Investigating the Relationship Between the Microbiome and Environmental Characteristics

Erin Beckman[1], Christine Chai[2], Jingjing Lyu[3], Shant Mahserejian[4], Hoang Tran[5], Shirin Yavari[6]

Faculty Mentors: Herman Mitchell[7], Agustin Calatroni[7], Emily Lei Kang[8]

Abstract

Recent research has shown that the microscopic communities found in people and in the environment play a large role in human health and different diseases. A microbial community is called a microbiome, and it is found using sequencing techniques to quantify the presence of operational taxonomic units (OTUs). The presence of allergies and asthma have been linked to the environmental microbiome during developmental years. This project investigates the interplay between environmental factors and the microbiome. To better understand this relationship, three data sets are studied: OTU count data, characteristics of the home, and diversity statistics. First, the dimension of the data was reduced through pre-processing. Several techniques, including analysis of variance, principal component analysis (PCA) and volcano plots, are used to explore and visualize the data. Correlation analysis on diversity statistics was also employed to explore the data. Finally, predictive modeling was implemented to reveal associations between different home characteristics and the microbiome. Site was determined to be the most significant factor in separating the OTU data.

1 Introduction

A new direction of research has emerged within the past several years that focuses on discovering and categorizing microbiomes. A microbiome is a full microbial community. Each human body has its own distinct microbiome; it is estimated that microbial cells make up around 90% of the cells in and on the human body. Recently, interest in the human microbiome has increased significantly due to findings suggesting that the microbiome is intricately related to diseases and the overall health of the human body [?]. For example, the microbiome of the gut has been suggested to be a predictor for infection by *Clostridium difficile* (C-diff) after taking a dose of antibiotics [?]. Other studies associate the human microbiome with obesity and diabetes [?].

Environments can also have distinguishing microbiomes. Dust samples from different places reveal a wide distribution of microbes, and the results vary from place to place. With this knowledge, researchers have begun to study the microbiomes of people's environments. Since around 69% of time is spent in the residence, the most important environment to study is the home [?]. A connection between the microbiome and asthma or allergies has been found through research of the home microbiome [?]. This research, done in part by Rho, Inc., has consisted of a large study of children as they develop and the microbiomes they live in at the early stages of life.

Despite this link, researchers still do not know very much about how the microbiome is related to the everyday characteristics of a home. For instance, having a dog changes the microbiome greatly, as does living in different cities. Therefore, the connection between the characteristics of a home and the microbiome is a question of great interest. If this connection could be elucidated, then recommendations could be made to new parents on how to create a beneficial microbiome and prevent their child from developing asthma or allergies. In this way, this type of research is looking for primary prevention of these diseases, rather than secondary.

An OTU is a working definition of a species or group of species when using DNA sequencing. This definition is necessary because defining a species can be difficult and using only sequence similarity, which is how OTUs are

[1] Department of Mathematics, Duke University
[2] Department of Statistical Science, Duke University
[3] Department of Mathematics, Clarkson University
[4] Department of Applied and Computational Mathematics and Statistics, University of Notre Dame
[5] Department of Statistics, Florida State University
[6] Department of Mathematics, California State University Fullerton
[7] Rho, Inc.
[8] University of Cincinnati

defined, can be much simpler [?]. The microorganisms analyzed in this paper belong to the kingdoms Archaea and Bacteria. Due to recent advances in technology and methodology, the dust samples can be analyzed for over 50,000 operational taxonomic units (OTUs) at a time without having to deal with culturing. Because the presence of so many OTUs can be distinguished, one of the main problems when dealing with the study of the microbiome will be the large dimensionality of the data. Many recent advances in this area can be seen in [?].

2 Data Description

The Urban Environment and Childhood Asthma (URECA) birth cohort collected all the data used in the report. Mothers were recruited while pregnant and the participants were tracked for many years by surveys and sample collection. The participants were located in four different areas: Boston, Baltimore, New York City, and St. Louis. The abbreviations of the city names are in Table 3. Dust was collected in each of the participants' home when the child was one year old, and the dust was then analyzed for the presence of 51,094 different OTUs. Though more participants were involved in the study, there was a sufficient amount of dust to analyze 277 participants. For each participant, nine additional pieces of information were recorded: house location, house type, whether the house had water problems, and whether or not the house had dogs, cats, rats, mice, other rodents, or cockroaches. This information will be referred to as the "home characteristics". The counts of each OTU in a home make up the "home microbiome" or just "microbiome".

There are several ecological measures of diversity that were applied to the microbiome of each participant. These are richness, Simpson index, inverse Simpson index, Shannon index, and evenness. These are referred to as "omnibus statistics". Richness is defined as the number of species of OTUs found in the microbiome (call this value S). Letting p_i be the proportion of the ith OTU in S, then the definitions for the other diversity measures are as follows:

Statistic	**Formula**
Simpson	$1 - \sum_{i=1}^{n} p_i^2$
Inverse Simpson	$\dfrac{1}{1 - \sum_{i=1}^{S} p_i^2}$
Shannon	$-\sum_{i=1}^{S} p_i \ln p_i$
Evenness	$\dfrac{\texttt{Shannon}}{\ln S}$

Table 1: Formulas for diversity measures

These statistics allow a participant's microbiome to be encompassed in a single integer. This lends itself well to comparison between participants and comparison between home characteristics. The details for the variable names and what they described are listed in Table 2.

The dust sample analysis included phylogenetic data as well. Whenever possible, the kingdom, phylum, class, order, family, genus, and species were recorded for each OTU. However, most of this data is missing because the OTUs could only be identified up to the level of phylum or class. For this reason, analysis was not done using this information.

3 Data Pre-Processing

Before conducting any analysis, the data was pre-processed to make the results as meaningful as possible. In the dataset `home`, 23 of 277 samples are incomplete – three have one or two missing values, and the remaining 20 have many missing values in a single row. Imputing too many missing values can reduce data accuracy, so the 20 highly-missing samples were removed. The three rows with few missing values theoretically could

Variable	Description	% Yes
`home_waterprobs`	Water problems in past 12 months	32
`home_mice`	Mice problems	39
`home_rats`	Rat problems	1
`home_roach`	Cockroaches in last 12 months	39
`dog_any`	Dog living at home in past year	16
`cat_any`	Cat living at home in past year	25
`home_rodents`	Problems with mice or rats	39
Variable	Description	# Unique
`house_type`	Type of dwelling	6
`site`	Study site	4

Table 2: Home characteristics code book

Abbreviation	City
Ba	Baltimore
Bo	Boston
NY	New York
StL	St. Louis

Table 3: Abbreviation of City Names

be imputed, but using the R package `mice` (Multivariate Imputation by Chained Equations) [?] was too computationally expensive (i.e. too slow), so the three samples were also removed. Therefore, all 23 samples containing missing values were not considered in any analysis performed on the `home` dataset.

The OTU data is considered high-dimensional because the number of OTU variables was much larger than the number of samples taken ($p >> n$), so dimensional reduction was also desired. This was done using the `nearZeroVar` function in the `caret` package with default parameter values for freqRatio and percentUnique. This method removed OTUs which had minimal variation across all 254 samples. Since the goal of the study is to determine how different home characteristics influence the microbiome and development of asthma and allergies, OTUs that are relatively constant among all microbiomes complicate the analysis and give very little information to this end. This cut the number of significant OTUs down to 20,402, a 60% reduction of the original data.

Next, a heat map of the original data was created to look at the distribution, seen in Figure 2. This revealed the largely right-skewed nature of the data. Another way to display this original data is through a heat map. Figure 1 shows a heat map of the data after near zero variance OTUs were removed, with a log scale to improve color contrast.

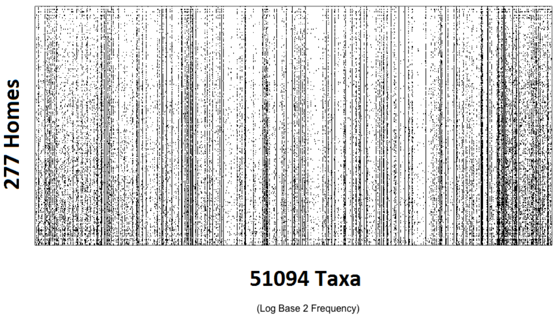
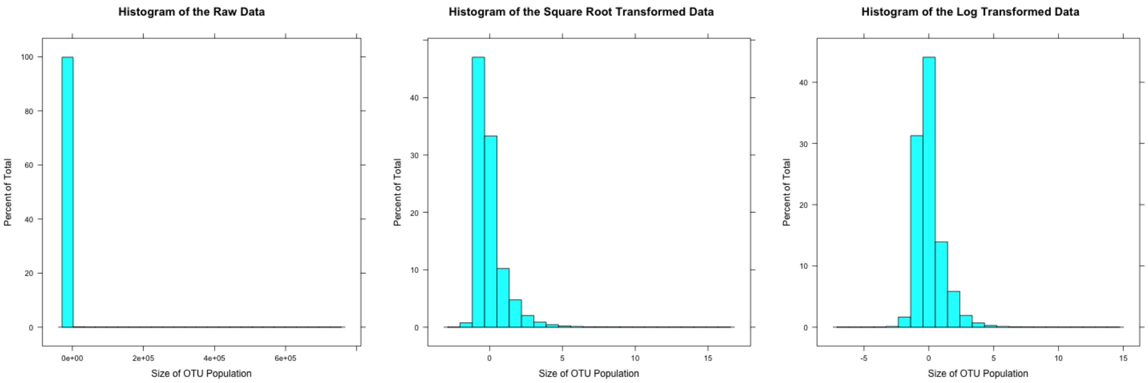


Figure 1: Heat map of OTU data. White indicates near zero values.

