

# The Nasal Microbiome and it's Contribution to Acute Respiratory Diseases

Srikanth Aravamuthan<sup>1</sup>, Briahnna Austin<sup>2</sup>, Kaitlin Dio<sup>3</sup>, David Perez-Suarez<sup>4</sup>, Yu Shi<sup>5</sup>, Qiyiwen Zhang<sup>6</sup>, Tianchi Zhang<sup>7</sup>

Problem Presenters:  
Agustin Calatroni, Petra LeBeau, and Hoang Tran  
Rho, Inc.

Faculty Mentor:  
Eric Chi  
North Carolina State University

## Abstract

This study analyzes data collected from an inner city asthma consortium to investigate the association between microbes in the environment and those found in the nose and their relation to allergic and respiratory diseases. The study of allergies and wheezing are of great importance. Approximately, 45 % of the US population is allergic to at least one allergen and respiratory illness are among the most expensive medical conditions. According to the World Health Organization<sup>8</sup>, many preventable respiratory diseases are inadequately discovered and/or treated. Respiratory diseases are especially prevalent in childhood and according to the CDC, *allergies can affect a child's physical and emotional health and can be life threatening.*<sup>9</sup> Our aim is to utilize statistical methods to recognize if certain environmental and/or nasal bacterial communities are related to the development of allergic diseases that lead to asthma.

## Keywords

Microbiome, Microbial Communities, Respiratory Disease, Asthma, Allergy, Atopic, Health Impacts, Immune system, Nasal Pathways, Environmental Sampling, Species Abundances, Operational Taxonomical Units Hypothesis, SVM, Random Forrest, Logistic , Birth Cohort, Repeated Assessments

---

<sup>1</sup>Western Michigan University

<sup>2</sup>California State University Long Beach

<sup>3</sup>University of Rhode Island

<sup>4</sup>The University of North Carolina at Greensboro

<sup>5</sup>University of California Los Angeles

<sup>6</sup>Washington University in St. Louis

<sup>7</sup>Virginia Commonwealth University

<sup>8</sup>NMH Fact Sheet June 2009 [http://who.int/nmh/publications/fact\\_sheet\\_respiratory\\_en.pdf](http://who.int/nmh/publications/fact_sheet_respiratory_en.pdf)

<sup>9</sup>CDC Trends in Allergic Conditions Among Children: United States, 1997?2011, 2003 <http://www.cdc.gov/nchs/products/databriefs/db121.htm>

# 1 Introduction

We are interested in analyzing the microbiome of the environment and nasal passages to understand its role in the development and prevention of human respiratory and allergic diseases. The human body can be seen as a set of environments, in which each environment has a unique but sometimes overlapping group of microorganisms. These environments are the human microbiomes which consist of both microorganisms like bacteria, archaea, and fungi and non-microorganisms such as viruses. Immune system responses are associated with the imbalance of a microbiota and can be indications of infectious (communicable) diseases.

Imbalance to the Human microbiota (our microbiome ecosystem) have been found to be associated with respiratory, neurological and endocrine disorders and diseases. Conditions such as asthma/allergies, depression, and diabetes have been found to be related to certain imbalances to the human microbiome. The microbial composition of individuals with specific phenotypes (observable characteristics) in comparison to others lacking those phenotypes can shed light on the connection between the microbiome and certain characteristics. The primary goal of this analysis is to identify how microbes in the dust and nasal cavity are associated with the development of allergy and recurrent wheezing. We are interested in understanding the childhood development of atopy or recurrent wheeze, which are risk factors of asthma.

To address our goals in understanding the association between microbes in the environment and in the nose, as well as the relation to later childhood development of respiratory illness, we will be looking into operational taxonomic units (OTUs). Operational taxonomic units are used to categorize microbes into taxa based on genetic sequence dissimilarities, where taxa are any named group of organisms. Likewise, we can think of OTUs representing the number of clusters of similar sequences called phylotypes; while there is no gold standard, 16S rRNA gene deep sequencing is the most common method of OTU grouping. For this study OTUs are linked to sample type (dust environment or nasal cavity) as well as phenotype (presence or absence of atopy/wheeze or asthma). The data is taken at four time points: environmental dust sample at 3 months, nasal cavity sample at 1 year, allergy/wheeze phenotype at 3 years, and asthma phenotype at 7 years. For this study, a phenotype sample is a observable characteristic measurements unlike the dust/nasal microbe samples are genetically sequenced.

## 1.1 Objectives

- 1.) Examine the association between microbes in the environment and those found in the nose
- 2.) Obtain an understanding of how microbes in the environment and nose relate to the development of allergic diseases and recurrent wheezing in early life

## 1.2 Biological Mechanisms

It is essential to discuss how biological or statistical biases can arise during an analysis, here we will highlight suspected biological biases. Since microbiome ecosystems impact each other, it is vital to understand how microbial adhesion can impact the microbial composition and distribution regardless of phenotype differences. Moreover, there exist interpersonal variation of microbial composition as well as underlying methods to OTUs grouping which can lead to different conclusion about the human microbiota. Moreover, Shukla, Budden et. al found that *the diversity of nasopharyngeal microbiota has a fundamental role in determining the host susceptibility to febrile lower respiratory infections and the development of asthma later*

*in life.*<sup>10</sup> Their findings imply the biodiversity of microbiota has some relation to development of respiratory illness which is the goal of this analysis. We should keep the aforementioned concepts in mind as we proceed through analysis.

Our analysis begins with analyzing the association between the nasal and dust microbiomes. This analysis includes correlation analysis and volcano plots. The analysis continues on to look at the relationship between the omnibus statistics and their relationship to the event utilizing directed cyclic graph modeling and mediation analysis. Finally we employ methods to identify individual OTUs related to the occurrence of the event. This is completed using both unsupervised and supervised learning methods consisting of correlation analysis, sparse principal component analysis, volcano plots, random forest, sparse support vector machines and elastic net. To identify the important OTUs, cross referencing the techniques created a list. These OTUs were then explored for their phylogenetic distribution and relationship to the event.

## 2 Data Description

### 2.1 URECA Study

The development of asthma predominantly occurs during childhood and is not distributed evenly throughout the population of the United States. Higher rates of asthma and morbidity due to asthma appear among ethnic minorities and children growing up in poor, urban neighborhoods. There are several environmental and lifestyle factors associated with the development of asthma, especially in the first years of life, that are especially predominant in urban environments. Among children with certain allergic conditions, the occurrence of virus-induced wheezing episodes in the first 2-3 years of life is a strong risk factor for asthma. This is especially important since asthma cannot typically be detected until several years later.

To begin to identify environmental and lifestyle factors related to the development of asthma in urban populations in the U.S., the Inner City Asthma Consortium began the Urban Environment and Childhood Asthma (URECA) study in 2004<sup>11</sup>. The URECA study is an ongoing longitudinal birth cohort that follows participants from before birth through adolescence with repeated clinical, nose and phone check-ups.

The URECA study consists of 560 subjects recruited from Baltimore, Boston, New York and St. Louis. To participate in the study, the participants had to live in neighborhoods where at least 20% of population had an income before the poverty line, and at least one of the parents had allergic diseases or asthma. The participants were recruited prenatally and must be born at 34 weeks gestation or later and without significant respiratory problems in the neonatal nursery.

### 2.2 Data Sample

The data examined in this analysis consists of a subset of 74 participants of the URECA study. The data includes the home dust samples for this subset taken at 3 months of age, nasal wash samples at age 1 year and subject level information including the presence of atopic or recurrent wheezing conditions at age 3 years and asthma at age 7 years.

The dust and nasal samples include the microbial composition in the form of 10,118 OTU (Operational Taxonomic Units) readings within each sample. The OTU composition is represented within each sample as a count, where each sample row adds to the same number. A partial phylogenetic tree was available for each OTU including the kingdom, phylum and class structure. Each sample also included alpha omnibus statistics including the Chao1 (Richness), Faith (Phylogenetic Diversity) and Pielou (Evenness). Chao1 includes the measure of the number of bacterial taxa detected in each sample, whereas Pielou includes the measure of the proportion of bacterial taxa detected in each sample. Conversely, Faith is an exact method to measure diversity by taking the sum of the number of branch lengths of a phylogenetic tree connecting all species.

