Investigating the Relationship Between the Microbiome and Environmental Characteristics

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

Problem Presenters: Herman Mitchell[7] and Agustin Calatroni[7]
Faculty Mentor: Emily Lei Kang[8]

July 22, 2015

[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

Outline

Background

What is a microbiome?

- A microbial community
- People and environments have distinctive microbiomes
- Recent links to human health and disease especially asthma/allergies

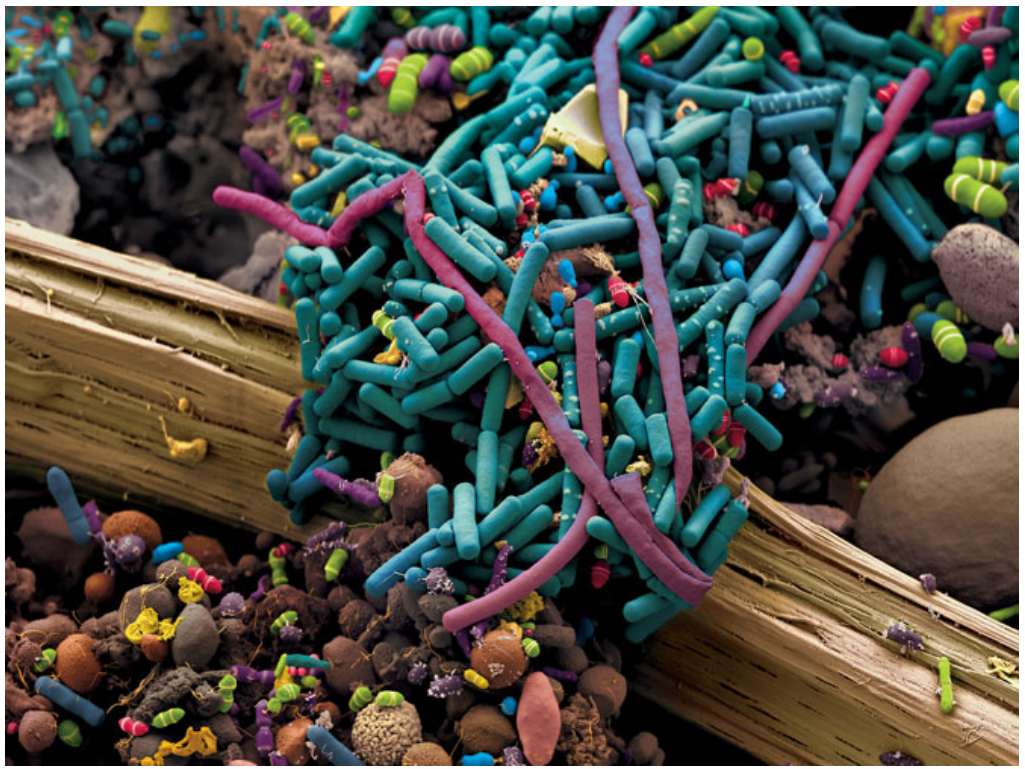


Figure: 3500:1 Magnification of Gut Bacteria (National Geographic 2013)

The Study

- Urban Environment and Childhood Asthma (URECA)
  - Birth Cohort Study
  - Dust samples collected annually
- Dust samples analyzed using 16S rRNA sequencing
  - Highly similar regions
  - Highly variable regions
- 97% similarity defines a unique taxon (bacteria)
- 277 homes sampled

The Study

Description of the data:

- Taxa frequency data

- Home characteristics

The Data

Taxa frequency Data: Frequencies are very sparse

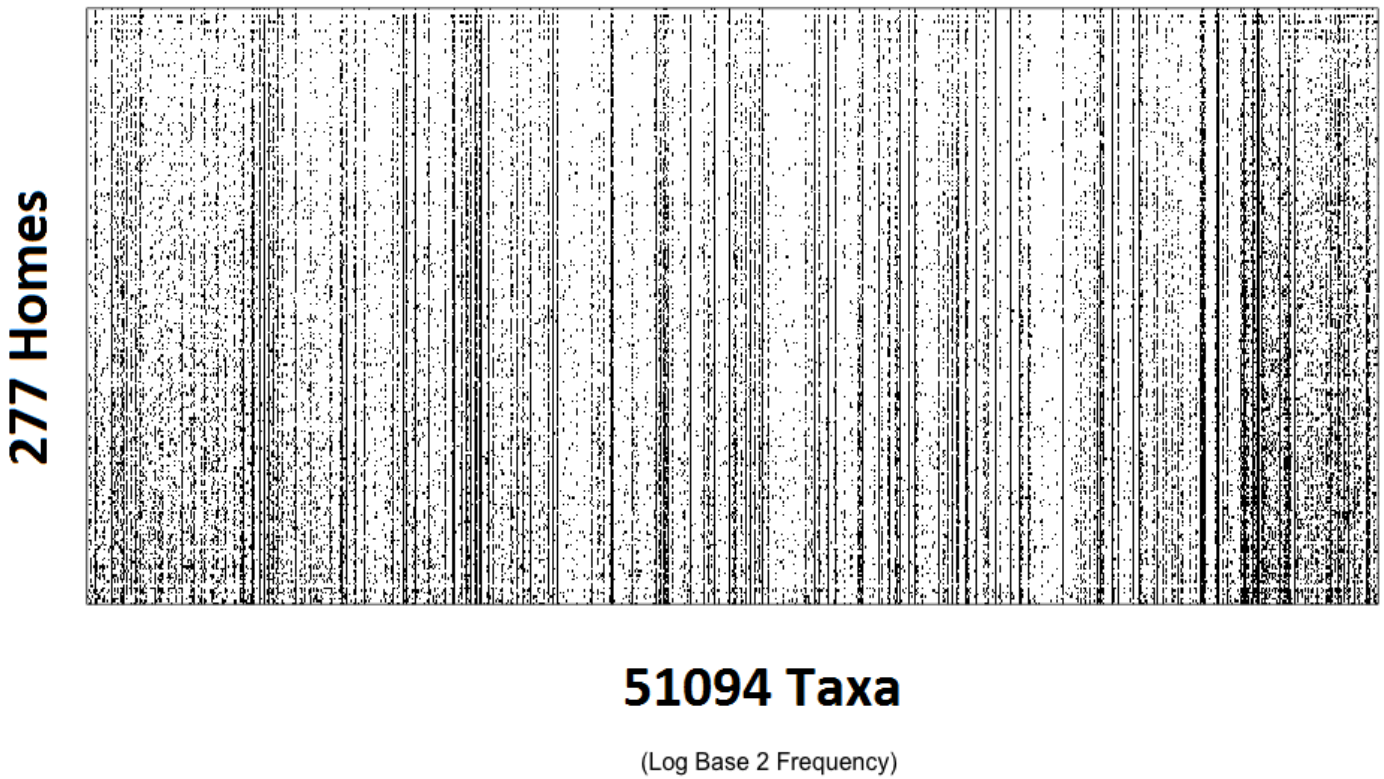


Figure: Heat Map of Taxa Data. White spaces are zero values.

The Data

Home Characteristics:

- Site (New York, St. Louis, Baltimore, Boston)
- House Type (7 types considered)
- Water Problems (Y/N)
- Dogs (Y/N)
- Cats (Y/N)
- Rats (Y/N)
- Mice (Y/N)
- Cockroaches (Y/N)

The Data

Home Characteristics:

- Site (New York, St. Louis, Baltimore, Boston)
- House Type (7 types considered)
- Water Problems (Y/N)
- Dogs (Y/N)
- Cats (Y/N)
- Rats (Y/N)
- Mice (Y/N)
- Cockroaches (Y/N)

The Problem

Can we identify:

- if the microbiomes are distinctly different between sites?
- which taxa of bacteria contribute to this difference?

Data Pre-Processing

Data Pre-Processing

Reducing the data

- Remove taxa variables with zero variance (one value for all observations) or near zero variance (very few unique values)
- $\sim 30,000$ taxa variables removed; $\sim 60\%$ reduction
- Reduced data dimension: $20,402$ taxa