Because much of the analysis requires the data to be as symmetric as possible, a transformation of the data was needed which would maximize the symmetry.

Many transformations were considered. The square root transformation and the modified log transformation, log(data + 1), turned out to make the data the most symmetric. After being transformed, the columns were scaled by subtracting the mean and dividing by the standard deviation. Figure 2 shows the histogram of the scaled and transformed data. Using the `boxcox` function in the `mass` packing in R, the modified log transformation was selected as the preferred transformation, and all further analysis was completed using this modified data set.



(a) Original data

(b) Square root transformed data

(c) Modified Log Transformed Data

Figure 2: Histograms of various data sets

4 Data Analysis

The analysis included two stages: exploratory data analysis, which helps illustrate general features of the microbiome and more complex data analysis to discover relationships between particular OTUs and home characteristics. To begin, exploratory data analysis (PCA) was done on the omnibus statistics and the home characteristics. Using notched box plots, it was possible to decipher which home characteristics have an impact on the omnibus statistics, and analysis of variance was done to verify statistical significance of these relationships. Correlation analysis was also done on the omnibus statistics. From there, principal component analysis was performed on the scaled data. This method allows visualization of the data in two or three dimensions to look for the presence of clustering. The exploratory data analysis provided insight about which home characteristics might be important before proceeding with more complex data analysis. Volcano plots were created to determine which OTUs were significant predictors of home characteristics. Further analysis done made use of clustering analysis, simple linear regression using particular OTUs, and predictive modeling. These steps are detailed in the following subsections.

4.1 Relationship between Omnibus Statistics and Home Characteristics

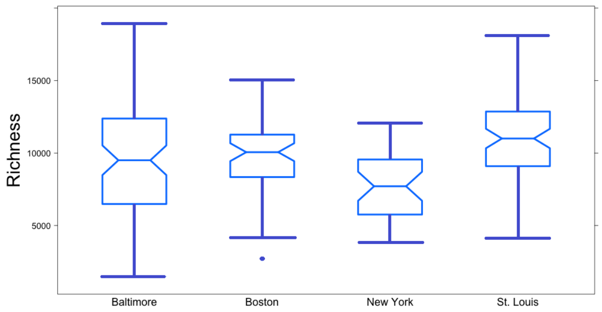


Figure 3: Richness by site

Descriptive statistics and linear models were used to explore the relationship between omnibus statistics and home characteristics. Notched box plots were used to obtain a preliminary view of differences in omnibus statistics at different levels of home characteristics. The line through the middle of each individual box is the median, and the notches are a 95% confidence interval about the median. For example, Figure 3 demonstrates some variation in richness at different sites because there is little overlap between the different confidence intervals. Of all omnibus statistics, richness conditioned on site showed the most significant differences.

Analysis of variance (ANOVA) was used to determine the statistical significance of relationships between home variables and the omnibus statistics. Five linear models were computed with each omnibus statistic as the response variable and the home characteristics as the predictors. After performing the ANOVAs for each omnibus statistic, it was determined that site was the only statistically significant variable at the 5% level, with the exception of Simpson, for which none of the variables were significant. Tukey's Honest Significant Differences (HSD) test was used on the ANOVAs to assess the statistical significance of mean differences at each home characteristic variable.

Site	Mean Diff.	p-value
Bo - Ba	163.67	0.99
NY - Ba	-1861.44	0.018
StL - Ba	1698.13	0.0041
NY - Bo	-2025.11	0.016
StL - Bo	1534.46	0.030
StL - NY	3559.57	0.00

(a) Richness

Variable	Mean Diff.	p-value
Bo - Ba	-7.81	0.71
NY - Ba	-11.49	0.52
StL - Ba	11.43	0.32
NY - Bo	-3.69	0.98
StL - Bo	19.24	0.048
StL - NY	22.93	0.037

(b) Inverse Simpson

Variable	Mean Diff.	p-value
Bo - Ba	-0.0072	0.91
NY - Ba	-0.013	0.76
StL - Ba	0.020	0.19
NY - Bo	-0.0053	0.98
StL - Bo	0.027	0.071
StL - NY	0.032	0.056

(c) Evenness

Table 4: Tukey's HSD results for site

4.2 Correlation Analysis of Omnibus Statistics

The correlations between omnibus statistics were analyzed. The correlation heat map is in Figure 4, and the corresponding table of values can be seen in the Appendix in Table 14.

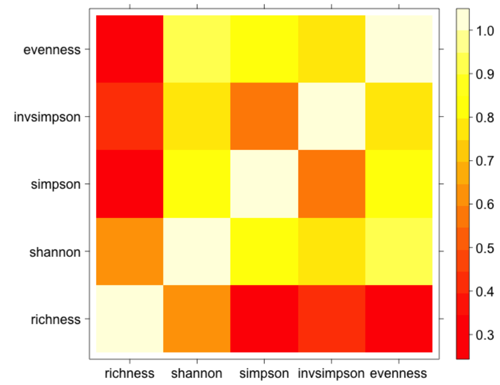


Figure 4: Heatmap comparing each pair of omnibus statistics

For each pair of omnibus statistics, a scatterplot of the ordered pairs was created. These can be seen in Figure 5, with the rows and columns named along the diagonal. The diagonal boxes plot the densities of each of the variables. Additionally, 3D plots were created for each combination of three omnibus statistics. One such figure is shown in Figure 16 in the Appendix. Finally, factor analysis was done on the diversity statistics. The results showed that even though the five diversity statistics were correlated, the dimension could not be reduced to one or two factors. Since the diversity statistics already encapsulate each participant's microbiome in only 5 variables, it is unsurprising that they could not be further reduced to fewer dimensions.

4.3 Principal Component Analysis

Principal component analysis (PCA) is a method that offers both dimensional reduction, and an alternative perspective to view data by representing the most salient features of a data set in a lower dimensional subspace. The method constructs linear combinations of observations with possibly correlated variables, which result in orthogonal vectors, or principal components (PCs). PCA chooses the orthogonal directions along which most of the variance in the data is captured. These PCs are essentially an orthogonal projection of the data into a subspace with dimension less than or equal to the number of variables, and they have a descending order of variance, such that the lower number PCs carry the largest variance of the data. The dimension of each PC is equal to the number of observations in the data since in fact they are a weighted average of the columns of the data matrix. Thus, only the first few PCs are needed to represent a

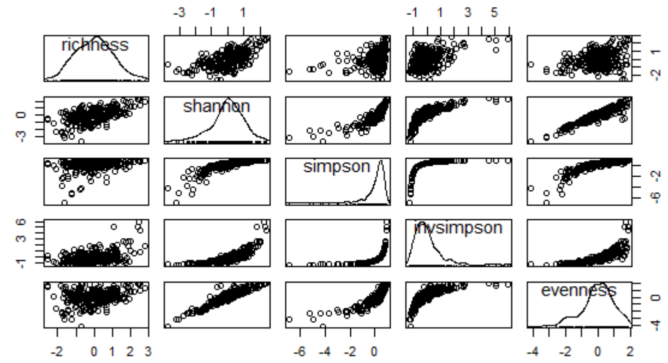
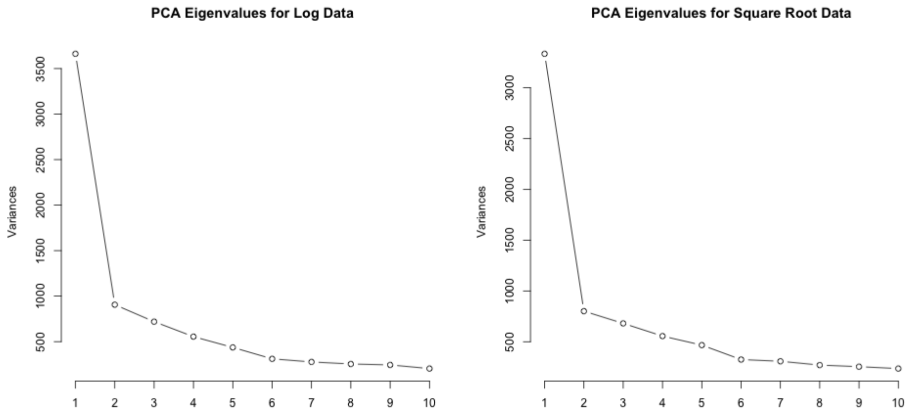


Figure 5: Scatterplot comparing each pair of omnibus statistics. Density plots are on the diagonal.

significant proportion of the variances in the data and therefore offer a reduced dimensional characterization of the data.