<sup>10</sup>Shukla, Shakti D., et al. "Microbiome effects on immunity, health and disease in the lung." Clinical & translational immunology 6.3 (2017): e133.

<sup>11</sup>Lynch Paper <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2860857/>

The subject level information includes the subjects asthma status in year 7, aeroallergen status at age 3 years and recurrent wheeze at year 3 and site. The de-identified site, a combined multinomial response (wheeze, atopy, both, neither) and allergic exposure values were also provided.

## 2.3 Data Cleaning

At the start of the analysis the data sets were combined into a master data set with two rows for each subject, one for dust and nasal, respectively. Columns with all zero values for the OTUs for the dust and nasal samples combined were removed before analysis as these OTUs had no variation that could be modeled. This resulted in the removal of 702 OTUs.

## 2.4 Defining the Response

A new variable, hereto referred to as "The Event", was defined in reference to the project aims. The literature states that the presence of wheezing and atopic symptoms are risk factors for the childhood development of asthma. With consideraion to the distribution of the multinomial response [make table of number of each outcome] "The Event" was defined as 1: child displays symptoms of wheezing, atopy or both and 0: child displays neither symptom.

## 2.5 Visualizing the Data

The large, high dimensional nature of the final data set airs special caution for both data visualization and modeling. The  $p > n$  condition necessitates the use of variable selection methods for identifying important OTUs before many regression techniques can be used. This also demands the use of selective and multiple visuals to understand the nature of the data.

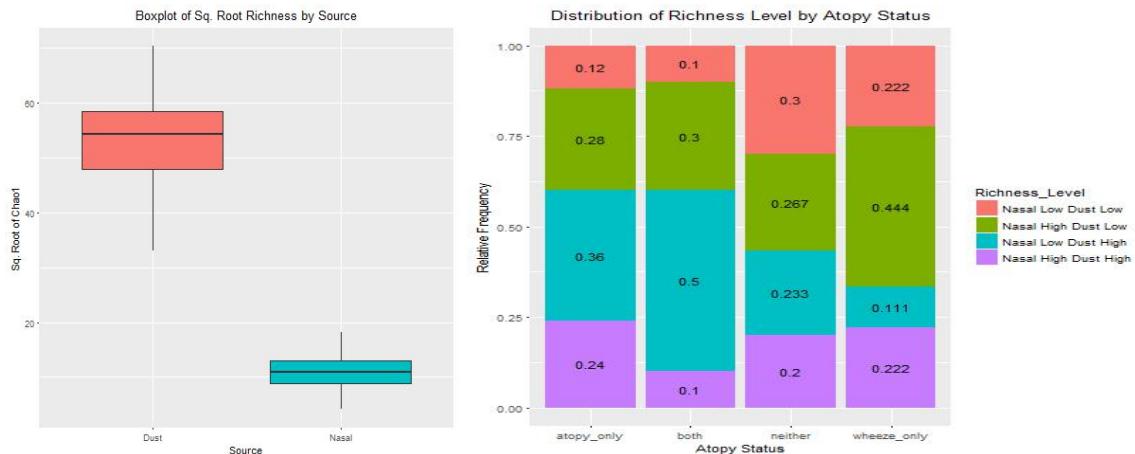


Figure 1: Box Plot of  $\sqrt{Richness}$  by Source(left); Distribution of Richness Level(right)

see (1)

To begin to visualize aspects of the data, bar and box plots were created. The first important visual is the difference between the rich ness (chao1) diversity measure between the nasal and dust samples. The biodiversity of the dust samples is much higher than the nasal wash samples, even on a square root transformed scale. The median richness for nasal is 120.35 while dust is 2939.64 on a non-transformed scale. This biodiversity difference is important when comparing and identifying the prevalence and impact of specific OTUs between samples, as the diversity of each type of sample differs drastically.

The main response of the analysis from the project goals is the wheezing or atopic status of the child at age three. The bar plot below displays the distribution of sample diversity among each outcome. As seen in the bar plot, the neither group displays a near even distribution of sample richness. The both group displays the greatest percentage (50%) of low diversity nasal samples and high diversity dust samples. The atopy only and both groups both show 60% high dust diversity and 40% low dust diversity, with the neither and Wheeze only showing the opposite distribution.

The technique utilized for visualizing the whole data set together was the use of circular plots. Circular plots are useful in representing detailed and complicated information. They are used frequently in two specific cases: when you have a long axis and numerous categories and when you want to show relationships between specific elements.

To begin to visualize our data, we used circular plots to identify the relationship between the microbiome type and phylum in terms of both the OTUs and the OTUs with significant p-values. On the left side of Figure 1, we see that the  $p_{22}$  phylum had the largest number of OTUs from both the dust microbiome and nasal microbiome since the chords that connect to this phylum from both microbiomes were the largest. Lastly, on the right side of Figure 1, we see that the  $p_{29}$  phylum had the largest number of OTUs with significant p-values since, again, the chords that connect this phylum from both microbiome types were the largest.

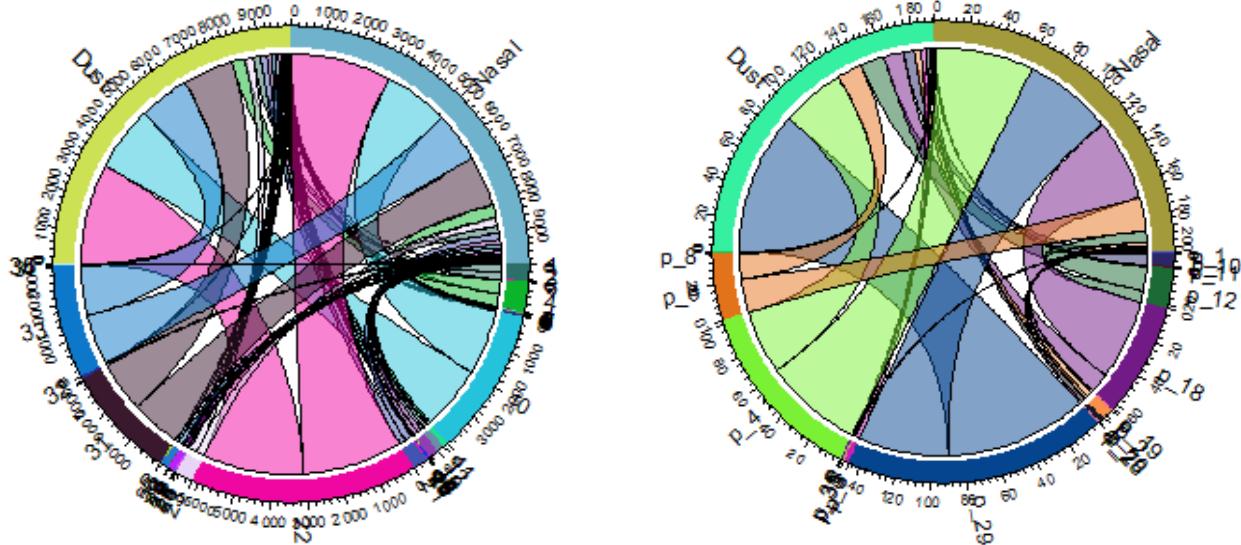


Figure 2:

Left: A circular plot that shows the relationship between the microbiome type and the phylum in terms of the OTU. The range of 1:36 represents the phylum. The "Dust" variable represents the dust microbiome and the "Nasal" variable represents the nasal microbiome. The chords represent the OTU values.

Right: A circular plot that shows the relationship between the microbiome type and the phylum in terms of the OTU with significant p-values. The  $\{p_1, p_4, \dots\}$  set represents the phylum. The "Dust" variable represents the dust microbiome and the "Nasal" variable represents the nasal microbiome. The chords represent the OTU values with significant p-values.