Data Pre-Processing

Transformation

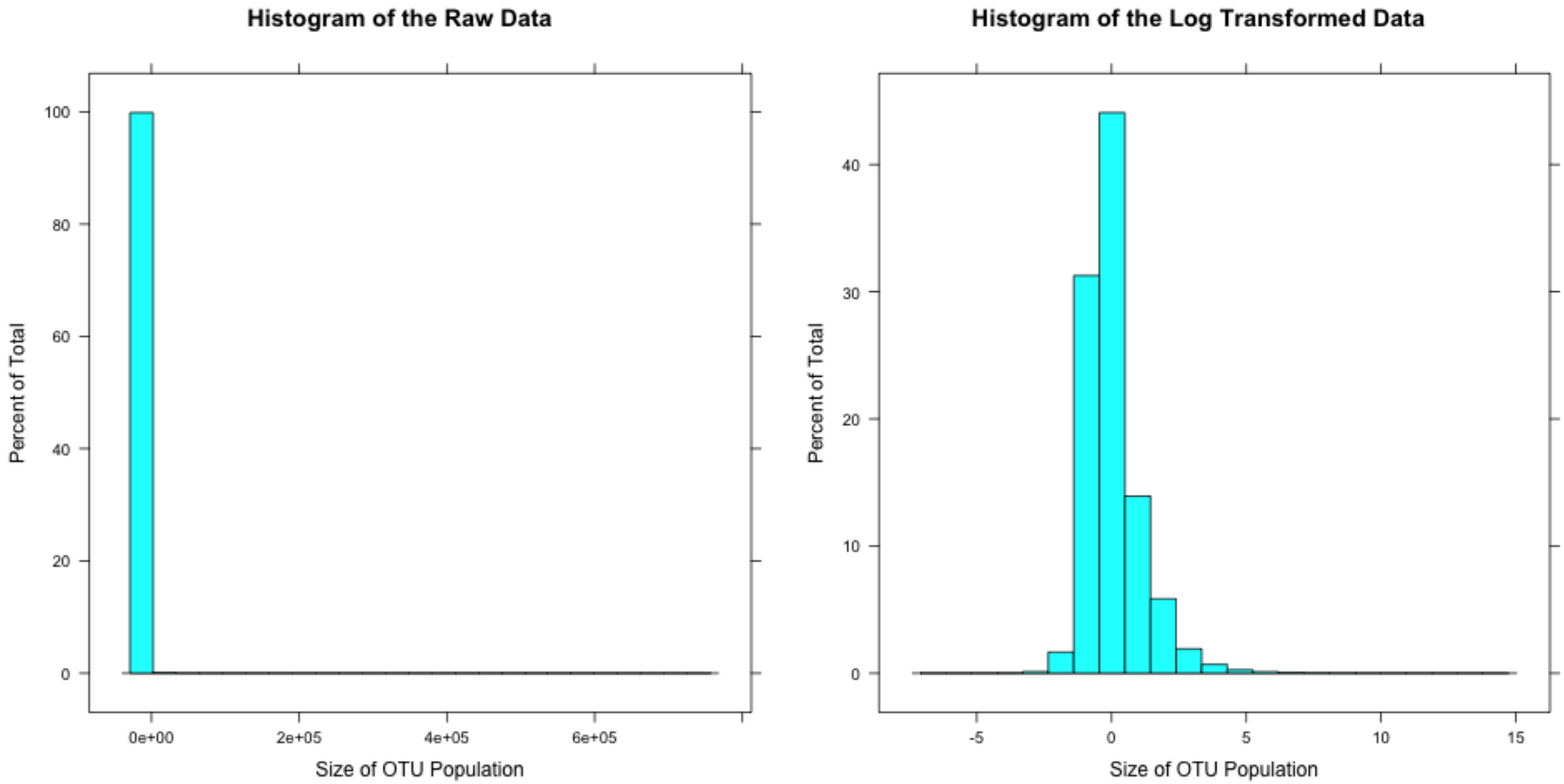


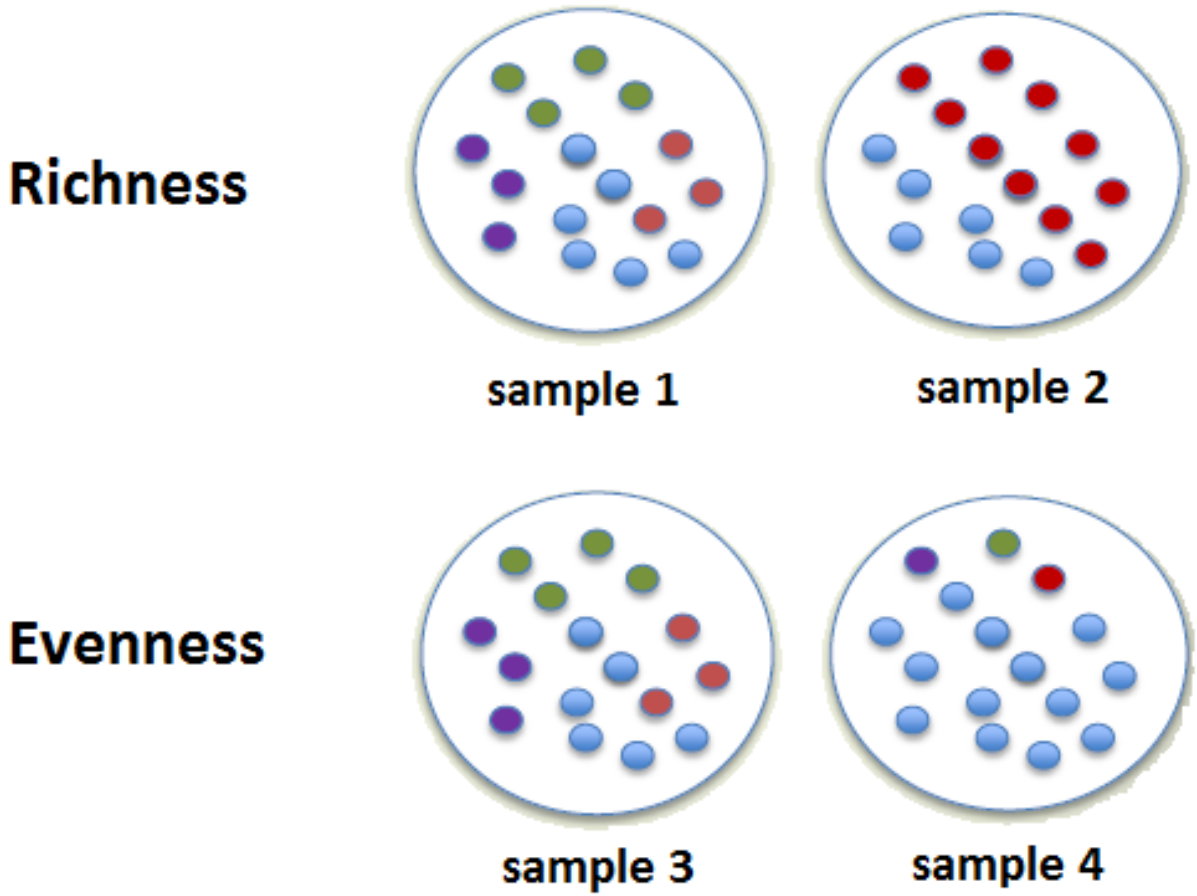
Figure: Histogram for the Raw Data & Scaled Modified Log Data

