INVESTIGATING THE USE OF CLASSIFICATION MODELS TO STUDY MICROBIAL COMMUNITY ASSOCIATIONS WITH BACTERIAL VAGINOSIS

A Dissertation
Presented in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
with a
Major in Bioinformatics and Computational Biology
in the
College of Graduate Studies
University of Idaho

by
Daniel Beck

May 2014

Major Professor: James A. Foster, Ph.D.

AUTHORIZATION TO SUBMIT DISSERTATION

This dissertation of Daniel Beck, submitted for the degree of Doctor of Philosophy with a major in Bioinformatics and Computational Biology and titled "Investigating the use of classification models to study microbial community associations with bacterial vaginosis," has been reviewed in final form. Permission, as indicated by the signatures and dates given below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor _____ Date _____
Dr. James A. Foster

Committee Member _____ Date _____
Dr. Larry Forney

Committee Member _____ Date _____
Dr. Mark McGuire

Committee Member _____ Date _____
Dr. Terence Soule

Program Administrator _____ Date _____
Dr. Eva Top

Discipline's College Dean _____ Date _____
Dr. Paul Joyce

Final Approval and Acceptance by the College of Graduate Studies

_____ Date _____
Dr. Jie Chen

ABSTRACT

Microbial communities are highly complex, often composed of hundreds or thousands of different microbe types. They are found nearly everywhere; in soil, water, and in close association with other organisms. Microbial communities are difficult to study. Many microbes are not easily grown in laboratory conditions. Interactions between microbes may limit the applicability of observations collected using isolated taxa. However, new sequencing technology is allowing researchers to study microbial communities in novel ways. Among these new techniques is 16S rRNA fingerprinting, which enables researchers to estimate the relative abundance of most microbes in the community.

These techniques are often used to study microbial communities living on or in the human body. These microbiomes are found at many different body sites and have been linked to the health of their human host. In particular, the vagina microbiome has been linked to bacterial vaginosis (BV). BV is highly prevalent with symptoms including odor, discharge, and irritation. While no single microbe has been shown to cause BV, the structure of the microbial community as a whole is associated with BV.

In this thesis, I explore methods that may be used to discover associations between microbial communities and phenotypes of those communities. I focus on associations between the vagina microbiome and BV. The first two chapters of this thesis describe software tools used to explore and visualize ecological datasets. In the last two chapters, I explore the use of machine learning techniques to model the relationships between the vagina microbiome and BV. Machine learning techniques are able to produce complex models that classify microbial communities by BV characteristics. These models may capture interactions that simpler models miss.

ACKNOWLEDGMENTS

I am deeply thankful to James Foster for constant support, guidance, and advice. This dissertation would have otherwise been impossible.

I would like to thank the other members of my committee, Larry Forney, Mark McGuire, and Terence Soule who provided critical discussions and advice.

I would also like to thank my fellow students and the broader IBEST community, for valuable conversations and discussions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Microbial communities thrive in many different environments and conditions. They are found in environments from rain forests [1] to deserts [2] and from acidic rivers [3] to basic lakes [4]. Nearly every environment on earth is home to microbes. In addition to their pervasiveness, microbial communities are important for a variety of reasons. Microbes play key roles in the treatment of wastewater [5]. Microbial communities in the soil drive nutrient cycling and influence the availability of nutrients to plants [6].

In addition to occurring throughout the environment, some microbial communities associate closely with other organisms. These microbial communities, known as microbiomes, are composed of microbes living on and in other organisms. Microbiomes often play key roles in the health of their host. Gut microbiomes aid in digestion and host energy uptake. This can be seen in cows and termites where microbes enable their hosts to metabolize cellulous [7, 8]. Similarly, in humans and other organisms, microbes break down complex carbohydrates to simpler forms their hosts can process [9, 10]. Microbial communities are associated with disease in complex ways. Human gut microbiomes are associated with obesity [11]. Microbial communities in the lungs may exacerbate diseases such as cystic fibrosis [12]. Certain vagina microbiomes are associated with bacterial vaginosis [13]. Other microbial communities may protect against invasions by pathogens or interact in complex ways with the immune system [14].

Past research into microbial communities has revealed many interesting characteristics and patterns. Many microbial communities are composed of hundreds or thousands of different microbe types. This high richness is often spread across phylogenetically diverse taxa [15, 16]. Gene-level views of microbial communities also show incredible levels of diversity. Functional gene richness has been shown to vary considerably across soil habitats [17].

In addition to this huge amount of largely unexplored taxa and gene richness, the ecological dynamics occurring in microbial communities are often completely unknown. Microbial composition can be highly variable over time [18] and environmental conditions [19]. Both microbial community richness and evenness can vary substantially [20]. Additionally the role of viruses and horizontal gene transfer in microbial communities is unclear.

In general, we know very little about the vast majority of microbes. Many microbes are difficult to culture on traditional media. Estimates of 'unculturable' microbes are often above 99% [21, 22]. While many of these microbes may have culturable relatives, it is not clear how ecologically similar closely related microbes may be. Even when microbes are easily grown in the lab, it is not immediately apparent what role they may play in natural communities. Interactions with other microbes and with variable environmental conditions may partly determine the ecological function of many microbes. The large number of microbes in most communities makes it difficult to replicate natural conditions in a controlled, replicable manner in the lab.

Advances in genetic sequencing have opened a new window into microbial community structure and function. Researchers can now determine the genetic sequence of millions of DNA fragments relatively cheaply. This has spurred the development of a variety of techniques for analyzing microbial communities. Two of these are metagenomic sequencing and meta-amplicon sequencing.

Metagenomic sequencing attempts to sequence a sample of all the genes in a microbial community. This allows researchers to determine which genes are present in the samples. In some cases, these genes can then be lumped into pathways or groups that reflect functional characteristics. One goal of metagenomic sequencing is to predict the genetic capabilities of the microbes present in the community.

Amplicon sequencing PCR amplifies relatively short regions of the microbial genome, which act as identifiers or barcodes for the microbe. Based on these gene regions, researchers can estimate the identity of the microbes present. For bacterial surveys,

portions of the 16S rRNA gene are typically used as the barcode. This sequencing often results in estimates of the relative abundance of the bacteria present in each microbial community. In this thesis I focus on datasets produced using 16S rRNA sequencing.

The bioinformatic processing of large sequencing datasets is an important and ongoing research area. In the case of 16S rRNA sequencing, processing includes removing low quality sequencing reads, chimera detection and removal, read trimming to remove low-quality ends, primers and barcodes, and data partitioning into projects and samples. This processing may include a clustering step that lumps similar sequences together to compensate for small sequencing errors. After this processing, the 16S rRNA sequences are grouped into bins or operational taxonomic units (OTUs). This may be done by clustering reads by sequence similarity or by comparing each read to a database of known bacteria. The result of this processing is generally a table of bacterial abundances. This abundance table includes the number of reads of each OTU found in every sample.

These processing steps include several choices and trade-offs. It is not immediately clear how decisions made during dataset processing affect downstream analyses. There is always a balance between removing sequences due to possible quality issues and keeping sequences that may inform hypotheses. Overly conservative processing may remove real sequence variability, while overly permissive processing may allow noise to distort or overwhelm real patterns. It is helpful, therefore, to have a tool that allows researchers to efficiently determine how changes in data processing may influence the results of downstream analyses. In Chapter 1 we present OTUbase, an R package that provides a framework for this type of data exploration. OTUbase is a tool that provides data structures and basic functions for manipulating and analyzing microbial community data [23].

After the reduction of sequencing datasets to abundance tables, several summary statistics may be calculated. These summary statistics may include richness and diversity measures or presence or absence of specific taxa. While these summary statistics allow rough comparisons between samples, researchers are often interested in exploring the microbial community data in more complex ways. It is often helpful to use plots and

figures to visualize microbial communities in order to find patterns and trends. Researchers may use principal component or coordinate analyses (PCA/PCoA), clustering dendrograms, scatterplots, and heatmaps to explore how sample data varies with the microbial community.

These analyses and visualizations may be produced using a wide variety of software. R is perhaps the most common of these tools, however custom pipelines and scripts are sometimes written in Python and other programming languages. These tools are generally command line based and require the user to have substantial programming abilities. While the format of these tools works well for incorporation into pipelines, the non-visual interface can hinder exploration of the data. Chapter 2 presents Seed, a visualization tool for microbial communities. Seed provides a visual interface for exploring microbial community structure. It includes tools that make it easy to look for patterns and trends in microbial datasets.

Simple analysis and visualization of these datasets may often be insufficient. Complex relationships and interactions between microbial taxa may not be readily discernible. Visualizations may miss patterns, pairwise correlations may not capture interactions, and richness measures may not be relevant for many questions. The huge number of taxa in many communities means narrowing down the parts of the community associated with environmental factors or phenotypes is difficult. This difficult problem is an important one, however, especially when the phenotype in question is a disease.

One example of a complex association between a microbial community and disease is the vagina microbiome and bacterial vaginosis. The vagina microbiome is often composed of hundreds of different microbe types, although frequently only a few are at high abundance levels. Common members of the vagina microbiome include various *Lactobacillus spp.* such as *L. iners*, *L. crispatus*, and *L. gasseri*. Some communities include *Prevotella*, *Atopobium*, and a variety of other genera [24].