### 3 Understanding the Different Microbiomes

#### 3.1 Nasal versus Dust: Correlation

To begin our analysis of the association between the nasal and dust microbiomes, correlation analysis was utilized on the omnibus statistics. The correlation was calculated using the Spearman rank-order correlation to evaluate the monotonic relationship between the omnibus statistics. [INSERT CORRELATION PLOT] The correlation between all three of the omnibus statistics within their own microbiome (dust to dust and nasal to nasal) represented by the red sections of the figure above were all significant at the fifth percentile level and shows a positive correlation. The correlation was the strongest between the faith and chao1 measures and stronger overall in the dust microbiome.

The dust and nasal microbiomes' diversity measures appear to be weakly negatively correlated. The correlation values between -0.18 and -0.13 were not significant at the fifth percentile level. This negative correlation indicates that the omnibus statistics are not increasing together, but rather that as one sample type is more diverse, the other is slightly less diverse. This leads to the practical conclusion that both samples are not displaying the same biodiversity information and may not influence outcomes in the same way.

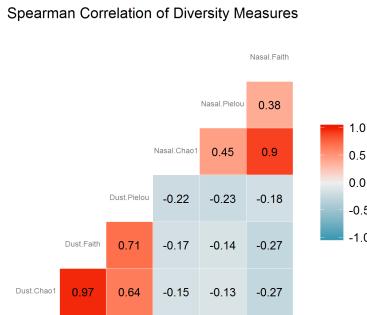


Figure 3: Correlation of Biodiversity between Dust and Nasal

#### 3.2 Nasal versus Dust: Volcano Plot

The OTU count data was analyzed for differential abundance. The differential abundance analysis used the generalized linear model:

$$\begin{aligned} K_{ij} &\sim NB(\mu_{ij}, \alpha_i) \\ \mu_{ij} &= s_j q_{ij} \\ \log_2(q_{ij}) &= x_j \beta_i \end{aligned}$$

where relative abundance counts  $K_{ij}$  for OTU  $i$  and sample  $j$  were modeled using a Negative Binomial distribution with mean  $\mu_{ij}$  and dispersion parameter  $\alpha_i$ . The mean was a product of the size factor  $s_j$  for sample  $j$  and the parameter  $q_{ij}$  was proportional to the relative abundance of OTUs for sample  $j$ . The coefficients  $\beta_i$  was the log 2 fold changes for OTU  $i$  for each column of the model matrix  $X$ .

The size factors were estimated using the median ratio method and the geometric means of the counts were provided. Dispersion estimates for Negative Binomial distributed data were fit using a parametric model. A Wald significance tests were used to determine the significance of coefficients in a Negative Binomial GLM, thereby difference in deviance between treatments.

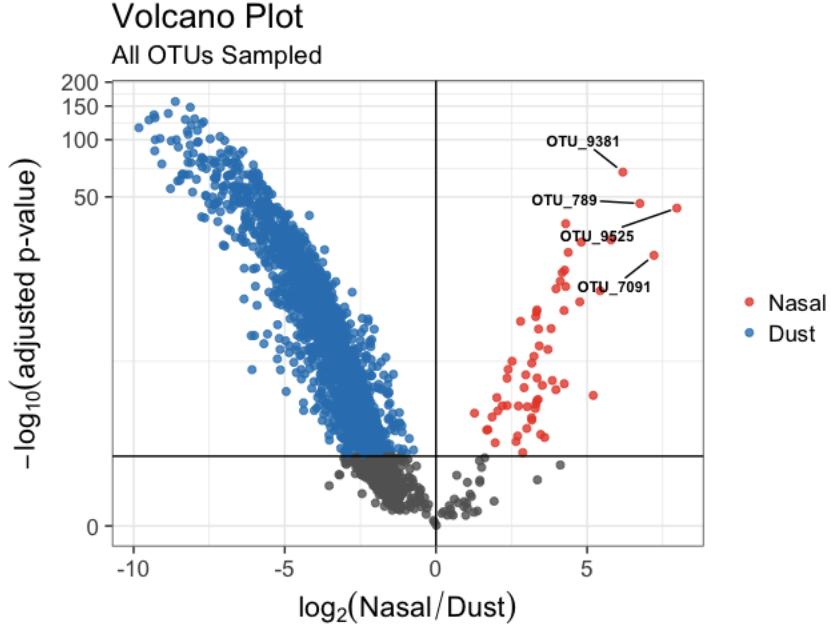


Figure 4: Volcano Plot identifying essential OTUs in Nasal (60 red points) and Dust(2107 blue points)

A volcano plot was used to organize OTUs by biological and statistical significance. The horizontal axis is the log 2 (fold change) between nasal versus dust samples and the vertical axis represents the -log 10 (p-value) for a Wald significance test of differences between samples. The horizontal axis indicates biological significance by the fold change whereas the vertical axis indicates the statistical significance of the fold change. From the 9416 OTUs with non-zero counts, 2167 OTUs were significant for differential abundance. Of the 2167 OTUs, 60 OTUs were prevalent in nasal samples versus dust samples with *OTU9381*, *OTU789*, *OTU9525*, and *OTU7091* were the top four differentially expressed OTUs prevalent in nasal samples. Note that adjusted p-value was used for the vertical axis to take into account the large number of multiple comparisons.

## 4 Biodiversity Information

### 4.1 Analysis of Microbiome Diversity on Development of Disease

To begin to explore the relationship between the omnibus statistics and the event, we first explored the effect of biodiversity indices for dust and nasal samples, as well as site differences on the outcome of events.

$$\begin{aligned} \log\left(\frac{P(Y=1)}{P(Y=0)}\right) = & 0.405341 - 0.048926 * X_{dust.Faith} + 0.249642 * X_{dust.Pielou} + 0.001822 * X_{dust.Chao1} \\ & + 0.044215 * X_{nasal.Faith} - 0.252760 * X_{nasal.Pielou} - 0.003202 * X_{nasal.Chao1} \end{aligned}$$

We use logistic regression to see the directions of the effects as well as whether they are statistically significant.

Across all logistic regressions, we did not find statistically significant effects of biodiversity on the event. Also, the site effects did not have a significant relationship with the outcome, so the site effects were not considered for the rest of the analysis.

To dig deeper into the data and build a comprehensive framework, we utilized a Directed Acyclic Graph (DAG) and Mediation Analysis (MA) to model the correlations between microbiome biodiversity in the dust and nasal environments and development of allergy and/or wheeze, and infer their causal relationships to

see if microbiome biodiversity affects the development of disease. If there is an association, we then try to compare pathways through which microbiome biodiversity affects the development of disease. Namely we try to compare if the dust microbiome directly affects the odds of disease or if the effect through the nasal microbiome acts as a mediator. In the following part, we briefly introduce the DAG and MA models and the application to the URECA data sample.

## 4.2 Model the Correlations-Directed Acyclic Graph

DAG is a special kind of directed graph that describes the relationships among components. We code DAG as seen in the above diagram.

The first vertex denotes the biodiveristy measures that describe the microbiome enviroment in the dust samples collected 3 months after births; the second vertex denotes the biodiveristy measures that describe the microbial environment in the nasal samples collected 1 year after births; the third vertex denotes children's development of allergy and/or wheeze 3 years after birth. The DAG's topological ordering must be consistent with the timeline to be interpretable. Therefore, the edges show all possible connections. Next, we fit the DAG with the data so that correlation coefficients for each path can be derived. The essential equations to model these correlations are:

$$\begin{aligned} Y_{dust} &\sim \text{Gaussian} \\ Y_{nasal}|Y_{dust} &\sim \text{Gaussian} \\ \text{logit}(event) = Y_{allergy}|Y_{dust}, Y_{nasal} &\sim \text{Gaussian} \end{aligned}$$

where  $Y_{dust}$  and  $Y_{nasal}$  are the continuous measures of the microbiome environment in the dust and nasal samples. The available measures include microbiome richness (Chao1), microbiome eveness (Pielou) and the summary of phylogenetic tree relationship (Faith). Also included are the important OTUs identified from the variable selection process or eigenvectors detected by PCA. The binary outcomes of allergy and/or wheeze are transformed to logit scale so that the residual is normally distributed.