Results from Exploration

Exploratory Data Analysis

- General Level
- Detailed Level

Exploratory Data Analysis



Exploratory Data Analysis

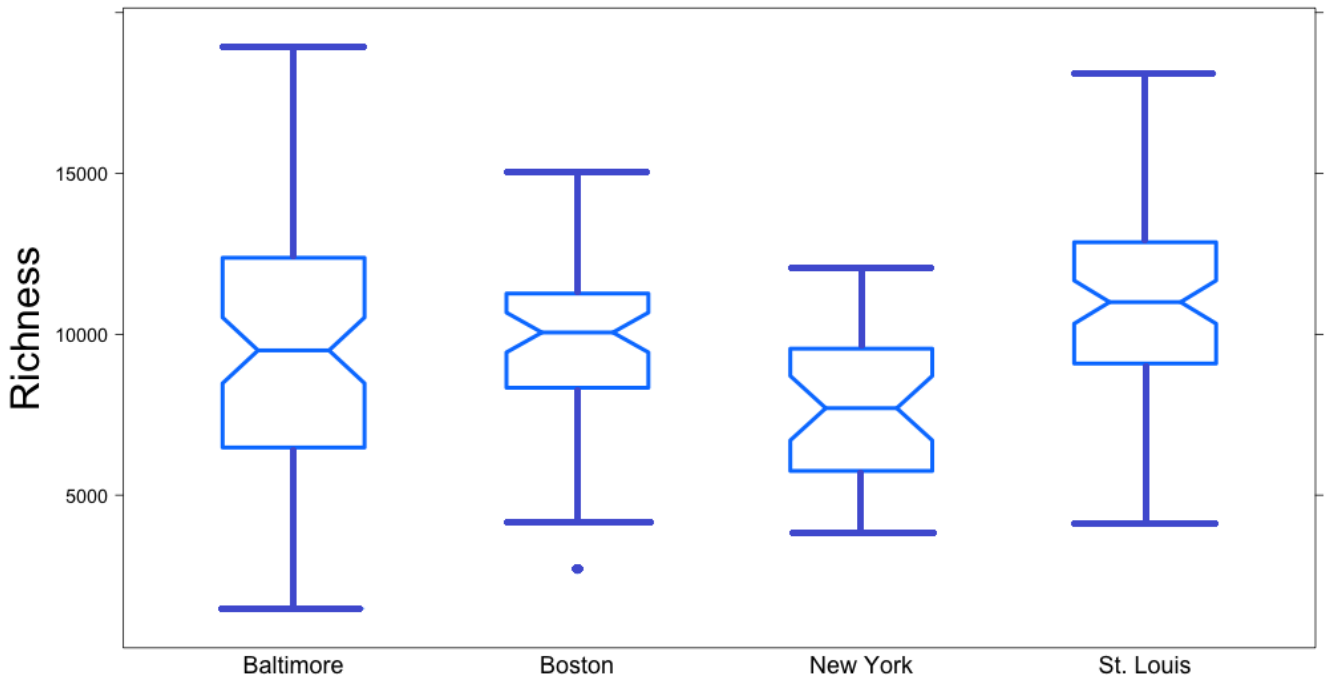


Figure: Box plots of Richness conditioned on Site

$\Longrightarrow$ New York is significantly different than the other sites in Richness.