More specifically, the PCs are the eigenvectors of the covariance matrix (or the correlation matrix for standardized data) attained by performing singular value decomposition (SVD) on the data matrix. Using the `screeplot` function in R reveals the variance captured by the eigenvalues corresponding to the PCs. In practice, the PCs up to the first "elbow" are considered, or those eigenvalues up to the point where the change in variance drops drastically. PCA was performed on both the modified log and square root transformed versions of the data after they were centered and scaled. As illustrated in Figure 6, the elbow occurs at the second eigenvalue, so only the first two PCs are needed to capture the best variance in the data. Alternatively, the summary of the PCA object in R displays the cumulative proportion of the variance captured by each PC, and the decision for how many PCs to keep can be made by using more quantitative approaches. In the modified log and square root transformation cases, the first two PCs capture 22.39% and 20.26% of the variance respectively, which is a relatively significant proportion taking into consideration that the data being represented had over 20,000 variables.

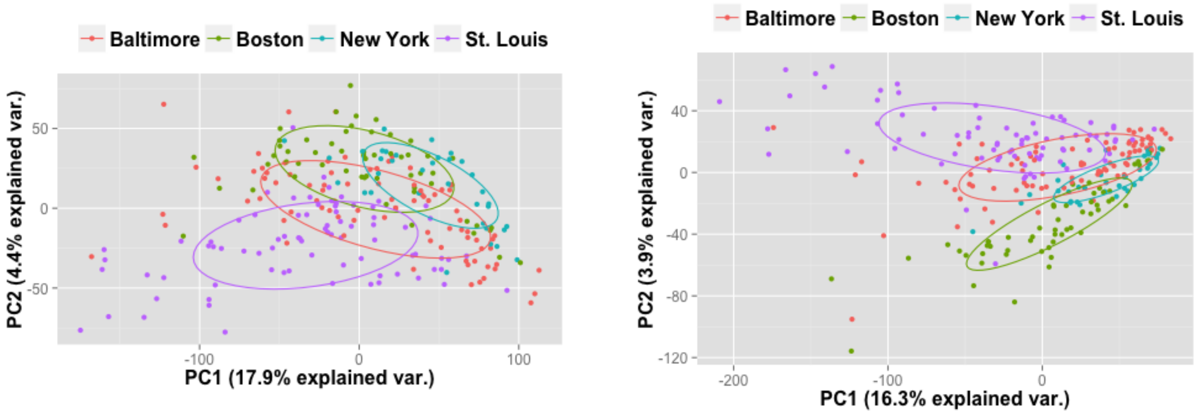


(a) Log transformed data

(b) Square root transformed data

Figure 6: Screeplots from PCA

A simpler visualization of the data is offered by plotting the order pairs given by the entries of the first two PCs. This results in the projection, or “shadow”, of the original data onto the subspace defined by the two orthogonal directions as illustrated in Figure 6 using `ggbiplot` in R. Furthermore, the points in these plots have been color coded corresponding to the site from where the samples originated, and a corresponding ellipse encompasses the 50% confidence region of the points located in the PC space. Note that PCA has revealed that samples sharing home site locations tend to have close relations when it comes to PC1 and PC2, which reinforces the significant change in variance seen in the box plots of Figure 3 in Section 4.1. Though the PCA made this revelation possible, color coding by other home characteristics did not display a clear separation in the projected subspace, and require further analysis to investigate how the OTUs affect their relations.



(a) Log transformed data

(b) Square root transformed data

Figure 7: Projection of the original data onto the subspace defined by the first two principle components.

4.4 Clustering

Through investigation, it was found that many unsupervised clustering methods, such as k-means clustering and random forest, do not work effectively. K-means clustering partitions data points into clusters to minimize the sum of squares within each cluster, but it only works under certain assumptions such as same variance across all variables [?], and this condition is not satisfied.

The variance of microbiomes across samples is extremely right-skewed, and in fact approximately 11% of the given 51,094 species have variance 1000 or higher. Hence it is no surprise that k-means gives a poor result, and when the number of clusters is set to 2, 3, 4, or 5, there is always one cluster of size larger than 200 elements. What is worse is that the total within-cluster sum of squares is higher than the between-cluster sum of squares, providing substantial evidence that k-means clustering does not work in this case.

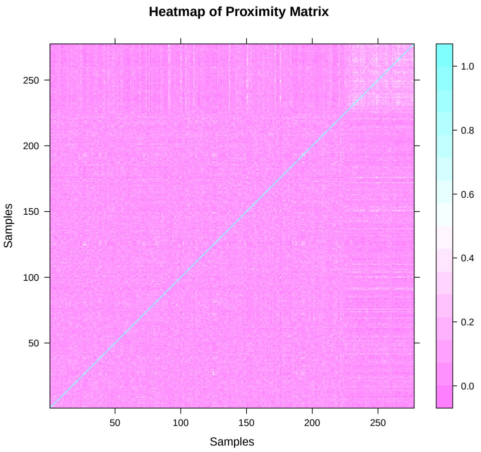


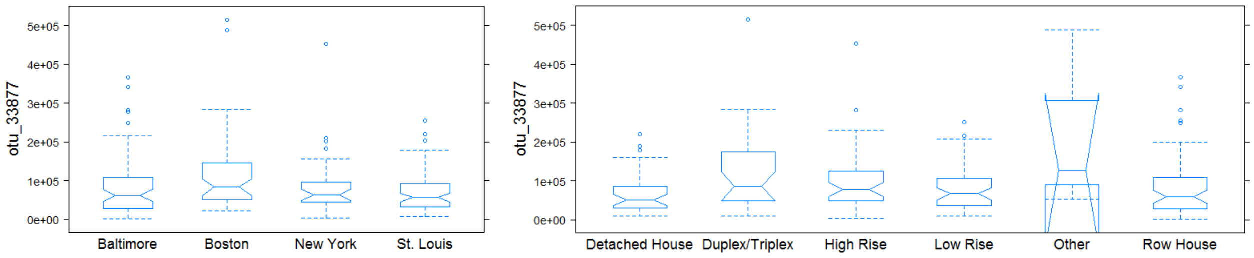
Figure 8: Proximity heat map for random forest method

The random forest method [?] constructs multiple decision trees, outputs the classification results by “voting”, and provides a proximity matrix based on how often data point pairs are classified in the same nodes. The proximity measure is from 0 to 1, so if data points are randomly assigned to two nodes, there is a 50% chance that they are in the same node. The highest proximity values for different pairs are approximately 0.53, so the data points are not highly correlated with each other. The matrix is plotted as a heat map in Figure 8; the white square at the top-right shows values close to 0.5, so no good clustering assignments can be inferred. Only the diagonal shows proximity equal to 1.0, and this is trivial because the diagonal values store the similarity between a point and itself. Therefore, random forest at this stage of analysis did not yield fruitful result.

4.5 Simple Linear Regression

Simple linear regression was performed to reveal associations between the presence of OTUs and different omnibus statistics or home characteristics. We selected two OTUs, OTU 33877 and OTU 27562, because they are particularly abundant in the homes. The analysis that follows could be repeated for any OTUs that are found to be significant in the future.

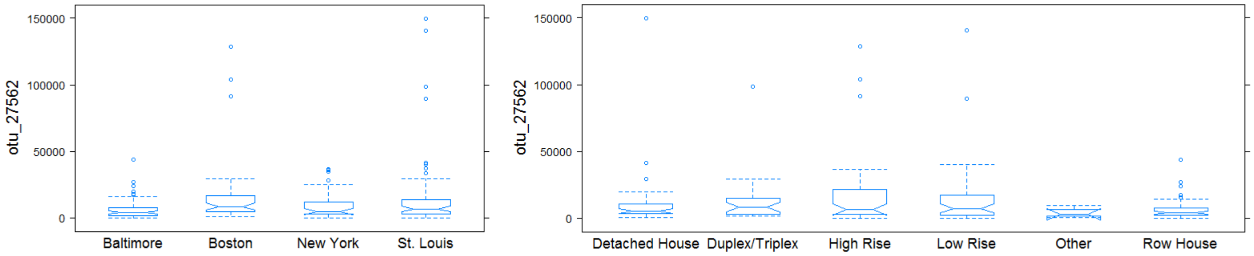
The median counts of the two species are 65,470 and 5,951, respectively. The boxplots for the two OTUs by site and house type are shown in Figures 9 and 10. These plots show that the number of these OTUs present in the samples is right-skewed, and many of the plots contain outliers.



