMP MGS HUMANn KEGG module abundances.

Bibliography

Abubucker, Sahar, Nicola Segata, Johannes Goll, Alyxandria M Schubert, Jacques Izard, Brandi L Cantarel, Beltran Rodriguez-Mueller, et al. 2012. “Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome.” *PLoS Computational Biology* 8 (6): e1002358. doi:10.1371/journal.pcbi.1002358.

Ball, Helen J, Angeles Sanchez-Perez, Silvia Weiser, Christopher J D Austin, Florian Astelbauer, Jenny Miu, James A McQuillan, Roland Stocker, Lars S Jermiin, and Nicholas H Hunt. 2007. “Characterization of an Indoleamine 2,3-Dioxygenase-like Protein Found in Humans and Mice.” *Gene* 396 (1): 203–13. doi:10.1016/j.gene.2007.04.010.

Baquero, F, and C Nombela. 2012. “The Microbiome as a Human Organ.” *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 18 Suppl 4 (July): 2–4. doi:10.1111/j.1469-0691.2012.03916.x.

Baumgart, Daniel C, and Simon R Carding. 2007. “Inflammatory Bowel Disease: Cause and Immunobiology.” *The Lancet* 369 (9573): 1627–40. doi:10.1016/S0140-6736(07)60750-8.

Benson, Andrew K, Scott A Kelly, Ryan Legge, Fangrui Ma, Soo Jen Low, Jaehyoung Kim, Min Zhang, et al. 2010. “Individuality in Gut Microbiota Composition Is a Complex Polygenic Trait Shaped by Multiple Environmental and Host Genetic Factors.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (44): 18933–38. doi:10.1073/pnas.1007028107.

Burns, Michael, Joshua Lynch, Timothy Starr, Dan Knights, and Ran Blekhman. 2015. “Virulence Genes Are a Signature of the Microbiome in the Colorectal Tumor Microenvironment.” *Genome Medicine* 7 (1): 55. doi:10.1186/s13073-015-0177-8.

Caporaso, Gregory J, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, et al. 2010. “QIIME Allows Analysis of High-Throughput Community Sequencing Data.” *Nature Methods* 7 (5). Nature Publishing Group: 335–36. doi:10.1038/nmeth.f.303.

Chu, Hiutung, and Sarkis K Mazmanian. 2013. “Innate Immune Recognition of the Microbiota Promotes Host-Microbial Symbiosis.” *Nature Immunology* 14 (7): 668–75. doi:10.1038/ni.2635.

Collins, Stephen M, Michael Surette, and Premysl Bercik. 2012. “The Interplay between the Intestinal Microbiota and the Brain.” *Nature Reviews. Microbiology* 10 (11): 735–42. doi:10.1038/nrmicro2876.

Coux, O, K Tanaka, and A L Goldberg. 1996. “Structure and Functions of the 20S and 26S Proteasomes.” *Annual Review of Biochemistry* 65: 801–47. doi:10.1146/annurev.bi.65.070196.004101.

Doan, N, A Contreras, J Flynn, J Slots, and C Chen. 2000. “Molecular Identification of Dialister Pneumosintes in Subgingival Plaque of Humans.” *Journal of Clinical Microbiology* 38 (8): 3043–47. http://www.ncbi.nlm.nih.gov/pubmed/10921975.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. “Least Angle Regression.” *Annals of Statistics* 32 (2). Institute of Mathematical Statistics: 407–99. http://projecteuclid.org/download/pdfview_1/euclid.aos/1083178935.

Evans, James M, Laura S Morris, and Julian R Marchesi. 2013. “The Gut Microbiome: The Role of a Virtual Organ in the Endocrinology of the Host.” *The Journal of Endocrinology* 218 (3): R37–47. doi:10.1530/JOE-13-0131.

Girish, K S, and K Kemparaju. 2007. “The Magic Glue Hyaluronan and Its Eraser Hyaluronidase: A Biological Overview.” *Life Sciences* 80 (21): 1921–43. doi:10.1016/j.lfs.2007.02.037.

Goodrich, Julia K, Jillian L Waters, Angela C Poole, Jessica L Sutter, Omry Koren, Ran Blekhman, Michelle Beaumont, et al. 2014. “Human Genetics Shape the Gut Microbiome.” *Cell* 159 (4): 789–99. doi:10.1016/j.cell.2014.09.053.

“Homo Sapiens Chondroitin Sulfate Degradation (metazoa).” 2015. Accessed July 21. http://biocyc.org/HUMAN/new-image?type=PATHWAY&object=PWY-6573.

Heuston, Sinéad, Máire Begley, Cormac G M Gahan, and Colin Hill. 2012. “Isoprenoid Biosynthesis in Bacterial Pathogens.” *Microbiology* 158 (Pt 6): 1389–1401. doi:10.1099/mic.0.051599-0.