Exploratory Data Analysis

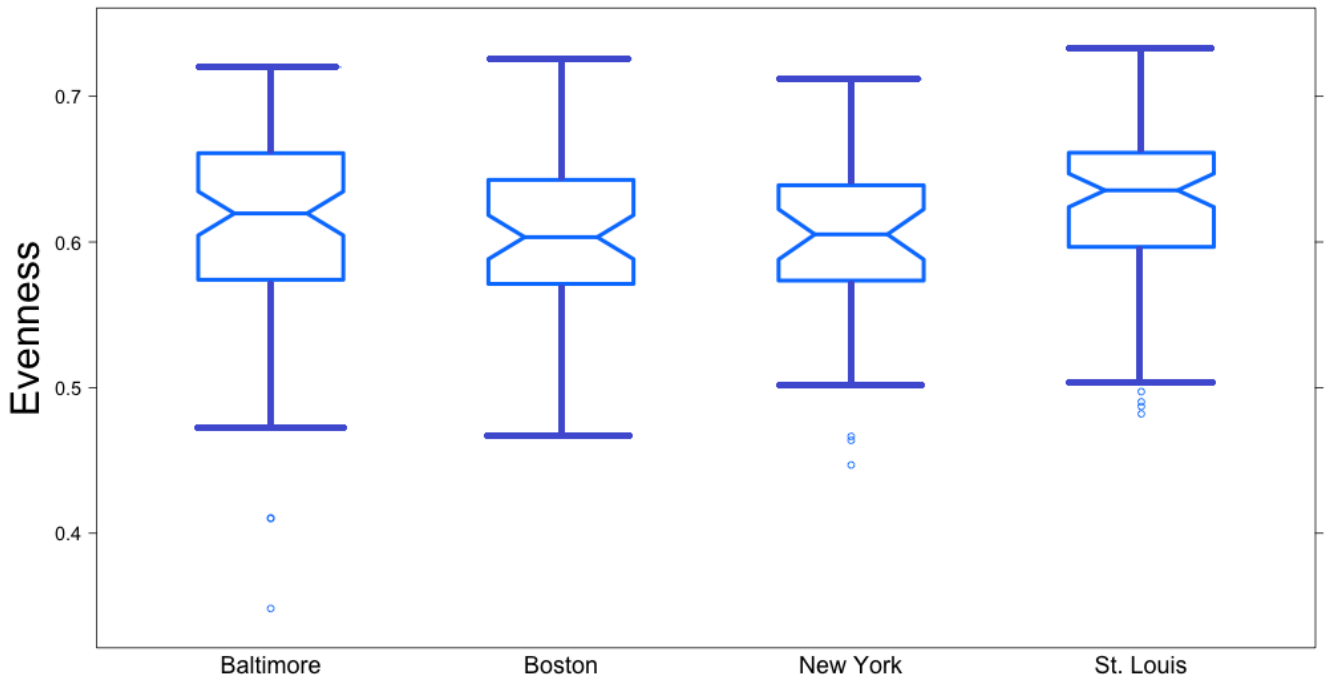


Figure: Box plots of Eveness conditioned on Site

$\Longrightarrow$ No significant difference between sites in Evenness.

Exploratory Data Analysis

Principal Component Analysis (PCA):

- Captures the common characteristics of the data via a weighted average of the variables
- Projects high-dimensional data in a low dimensional subspace

Exploratory Data Analysis

Only a few principal components are needed to explain a significant proportion of the variance. How many PCs to consider?