(a) Boxplot by sites

(b) Boxplot by house types

Figure 9: OTU 33877



(a) Boxplot by sites

(b) Boxplot by house types

Figure 10: OTU 27562

Each OTU was further analyzed to determine whether the amount of the OTU could be predicted by a combination of the other home statistics.

4.5.1 OTU 33877

An ordinary linear model including all the home characteristics and omnibus statistics as predictors produces aliased coefficients due to collinearity; the correlation values are shown in Table 14 in the Appendix. Therefore, backward selection was performed to reduce the number of predictors, and the remaining variables were found to be `site`, `house_type`, `richness`, and `invsimpson`. Additionally, the variable `richness` showed a funnel shape when plotted against the number of OTU 33877, so a quadratic term is added. The case of `invsimpson` is similar.

The four linear models in Table 5 are compared using ANOVA. By adding site information to Model 1, the explanation power significantly increases. With this change, the house type variance is also accounted for, and the residual sum of squares decreases by 8%. It also appeared that the quadratic term of `richness` was insignificant, so another model was done without this term (Model 4). The ANOVA comparison results show little difference in model results between Model 3 and 4 (p-value = 0.565). Therefore, the 4th model is the best. Not all sites or house types have meaningful (converging) coefficients, but it was seen that the number of OTU 33877 increases by 11.379 ± 1.621 (95% confidence interval) when `richness` increases by one unit.

Models	Equations	Adjusted R^2
Model 1	OTU 33877 $\sim$ richness + richness^2 + invsimpson + invsimpson^2	0.2022
Model 2	OTU 33877 $\sim$ site + richness + richness^2 + invsimpson + invsimpson^2	0.2402
Model 3	OTU 33877 $\sim$ site + house_type + richness + richness^2 + invsimpson + invsimpson^2	0.2878
Model 4	OTU 33877 $\sim$ site + house_type + richness + invsimpson + invsimpson^2	0.2900

Table 5: Model results

4.5.2 OTU 27562

The simple linear model does not converge, and the adjusted R^2 for all models attempted is less than 5%. It was shown however that `home_mice` is significant in predicting the amount of OTU 27562, according to the individual t-test (p-value 0.0052). The boxplot in Figure 11 illustrates that the presence of mice is associated with lower numbers of OTU 27562.

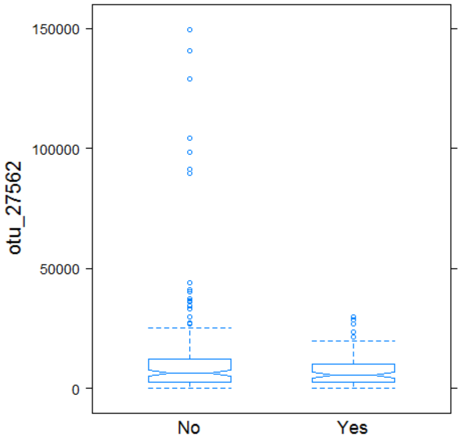


Figure 11: Boxplot by presence of mice for OTU 27562

4.6 Volcano Plots

Volcano plots are scatterplots that show the statistical significance versus the mean difference. Each OTU is represented by points on the plot, and each plot is associated with one binary variable.

In this section, volcano plots of home characteristics and sites are explored. The x-axis is the log odds associated with a binary variable, and the log odds are defined as below for a certain OTU species:

$$\text{log odds} = \log\left(\frac{\text{average OTU counts with variable = “yes”}}{\text{average OTU counts with variable = “no”}}\right)$$

The y-axis is the negative log of p-values, where the p-values are obtained from individual t-tests which compare the average OTU counts (based on the binary variable). Note that log refers to the natural logarithm.

4.6.1 Home Characteristics

There is a volcano plot for each binary variable – `cat_any`, `dog_any`, `home_mice`, `home_roach`, `home_rodents`, `home_waterprobs` (water problems). The variable `home_rats` was excluded because only three "yes" responses are present, compared with 251 "no" responses.

Figure 12 shows all six volcano plots corresponding to the binary variables. Above the red line indicates p-value < 0.05, i.e. statistical significance. Outside the region formed by the two blue lines implies the absolute value of log odds was larger than 2, i.e. at least $e^2 \approx 7.387$ times different in the average counts. Points (OTUs) that meet both criteria are considered "of interest" because they show statistically and practically significant difference with respect to the binary variable. These points are located in either the top-right or the top-left of the figures, and they are marked as solid green circles.

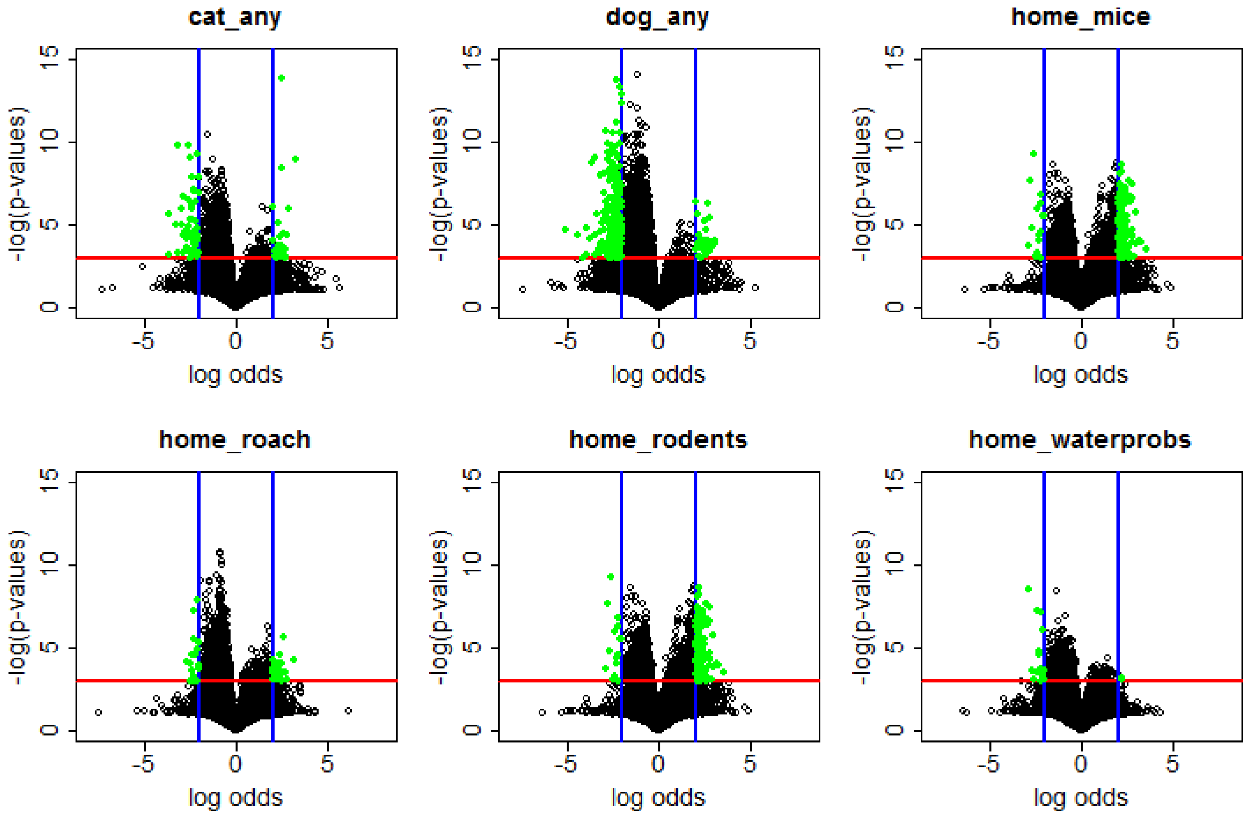


Figure 12: Volcano plots for each binary variable

The intersection matrix of OTUs for each variable is given in Table 6. Each variable has a corresponding group of OTUs "of interest", and the intersections of groups show how many OTUs they share. The OTUs of interest are the same for mice and rodents, presumably because mice are a type of rodent. All other variables share few OTUs, showing that most binary variables are dissimilar.

4.6.2 Sites

The volcano plots with respect to all four sites (Baltimore, Boston, New York, and St. Louis) are shown in Figure 13, and the format is similar to the one described in the previous section, but above the red line means p-value $< 10^{-6}$. Each site is treated as a binary variable, so the OTU counts are compared at one site versus all other sites. Since count data are used, the OTU data are not transformed.