Bacterial vaginosis (BV) is a disease associated with the vagina microbiome. Symptoms of BV include odor, discharge, and irritation. BV prevalence is high, with

estimates of affected women as high as nearly 30% [25]. BV has been linked to increased rates of preterm birth [26] and increased susceptibility to some STDs [27, 28]. No single microbe has been shown to cause BV, however, BV has been linked to the microbial community as a whole [29]. Notably, communities with high abundances of *Lactobacillus* species seem to be indicative of a healthy community [30].

The apparently complex relationship between BV and the vagina microbiome makes it an ideal system in which to study methods that may identify important parts of the microbial community. In Chapter 3, we explore methods that use machine-learning classifiers to identify possible links between the vagina microbiome and BV. These methods involve two general steps. In the first step, a classification model is generated using some machine learning method. The model accuracy is then measured to determine how well the model sorts samples by BV characteristics. In the second step, the model is deconstructed to determine what microbial community features are important to the classifier accuracy. These features can reasonably be hypothesized to be associated with BV. Chapter 3 compares the models generated using three machine-learning techniques, genetic programming (GP), logistic regression (LR), and random forests (RF) [31].

Two key results are discussed in detail in Chapter 3. First, all three machine-learning algorithms produce classification models with a high accuracy of between 80 and 90%. Second, the important features identified by each method are largely unique. There is little overlap in the top fifteen features identified using GP, LR, and RF. Additionally, the study presented in Chapter 3 ranks features by importance, but does not calculate an effect size for the features. Consequently, it is difficult to determine how many features are actually important for the classification model. This feature importance ambiguity limits the classification models' usefulness. While it is clear from their high classification accuracy that the models are capturing a link between the microbial community and BV, the model features responsible for this accuracy are unknown.

Chapter 4 extends the results of Chapter 3 by looking more closely at the important features identified by each model. It attempts to determine why each method results in

different important features and estimates how important each feature is to the overall accuracy. Chapter 4 uses three different types of feature subsets to answer these questions. The first subset type selects the top N features from each feature ranking. The second subset type selects features using a sliding window across the feature rankings. The third subset type selects features randomly.

These subsets illustrate key classification model characteristics. First, only a few features are necessary to obtain high classification accuracy. Second, the importance measures for the different machine learning techniques are often not ideal. Third, there appears to be substantial redundancy in the microbial community features. Chapter 4 explores these results in detail.

In summary, the first two chapters introduce OTUbase and Seed, respectively. These software packages provide tools for manipulating, analyzing and visualizing microbial community data. The third chapter explores using machine learning techniques to generate models for classifying microbial communities by BV characteristics. It compares the accuracy of models generated using three different types of machine learning algorithms and the important features identified by those algorithms. The fourth chapter extends the results of the third chapter by using feature subsets to validate the identified important features.

1.1 References

[1] J. L. Rodrigues, V. H. Pellizari, R. Mueller, K. Baek, E. d. C. Jesus, F. S. Paula, B. Mirza, G. S. Hamaoui, S. M. Tsai, B. Feigl, *et al.*, "Conversion of the amazon rainforest to agriculture results in biotic homogenization of soil bacterial communities," *Proceedings of the National Academy of Sciences*, vol. 110, no. 3, pp. 988–993, 2013.

[2] K. P. Drees, J. W. Neilson, J. L. Betancourt, J. Quade, D. A. Henderson, B. M. Pryor, and R. M. Maier, "Bacterial community structure in the hyperarid core of the atacama desert, chile," *Applied and environmental microbiology*, vol. 72, no. 12, pp. 7902–7908, 2006.

[3] A. I. López-Archilla, I. Marin, and R. Amils, "Microbial community composition and ecology of an acidic aquatic environment: the tinto river, spain," *Microbial ecology*, vol. 41, no. 1, pp. 20–35, 2001.

[4] B. E. Jones, W. D. Grant, A. W. Duckworth, and G. G. Owenson, "Microbial diversity of soda lakes," *Extremophiles*, vol. 2, no. 3, pp. 191–200, 1998.

[5] M. Wagner, A. Loy, R. Nogueira, U. Purkhold, N. Lee, and H. Daims, "Microbial community composition and function in wastewater treatment plants," *Antonie Van Leeuwenhoek*, vol. 81, no. 1-4, pp. 665–680, 2002.

[6] M. G. Van Der Heijden, R. D. Bardgett, and N. M. Van Straalen, "The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems," *Ecology letters*, vol. 11, no. 3, pp. 296–310, 2008.

[7] M. Hess, A. Sczyrba, R. Egan, T.-W. Kim, H. Chokhawala, G. Schroth, S. Luo, D. S. Clark, F. Chen, T. Zhang, *et al.*, "Metagenomic discovery of biomass-degrading genes and genomes from cow rumen," *Science*, vol. 331, no. 6016, pp. 463–467, 2011.

[8] D. G. Boucias, Y. Cai, Y. Sun, V.-U. Lietze, R. Sen, R. Raychoudhury, and M. E. Scharf, "The hindgut lumen prokaryotic microbiota of the termite reticulitermes flavipes and its responses to dietary lignocellulose composition," *Molecular ecology*, vol. 22, no. 7, pp. 1836–1853, 2013.

[9] F. Bäckhed, R. E. Ley, J. L. Sonnenburg, D. A. Peterson, and J. I. Gordon, "Host-bacterial mutualism in the human intestine," *science*, vol. 307, no. 5717, pp. 1915–1920, 2005.

[10] B. L. Cantarel, V. Lombard, and B. Henrissat, "Complex carbohydrate utilization by the healthy human microbiome," *PloS one*, vol. 7, no. 6, p. e28742, 2012.

[11] P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, *et al.*, "A core gut microbiome in obese and lean twins," *Nature*, vol. 457, no. 7228, pp. 480–484, 2009.

[12] D. Willner, M. R. Haynes, M. Furlan, R. Schmieder, Y. W. Lim, P. B. Rainey, F. Rohwer, and D. Conrad, "Spatial distribution of microbial communities in the cystic fibrosis lung," *The ISME journal*, vol. 6, no. 2, pp. 471–474, 2012.

[13] Z. Ling, J. Kong, F. Liu, H. Zhu, X. Chen, Y. Wang, L. Li, K. E. Nelson, Y. Xia, and C. Xiang, "Molecular analysis of the diversity of vaginal microbiota associated with bacterial vaginosis," *BMC genomics*, vol. 11, no. 1, p. 488, 2010.

[14] J. L. Round and S. K. Mazmanian, "The gut microbiota shapes intestinal immune responses during health and disease," *Nature Reviews Immunology*, vol. 9, no. 5, pp. 313–323, 2009.

[15] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, *et al.*, "Environmental genome shotgun sequencing of the sargasso sea," *science*, vol. 304, no. 5667, pp. 66–74, 2004.

[16] C. J. Castelle, L. A. Hug, K. C. Wrighton, B. C. Thomas, K. H. Williams, D. Wu, S. G. Tringe, S. W. Singer, J. A. Eisen, and J. F. Banfield, "Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment," *Nature communications*, vol. 4, 2013.

[17] N. Fierer, J. W. Leff, B. J. Adams, U. N. Nielsen, S. T. Bates, C. L. Lauber, S. Owens, J. A. Gilbert, D. H. Wall, and J. G. Caporaso, "Cross-biome metagenomic analyses of soil microbial communities and their functional attributes," *Proceedings of the National Academy of Sciences*, vol. 109, no. 52, pp. 21390–21395, 2012.

[18] S. Schmidt, E. Costello, D. Nemergut, C. Cleveland, S. Reed, M. Weintraub, A. Meyer, and A. Martin, "Biogeochemical consequences of rapid microbial turnover and seasonal succession in soil," *Ecology*, vol. 88, no. 6, pp. 1379–1385, 2007.

[19] P.-Y. Qian, Y. Wang, O. O. Lee, S. C. Lau, J. Yang, F. F. Lafi, A. Al-Suwailem, and T. Y. Wong, "Vertical stratification of microbial communities in the red sea revealed by 16s rdna pyrosequencing," *The ISME journal*, vol. 5, no. 3, pp. 507–518, 2011.

[20] A. Shade, J. G. Caporaso, J. Handelsman, R. Knight, and N. Fierer, "A meta-analysis of changes in bacterial and archaeal communities with time," *The ISME journal*, 2013.

[21] W. Wade, "Unculturable bacteria-the uncharacterized organisms that cause oral infections," *Journal of the Royal Society of Medicine*, vol. 95, no. 2, pp. 81–83, 2002.

[22] R. I. Amann, W. Ludwig, and K.-H. Schleifer, "Phylogenetic identification and in situ detection of individual microbial cells without cultivation.," *Microbiological reviews*, vol. 59, no. 1, pp. 143–169, 1995.

[23] D. Beck, M. Settles, and J. A. Foster, "Otubase: an r infrastructure package for operational taxonomic unit data," *Bioinformatics*, vol. 27, no. 12, pp. 1700–1701, 2011.

[24] J. Ravel, P. Gajer, Z. Abdo, G. M. Schneider, S. S. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, *et al.*, "Vaginal microbiome of reproductive-age women," *Proceedings of the National Academy of Sciences*, vol. 108, no. Supplement 1, pp. 4680–4687, 2011.