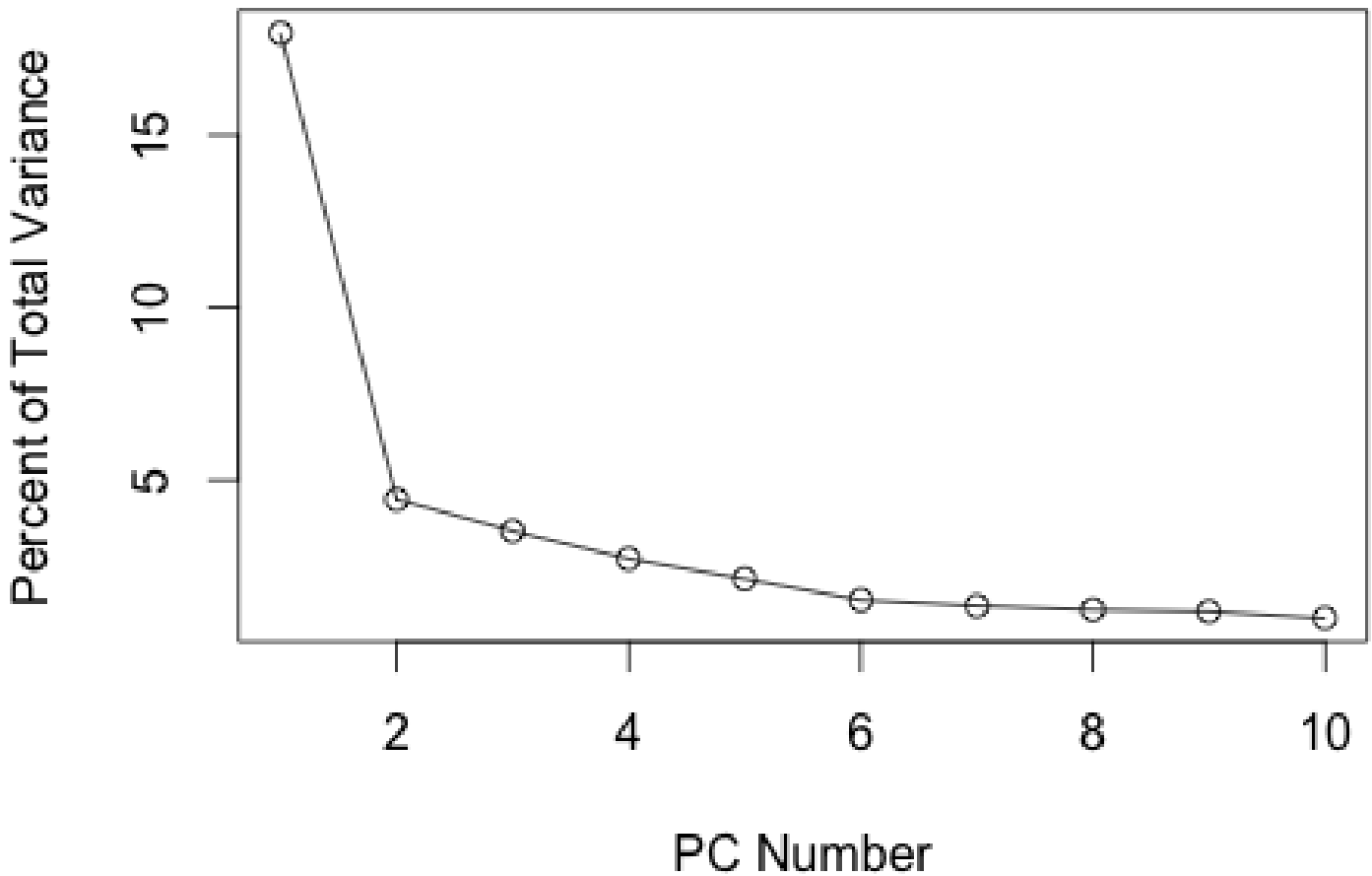


Figure: Eigenvalues numbers corresponding to PCs vs Captured variance

Exploratory Data Analysis

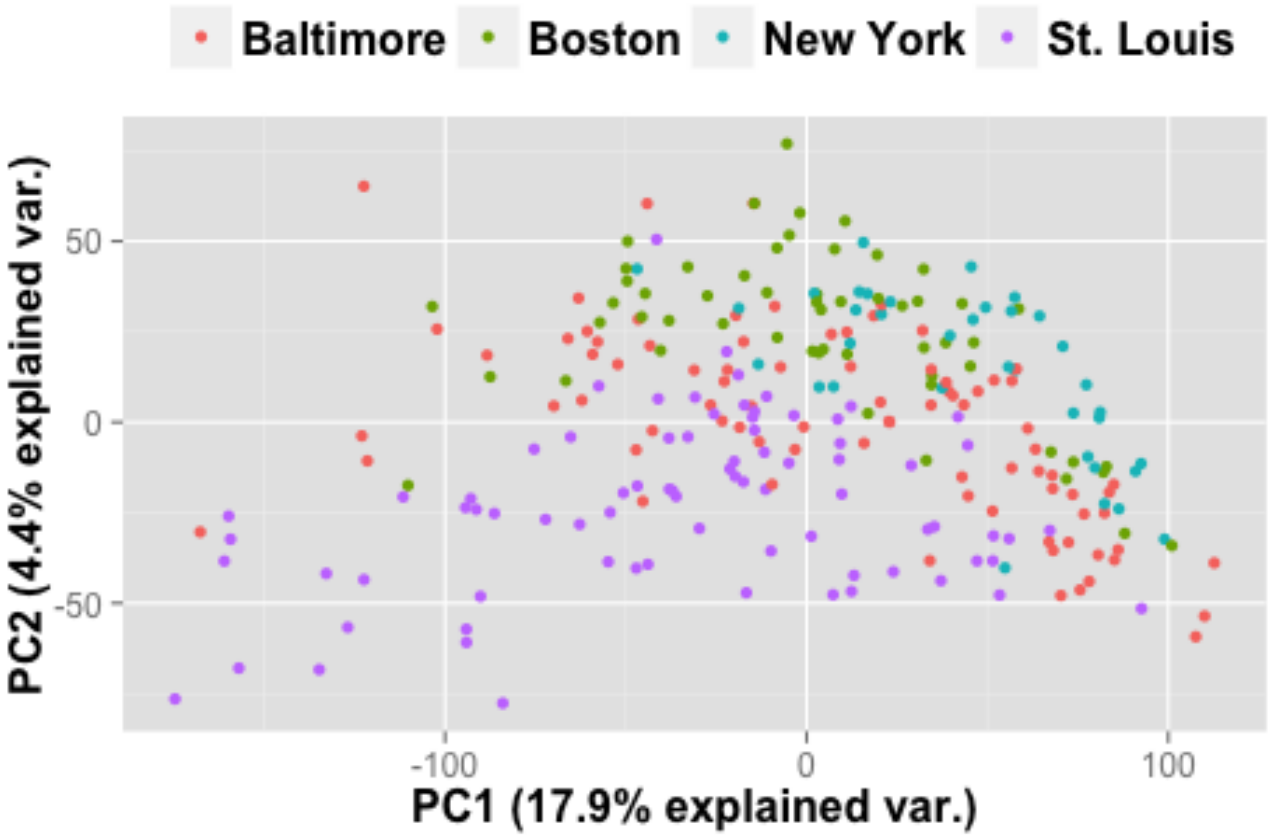


Figure: Data projected onto the subspace defined by PC1 and PC2. 22.4% of the variance captured