	Cat	Dog	Mice	Roach	Rodents	Water
Cat	84	13	2	0	2	1
Dog	13	322	3	3	3	4
Mice	2	3	138	1	138	1
Roach	0	3	1	51	1	1
Rodents	2	3	138	1	138	1
Water	1	4	1	1	1	18

Table 6: Intersection matrix of OTUs for each binary variable

The intersection matrix is given in Table 7. The New York site has more than 1,000 OTU species associated with it; Boston and St. Louis sites have between 100 and 500; Baltimore has less than 10 OTU species that are statistically of interest. A large proportion of the OTUs at each site are shared with other locations. For example, 35 of the 47 total OTUs associated with St. Louis are shared with New York.

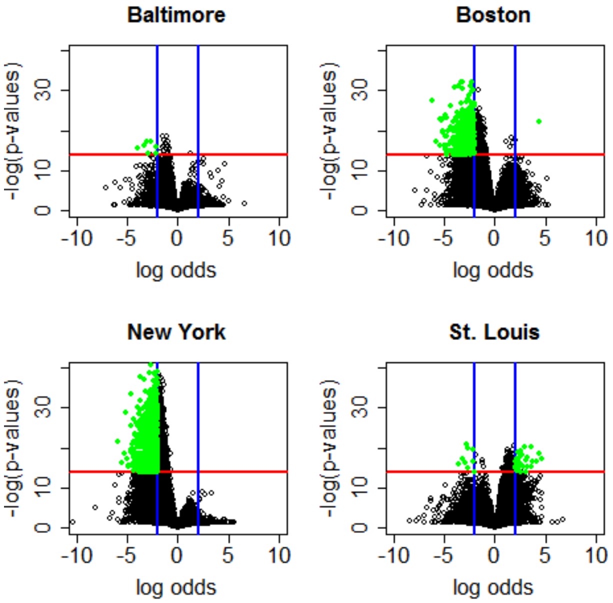


Figure 13: Volcano plots for each site vs OTUs

	Baltimore	Boston	New York	St. Louis
Baltimore	7	5	3	5
Boston	5	311	180	36
New York	3	180	1231	35
St. Louis	5	36	35	47

Table 7: Intersection matrix of OTUs for each site

5 Predictive Modeling

The high dimensional nature of this data set lends itself well to the elastic net method. An l_1 regularized logistic regression would be a natural choice for predicting the binary outcome variables, but one issue with this penalty is its treatment of correlated predictors. For example, if a group of predictors is highly correlated, an l_1 penalized regression might only select one variable from this group. Elastic net blends l_1 and l_2 penalties, which allows for variable selection along with the de-correlating properties of an l_2 penalty. For a penalized logistic regression, elastic net solves the following optimization problem [?]:

$$\min_{\beta_0,\beta} -\sum_{i=1}^{N}\left[y_i(\beta_0+\beta^T x_i)+\log(1+\exp\{\beta_0+\beta^T\})\right]+\lambda\left[(1-\alpha)\|\beta\|_2^2/2+\alpha\sum_{j=1}^{p}|\beta_j|\right] \tag{1}$$

Before fitting any linear models, the `findCorrelation` function in the `caret` package [?] was used to remove 696 OTUs with high pairwise correlations. Seven elastic net logistic regressions were fit with each home characteristic variable as the response and all 19,560 remaining OTUs as the predictors. A multinomial logistic regression with an elastic net penalty was performed on the site variable. A categorical variable G with $K > 2$ levels has a multinomial distribution. Instead of choosing a reference class and calculating multiple logistic regressions, the probability that G equals a particular class l is modeled as [?]:

$$\Pr(G = l|x) = \frac{\exp\{\beta_{0l} + x^T\beta_l\}}{\sum_{k=1}^{K}\exp\{\beta_{0k} + x^T\beta_k\}} \tag{2}$$

The reason for the parameterization is that coefficients can be calculated for each class. The objective function has two parts: a log likelihood and a penalty function.

$$\min_{\beta_0,\beta} -\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K}y_{ik}(\beta_{0k}+x_i^T\beta_k)-\log\left(\sum_{k=1}^{K}\exp\{\beta_{0k}+x_i^T\beta_k\}\right)\right)+\lambda\left[(1-\alpha)\|\beta\|_2^2/2+\alpha\sum_{j=1}^{p}|\beta_j|\right] \tag{3}$$

The α and λ elastic net parameters were tuned over a grid using the `train` function in `caret` [?]. Higher values of α correspond to higher weights on the l_1 penalty and lower values correspond to higher weights on the l_2 penalty. Five-fold cross validation was repeated three times for each regression. Table 8 presents results for binary home characteristics and site. The positive and negative columns in Table 8 (b) are the number of positive and negative coefficients, respectively. The values of these coefficients are in Tables 18 - 21 in the Appendix. All counts for the number of OTUs selected include the intercept. The elastic net algorithm did not converge for house type, possibly due to too few class counts for certain house types during cross validation. The rats home characteristic variable only had three "Yes" responses to 251 "No" responses, so it was not used as a predictor.

Variable	α	λ	# OTUs Selected
Dog	0.55	0.037	106
Roach	0.55	0.14	41
Water	1.0	0.076	23
Mice	0.55	0.14	36
Cat	0.55	0.017	166
Rodents	0.10	0.014	1289

(a) Binary variables

Site	# OTUs Selected	# Positive	# Negative
Baltimore	49	38	11
Boston	36	32	4
New York	28	20	8
St. Louis	36	34	2
$\alpha = 0.55$ & $\lambda = 0.024$			

(b) Site variable

Table 8: Elastic net results

No OTU was selected across all sites. However, OTU 22595 was selected for both Baltimore and St. Louis. Table 9 shows the number of intersected OTUs between all pairwise combinations of the binary home characteristics. All counts have the intercept removed.

The counts in Table 9 were compared to the results from computing the individual simple linear regressions with each OTU as the response variable and the various home characteristics as the predictor. Table 10 presents the number of OTUs that were found to be significant for both methods, for each home characteristic variable.

The predictive ability of elastic net was assessed by splitting the data into training and test sets then evaluating precision, recall and the F_1 score. For each home characteristic, 70% of the data was randomly permuted into the training set with the remaining in the test set. The coefficients found by training the model on 70% of the

	Cat	Dog	Mice	Roach	Rodents	Water
Cat	165	0	0	0	22	0
Dog	0	105	0	0	10	0
Mice	0	0	35	0	35	0
Roach	0	0	0	40	4	0
Rodents	22	10	35	4	1288	3
Water	0	0	0	0	3	22

Table 9: Number of intersected OTUs from elastic net

	Cat	Dog	Mice	Roach	Rodents	Water
Count	9	23	7	4	47	2

Table 10: Number of intersected OTUs from elastic net and simple linear regressions

data were used to predict outcomes in the test set. Classes were chosen by selecting the site with the highest probability. Precision, recall and F_1 are defined as:

$$\text{Precision} = \frac{tp}{tp + fp}$$
$$\text{Recall} = \frac{tp}{tp + fn}$$
$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

where tp is the number of true positives (predict “Yes” when true outcome is “Yes”), fp is the number of false positives (predict “Yes” when true outcome is “No”) and fn is the number of false negatives (predict “No” when true outcome is “Yes”). The results for precision, recall F_1 as well as the confusion table for site are presented in Table 11. Predicted outcomes in the confusion table are in **bold** text.

Variable	Precision	Recall	F_1
Cat	0.69	0.50	0.58
Dog	0.63	0.42	0.50
Mice	0.61	0.63	0.62
Roach	0.50	0.23	0.32
Rodents	0.60	0.40	0.48
Water	0.42	0.33	0.37

(a) Binary variable accuracy measures

	Ba	Bo	NY	StL
Ba	22	0	0	0
Bo	0	15	0	0
NY	0	0	10	0
StL	1	0	0	23

(b) Confusion table for site

Table 11: Elastic net prediction results

Logistic regression with an elastic net penalty provides predictive accuracy and variable selection, but other methods could outperform elastic net with regards to prediction. The support vector machine (SVM) is also a shrinkage method that can be used with high dimensional data, but its coefficients can be difficult to interpret. The predictive power of SVM was assessed by once again dividing the data into 70% training and 30% test sets, then calculating precision, recall and the F_1 score.

Table 12 shows that the performance of SVM is unsatisfactory for the binary variables. For three of the home characteristic variables, all observations were predicted to be “No”, which resulted in division by zero when calculating precision and F_1. “NaN” in this table stands for not a number. The misclassification error for site is about 21% and the confusion table is reported in Table 12. Predicted outcomes are in **bold** text. The predictive ability of SVM is worse than elastic net but is still acceptable.