To follow these ideas, we fit the following regression models/equations:

$$\begin{cases} Y_{dust} = \mu_1 + \eta_1, & \eta_1 \sim N(0, e_1) \\ Y_{nasal} = \mu_2 + a_{21}(Y_{dust} - \mu_1) + \eta_2, & \eta_2 \sim N(0, e_2) \\ Y_{event} = \mu_3 + a_{31}(Y_{dust} - \mu_1) + a_{32}(Y_{nasal} - \mu_2) + \eta_3, & \eta_3 \sim N(0, e_3) \end{cases}$$

the second model is an ordinary least squares (OLS) regression and the third model is a logistic regression. Upon fitting the three equations above, we get estimates for  $a_{21}, a_{31}, a_{32}$  from regression coefficients and estimates of  $e_1$  from empirical variance of  $Y_{dust}$ ,  $e_2$  from residual sum of squares from OLS regression and  $e_3$  from residual sum of squares from logistic regression. We write the three equations in matrix form

$$Y - \mu = A(Y - \mu) + D$$

where  $Y - \mu = \begin{pmatrix} Y_{dust} - \mu_1 \\ Y_{nasal} - \mu_2 \\ Y_{allergy} - \mu_3 \end{pmatrix}$ , the coefficient matrix  $A = \begin{pmatrix} 0 & 0 & 0 \\ a_{21} & 0 & 0 \\ a_{31} & a_{32} & 0 \end{pmatrix}$ , and residual variance matrix  $D = \begin{pmatrix} \eta_1 & 0 & 0 \\ 0 & \eta_2 & 0 \\ 0 & 0 & \eta_3 \end{pmatrix}$ . This means  $(I - A)$  is a lower triangular matrix and thus invertable, so we have

$$Y - \mu = (I - A)^{-1}D$$

and

$$Var(Y) = Var(Y - \mu) = (I - A)^{-1} \begin{pmatrix} e_1 & 0 & 0 \\ 0 & e_2 & 0 \\ 0 & 0 & e_3 \end{pmatrix} (I - A)^{-T}$$

The program for generating the correlation coefficients matrices using three biodiversity measures is as follows:

We test the Gaussian assumptions by showing the histogram of  $Y_{dust}$  and QQ plot of OLS regression as follows:

the two plots above indicate that the normality assumptions are reasonable in our case.

From the three result tables built from three biodiversity indices we consistently observe contradictory signs on the two pathways with negative correlation coefficients on mediation pathway and a positive correlation coefficient on direct pathway. In general, these correlation coefficients are small in general. We did not find statistical support for either pathway and we further conducted causal inference analysis in the next section.

#### 4.3 Causal Inference-Mediation Analysis:

Mediation analysis (MA) is commonly used to assess causal mechanisms. We use this analysis to examine the causal mediation effects of the nasal microbiome's contribution to the development of disease. We consider the dust microbiome as "treatment" and nasal microbiome as a "mediation effect" on the pathway. We code MA as follows:

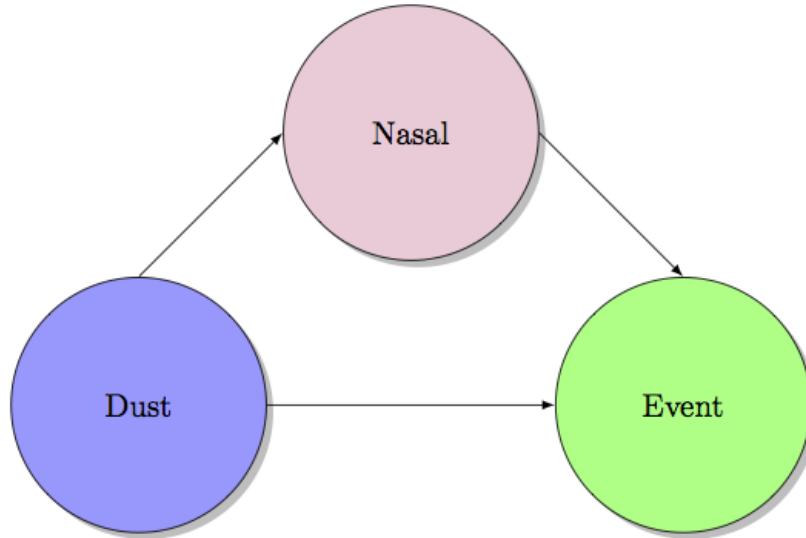


Figure 5: Mediation Pathways between Dust,Nasal and Event

We still try to compare the two pathways: the "direct pathway" goes from dust to event directly whereas the "causal mediation pathway" goes from dust to event using the nasal environment as a mediator.

We use the mean as the cutoff value for dust microbiome bio-diversity and define the value beyond the cutoff as "treatment" and below as "control"; Let  $M_i(t)$  denote the potential value of a mediator for subject  $i$  under the dust microbiome status ( $t=1$  for treatment and  $t=0$  for control) and let  $Y_i(t, m)$  denote the potential outcome that would result under treatment and mediating status. The total unit treatment effect can be written as,

$$\tau_i = Y_i(1, M_i(1)) - Y_i(0, M_i(1))$$

The total unit treatment effect can be decomposed into causal mediation effects and direct effects. The causal mediation effects are defined as,

$$\tau_{i,mediation}(t) = Y_i(t, M_i(1)) - Y_i(t, M_i(0))$$

and the direct effects are defined as

$$\tau_{i,direct}(t) = Y_i(1, M_i(t)) - Y_i(0, M_i(t))$$

we then have the relationship as

$$\tau_i = \tau_{i,mediation}(t) + \tau_{i,direct}(1 - t)$$

Then the average causal mediation effects (ACME)  $\overline{\tau_{mediation}}$  and the average direct effects (ADE)  $\overline{\tau_{direct}}$  representing the population averages of these causal mediation and direct effects become our estimates of interest. Assumptions of the analysis include randomization of treatment and sequential ignorability of treatment (Imai, Kosuke, Luke Keele, and Teppei Yamamoto. "Identification, inference and sensitivity analysis for causal mediation effects." Statistical science 25.1 (2010): 51-71.).

We did not observe significant ACME or ADE for the two paths. There seems to be a positive total effect between microbiome biodiversity and odds of disease, but the correlation is not statistically significant.

From DAG and MA modeling analysis, we observe weak correlations which are not statistically significant. We thus say that there is a lack of statistical evidence supporting the existance of connections among dust microbiome biodiversity, nasal microbiome biodiversity and development of disease from our dataset.

Instead of looking at the single values of the biodiversity in the microbiome environment, we instead try to examine individual OTUs and see if they affect the development of disease in the next part.

## 5 Identifying Vital Operational Taxonomic Units

### 5.1 Unsupervised Learning Methods

#### 5.1.1 Analysis

The OTUs were tested for association using Spearman's  $\rho$ . For Spearman's test, p-values are computed via the asymptotic t approximation. Individual OTUs were determined to be significant with a p-value less than 0.01. From the 1052 OTUs with non-zero counts in both dust and nasal samples, 32 OTUs were determined to be significant with four OTUs significantly prevalent in dust samples and three OTUs significantly prevalent in nasal samples previously determined by differential abundance analysis.

#### 5.1.2 Sparse Principal Component Analysis

Principal component analysis(PCA) is a widely used unsupervised learning technique for data processing and dimension reduction. It is a method used to reduce a data set into a smaller number of "principal components". These principal components explain a portion of the total variation of the data set and are a linear combination of all of the original variables. While PCA handles dimension reduction, it also has some drawbacks in real applications. In the PCA, one drawback is while each principal component is a linear combination of all the original variables, it is difficult to interpret the results. In our data, the OTUs are categorical variables, which add to the difficulty in explanation. To handle the sparse nature of the data, we used a new method called sparse principal component analysis using a lasso (elastic net) to produce the modified principal component with spares loadings. Also, in our data the  $p > n$ , the SPCA handles this condition much better than the PCA.

Typically, the PCs are the eigenvectors of the covariance matrix attained by performing single value decomposition (SVD) on the data matrix. However, with the SPCA, the eigenvectors of the covariance matrix are attained by penalized matrix decomposition(PMD). The advantage of applying PMD to a data matrix L1-constraints on the column, but not the rows yields an efficient algorithm for the SCOTLASS (Simplified Component Technique-Lasso) method for finding sparse principal components. As a principled procedure, SPCA enjoys advantages in several aspects, including computational efficiency, high explained variance and an ability in identifying important variables.