[25] E. H. Koumans, M. Sternberg, C. Bruce, G. McQuillan, J. Kendrick, M. Sutton, and L. E. Markowitz, "The prevalence of bacterial vaginosis in the united states, 2001-2004; associations with symptoms, sexual behaviors, and reproductive health," *Sexually transmitted diseases*, vol. 34, no. 11, pp. 864–869, 2007.

[26] S. L. Hillier, R. P. Nugent, D. A. Eschenbach, M. A. Krohn, R. S. Gibbs, D. H. Martin, M. F. Cotch, R. Edelman, J. G. Pastorek, A. V. Rao, *et al.*, "Association between bacterial vaginosis and preterm delivery of a low-birth-weight infant," *New England Journal of Medicine*, vol. 333, no. 26, pp. 1737–1742, 1995.

[27] H. C. Wiesenfeld, S. L. Hillier, M. A. Krohn, D. V. Landers, and R. L. Sweet, "Bacterial vaginosis is a strong predictor of neisseria gonorrhoeae and chlamydia trachomatis infection," *Clinical Infectious Diseases*, vol. 36, no. 5, pp. 663–668, 2003.

[28] L. Myer, L. Denny, R. Telerant, M. de Souza, T. C. Wright, and L. Kuhn, "Bacterial vaginosis and susceptibility to hiv infection in south african women: a nested case-control study," *Journal of Infectious Diseases*, vol. 192, no. 8, pp. 1372–1380, 2005.

[29] B. B. Oakley, T. L. Fiedler, J. M. Marrazzo, and D. N. Fredricks, "Diversity of human vaginal bacterial communities and associations with clinically defined bacterial vaginosis," *Applied and environmental microbiology*, vol. 74, no. 15, pp. 4898–4909, 2008.

[30] S. Srinivasan, N. G. Hoffman, M. T. Morgan, F. A. Matsen, T. L. Fiedler, R. W. Hall, F. J. Ross, C. O. McCoy, R. Bumgarner, J. M. Marrazzo, *et al.*, "Bacterial

communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria," *PloS one*, vol. 7, no. 6, p. e37818, 2012.

[31] D. Beck and J. A. Foster, "Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics," *PloS one*, vol. 9, no. 2, p. e87830, 2014.

CHAPTER 2
OTUBASE: AN R INFRASTRUCTURE PACKAGE FOR OPERATIONAL TAXONOMIC UNIT DATA

2.1 Notes

Chapter 1 describes OTUbase, an R package that provides data structures and functions for operational taxonomic unit data manipulation and analysis. The goal of OTUbase is to provide a framework that allows researchers to incorporate typical microbiome data into analyses and visualizations. OTUbase organizes and links together data types that include sequencing reads, read quality, taxonomic information, and sample metadata. OTUbase is open source and available at http://www.bioconductor.org/packages/release/bioc/html/OTUbase.html. OTUbase is published in the journal Bioinformatics. The published manuscript is reprinted here. License information and reprinting permission from Oxford Press is shown in the appendix. Citation information is shown below.

Daniel Beck, Matt Settles, and James A. Foster (2011) OTUbase: an R infrastructure package for operational taxonomic unit data. *Bioinformatics* 27(12): 1700-1701. PMC3106189

2.2 Abstract

OTUbase is an R package designed to facilitate the analysis of operational taxonomic unit (OTU) data and sequence classification (taxonomic) data. Currently there are programs that will cluster sequence data into OTUs and/or classify sequence data into known taxonomies. However, there is a need for software that can take the summarized output of these programs and organize it into easily accessed and manipulated formats. OTUbase

provides this structure and organization within R, to allow researchers to easily manipulate the data with the rich library of R packages currently available for additional analysis.

Availability

OTUbase is an R package available through Bioconductor. It can be found at http://www.bioconductor.org/packages/release/bioc/html/OTUbase.html.

2.3 Introduction

New sequencing technologies, such as Roche's 454 pyrosequencing, have made it possible to sequence large amounts of DNA quickly. These new capabilities are being used to sequence portions of marker genes that allow investigators to determine which organisms are present within a sample. This is especially useful for the study of microbes. Microbes are often difficult to culture. Therefore, culture-independent methods are used to determine the composition of microbial communities. 454 16s amplicon sequencing is one such culture-independent method [1].

There are currently two primary approaches used to analyze environmental community amplicon data. The first approach uses a database of identified sequences to train an algorithm to classify a set of unknown sequences, such as 454 sequence reads. A widely used example of this is the RDP classifier [2]. This type of analysis labels each sequence with a taxonomic classification. A classification approach has the advantage of producing taxonomic names for each amplicon. However, it is dependent on the accuracy of the algorithm and the quality of the database used to train the algorithm.

A second approach to the analysis of amplicon data is to cluster the sequences based on overall sequence similarity [3]. These clusters represent operational taxonomic units, or OTUs. The OTU approach potentially decreases the biases of using an incomplete database to classify the sequences. However, the number of OTUs may be inflated by sequencing error and the taxonomic identity of the microbes in the sample remains unresolved [4].

Programs are available to perform both of these types of data analysis. In particular, some programs or pipelines are able to both cluster and classify sequences. These include Mothur [3], the RDP pipeline [5], QIIME [6], and PANGEA [7]. These programs are also able to perform some limited downstream analysis of the data.

OTUbase is an R infrastructure package that imports OTU or classification files produced by these programs along with the experiment metadata and sequence data. The imported data is stored in a structure that is easy to access and manipulate. The researcher is able to quickly summarize and visualize the data. More importantly, the data is abstracted from the raw data. This abstraction allows for the development of novel analysis techniques using the power of the R statistical programming environment. The structure of OTUbase is general enough to be applicable to all amplicon types, including 16s, 18s and IGS regions, or other targeted genes. Additionally, OTUbase is not restricted to any specific sequencing technique. Any data that can be clustered or classified into OTUs may be explored using OTUbase. This may include data from older Sanger sequencing and newer approaches such as Illumina sequencing.

2.4 Features

OTUbase is an R package available through Bioconductor (www.bioconductor.org). The package includes two types of S4 classes, OTUset and TAXset, that organize and structure OTU and classification (taxonomic) data respectively. S4 classes are a versatile data structure available in R. These classes include slots, or compartments within the class that hold distinct parts of the data. There are many types of data associated with typical community amplicon experiments. These data may include sample metadata, sequencing reads and quality, sequence memberships in OTUs, and sequence classifications. The classes provided by OTUbase organize these data into arrays.

The OTUset class type in OTUbase is designed for OTU related analyses. It includes slots for: the read identifier, the read's DNA sequence, the DNA sequence quality of the read, the sample identifier the read comes from, and the OTU identifier the read belongs

to. These slots are linked together by position. For example, the first sequence ID is linked to the first sequence in the sequence slot and the first sample ID in the sample ID slot. In addition to these slots, two slots hold metadata. The first metadata slot is for sample metadata that is linked to a corresponding sample identifier. The last slot holds OTU metadata and is linked to the OTU identifier slot. The OTU metadata slot can be used to hold OTU classification information, for example.

The TAXset class type provided in OTUbase is designed for classification data. This class is similar to the OTU class with the key difference being that the slot for the OTU identifier is replaced by a slot for the classification. The OTU metadata is replaced by a taxonomy metadata slot, which may hold additional information about specific taxonomic categories, such as family, order, etc.

In addition to providing structure and organization for amplicon data, OTUbase makes it easy to import data into OTUbase objects. OTUbase is able to read the output files produced by both the RDP classifier and Mothur. In general OTUbase is able to import large datasets quickly. Memory and computation limitations that arise in the analysis of amplicon data are much more important during the classification or clustering of sequences into OTUs then in the downstream analysis. Consequently, OTUbase can handle large datasets quickly and efficiently when the optional sequence and quality data is not included. It is expected that the majority of users will not find sequence data necessary. This capability is included, however, due to the possible use of this data in the development of new analysis techniques.

The key strength of OTUbase is that it interfaces easily with the rich library of packages already available in R for data exploration, visualization and statistical analyses. For example, OTUbase can produce abundance data that is accepted as input to packages that provide ecological data analyses (e.g. vegan [8]). By providing a structure and organization for amplicon data within R, OTUbase enables the efficient development of new analyses and visualization techniques.