Variable	Precision	Recall	F_1
Cat	NaN	0	NaN
Dog	NaN	0	NaN
Mice	0.63	0.17	0.26
Roach	0.60	0.20	0.30
Rodents	0.88	0.23	0.37
Water	NaN	0	NaN

(a) Binary variable accuracy measures

	Ba	Bo	NY	StL
Ba	19	2	5	3
Bo	0	14	0	0
NY	0	0	5	0
StL	5	0	0	20

(b) Confusion table for site

Table 12: Support vector machine prediction results

The random forest algorithm was used as a classification method to model the home characteristics such as the dog ownership versus the OTU counts. Random forests is an ensemble learning method that produces many classification trees rather than a single tree. One benefit of this method is its ability to handle large numbers of variables without deletion and to give a measure of variable importance which is useful for model selection.

This type of algorithm is especially useful when backward variable selection is not appropriate. When we fit random forests with the training data set (approximately two-thirds of the data), we use bootstrapping to draw samples with replacement from the data set to grow the current tree. The remaining one-thirds of the data is used to get an unbiased estimate of classification error as trees are added to the forest. For our random forest classification we created 254 trees with 70% of the data as training data and 30% of the data as test data. The results for precision, recall, and the F_1 score are listed below. The misclassification error rate for the confusion table is about 10%.

Variable	Precision	Recall	F_1
Cat	0.86	0.67	0.75
Dog	0.80	0.33	0.47
Mice	0.64	0.53	0.58
Roach	0.67	0.60	0.63
Rodents	0.81	0.57	0.67
Water	0.50	0.17	0.25

(a) Binary variable accuracy measures

	Ba	Bo	NY	StL
Ba	21	0	2	1
Bo	0	15	0	0
NY	0	1	8	0
StL	3	0	0	22

(b) Confusion table for site

Table 13: Random forest prediction results

The variable importance plot is a critical output of the random forest algorithm. It shows the importance of each OTU in the data matrix in classifying the data. Figure 17 in the Appendix shows each OTU on the y-axis, and their importance on the x-axis. They are ordered top-to-bottom as most- to least-important. Therefore, the most important OTUs are at the top and an estimate of their importance is given by the position of the dot on the x-axis. OTU 8680 was the most important OTU for sites according to Figure 17.

The function `partialPlot` in `randomForest` [?] package in R was used to create partial dependence plot which gives a graphical depiction of the marginal effect of a specific OTU on the site probability. OTU 8680 was chosen from variable importance plot to be the OTU of choice. This function was run for the four different sites, Baltimore, Boston, New York, and St. Louis; the plots are depicted in Figure 14. Random forest is a non-parametric model and as it is apparent from Figure 14 a non linear relationship exists between the probability of a specific site with respect to the log transformed OTU 8680 counts.

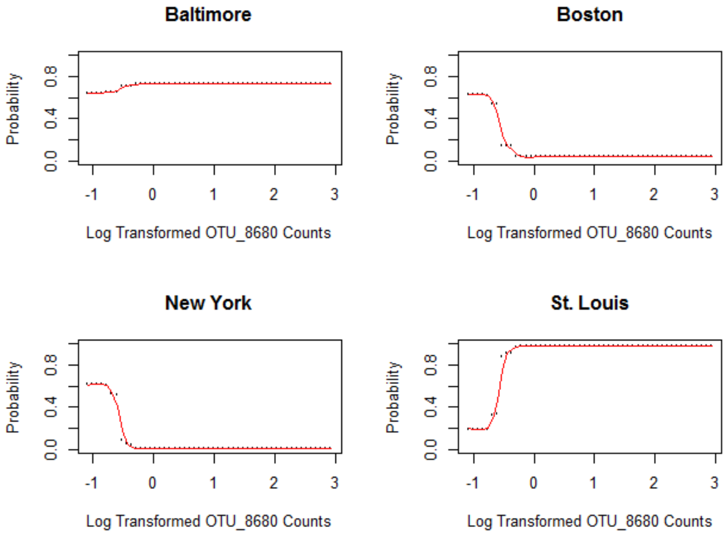


Figure 14: Random forest partial plots

6 Summary and Future Work

6.1 Results

The omnibus statistics provided information as an overview of the OTU data, and they shared common traits throughout the sample data as they showed a high correlation within the different statistical measurements. The `richness` statistic inferred that the site location carried the largest difference in variance. Exploration of the data made use of PCA to probe deeper into the details of the OTU data, and to offer an alternative perspective to view the high dimensional data set. Doing so confirmed that the site locations to be the home characteristic that most captured the largest differences in the microbiome composition. This means that the city location of a household most greatly impacts the presence or absence of particular OTUs in the microbiome. To explore the other home characteristics, simple linear regression was applied to two OTUs (OTU 33877 and OTU 27562) that were particularly abundant throughout the sampled data, and the diversity statistics were used as predictors. Though these models did not converge, it was shown that the presence of mice in a home was associated with lower numbers of OTU 27562.

Using volcano plots revealed those OTUs that had statistically and practically significant differences with respect to the binary variables, and that indeed each home characteristic is associated with particular subsets of OTUs. Furthermore, the data intersections among the OTUs was highest among mice and rodents, meaning that samples with mice and/or rodents share similar OTUs. This was not the case for the other binary variables associated with home characteristics, and so the OTUs associated with dogs, cats, roaches, and water damage were not shared among samples. Binary representations of the home site locations were also considered for making volcano plots by taking each city versus the other three cities. Focusing on the sites highlighted the number of OTUs associated to each city, while indicating that the same OTU sets were not shared among cities. New York and Boston were found to be most significantly affected by a lack of OTUs, further reinforcing previously claims on the variance of OTUs between different sites.

Predictive models were fit to the OTU data in order to test the ability of the microbiome configuration to infer home characteristics. To create a test case scenario, 70% of the data was use to train the models and tune parameters, while their prediction capability was tested with the remaining 30%. Elastic net logistic regression was fit using five-fold cross validation, where seven home characteristics were set as responses (rats home characteristics were omitted, and the approach did not converge for house types), and 19,560 OTUs were set as predictors (696 OTUs were omitted due to high pairwise correlation). By doing so, OTUs associated with each home characteristic were successfully identified. As motivated by the data exploration analysis, further predictive analysis revealed more detail about positive or negative affect of OTUs in selecting the home site location. The predictive modeling also shed light on the uniqueness of the OTUs associated with each home characteristic, in that no OTUs were selected across all sites, however one OTU was selected for both Boston and St. Louis. The number of intersected OTUs via elastic net was also compared to those from the simple linear regression, . Finally, the predictive capability of the elastic net model yielded only one OTU being misclassified to the wrong site location. The same procedure was repeated using SVM, which yielded acceptable results, though not as good as elastic net.

The random forest classification method was used to model different home characteristics against the OTU frequency counts. This approach suggested that non-linear relationships exist between the site probabilities and the most statistically significant log transformed OTU. Also, the prediction results from random forest are similar to elastic net results, which is further confirmation of strong association between sites and OTUs.

6.2 Conclusion

OTU data originating from the URECA birth cohort study was analyzed to study the connections between the microbiome found in bedroom dust samples and their corresponding home characteristics with the ultimate intention of identifying those members in the microbiome responsible for asthma and allergy development or prevention. For this stage of the study, the objective was to make connections between OTUs and the home characteristics of the collected samples (e.g. which OTUs are associated with a home having dogs, etc.). The omnibus statistics were first considered to achieve an understanding of the samples with an overall representation of the OTU data, and correlation analysis on this portion of the data directed the focus towards attempts at separating the data by home characteristics.

Next, the detailed OTU data was considered, though it first required some processing. Samples with missing home characteristics OTUs displaying near zero variance were removed, leaving 277 homes and 20,402 OTUs to be considered for analysis. This offered a near 60% reduction from the original OTU data set, which aided the high-dimensional nature of the problem, though it did not completely cure the fact that we have far more variables than samples. In order to improve the right-skewed nature of the data, a modified log transformation was selected for improving the symmetry of the data, and to achieve a resemblance closer to a normal distribution. This reduced and transformed data was then appropriate to be used for the analysis that followed.

Upon conducting exploratory analysis through correlation analysis on the omnibus statistics, the site locations were the home characteristic that stood out the most in terms of differences in microbiome configurations. PCA confirmed this, and showed promise for separability when considering some sites, though there may be some non-orthogonal relations that would explain the separations better. Simple linear regression analysis also reinforced the idea of uniqueness in OTUs between site locations.

Finally, attempts for predictive modeling returned promising results with encouraging success rates at identifying the site location when using microbiome configuration as predictors. Using the elastic net approach also suggested that none of the considered OTUs were shared among sites, which provided further support to the idea of sites associating with separate groups of OTUs in the microbiome. Using other approaches such as SVM and random forest returned similar results that were acceptable, but not as successful as elastic net, which was identified as most suitable for the high-dimensional data being analyzed.