### 5.2 Supervised Learning Methods

#### 5.2.1 Volcano Plot

To begin to identify important OTUs from each microbiome in relation to the event of 1=both, atopic, wheezing and 0=neither, additional volcano plots were employed. The horizontal now represents the log 2 (fold change) between event and neither. The vertical axis is computed using a Wald significance test of the difference between samples.

The volcano plot for the nasal OTUs and the event, only 13 OTUs were identified as having a strong biological and statistically significant difference between outcomes. The volcano plot for the dust OTUs identified 320 OTUs that were differentially expressed between outcomes, with the top ten OTUs labeled on the plot.

#### 5.2.2 Random Forest

Random Forest is a widely used machine learning algorithm, and one of the features random forest gives out is the relative importance ranking of variables. We use the usual "event" (allergy or wheeze or both) variable as outcome and use OTU counts as input to mine the information from OTUs. The random forest can be

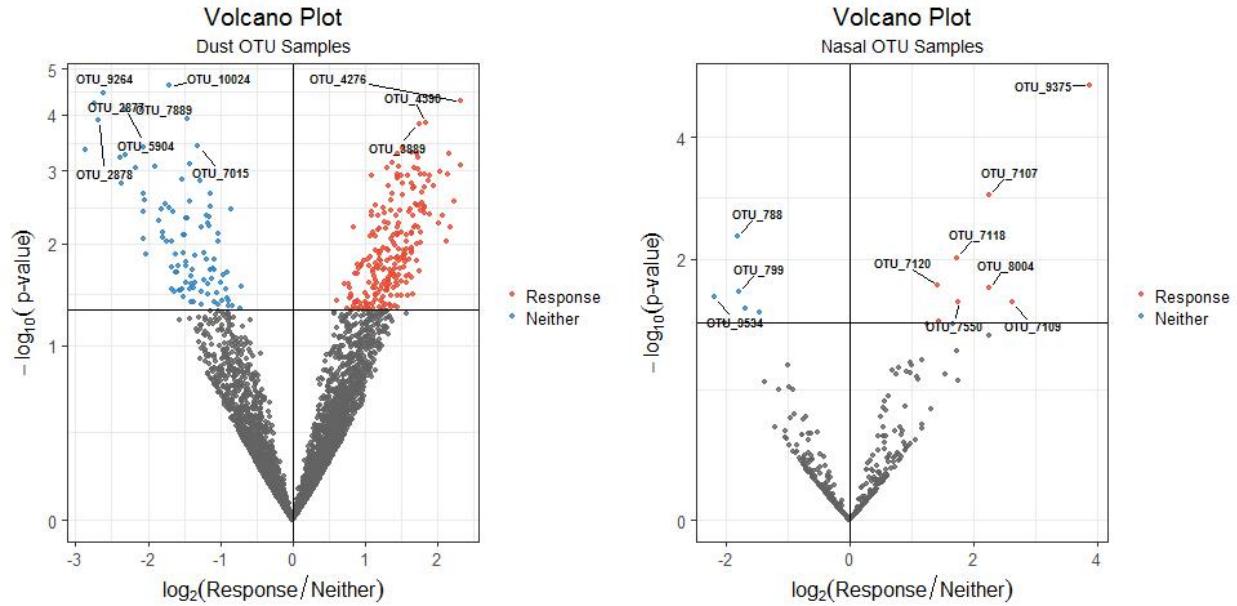


Figure 6:

Left: Super volcano plot for the Dust microbiome.  
Right: Super volcano plot for the Nasal microbiome

built differently using different depth, number of trees, proportions of sub-sample and so on and we adopt the commonly used set up: 500 number of parallel trees, 20 maximum depth the tree can grow and 0.6 as the subsample proportion. Using xgboost package in R, we are able to process the high dimensional sparse matrix (74 by 20272) within seconds.

The importance metric is based on measure of "Gain". Gain is the improvement in accuracy brought by a feature to the branches it is on. Based on the importance ranking, we draw the importance plot as above and we identify key OTUs in dust and nasal samples that contributed most to the prediction power. The top OTUs from dust samples are *OTU430*, *OTU664*, *OTU695*, *OTU769*; the top OTUs from nasal samples are *OTU789*, *OTU788*, *OTU799*, *OTU792*.

### 5.2.3 Sparse Support Vector Machine

Variable selection refers to the problem of selecting a subset of independent variables that are most predictive of a given outcome. Appropriate variable selection can improve the accuracy and interpreting of an inference model. The sparse support vector machine (ssvm) is an supervised regression analysis with the inputs in the form of attribute vectors. SSVM constructs a hyperplane that separates two classes to achieve maximum separation between the classes. By separating the classes with a large margin, generalization error is minimized. The objective of achieving the minimum generalization error is to predict the correct class of data without any error, when it arrives for classification (Soman, 2009). This method uses a subset of nonzero weighted variables found by the linear models to produce a final nonlinear model. The response for our SSVM was the outcome of interest and the predictors were the dust and nasal OTU, separately.

### 5.2.4 Elastic Net

One of the most important problems when researching the association between microbes in the nose and those in the environment and their relationship to the development of allergy is to identify some essential

variables (OTUs) among tens of thousands OTUs. After the data cleaning step, some of non-informative OTUs have already been dropped off. Variable selection is the following step that allows us concentrate more on those information-rich OTUs and identify a smaller subset.

For larger  $p$  ( 9431 ) small  $n$  (74) problems, elastic net is a regularized regression method commonly employed. The Elastic Net method linearly combines the  $\mathcal{L}_1$  and  $\mathcal{L}_2$  penalties of the lasso and ridge methods. It overcomes the limitations of the LASSO method whose basic form is:

$$\min_{\beta_0, \beta} \|y - \beta_0 - X\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_1 = \sum_{i=1}^p |\beta_i| \leq t$$

LASSO has two major limitations:

- If  $p > n$ , the lasso method selects at most  $n$  variables. The number of selected genes is bounded by the number of samples.
- If there is a group of highly correlated variables, then the LASSO tends to select one variable from a group and ignore the others.

However, both of limitations would conflict with nature of our data, which motivates us to choose the elastic net method. Specifically, it is common to see correlation between covariates in the biological study and the number of observations is much smaller than that of the covariates. To overcome these limitations, the elastic net adds a quadratic part to the penalty (which is related to ridge regression), and estimates of the elastic net method are in the form of :

$$\hat{\beta} = \operatorname{argmin}_{\beta} (\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1)$$

Here, logistic regression is considered since the response is binomial (0=neither, 1=response). For the binomial model:

$$\log \frac{Pr(y_i = 1)}{Pr(y_i = 0)} = \beta_0 + \beta^T x$$

The objective function for the penalized logistic regression uses the negative binomial log-likelihood, and is:

$$\min_{(\beta_0, \beta) \in \mathbb{R}^{p+1}} - \left[ \frac{1}{N} \sum_{i=1}^N y_i \cdot (\beta_0 + x_i^T \beta) - \log(1 + e^{(\beta_0 + x_i^T \beta)}) \right] + \lambda [(1 - \alpha) \|\beta\|_2^2 / 2 + \alpha \|\beta\|_1].$$

Logistic regression is often plagued with degeneracies when  $p > N$  and exhibits wild behavior even when  $N$  is close to  $p$ ; the elastic-net penalty alleviates these issues while also selecting variables.

We first address the variable selection for the dust data. To show the benefit of the elastic net method over LASSO in the current case, we compare the n-fold( $=5$ ) cross validation of two methods as graphs below. Note that the results of cv.glmnet are random, since the observations are assigned at random. To reduce variability,  $n = 5$  folds was chosen due to the "smaller" number of observations. Top left is corresponding to the elastic net, which results in around 70 variables; while the top right corresponding to the LASSO shows a wild behavior with variables selected, as the error decreases with less variable. This reveals a general superiority of elastic net method over LASSO.