A detailed example workflow can be found in the vignette included with the OTUbase

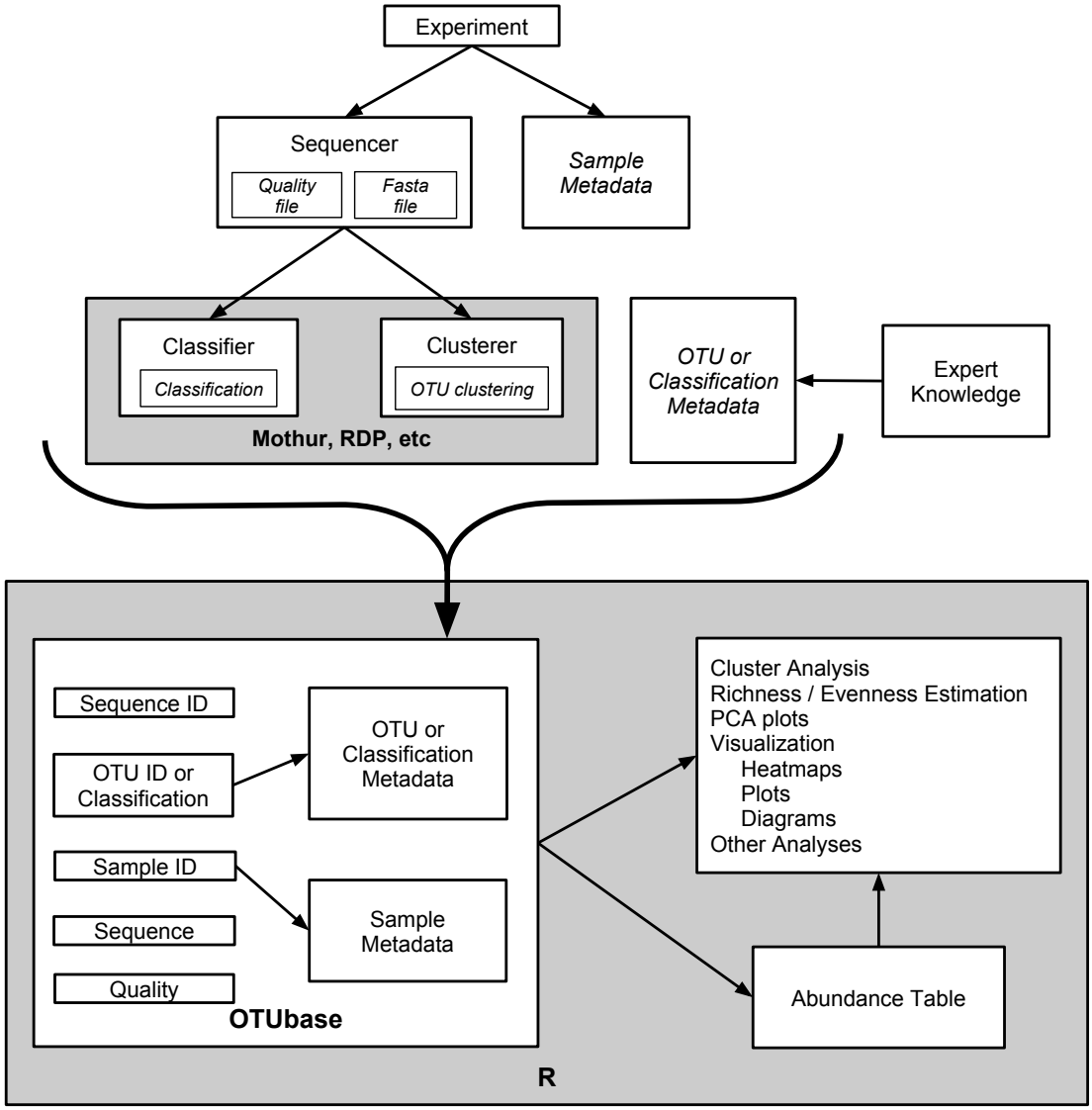

Figure 2.1: Organization of data within OTUbase. Data analysis starts by either clustering or classifying the reads obtained from the sequencer with an external application. The clustering or classification results are collected by OTUbase along with read and quality data and sample metadata. OTUbase then organizes the data into an R object. The library of R tools can then be used to analyze and visualize the data.

package. This example uses data originally presented in Sogin *et al.* in 2006 [9]. A typical workflow will include data import and the generation of an abundance table using the functions *readOTUset* and *abundance*. This abundance table can then be used in many analyses such as richness calculations (*estimateR* from the vegan package) and cluster analyses (*clusterSamples*). The data may be visualized using R commands such as *heatmap*. OTUbase allows an abundance table to be generated based on any column in the metadata. This makes OTUbase a powerful tool for data exploration. OTUbase also includes functions to manipulate OTU data. The subsetting function *subOTUset* is able to quickly generate subsets of the complete dataset. This allows the user to focus analyses on interesting subsets of the data.

2.5 Conclusion

OTUbase organizes and structures data associated with community amplicon experiments into R classes, which allows the researcher to easily visualize and manipulate the experimental data. OTUbase has a number of functions that allow OTUbase to import data from the commonly used Mothur package and the RDP classifier. It also includes functions that enable the user to easily use existing R packages to perform complex analyses and visualizations.

2.6 Funding

Bioinformatics facilities and some student support were provided by National Institutes of Health/National Center for Research Resources [P20RR16448, P20RR016454]

2.7 References

[1] P. Hugenholtz *et al.*, “Exploring prokaryotic diversity in the genomic era,” *Genome Biol*, vol. 3, no. 2, pp. 1–0003, 2002.

[2] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy," *Applied and environmental microbiology*, vol. 73, no. 16, pp. 5261–5267, 2007.

[3] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, *et al.*, "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and environmental microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[4] V. Kunin, A. Engelbrektson, H. Ochman, and P. Hugenholtz, "Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates," *Environmental microbiology*, vol. 12, no. 1, pp. 118–123, 2010.

[5] J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. Kulam-Syed-Mohideen, D. McGarrell, T. Marsh, G. M. Garrity, *et al.*, "The ribosomal database project: improved alignments and new tools for rrna analysis," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D141–D145, 2009.

[6] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, *et al.*, "Qiime allows analysis of high-throughput community sequencing data," *Nature methods*, vol. 7, no. 5, pp. 335–336, 2010.

[7] A. Giongo, D. B. Crabb, A. G. Davis-Richardson, D. Chauliac, J. M. Mobberley, K. A. Gano, N. Mukherjee, G. Casella, L. F. Roesch, B. Walts, *et al.*, "Pangea: pipeline for analysis of next generation amplicons," *The ISME journal*, vol. 4, no. 7, pp. 852–861, 2010.

[8] J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner, *vegan: Community Ecology Package*, 2013. R package version 2.0-10.

[9] M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl, "Microbial diversity in the deep sea and the underexplored "rare biosphere"," *Proceedings of the National Academy of Sciences*, vol. 103, no. 32, pp. 12115–12120, 2006.

CHAPTER 3

SEED: A MICROBIAL COMMUNITY VISUALIZATION TOOL

3.1 Notes

Chapter 2 describes Seed, an R package that provides a visual user interface for exploring ecological datasets. The goal of Seed is to remove coding and parameter adjustment requirements from the production of plots and figures. This allows researchers to focus on data exploration and hypothesis generation. Seed is open source and available at https://github.com/danlbek/Seed.

An application note describing Seed is currently being prepared for publication.

3.2 Abstract

In this paper we present Seed, a data exploration tool for microbial communities. Seed is written in R using the Shiny library. This provides users access to powerful R based functions and libraries through a simple user interface. Seed allows users to explore ecological datasets using principal coordinate analyses, scatter plots, bar plots, hierarchal clustering, and heatmaps. Seed is open source and available at https://github.com/danlbek/Seed.

3.3 Introduction

The proliferation of microbial community profiling is allowing researchers to study microbial communities in new ways. Increasingly, researchers in diverse fields are asking questions relating to how microbial communities vary across samples. For example, researchers studying the human microbiome are interested in how microbial composition changes across body sites [1] and through time [2]. Researchers studying disease look at how microbial communities differ between samples from healthy and unhealthy

individuals [3]. The answers to these questions are often explored using high throughput sequencing technology that allows researchers to identify the microbial composition of a large number of samples. This produces a wealth of data about microbial composition in many different environments and conditions.

In conjunction with advances in sequencing resources, researchers have developed a number of powerful software tools to analyze and visualize this wealth of data. Tools such as mothur [4] and Qiime [5] aggregate many tools to allow researchers to quickly and efficiently process large sequencing datasets. These programs are in a format that allows them to be easily incorporated into analysis pipelines that can be run with minimal user interaction. Many of these programs, notably mothur and Qiime, include many programs that perform, among others, the following tasks:

1. Filter and trim raw sequencing reads by quality characteristics
2. Separate large datasets by projects or samples
3. Classify reads into OTUs (using clustering and/or taxonomic classification)
4. Perform statistical analyses and generate visualizations

The output of these programs generally includes a table listing the abundance of each microbial taxon in every sample. This table is then used to calculate richness and diversity metrics, perform PCA/PCoA analyses, and visualize differences among samples.

The focus of these currently available packages is the efficient processing of large datasets, but they are not designed for open-ended data exploration. They excel at performing robust, computationally intensive calculations that attempt to minimize the effects of noise and sequencing artifacts on downstream analyses. They often use a non-visual interface for analysis, even when they provide a GUI for their own functions, requiring the user to know specific command and parameter combinations. While this setup is ideal for pipeline development, it is often a hindrance for data-exploration. There is a need for a tool that allows researchers to quickly and easily visualize and explore the data that results from these pipelines.

In this paper, we present Seed (Simple Exploration of Ecological Data), a software package that focuses on data exploration and visualization of microbial community data derived from next generation sequencing.

3.4 Seed Description

Seed is an open source application that allows researchers to visually explore microbial community data. It is designed to allow many different analyses and visualizations including principal component and coordinate analysis (PCA/PCoA), hierarchal clustering, scatter plots, bar plots, and heatmaps.

Seed is written in the R programming language [6] using RStudio's Shiny library [7]. R is open source and available for Linux, MacOS, and Windows operating systems. It is also one of the most functionally complete tools for analyzing ecological datasets. The use of R allows us to take advantage of the wealth of R packages available for complex analyses and visualizations. Notably, Seed uses functions from R packages including vegan [8], Heatplus [9], gplots [10], and WGCNA [11, 12].