6.3 Future Work

For each observation in this study, the count of each OTU was reported. Methods have been developed for variable selection in high dimensional compositional data, which treats the variables in each observation as proportions of a whole [**?**]. Future work should explore the application of these methods to this data set.

Principal component analysis was used to visualize the data in lower dimensions and assess separability in the site variable. The principal components can be used in conjunction with prediction techniques such as logistic regression and support vector machine to perform prediction in a lower dimensional subspace. In addition, canonical correlation analysis and its sparse analog can also be used to further investigate the correlation structure of the OTUs.

The analysis in this paper did not utilize the OTU class characteristics such as the phylum. Future work could investigate the commonalities in phylogenetic traits between the OTUs selected by elastic net. In addition, variants on group lasso could be used to perform variable selection along the OTU class characteristics.

7 Appendix

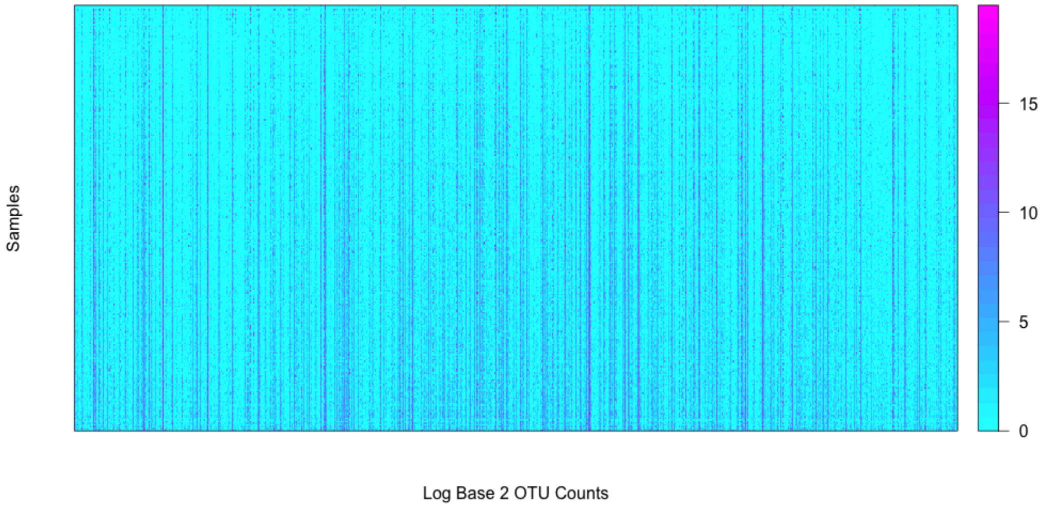


Figure 15: Heat map of OTU data with near zero variance OTUs removed

	richness	shannon	simpson	invsimpson	evenness
richness	1.00	0.60	0.31	0.43	0.29
shannon	0.60	1.00	0.81	0.79	0.93
simpson	0.31	0.81	1.00	0.57	0.83
invsimpson	0.43	0.79	0.57	1.00	0.76
evenness	0.29	0.93	0.83	0.76	1.00

Table 14: Correlation between diversity statistics

Site	Mean Diff.	*p*-value
Bo - Ba	-0.032	0.99
NY - Ba	-0.22	0.35
StL - Ba	0.32	0.011
NY - Bo	-0.18	0.55
StL - Bo	0.35	0.012
StL - NY	0.54	0.00

(a) Shannon

Site	Mean Diff.	*p*-value
Bo - Ba	0.0042	0.81
NY - Ba	0.0012	0.99
StL - Ba	0.010	0.087
NY - Bo	-0.0030	0.95
StL - Bo	0.0059	0.60
StL - NY	0.0088	0.36

(b) Simpson

Table 15: Tukey's HSD results for site

Variable	Mean Diff.	*p*-value
Water	1036.71	0.64
Mice	265.53	0.51
Rats	-176.53	0.92
Roach	237.65	0.56
Dog	-188.95	0.73
Cat	151.31	0.74

(a) Richness

Variable	Mean Diff.	*p*-value
Water	0.12	0.16
Mice	-0.0042	0.96
Rats	-0.96	0.80
Roach	0.12	0.15
Dog	0.019	0.87
Cat	0.13	0.18

(b) Shannon

Table 16: Tukey's HSD results

Variable	Mean Diff.	p-value
Water	0.0070	0.054
Mice	-0.00060	0.86
Rats	0.00024	0.99
Roach	0.0039	0.26
Dog	0.0027	0.56
Cat	0.0042	0.28

(a) Simpson

Variable	Mean Diff.	p-value
Water	2.79	0.62
Mice	-0.17	0.98
Rats	5.34	0.83
Roach	6.68	0.22
Dog	-5.33	0.47
Cat	11.40	0.065

(b) Inverse Simpson

Variable	Mean Diff.	p-value
Water	0.013	0.14
Mice	-0.0021	0.80
Rats	-0.013	0.71
Roach	0.012	0.14
Dog	0.0013	0.90
Cat	0.014	0.13

(c) Evenness

Table 17: Tukey's HSD results

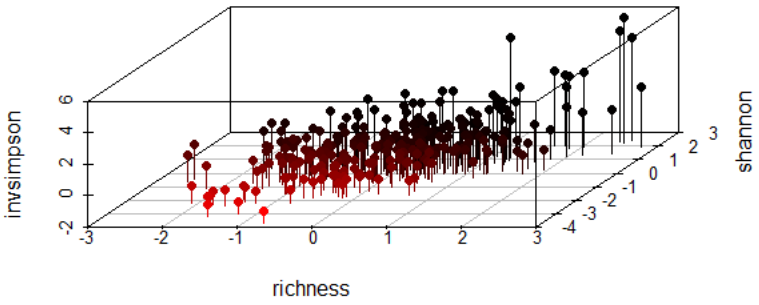


Figure 16: 3D Plot comparing Shannon index, richness, and inverse Simpson index

	Variable	Coefficient
1	int	0.71
2	otu_38331	0.43
3	otu_36477	0.37
4	otu_40751	0.29
5	otu_25773	0.29
6	otu_43266	0.23
7	otu_30983	0.23
8	otu_23267	0.16
9	otu_551	0.16
10	otu_8680	0.15
11	otu_208	-0.14
12	otu_2203	0.14
13	otu_3510	0.11
14	otu_16727	0.11
15	otu_17253	-0.11
16	otu_46873	0.10
15	otu_17253	-0.11
16	otu_46873	0.10
17	otu_20483	0.09
18	otu_39826	0.09
19	otu_46410	0.09
20	otu_41940	0.08
21	otu_32717	0.07
22	otu_36748	0.05
23	otu_23364	0.05
24	otu_36899	0.05
25	otu_30475	-0.05
26	otu_47865	0.05
27	otu_25987	-0.04
28	otu_47479	0.04
29	otu_44150	0.04
30	otu_5837	0.04
31	otu_50722	0.04
32	otu_34904	0.03
33	otu_1478	-0.03
34	otu_4959	0.03
35	otu_29222	-0.02
36	otu_19118	-0.02
37	otu_23509	0.01
38	otu_31600	0.01
39	otu_29300	0.01
40	otu_22595	0.01
41	otu_46140	-0.01
42	otu_41417	0.01
43	otu_40582	0.01
44	otu_47867	-0.00
45	otu_33873	-0.00
46	otu_22781	0.00
47	otu_45119	-0.00
48	otu_41493	0.00
49	otu_3574	0.00

Table 18: Elastic net coefficients for Baltimore

	Variable	Coefficient
1	otu_32505	0.42
2	int	-0.36
3	otu_5955	0.19
4	otu_47462	0.19
5	otu_19602	0.18
6	otu_44431	0.18
7	otu_24883	0.17
8	otu_50697	0.15
9	otu_30298	0.14
10	otu_36013	0.11
11	otu_50764	0.10
12	otu_28823	0.10
13	otu_32235	0.09
14	otu_40942	-0.09
15	otu_37958	-0.08
16	otu_45219	0.07
17	otu_8642	0.06
18	otu_34021	0.06
19	otu_24863	0.05
20	otu_42163	0.03
21	otu_35547	0.03
22	otu_138	0.02
23	otu_22517	0.02
24	otu_42931	0.02
25	otu_11863	0.02
26	otu_17892	-0.02
27	otu_32954	0.02
28	otu_38943	0.02
29	otu_1862	0.01
30	otu_42373	0.01
31	otu_16188	0.01
32	otu_10134	0.01
33	otu_581	0.01
34	otu_32800	0.00
35	otu_14350	0.00
36	otu_49776	0.00