The graphs on the bottom line plot coefficients against the deviance explained. Since there is a large amount of variables, one cannot identify the variables efficiently. The differences between models will still appear. It can be observed that more coefficients will shrinkage to 0 via elastic net while for LASSO, there is no such sparsity.

standard deviation of misclassification error	0.04(EN)	0.06(LASSO)
minimum mean of misclassification error	0.40(EN)	0.45(LASSO)

Table 1: Standard Deviation and Minimum of Misclassification Error for Dust data

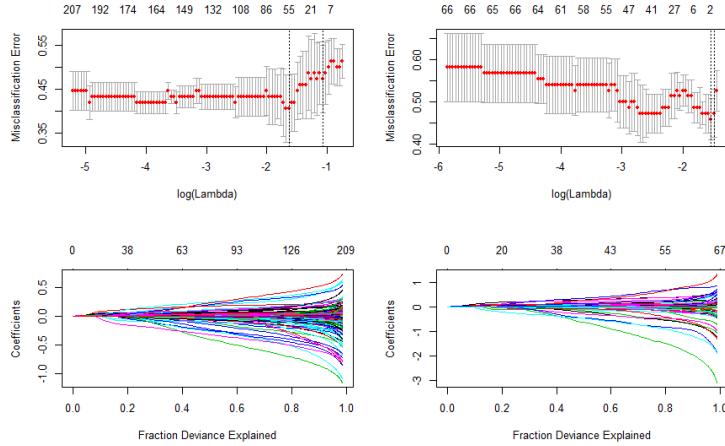


Figure 7: Cross Validation Plot of Dust Data

From the table above, the minimum misclassification error for LASSO is a bit higher than that for elastic net. This is reasonable since there are fewer variables in LASSO, following the known fact that the more variables, the lower the misclassification error.

The following plot shows the number of nonzero variables against value of the lambda penalty. From the right figure, you can see that the range of variables selected by LASSO is strictly limited by the number of observations. Given the same level of lambda, for instance lambda=0.1, we can see number of variables selected by elastic net is over 100 while number of variables selected by LASSO is only around 40. One explanation of this could be the penalty in the LASSO would be more severe compared to that in the elastic net method. The plot on the top of next page shows misclassification error against lambda in log scale for

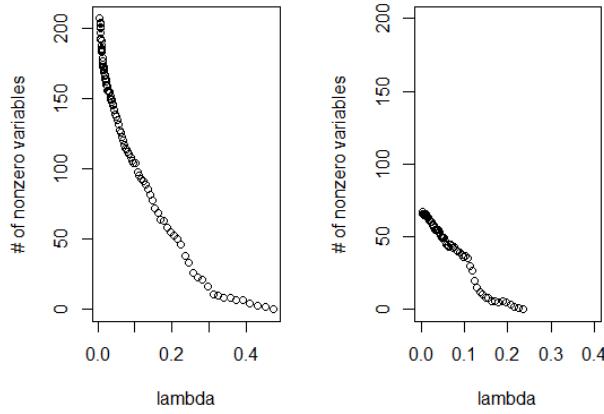


Figure 8: Number of Nonzero Coefficient against Lambda for Dust data

two methods. Given the same penalty level, the misclassification error for elastic net is always smaller than that for LASSO.

The next step is the analysis of the nasal data under the same conditions to see whether a similar result is found. Top left is corresponding to the elastic net, which results in around 80 variables selected; while the top right corresponding to the LASSO results in around 50 variables selected. The standard deviation of

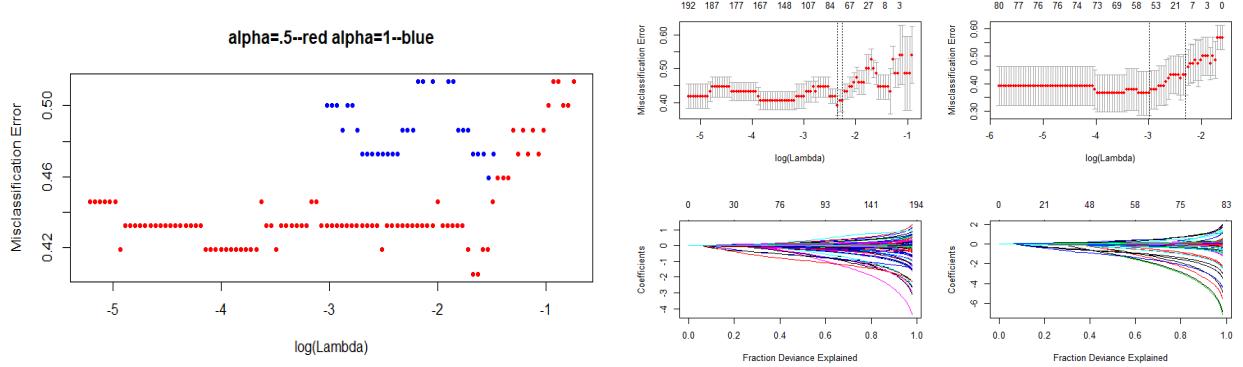


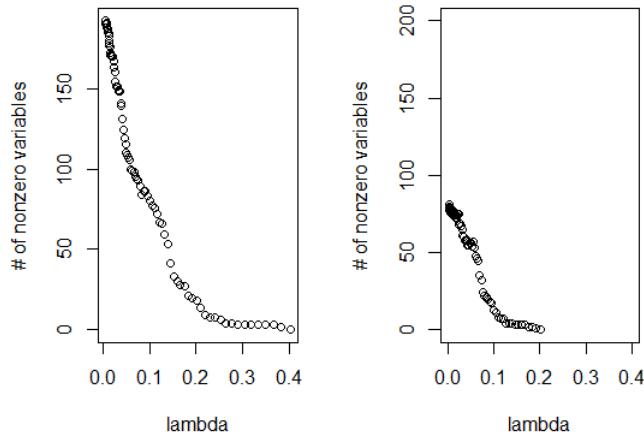
Figure 9: Misclassification against  $\log(\lambda)$ (left); Cross Validation Plot of Nasal Data

misclassification (cross validation) error and the minimum of misclassification(cross validation) error roughly is following:

standard deviation of misclassification error	0.035(EN)	0.070(LASSO)
mininum mean of misclassification error	0.39(EN)	0.36(LASSO)

Table 2: Standard Deviation and Minimum of Misclassification Error for Nasal data

Both minimum misclassification errors are close to each other while the standard deviation of the elastic net is only half of that from LASSO. When considering the number of nonzero variables against the value of penalty lambda, it shows the same pattern appears in the dust.



To explore the difference between these two methods,it is important to look at the specific variables selected.

From the table in the Appendix,we see that each variable selected via elastic net is also selected by LASSO. There are 56 overlapping variables (set A) and 27 variables(set B) only appearing in the elastic net. Correlation analysis shows that 22 variables out of set B have essential correlation (either greater than 0.5 or smaller than -0.5) with 20 variables out of set A. The grouped effect is revealed in this way. The two methods will not result in the same order importance of OTUs.

Generally, the results from both the dust and nose samples support the fact that elastic net outperforms LASSO in three aspects: 1) misclassification error; 2) number of variables should not be limited by the number of observations, which is relatively small and 3) grouped effect: it should be able to include whole groups into the model automatically once one among them is selected.

## 6 Modeling the Important Operational Taxonomic Units

### 6.1 Cross Referencing

The OTUs were cross-referenced between learning techniques not using and using the event response variable. Learning techniques not using the event response variable included differential abundance analysis of OTUs in dust and nasal samples, correlation of OTUs in dust and nasal samples, and sparse PCA of OTUs in both samples. Learning techniques using the event response variable included differential abundance analysis of OTUs, random forest, sparse SVM, and variable selection using elastic net: where all methods for subjects with either aeroallergen sensitization or recurrent wheezing versus neither. From the 10118 OTUs sampled, 397 OTUs were present in at least one learning technique not using and one learning technique using the event response variable, 20 OTUs were present in at least three pairwise cross-references, whereas only six OTUs were contained at least four pairwise cross-references: *OTU689*, *OTU695*, *OTU788*, *OTU8731*, *OTU799*, and *OTU9364*.