The use of Shiny allows Seed to be a web-based application, which may be installed locally or hosted on a remote server. When running Seed from a central server, users can access it through a web browser and are not required to install it locally. This means non-expert users can quickly and easily begin using Seed, even without local installations of R. Additionally, updates to R, Shiny, Seed, and underlying packages can be done seamlessly and invisibly to the end user. The use of a web browser also provides a familiar interface to most users, allowing them to quickly and easily learn to use Seed. Figure 3.1 is an example of the user interface.

Currently Seed requires two types of data, microbial abundance data and sample metadata. The microbial abundance data contain counts or abundances of each microbial taxon in each sample. The sample metadata contain information about each sample. Both files must be plain text files. Seed can accept comma, tab, or semicolon-separated values. Seed automatically associates the microbe abundance data with the sample information

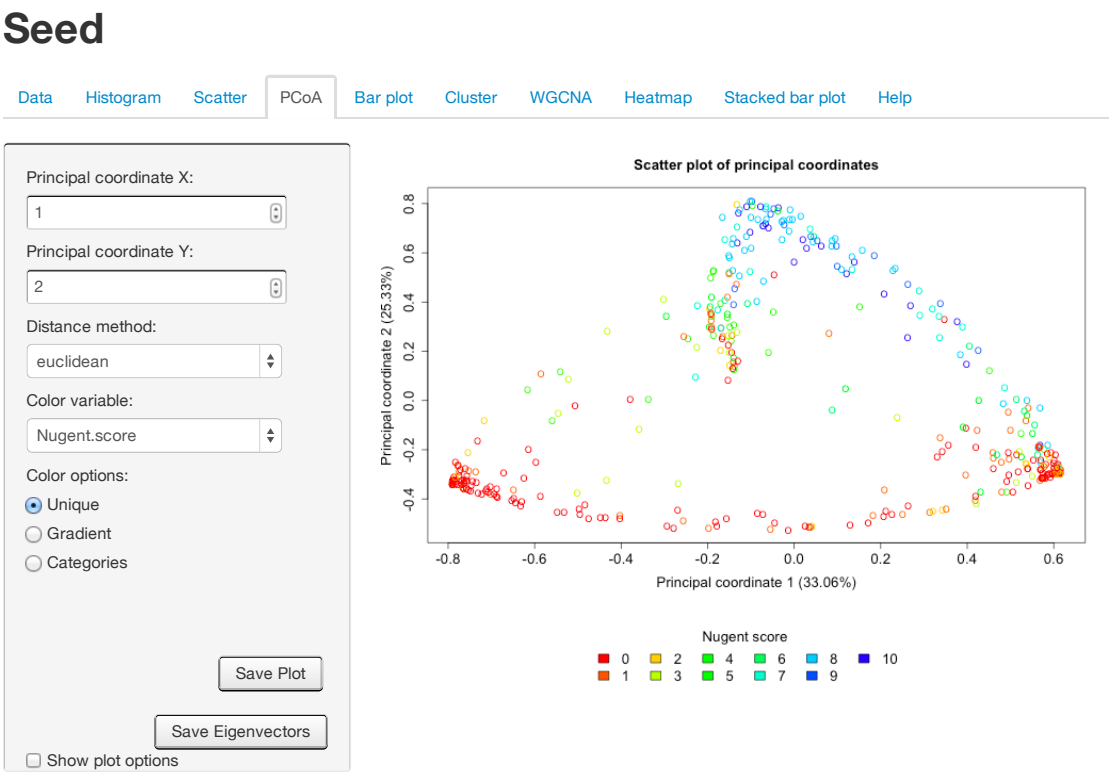

Figure 3.1: This figure shows Seed's user interface. The plot shown here is based on data originally published by Ravel *et al.* [13].

using the sample name. The user is also shown a subset of the imported data for verification purposes. After the abundance data are loaded, they may be modified using a number of common transformations including presence/absence, relative abundance, and Hellinger transformations. Seed is not limited to microbial data, though that was our primary research domain. It can be used to explore any data that include both feature counts and values for response variables.

After the user imports and verifies their dataset, they may easily explore their data with several plots. These plot options can be seen in Figure 3.1 and include histograms, scatter plots, PCA/PCoA plots, bar plots, cluster dendrograms, heatmaps, and stacked bar plots. All plots may be saved in either PDF or PNG formats. Examples of some of the plots generated by Seed are shown in Figure 3.2. Many of the plots include options to incorporate sample information by coloring points or bars according to metadata values. This allows users to easily visualize the relationship between the sample metadata and the structure of the microbial communities present in the samples.

The design of Seed emphasizes simplicity over exhaustive inclusion of parameters. In many or most cases, researchers will use Seed to understand general trends in the data, which may then inform more specialized analyses. Seed is designed to quickly explore ecological datasets and to act as a hypothesis-generating or brainstorming tool. Publication quality figures and polished analyses are beyond the current scope of this project. Additionally, large dataset analysis may be too slow for a comfortable user experience. The upper limit for dataset size will depend on a number of factors including CPU speed, memory size, and user patience. In general, computationally intensive analyses using large datasets are not ideal for interactive user interfaces. These types of analyses will likely require different tools. As with any software package, not all analyses have been implemented in Seed. We encourage users to also consider other visualization tools including phyloseq [14] for analyses incorporating phylogenetic relationships and EMPeror [15] for PCoA analyses of very large datasets. Additionally, Seed does not provide guidance on which tools are appropriate for any given analyses, that still relies on

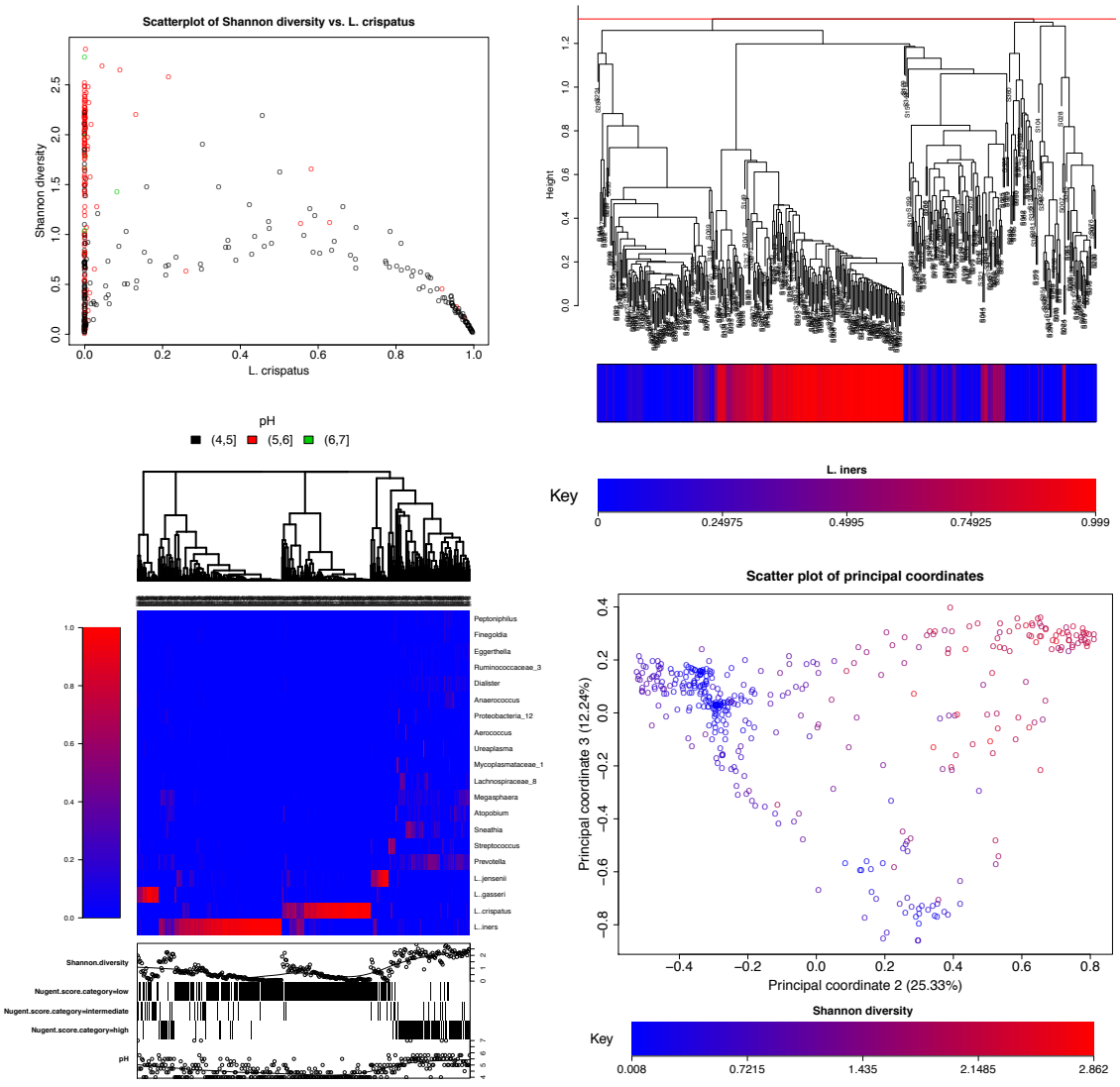

Figure 3.2: A sample of the plots generated using Seed. Clockwise from the upper left, a scatterplot, a cluster dendrogram, a PCoA plot, and a heatmap are shown here. The plots are based on data originally published by Ravel *et al.* [13].

user expertise.