Table 19: Elastic net coefficients for Boston

	Variable	Coefficient
1	int	-0.87
2	otu_29154	0.57
3	otu_28857	0.31
4	otu_12633	0.30
5	otu_4203	0.18
6	otu_28230	0.16
7	otu_9102	0.09
8	otu_12599	0.08
9	otu_17168	0.07
10	otu_12257	0.07
11	otu_34017	-0.07
12	otu_36973	0.06
13	otu_33940	0.06
14	otu_47167	0.06
15	otu_7108	-0.06
16	otu_30394	0.05
17	otu_32946	0.05
18	otu_14742	0.04
19	otu_31200	-0.04
20	otu_13752	0.03
21	otu_41586	0.03
22	otu_26370	-0.03
23	otu_26085	0.02
24	otu_4251	0.02
25	otu_6165	0.01
26	otu_31	-0.01
27	otu_44358	-0.01
28	otu_31831	-0.01

Table 20: Elastic net coefficients for New York

	Variable	Coefficient
1	int	0.52
2	otu_10852	0.46
3	otu_46521	0.33
4	otu_30194	0.26
5	otu_22120	0.25
6	otu_15105	0.25
7	otu_9940	0.19
8	otu_32493	0.17
9	otu_801	0.16
10	otu_21965	0.15
11	otu_19195	0.14
12	otu_5862	0.14
13	otu_49957	0.13
14	otu_14651	0.13
15	otu_5764	0.11
16	otu_32052	-0.11
17	otu_885	0.11
18	otu_39473	0.10
19	otu_22595	-0.09
20	otu_48213	0.09
21	otu_25876	0.07
22	otu_259	0.06
23	otu_48534	0.06
24	otu_42860	0.05
25	otu_47935	0.04
26	otu_38852	0.04
27	otu_34228	0.04
28	otu_27147	0.03
29	otu_32832	0.03
30	otu_20546	0.03
31	otu_14948	0.02
32	otu_19183	0.02
33	otu_39834	0.02
34	otu_46551	0.01
35	otu_33140	0.01
36	otu_10561	0.00

Table 21: Elastic net coefficients for St. Louis

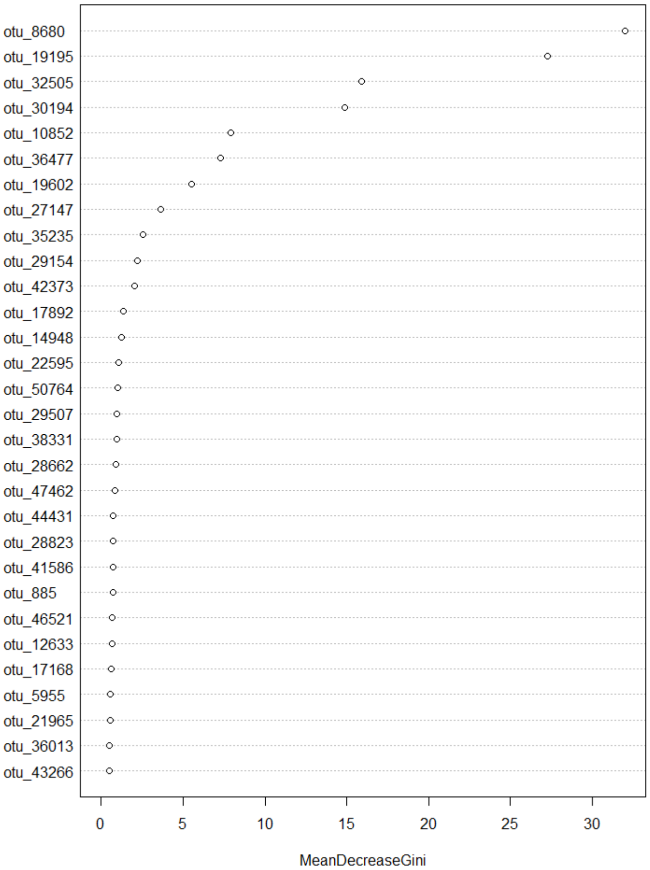


Figure 17: Variable importance plot

Listing 1: Function for training the model

```
trainfit <- function(var_name, x, y, folds, repeats, std){
  #x is the matrix with predictors
  #y is the matrix with response(s)
  #select a response by setting var_name = "response", for example,
     var_name = "dog_any"
  #folds is the number of folds for cross validation
  #repeats is the number of times to repeat CV
  #set std = 1 for standardization, std = 0 for no standardization

  #x matrix with intercept for selecting column names of nonzero
     predictors
  x_int <- cbind(int = rep(1, dim(x)[1]), x)

  nlevels <- length(levels(y[,var_name]))

  if(nlevels > 2){
    fam <- "multinomial"
  }
  else{
    fam <- "binomial"
  }

  fitControl <- trainControl(method = "repeatedcv", number = folds,
     repeats = repeats)

  glmfit <- train(x = x, y = y[,var_name], method = "glmnet", trControl =
     fitControl, standardize = std, family = fam)

  fit.std <- glmfit

  if(nlevels <= 2){
  coef.std <- coef(fit.std$finalModel, s = fit.std$finalModel$lambdaOpt)
  nzero.std <- coef.std != 0
  totnzero.std <- sum(nzero.std)
  indnzero.std <- colnames(x_int)[which(nzero.std)]
  }
  else{
    totnzero.std <- NULL
    indnzero.std <- NULL
  }


  #tot is the total number of non-zero predictors
  #ind is the column names of non-zero predictors
  #model.fit is the object returned by train
  return(list(tot = totnzero.std, ind = indnzero.std, model.fit  = fit.std
     , optimal_param = glmfit$bestTune))

}
```

```
#example use
#dog_caret <- trainfit("dog_any", x, y, folds = 5, repeats = 3)
```

Listing 2: Function for prediction

```
source("trainfit.R")
prediction <- function(var_name, x, y){

  cdp <- createDataPartition(y = y[,var_name], times = 1, p = .7, list =
     FALSE)

  train_y <- y[cdp,]
  train_x <- x[cdp,]

  test_y <- y[-cdp,]
  test_x <- x[-cdp,]

  train_roach <- trainfit(var_name, x = train_x, y = train_y, folds = 5,
     repeats = 3, std = 1)
  pred_roach <- predict(train_roach$model.fit$finalModel, newx = as.matrix
     (test_x),
                        s = train_roach$model.fit$finalModel$lambdaOpt,
                           type = "response")
  pred_val <- as.matrix(round(pred_roach))

  miscl_table <- as.matrix(table(pred_val,test_y[,var_name]))

  prec <- miscl_table[2,2]/(miscl_table[2,2] + miscl_table[2,1])
  rec <- miscl_table[2,2]/(miscl_table[2,2]+miscl_table[1,2])

  ff <- 2*(prec*rec)/(prec+rec)

  return(list(miscl_table = miscl_table, precision = prec, recall = rec,
     f_score = ff))

}
```

References

[1] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[2] Michael Hahsler, Kurt Hornik, and Christian Buchta. Getting things in order: An introduction to the r package seriation. *Journal of Statistical Software*, 25(3):1–34, 2008.

[3] Ranjit Kumar. Bioinformatics analysis and interpretation of microbiome data. `http://www.uab.edu/medicine/camac/images/10-A-Kumar_slide_show.pdf`. Online; accessed July 2015.

[4] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Applications*, 2:73–94, 2015.

[5] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[6] Wei Lin, Pixu Shi, Rui Feng, and Hongzhe Li. Variable selection in regression with compositional covariates. *Biometrika*, pages 1–13, 2014.

[7] Susan Lynch. Urban microbes, allergens and cytokine patterns in the development of asthma. In *AAAAI Annual Meeting*, February 2015.

[8] et al. Max Kuhn. caret: Classification and regression training. `http://CRAN.R-project.org/package=caret`, 2015. R package version 6.0-52.

[9] David Robinson. K-means clustering is not a free lunch. `http://varianceexplained.org/r/kmeans-free-lunch/`. Online; accessed 2015.

[10] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011.

[11] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

[12] Vincent Q. Vu. ggbiplot: A ggplot2 based biplot. `http://github.com/vqv/ggbiplot`, 2011. R package version 0.55.

[13] Daniela M Witten, Ali Shojaie, and Fan Zhang. The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics*, 56(1):112–122, 2014.

[14] Nathan Wolfe. Small, small world. `http://ngm.nationalgeographic.com/2013/01/125-microbes/oeggerli-photography`. Online; accessed July 2015.

[15] Ed Yong. How to make better predictions from our gut microbes. `http://phenomena.nationalgeographic.com/2015/07/15/how-to-make-better-health-predictions-from-our-gut-microbes/`. Online; accessed July 2015.

[16] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.