Exploratory Data Analysis

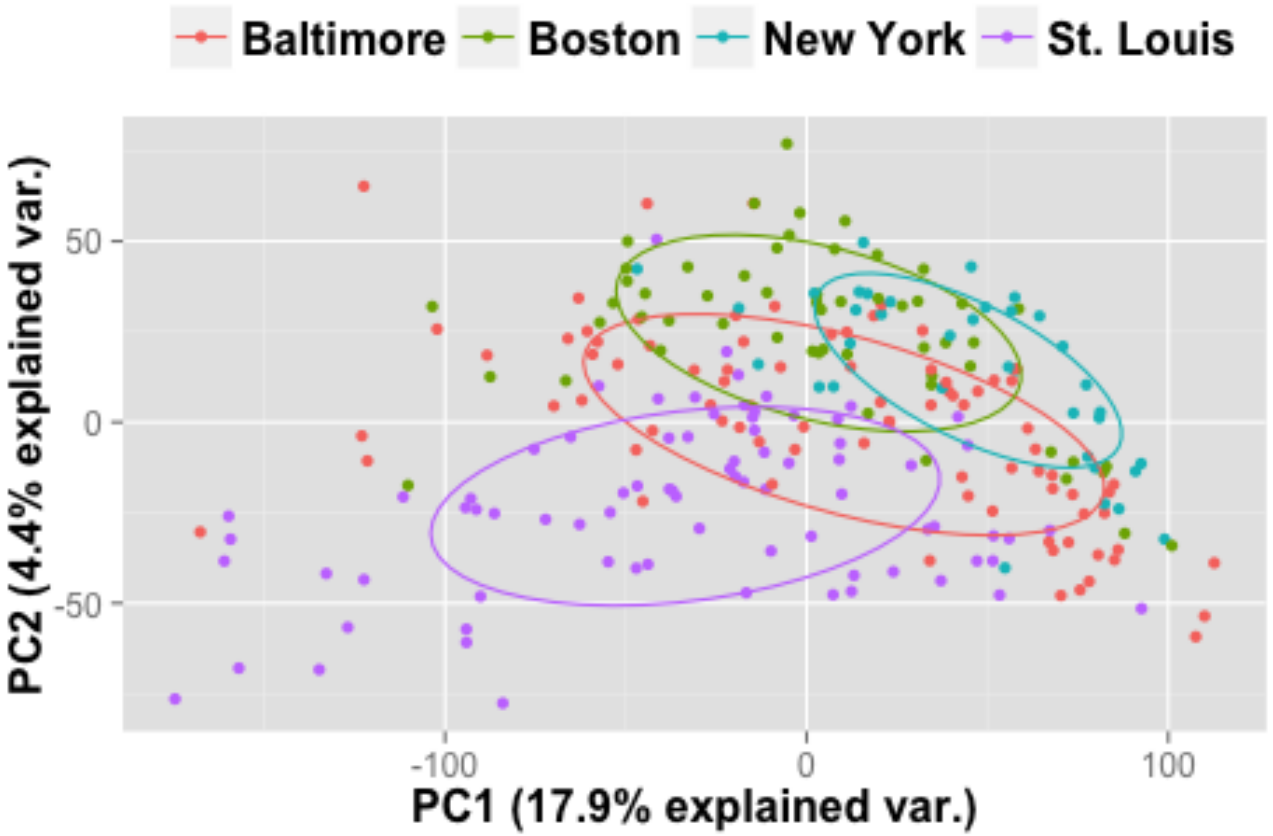


Figure: Data projected onto the subspace defined by PC1 and PC2. 22.4% of the variance captured

Exploratory Data Analysis

 -Baltimore;  -Boston;  -New York;  -St. Louis;

Figure: Data projected onto the subspace defined by PC1, PC2, & PC3. 25.9% of the variance captured

Volcano Plot

Volcano Plots:

- Identifies taxa with statistically and practically significant differences
- Helpful to explore binary variables
- Site factors were viewed as binary by considering each city vs. the other three.