3.5 Future Directions

Seed is freely available at https://github.com/danlbek/Seed. Development of Seed is ongoing. We are continuing to add new visualizations and to improve existing ones. Future development will focus on adding phylogenetic and taxonomic data structures, which will allow for analyses that take microbial relationships into account. We welcome user contributions to the project and encourage labs to copy and modify the code to suit their own needs.

3.6 Acknowledgements

We would like to thank Larry Forney, Roxana Hickey, Janet Williams, and other users for helpful conversations, recommendations and bug reports, and for the datasets used for the figures herein. We also thank Christopher Dennis who helped with the programming and development.

3.7 Funding

Funding for this project was provided by the NIH INBRE award P20GM016454 and by the NSF STC award DBI0939454. Computational support provided by NIH COBRE award P20GM16448.

3.8 References

[1] H. M. P. Consortium *et al.*, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[2] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, *et al.*, "Moving pictures of the human microbiome," *Genome Biol*, vol. 12, no. 5, p. R50, 2011.

[3] S. Srinivasan and D. N. Fredricks, "The human vaginal bacterial biota and bacterial vaginosis," *Interdisciplinary perspectives on infectious diseases*, vol. 2008, 2009.

[4] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, *et al.*, "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and environmental microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[5] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, *et al.*, "Qiime allows analysis of high-throughput community sequencing data," *Nature methods*, vol. 7, no. 5, pp. 335–336, 2010.

[6] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[7] RStudio and Inc., *shiny: Web Application Framework for R*, 2013. R package version 0.8.0.

[8] J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner, *vegan: Community Ecology Package*, 2013. R package version 2.0-10.

[9] A. Ploner, *Heatplus: Heatmaps with row and/or column covariates and colored clusters*, 2012. R package version 2.6.0.

[10] G. R. Warnes, B. Bolker, L. Bonebakker, R. Gentleman, W. H. A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz, and B. Venables, *gplots: Various R programming tools for plotting data*, 2013. R package version 2.12.1.

[11] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[12] P. Langfelder and S. Horvath, "Fast r functions for robust correlations and hierarchical clustering," *Journal of statistical software*, vol. 46, no. 11, 2012.

[13] J. Ravel, P. Gajer, Z. Abdo, G. M. Schneider, S. S. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, *et al.*, "Vaginal microbiome of reproductive-age women," *Proceedings of the National Academy of Sciences*, vol. 108, no. Supplement 1, pp. 4680–4687, 2011.

[14] P. J. McMurdie and S. Holmes, “phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data,” *PloS one*, vol. 8, no. 4, p. e61217, 2013.

[15] Y. Vázquez-Baeza, M. Pirrung, A. Gonzalez, and R. Knight, “Emperor: a tool for visualizing high-throughput microbial community data,” *Structure*, vol. 585, p. 20, 2013.

CHAPTER 4
MACHINE LEARNING TECHNIQUES ACCURATELY CLASSIFY MICROBIAL COMMUNITIES BY BACTERIAL VAGINOSIS CHARACTERISTICS

4.1 Notes

Chapter 3 describes using machine learning techniques to generate BV classification models. We compare models generated using genetic programming, random forests, and logistic regression. This study had two primary goals. The first goal was to determine if these methods could generate accurate models classifying microbial communities by BV characteristics. The second goal was to determine which microbial community features contributed to the high classification accuracy. This study has been published in PLoS ONE and is available under the terms of the Creative Commons Attribution License. The published manuscript is reprinted here. Citation information is shown below.

Daniel Beck and James A. Foster (2014) Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. PLoS ONE 9(2): e87830. PMC3912131

4.2 Abstract

Microbial communities are important to human health. Bacterial vaginosis (BV) is a disease associated with the vagina microbiome. While the causes of BV are unknown, the microbial community in the vagina appears to play a role. We use three different machine-learning techniques to classify microbial communities into BV categories. These three techniques include genetic programming (GP), random forests (RF), and logistic regression (LR). We evaluate the classification accuracy of each of these techniques on two different datasets. We then deconstruct the classification models to identify important

features of the microbial community. We found that the classification models produced by the machine learning techniques obtained accuracies above 90% for Nugent score BV and above 80% for Amsel criteria BV. While the classification models identify largely different sets of important features, the shared features often agree with past research.

4.3 Introduction

4.3.1 Microbial communities and disease

Microbial communities play critical roles in human health and disease. For example, gut microbial communities have been linked to obesity [1, 2], lung communities to pulmonary infections [3], and vaginal communities to bacterial vaginosis [4, 5, 6]. The complexity of these communities, however, makes determining specific causes of disease difficult.

In many natural environments, next-generation 16S rRNA sequencing unveils hundreds to thousands of microbe types. Everything from the physiology to the ecological roles of most of these microbes remains unknown. These microbes are difficult to study, both due to their large numbers and our inability to culture many of them in the lab [7]. The composition of these communities may fluctuate widely with environmental factors or as a result of microbial interactions.

4.3.2 Vaginal microbiome and bacterial vaginosis

The vagina microbiome is complex, with microbial composition varying between women and over time. This variation may be caused by immune factors, environmental variables, or dynamic microbial interactions. In some women the microbial community includes hundreds of microbe types, while in other women, the microbial community is dominated by a single species, often in the *Lactobacillus* genus [8, 9, 10]. Across women, the communities appear to cluster into distinct community types.

Bacterial vaginosis (BV) is a common condition, affecting up to 29% of all women [11]. BV is associated with increased risk for some STDs and preterm birth. Researchers have defined BV in two common ways. In clinical settings, Amsel criteria are often used.

Amsel criteria include the presence of discharge, a positive whiff test, the presence of clue cells, and a pH greater than 4.5. Amsel criteria BV is defined by the presence of at least three of these criteria [12]. Nugent score is a second way to define BV. The Nugent score relies primarily on counting gram-positive cells with morphologies similar to some *Lactobacillus sp.* (large rods) [13]. Nugent scores range from 0 to 10, with BV defined as a score greater than or equal to 7. The two definitions for BV lead to some interesting results. Using Nugent score BV definitions, up to 30% of all BV diagnoses are "asymptomatic", meaning that the woman in question has no symptoms though her microbiome elicits a high Nugent score, perhaps because her "normal" microbiome happens to contain more species with large rods than most other women. The significance of this phenomenon is uncertain.

It is difficult to identify a single cause of BV, even though the microbial community and BV are correlated. The number of microbe types found within the vagina microbiome is very large and the number of possible interactions between these microbes is even larger. In addition, noise in the data may obscure relationships between the microbial community and BV. Different bacterial consortia may also provide very similar functionality.

4.3.3 Machine learning and models

These difficulties are analogous to a problem faced by genetic epistasis researchers, where there are so many possible genetic interactions that may be linked to disease that it is difficult to determine the few that really matter. In this study, we applied three machine learning algorithms that have successfully discovered genetic interactions associated with disease to uncover possible microbial interactions associated with BV. In particular, we build models of BV diagnosis in the form of classifiers that were discovered with genetic programming (GP) [14, 15], random forests (RF) [16, 17], and logistic regression (LR) [17].

Genetic programming uses computational analogs of evolutionary processes to search for highly fit models. In our case, these models are decision trees where the leaves are

features that may be relevant to diagnosing BV, and where internal nodes are functions that operate on data passed on from their dependent nodes. GP transforms a population of candidate models by combining substructures from multiple "parent" models, modifying individual models randomly, and retaining only those models that are better at classifying BV from our input datasets for the next iteration. When the algorithm is stopped, the best model in the final population tends to be a very good predictor of BV. To determine which microbial populations or patient behaviors were most closely associated with BV, we analyzed which features were in the best GP classifiers and how they were used.

GP is very flexible and allows nearly unlimited model complexity. However, it searches for models stochastically and does not exhaustively search all possible models. In addition, the models produced by GP can be very large, and are often difficult to interpret. Also, computation costs tend to be high.

Random forests is an ensemble technique that builds a population of tree classifiers, where the final classification of a given set of features is its most frequent classification by the team members. RF is computationally efficient but may not be as flexible as GP. It is easier to extract important model features from RF models than from GP models, but not as easy as with logistic regression.

Logistic regression fits a linear model to the data, producing a linear combination of features and regression coefficients whose value for a given set of microbial communities and patient behaviors (in our case) quantifies the likelihood that the patient had BV. There are many ways to build the LR model. We use a maximum likelihood method implemented in the R package glmnet [18], in which the final model was parameterized in such as way as to maximize the probability that this set of features was associated with BV. Features were selected for inclusion in the model by glmnet using the lasso [19]. It was then straightforward to determine which features were most useful in BV diagnosis: the magnitudes of the regression coefficients indicate the weight given to the corresponding feature.

LR is computationally very efficient. And the fitted model is easy to interpret.

However, the structure of the final model is dependent on how terms are added to the regression equation, and LR may not be appropriate for non-linear phenomena. LR models are the easiest to interpret of the three in this study.