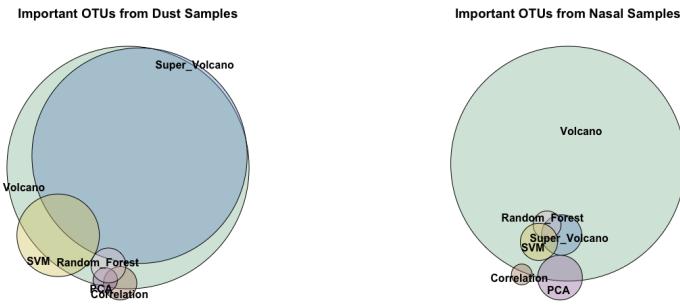


Figure 10: Venn Diagram for Dust Sample(left)    Venn Diagram for Nasal Sample(right)

### 6.2 Phylogenetic Distribution

There were 494 OTUs identified using the supervised learning techniques (super volcano, random forest, support vector machines and elastic net) from the dust samples and 92 from the nasal samples. To further explore the biological composition of these selected OTUs, the distribution of the phylum and class structure of the OTUs that appeared in the nasal and dust samples originally and the subset were compared. The relative frequency of each phylum or class was computed among the original OTUs that appeared in each type of sample (nasal and dust) and the subset identified by supervised learning techniques. The dust phylum distribution identifies p\_29 and p\_4 as having a greater relative frequency among the supervised subset than the original dust OTU representation, as well as having the largest relative frequency within the subset of OTUs. The dust class distribution identifies c\_17, c\_16 and c\_46 as having the largest relative frequency within the supervised subset, and a larger presence in the supervised subset versus the overall dust OTU distribution.

The nasal phylum distribution also identifies p\_29 as having a larger relative frequency in the subset compared to the original nasal OTU distribution. The other phyla appear to be fairly evenly distributed between the data sets, with some with a very low relative frequency in the original data showing a slight increase in the

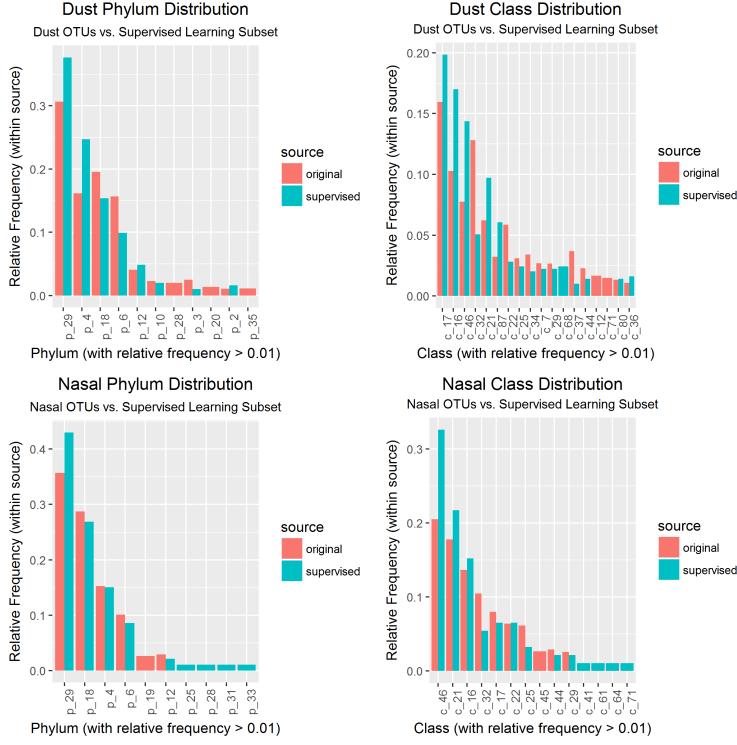


Figure 11: Distribution of Dust Phylum (top left); Distribution of Dust Class (top right); Distribution of Nasal Phylum (bottom left); Distribution of Nasal Class (bottom right)

supervised subset. The class distribution identifies  $c_{46}$ ,  $c_{21}$  and  $c_{16}$  as representing the majority of the subset of OTUs and having a larger presence in the subset than the original nasal OTUs.

### 6.3 Logistic Regression

We do the logistic regression in two ways in terms of different type of predictor: we consider the absence and presence of OTUs as the first kind of predictor while the raw count of OTUs is the second one.

$$\log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 \Phi(OTU\_dust) + \beta_2 \Phi(OTU\_nas) + \epsilon_\beta$$

$$\log\left(\frac{P(y=1)}{P(y=0)}\right) = \alpha_0 + \alpha_1 \Phi(OTU\_dust) + \alpha_2 \Phi(OTU\_nas) + \epsilon_\alpha$$

in which  $\Phi(\cdot)$  is referred to the indicator function.

OTU	Nasal(absence-presence)	Dust(absence-presence)	Nasal( raw count)	Dust( raw count)
OTU_1304	NA	0.147	NA	0.021*
OTU_3176	0.774	0.395	0.990	0.074
OTU_3694	0.991	0.027*	0.991	0.012*
OTU_430	NA	0.002*	NA	0.006*
OTU_4372	NA	0.012*	NA	0.018*
OTU_5194	NA	0.010*	NA	0.059
OTU_5424	NA	0.010*	NA	0.014*
OTU_7107	0.256	NA	0.151	NA
OTU_7550	0.753	0.824	0.150	0.724
OTU_757	NA	0.031*	NA	0.005*
OTU_7889	0.992	0.995	0.992	0.061
OTU_9264	NA	0.078	NA	0.028*
OTU_9526	0.767	NA	0.094	NA
OTU_9608	NA	0.025*	NA	0.068
OTU_689	NA	0.954	NA	0.011*
OTU_695	NA	0.158	NA	0.005*
OTU_788	0.089	0.818	0.060	0.849
OTU_8731	NA	0.006*	NA	0.016*
OTU_799	0.154	NA	0.030*	NA
OTU_9364	0.128	0.741	0.141	0.789

Table 3: P-Value of Different Models

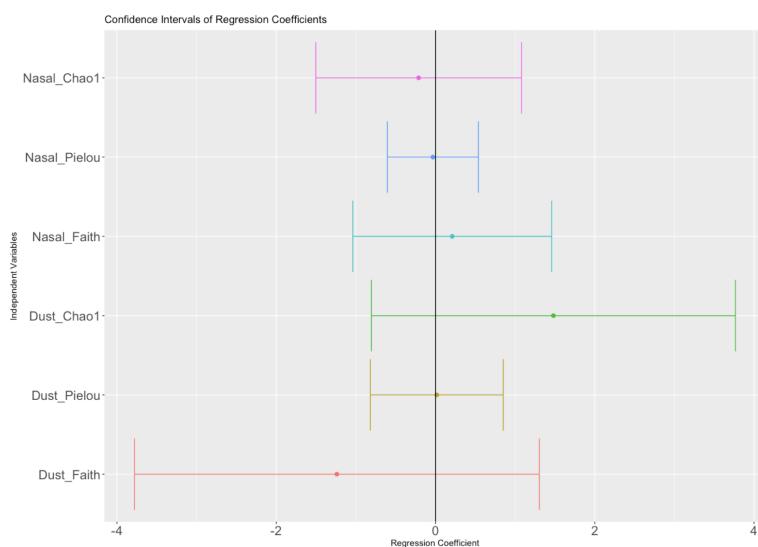


Figure 12: Confidence Interval for Regression Coefficient

OTU1374	$\overset{1.00}{\sim}$	OTU1386	OTU5644	$\overset{0.80}{\sim}$	OTU8409	OTU2699	$\overset{0.60}{\sim}$	OTU1374
OTU7014	$\overset{1.00}{\sim}$	OTU4362	OTU3734	$\overset{0.97}{\sim}$	OTU5743	OTU7017	$\overset{0.51}{\sim}$	OTU6449
OTU7084	$\overset{0.60}{\sim}$	OTU7080	OTU7099	$\overset{0.56}{\sim}$	OTU7080	OTU7086	$\overset{0.58}{\sim}$	OTU7090
OTU7107	$\overset{0.99}{\sim}$	OTU7109	OTU5812	$\overset{0.53}{\sim}$	OTU7126	OTU7108	$\overset{1.00}{\sim}$	OTU7126
OTU7125	$\overset{0.97}{\sim}$	OTU7126	OTU659	$\overset{0.53}{\sim}$	OTU738	OTU680	$\overset{0.53}{\sim}$	OTU738
OTU9710	$\overset{0.53}{\sim}$	OTU738	OTU7569	$\overset{0.84}{\sim}$	OTU7889	OTU5829	$\overset{0.51}{\sim}$	OTU7907
OTU659	$\overset{0.56}{\sim}$	OTU8030	OTU680	$\overset{0.56}{\sim}$	OTU8030	OTU8032	$\overset{0.91}{\sim}$	OTU8040
OTU8035	$\overset{0.64}{\sim}$	OTU8040	OTU8051	$\overset{0.82}{\sim}$	OTU8040	OTU5829	$\overset{0.75}{\sim}$	OTU8699
OTU659	$\overset{0.62}{\sim}$	OTU9106	OTU680	$\overset{0.62}{\sim}$	OTU9106	OTU9710	$\overset{0.62}{\sim}$	OTU9106
OTU4283	$\overset{0.68}{\sim}$	OTU9432	OTU4283	$\overset{0.67}{\sim}$	OTU9480	OTU10044	$\overset{0.59}{\sim}$	OTU9495
			OTU7107	$\overset{0.66}{\sim}$	OTU9526			