Volcano Plot

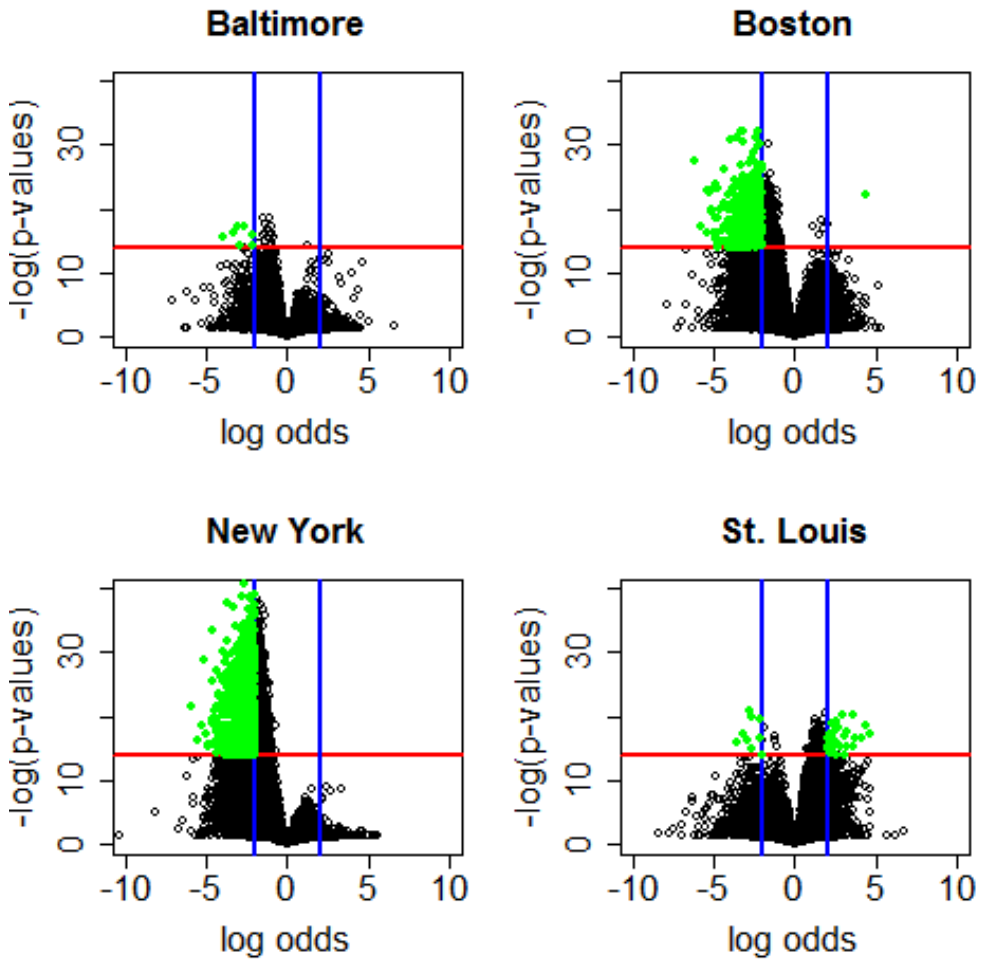


Figure: Volcano Plot for Sites

$\Longrightarrow$ Again, New York stands out for its lack of bacteria.

Results from Prediction

Predictive Modeling

- Elastic net technique
- Taxa selection based on site
- Predictive ability of selected taxa

Predictive Modeling

- Evidence of microbiome discrepancies at different sites. Which taxa are contributing?

- Use a multinomial logistic regression with an elastic net penalty with site as a response variable and the taxa as predictors

Predictive Modeling

G (site) has a multinomial distribution with $K = 4$ levels.The probability that G equals a particular class l is modeled as:

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

The objective function will have two parts: a **log likelihood** and a **penalty** function.

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

Predictive Modeling

- 70% training set, 30% test set. For the training set, 5-fold cross validation was repeated 3 times to tune α and λ.

Site	# Taxa Selected
Baltimore	39
Boston	38
New York	26
St. Louis	30
$\alpha = 0.55$ and $\lambda = 0.03$	

- No selected taxa were shared between all sites.
- 1 taxa was selected for both Baltimore and Boston
  - *Clostridiaceae Clostridium*

Predictive Modeling

- The location of each home in the test set was predicted by selecting the site with the highest probability.
- *Only one home in the test set was misclassified!*
- Obtained 98% accuracy using $< 1\%$ of the taxa

		Actual			
		Ba	Bo	NY	StL
Predicted	**Ba**	24	0	0	0
	Bo	0	15	0	1
	NY	0	0	10	0
	StL	0	0	0	22

Table: Elastic Net Confusion Table

Conclusions

Conclusions and Discussion

- Evidence of separability in microbiome characteristics at different sites
- At each site, elastic net selects between 26 and 39 "important" taxa

Future Work

- Apply supervised techniques to a reduced dimension data set
- Identify the home characteristics which correspond to site distinctions
- Analysis based on asthma/allergy diagnosis


Thank You

References I

[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.

References II

[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.

References III

[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.