4.3.4 BV diagnosis as a classification problem

In this paper we apply these machine learning methods to classifying microbial communities into BV+ and BV- categories. We show that the methods accurately classify women by BV status based on their vagina microbiome and associated environmental factors. Additionally, we identify the parts of the microbial community that seem to play important roles in determining BV status.

We are interested in two aspects of the classification models, classification accuracy and feature usage. The accuracy of the models is a measure of how well they partition samples into diseased and non-diseased categories. We measure accuracy as the percentage of correctly classified samples. Different machine learning algorithms have different ways of selecting and weighting features, so our analysis of feature usage was algorithm specific.

4.4 Results

Before generating classification models, we first collapsed many of the microbes into groups based on correlations. We did this to both reduce the number of factors and to increase the interpretability of the classification models. The groups of correlated microbes are shown in Figure 4.1. We used two different datasets to train and evaluate the models, one from Srinivasan *et al.* [9] and one from Ravel *et al.* [8]. The two datasets produced different correlated microbe groups. There is some similarity in the groups, for example CG1 in the Srinivasan *et al.* dataset shares many microbes with CG4 in the Ravel *et al.* dataset.

After obtaining classification models using GP, LR, and RF, we evaluated the accuracy of the models with receiver operator curves (ROCs). ROCs show the performance of the

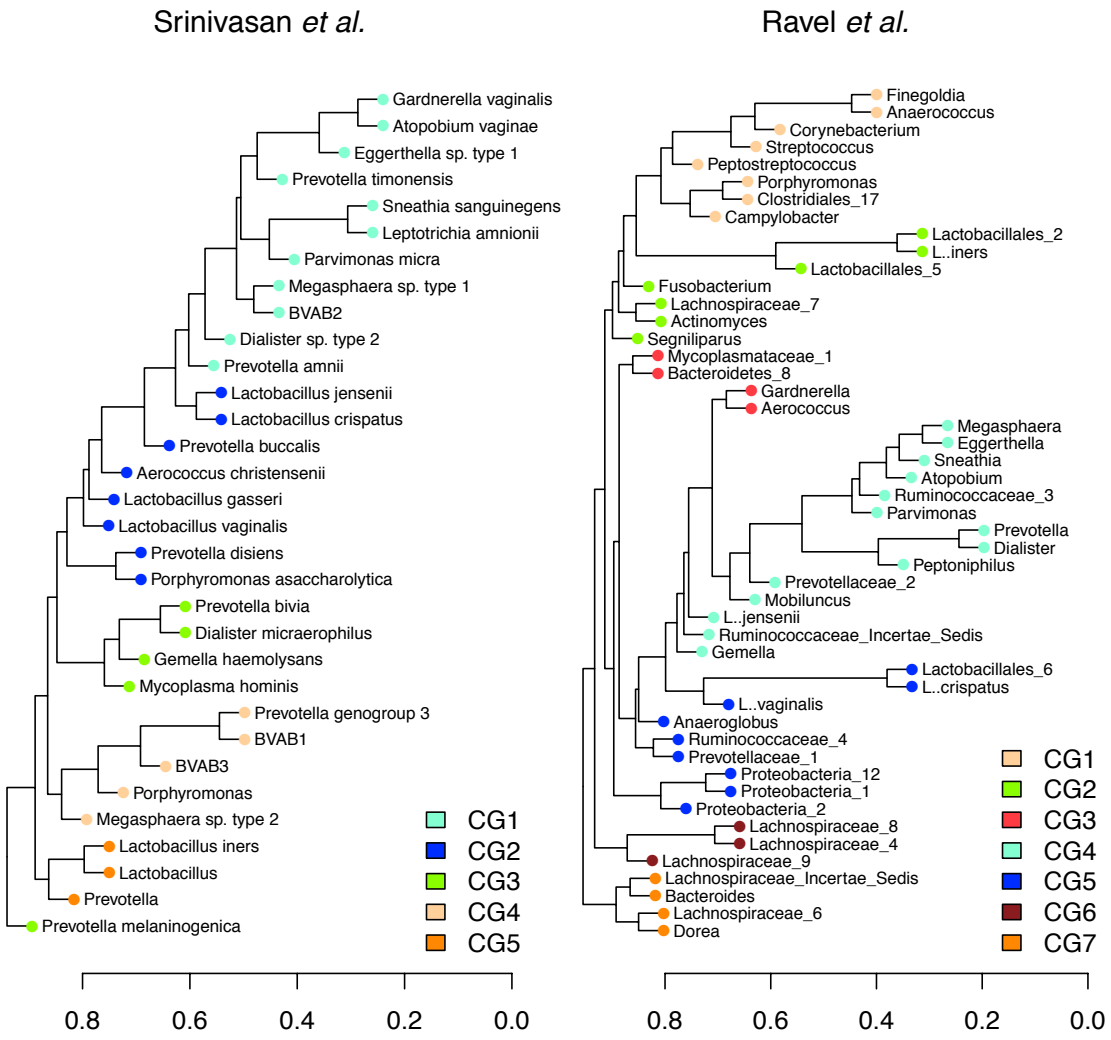

Figure 4.1: This figure shows the correlated microbe groups. We converted the sparCC correlations between microbial taxa to distances by subtracting the absolute value of the correlation from one. We then clustered the taxa and defined correlated groups using a dynamic tree-pruning algorithm (from the R library dynamicTreeCut). Microbial taxa not falling into these groups are not shown.

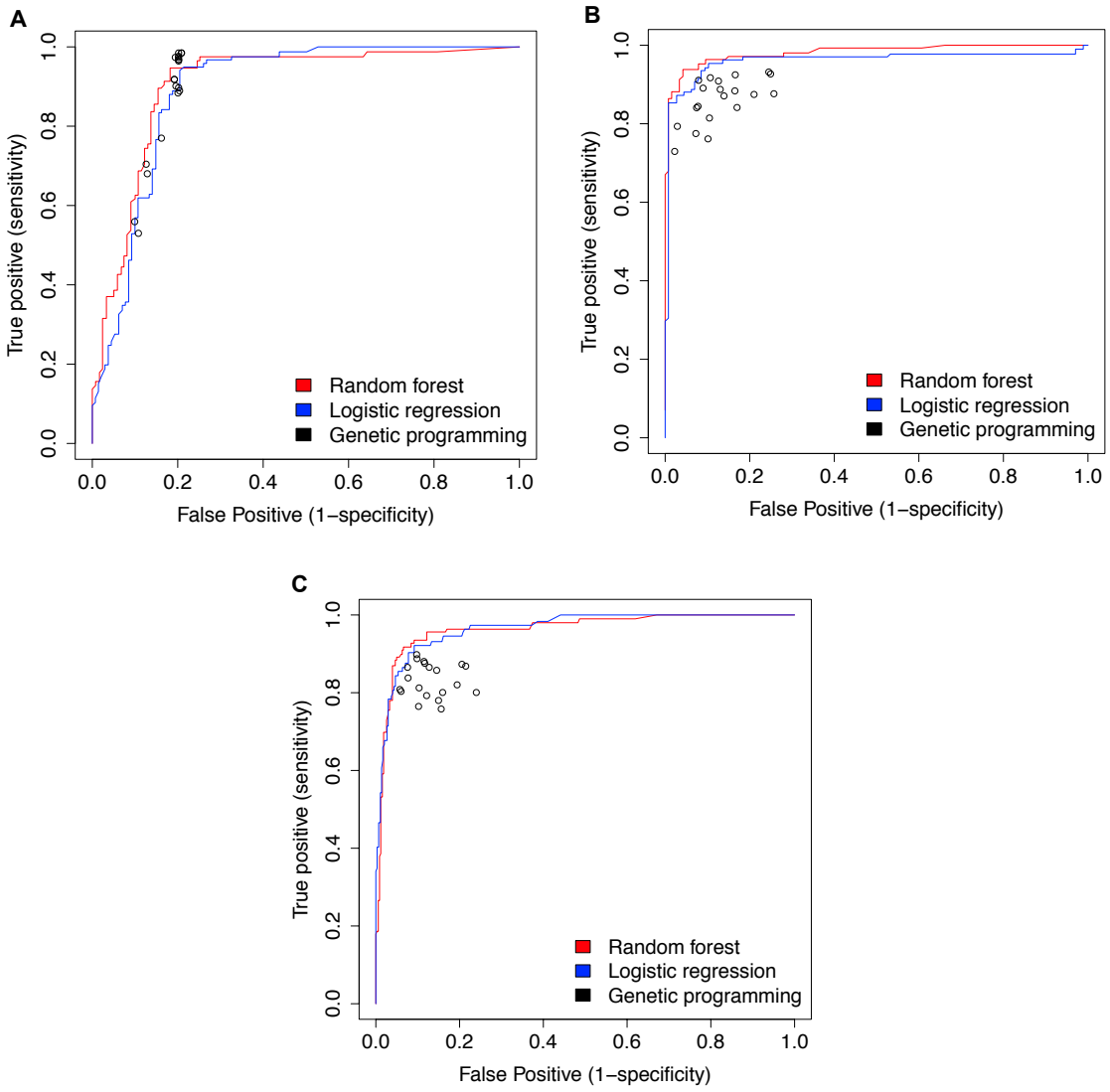

Figure 4.2: This figures shows the accuracy of different classifiers at classifying microbial communities into BV categories. The red and blue lines show the accuracy of random forest and logistic regression classifiers respectively. The black dots are different genetic programming models. Panel A shows the results using the Srinivasan *et al.* dataset and Amsel BV. Panel B uses the Srinivasan *et al.* dataset and Nugent score BV. Panel C uses the Ravel *et al.* dataset and Nugent score BV.

model at classifying both BV+ and BV- samples. This allows us to simultaneously compare the type 1 and type 2 errors for each model. Figure 4.2 shows the ROCs for each of the analyses. A perfect model would have a curve that forms a right angle in the upper left of the ROC. More accurate models have a true positive rate closer to 1 and a false positive rate closer to 0.