Table 4: Correlation between variables selected by LASSO and variables not selected by LASSO

## Appendix

Predictor	Estimate	P-Value
Intercept	0.405	0.901
Y_dust.Faith	-0.049	0.340
Y_dust.Pielou	0.250	0.969
Y_dust.Chao1	0.002	0.204
Y_nasal.Faith	0.044	0.741
Y_nasal.Pielou	-0.253	0.911
Y_nasal.Chao1	-0.003	0.748

## Conclusion

## Main Findings

Our analysis had two main goals: (1) examine the association between the nasal and dust microbiomes and (2) obtain an understanding of how microbes in the environment and nose relate to the development of allergic diseases and asthma in early life. To begin with goal one, we started with exploring the correlation between the nasal and dust microbiomes. This analysis showed a negative correlation between the diversity indices of opposing samples within a range from -0.27 to -0.13. To continue with the goal one analysis, we explored the differential abundance of OTUs in the dust and nasal samples. This analysis identified 60 OTUs for the nasal microbiome and 2107 OTUs from the dust microbiome that were differentially expressed. Overall, the analysis of the association between the dust and nasal microbiomes found a difference in the diversity and abundance of OTUs between the environments. To begin to address goal two, we started our analysis with the diversity indices provided. These indices summarize the 10,118 OTU readings into a single value from each of the 74 dust and nasal samples. We started with the simplest method, logistic regression, to analyze the relationship between these indices and the outcome of interest and found no statistically significant predictors. Next, a DAG model and mediation analysis were fit using the dust and nasal indices in relation to the outcome of interest. Here the correlations had opposing signs and the p-values of the mediation analysis

Variables	Description
Study_id	Anonymized participant identifier(dust/nasal); 1~148
Sample_id	Subjects followed by this study:1~74
asthma_ägey7	Whether detecting asthma or not at age 7: YES=1; NO=0
allergic_aero_y3	Whether detecting allergy or not at age 3: YES=1; NO=0
rec_wheeze_y3	Whether detecting recurrent wheezing or not at age 3: YES=1; NO=0
wheeze_atopy_grp	Combine the results of fourth and fifth row: atopy_only;wheeze only; neither; both
nasal_score	Value evaluating the nasal situation:0~10
nasal_score_ge5	Binary outcome of nasal score: if nasal_score >5 ,then TRUE; if nasal_score <5 ,then FALSE
site	4 different sites:New York,St.Louis, Baltimore,Boston
data_source	Two sources: dust and nasal
Faith	Phylogenetic diversity: one of summary statistics for biodiversity indices
Chao1	Richness:one of summary statistics for biodiversity indices
Pielou	Evenness:one of summary statistics for biodiversity indices
OTU_1 ... OTU_10118	Operational Taxonomic Units:1~10118

Table 5: Variable Description Table

were not significant. This lead to the necessity to understand the individual OTU samples better. We utilized both unsupervised learning techniques and supervised learning techniques to perform variable selection and dimension reduction. Two subsets of important OTUs were identified from these techniques: one from the supervised learning methods that identified OTUs in connection to the response and one from the cross reference between methods using and not using the outcome of interest. The phylogenetic distribution of the supervised learning subset displayed a difference in relative frequency of certain phyla and classes. The logistic regression of the cross referenced OTUs identified certain individual OTUs that had a significant relationship with outcome of interest.

## Limitations

Certain limitations were identified during the analysis and are important to take into consideration when interpreting the main findings. Due to only having a subset of de-identified data were available for this analysis due to health data privacy restrictions. This restriction lead to having de-identified cities which restricts the amount of knowledge available about this variable. Individual traits or known differences between the cities were unknown to us. Another restriction of having a subset of the data is that the sampling method for choosing this subset of the study population has an unknown influence on the results of the methods employed.

## Future Work

The most important extension of this analysis is to investigate the relationship between the microbiomes and the other responses. Looking at the relationship between the dust and nasal microbiomes and the multinomial response of wheezing, atopy, both or neither as separate groups, as well as looking at the ultimate response of presence or absence of asthma would be a clear next step for the analysis. This is important due to the many insights that could be made about how the environmental and nasal microbiomes effect the development of certain diseases.

## References

- [1] Jinbo Bi, Kristin Bennett, Mark Embrechts, Curt Breneman, and Minghu Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3(Mar):1229–1243, 2003.
- [2] ST Buckland. On the variable circular plot method of estimating animal density. *Biometrics*, pages 363–384, 1987.
- [3] Benjamin Callahan, Diana Proctor, David Relman, Julia Fukuyama, and Susan Holmes. Reproducible research workflow in r for the analysis of personalized human microbiome data. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 21, page 183. NIH Public Access, 2016.
- [4] Daniel B DiGiulio, Benjamin J Callahan, Paul J McMurdie, Elizabeth K Costello, Deirdre J Lyell, Anna Robaczewska, Christine L Sun, Daniela SA Goltzman, Ronald J Wong, Gary Shaw, et al. Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences*, 112(35):11060–11065, 2015.
- [5] James E Gern. The urban environment and childhood asthma study. *Journal of Allergy and Clinical Immunology*, 125(3):545–549, 2010.
- [6] Finn V Jensen. *An introduction to Bayesian networks*, volume 210. UCL press London, 1996.
- [7] Sung-Kwon Kim, Douglas S Reed, Sara Olson, Matthias J Schnell, John K Rose, Phillip A Morton, and Leo Lefrançois. Generation of mucosal cytotoxic t cells against soluble protein by tissue-specific environmental and costimulatory signals. *Proceedings of the National Academy of Sciences*, 95(18):10814–10819, 1998.
- [8] Sharon M Lutz. Mediation analysis in genome-wide association studies: current perspectives. 2015.
- [9] Susan V Lynch, Robert A Wood, Homer Boushey, Leonard B Bacharier, Gordon R Bloomberg, Meyer Kattan, George T O’Connor, Megan T Sandel, Agustin Calatroni, Elizabeth Matsui, et al. Effects of early-life exposure to allergens and bacteria on recurrent wheeze and atopy in urban children. *Journal of Allergy and Clinical Immunology*, 134(3):593–601, 2014.
- [10] David Peter MacKinnon. *Introduction to statistical mediation analysis*. Routledge, 2008.
- [11] Paul J McMurdie and Susan Holmes. phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8(4):e61217, 2013.
- [12] Roberto Romero, Sonia S Hassan, Pawel Gajer, Adi L Tarca, Douglas W Fadrosh, Lorraine Nikita, Marisa Galuppi, Ronald F Lamont, Piya Chaemsaiithong, Jezid Miranda, et al. The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome*, 2(1):4, 2014.
- [13] Peter D Sly, Attilio L Boner, Bengt Björksten, Andy Bush, Adnan Custovic, Philippe A Eigenmann, James E Gern, Jorrit Gerritsen, Eckard Hamelmann, Peter J Helms, et al. Early identification of atopy in the prediction of persistent asthma in children. *The Lancet*, 372(9643):1100–1106, 2008.
- [14] 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.
- [15] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.