Human Microbiome Project Consortium. 2012. “A Framework for Human Microbiome Research.” *Nature* 486 (7402): 215–21. doi:10.1038/nature11209.

Human Microbiome Project Consortium. 2012. “Structure, Function and Diversity of the Healthy Human Microbiome.” *Nature* 486 (7402): 207–14. doi:10.1038/nature11234.

Kachrimanidou, Melina, and Nikolaos Malisiovas. 2011. “Clostridium Difficile Infection: A Comprehensive Review.” *Critical Reviews in Microbiology* 37 (3): 178–87. doi:10.3109/1040841X.2011.556598.

Khoruts, Alexander, Johan Dicksved, Janet K Jansson, and Michael J Sadowsky. 2010. “Changes in the Composition of the Human Fecal Microbiome after Bacteriotherapy for Recurrent Clostridium Difficile-Associated Diarrhea.” *Journal of Clinical Gastroenterology* 44 (5): 354–60. doi:10.1097/MCG.0b013e3181c87e02.

Ley, Ruth E. 2010. “Obesity and the Human Microbiome.” *Current Opinion in Gastroenterology* 26 (1): 5–11. doi:10.1097/MOG.0b013e328333d751.
Meinshausen, Nicolai, and Peter Bühlmann. 2010. “Stability Selection.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 72 (4). Blackwell Publishing Ltd: 417–73. doi:10.1111/j.1467-9868.2010.00740.x.

"MetaCyc L-Histidine Degradation VI." 2015. Accessed July 21. http://www.biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=HISHP-PWY.

"MetaCyc L-Tryptophan Degradation to 2-Amino-3-Carboxymuconate Semialdehyde." 2015. Accessed July 21. http://biocyc.com/META/NEW-IMAGE?type=PATHWAY&object=PWY-5651&detail-level=%203.

Opitz, C A, W Wick, L Steinman, and M Platten. 2007. "Tryptophan Degradation in Autoimmune Diseases." *Cellular and Molecular Life Sciences: CMLS* 64 (19-20): 2542–63. doi:10.1007/s00018-007-7140-9.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research: JMLR* 12 (Oct): 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.

Spor, Aymé, Omry Koren, and Ruth Ley. 2011. "Unravelling the Effects of the Environment and Host Genotype on the Gut Microbiome." *Nature Reviews. Microbiology* 9 (4): 279–90. doi:10.1038/nrmicro2540.

Stewart, Jessica A, Vinton S Chadwick, and Alan Murray. 2005. "Investigations into the Influence of Host Genetics on the Predominant Eubacteria in the Faecal Microflora of Children." *Journal of Medical Microbiology* 54 (Pt 12): 1239–42. doi:10.1099/jmm.0.46189-0.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 58 (1). Wiley for the Royal Statistical Society: 267–88. doi:10.2307/2346178.

Tremaroli, Valentina, and Fredrik Bäckhed. 2012. "Functional Interactions between the Gut Microbiota and Host Metabolism." *Nature* 489 (7415): 242–49. doi:10.1038/nature11552.

Trude, Handal, Olsen Ingar, Clay B Walker, and Dominique A Caugant. 2005. “Detection and Characterization of Î²-Lactamase Genes in Subgingival Bacteria from Patients with Refractory Periodontitis.” *FEMS Microbiology Letters* 242 (2): 319–24. doi:10.1016/j.femsle.2004.11.023.

Turnbaugh, Peter J, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. 2007. “The Human Microbiome Project.” *Nature* 449 (7164): 804–10. doi:10.1038/nature06244.

Turnbaugh, Peter J, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, et al. 2009. “A Core Gut Microbiome in Obese and Lean Twins.” *Nature* 457 (7228): 480–84. doi:10.1038/nature07540.

Xiong, Hui Y, Babak Alipanahi, Leo J Lee, Hannes Bretschneider, Daniele Merico, Ryan K C Yuen, Yimin Hua, et al. 2015. “RNA Splicing. The Human Splicing Code Reveals New Insights into the Genetic Determinants of Disease.” *Science* 347 (6218): 1254806. doi:10.1126/science.1254806.

Yatsunenko, Tanya, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, et al. 2012. “Human Gut Microbiome Viewed across Age and Geography.” *Nature* 486 (7402): 222–27. doi:10.1038/nature11053.

Zackular, Joseph P, Mary A M Rogers, Mack T Ruffin 4th, and Patrick D Schloss. 2014. “The Human Gut Microbiome as a Screening Tool for Colorectal Cancer.” *Cancer Prevention Research* 7 (11): 1112–21. doi:10.1158/1940-6207.CAPR-14-0129.