As can be seen in Figure 4.2, both LR and RF tend to outperform GP. However, the accuracy of all the machine-learning techniques was remarkably similar. RF and LR models obtained accuracies consistently between 90% and 95% when classifying on Nugent score BV. GP models often classified samples with similar accuracies, but high variation between GP models reduced the average GP accuracy. The models perform slightly worse when classifying on Amsel criteria BV. However, all three techniques obtained accuracies above 80%.

After determining the accuracy of the models, we deconstructed the models to determine which features were most useful. We ranked the features by their apparent importance to each model. The top fifteen features for each classification technique are shown in Table 4.1. There is little overlap between the important factors used by the classification models. For the Srinivasan *et al.* dataset, when classifying by Amsel criteria BV, the Nugent score is the only important feature shared by all classification techniques. Four other features are shared between GP and either RF or LR. The results are similar when classifying by Nugent score BV. For the Srinivasan *et al.* dataset, CG2 and the whiff test results are identified by all three techniques. For the Ravel *et al.* dataset, all techniques identify CG4.

The different classification techniques varied widely in computational time. LR and RF were relatively quick, usually completing in less than an hour on a single laptop. GP, on the other hand, took several hours longer.

Table 4.1: This table shows the fifteen most important features identified by the different classifiers. Features common to all three techniques are labeled 'a'. Features common to two techniques are labeled 'b'. The features are listed in order of importance.

Srinivasan *et al.* - Amsel			**Srinivasan *et al.* - Nugent score**			**Ravel *et al.* - Nugent score**		
Genetic Programming	**Random Forests**	**Logistic Regression**	**Genetic Programming**	**Random Forests**	**Logistic Regression**	**Genetic Programming**	**Random Forests**	**Logistic Regression**
nugent[a]	nugent[a]	Prevotella genogroup 7	pH[b]	CG2[a]	clue[b]	CG4[a]	CG4[a]	CG1[b]
Moryella indoligenes	CG1	Streptococcus agalactiae[b]	CG2[a]	CG1[b]	CG1[b]	pH[b]	pH[b]	Lactobacillus 2
Mycoplasma[b]	Uncorrelated microbes	Peptoniphilus lacrimalis	whiff[a]	CG5	race	GpI	CG3	Sutterella
Fusobacterium	CG3	Bifido-bacteriaceae	Sutterella wadsworthensis	vag_fluid	whiffa	Corio-bacteriaceae 3	Ureaplasma	CG6
Sutterella wadsworthensis	CG4	Raoultella planticola	Neisseria gonorrhoeae	pH[b]	Finegoldia magna[b]	Peptostrepto-coccaceae Incertae Sedis	CG5	Total number reads[b]
Bacteroides Porphyromonasb	Anaerococcus prevotii tetradius	Mycoplasma[b]	Bacteroides	whiff[a]	CG2[a]	Flexi-bacteraceae 5	Total number reads[b]	Bulleidia
CG5[b]	CG2	nugent[a]	Haemophilus	clue[b]	Streptococcus anginosus	Moryella	CG2[b]	Proteobacteria 3
Streptococcus agalactiae[b]	CG5[b]	Peptoniphilus harei	Bacteroides xylanisolvens	CG4	Peptostrepto-coccus	Megamonas	CG1[b]	Bilophila
Veillonella montpellierensis	Peptostrepto-coccus	Dialister propionicifaciens	Eubacteriaceae Lachnospiraceae	Streptococcus agalactiae[b]	Streptococcus agalactiae[b]	Entero-bacteriaceae 2	Ethnic Group	CG2[b]
Candidatus Peptoniphilus massiliensis	Arcano-bacterium phocae	Fuso-bacteriaceae	Arcano-bacterium phocae	CG3	Raoultella planticola	Chryseo-bacterium	Uncorrelated microbes	Lactobacillales 1[b]

Continued...

Srinivasan *et al.* - Amsel			Srinivasan *et al.* - Nugent score			Ravel *et al.* - Nugent score		
Genetic Programming	**Random Forests**	**Logistic Regression**	**Genetic Programming**	**Random Forests**	**Logistic Regression**	**Genetic Programming**	**Random Forests**	**Logistic Regression**
Clostridiales	race	Megasphaera micronuciformis	Eubacteriaceae Ruminococcaceae	Finegoldia magna[b]	Streptococcus parasanguinis	Patulibacter	Lactobacillus gasseri	CG4[a]
Bifidobacterium breve	Coriobacteriaceae	Porphyromonas sp. type 1	Mobiluncus curtisii	Uncorrelated microbes	Megasphaera	Haemophilus	Clostridiales 15	Salmonella
Haemophilus	Actinomyces	Haemophilus pittmaniae	Campylobacter ureolyticus	Lactobacillus coleohominis	Prevotella genogroup 4	Clostridia 2	Community group	Dermabacter
Streptococcus salivarius thermophilus	Bacteroides Porphyromonas[b]	Neisseria gonorrhoeae	Delftia tsuruhatensis	Anaerococcus vaginalis	Streptococcus salivarius thermophilus	Rothia	Lactobacillales 1[b]	Flexibacteraceae 2
Fusobacterium periodonticum	Finegoldia magna	Asticcacaulis excentricus	Ruminococcaceae	Streptococcus mitis oralis	Pseudomonadaceae	Bacillus c	Staphylococcus	Exiguobacterium

4.5 Discussion

This study demonstrates the feasibility of using classification models to identify important microbial community features related to BV. However, the results of this study also show many complications that must be taken into account when designing future studies.

First, we can look at the results of the classification techniques within a single dataset. Classifier accuracy is similar between the three techniques. The accuracy obtained by each classification method is high, often exceeding 80% regardless of the dataset or classification technique. The strength of the classification accuracy indicates the presence of some signal of BV in the dataset. A better than random classification accuracy indicates the presence of some feature in the dataset that is associated with BV.

The GP results show wider variation between models when the classification phenotype is Nugent score BV. This variation is not seen when classifying based on Amsel criteria BV. While the cause of this variation is unclear, there are a number of possible explanations. GP can theoretically explore a much larger set of possible models than RF and LR. This wider exploration, in combination with a large stochastic component, may increase the variation in the GP model accuracy. Additionally, the specific GP implementation we use may not efficiently avoid local optima. Further optimization of GP methods may increase overall accuracy and decrease its variation between models.

The high accuracy of the classification techniques indicates the presence of some association between some dataset features and BV. The top fifteen important features for each technique are shown in Table 4.1. These results are interesting for many reasons. The few features that overlap differ between the three analyses. In the Srinivasan *et al.* dataset, when classifying on Amsel criteria BV, Nugent score is the only important feature shared by all classification techniques. When classifying on Nugent score BV, the whiff test and CG2 were important to all three techniques. Similarly, the Ravel *et al.* dataset resulted in CG4 and pH selected by all three techniques. These factors have often been identified by previous studies as correlates with BV [4, 5]. BV defined by Nugent score overlaps with BV defined by Amsel criteria. This may explain the apparent importance of Nugent score

when classifying by Amsel criteria BV. Similarly, the presence of Amsel criteria such as vaginal discharge and odor, clue cells, and pH when classifying by Nugent score BV likely reflects the overlap between Amsel criteria BV and Nugent score BV. Ravel *et al.* identified a group of microbes that overlaps substantially with CG4, which all three techniques identified as important. This group includes *Megasphaera*, *Eggerthella*, *Sneathia*, *Prevotella*, and *Dialister*, among others.

While the important features identified by all classification techniques seem to agree with previous research, there are many features that are shared by only two techniques, or are unique to a single technique. In fact, the majority of the first fifteen important features are unique to a single classification technique. This lack of overlap has many possible explanations. Using the first fifteen most important features identified by each technique is an arbitrary choice. Our analysis doesn't determine how each feature affects the overall classification accuracy. Additionally, the use of one feature may change the relative importance of the remaining features. This may amplify small differences in the classification techniques, resulting in very different sets of important features.

Our analysis highlights an important aspect of using classification models to detect parts of the microbial community that are associated with BV. The features included in the analysis are likely important to the outcome. This can be seen in the important features identified by the techniques. Amsel criteria are found to be important when classifying by Nugent score BV and Nugent score is identified as important when classifying by Amsel criteria BV. These findings may be unsurprising, as both the Amsel criteria and Nugent score attempt to diagnose BV. It may be more informative to remove these features from the dataset before applying the classification techniques.

While each technique obtained similar classification accuracy, technical characteristics of the techniques differentiate them in important ways. A key consideration for these techniques is the easy extraction of important parameters. This is a difficult problem for large and complex GP models. The approach we take in this study is to determine how varying the value of each feature independently affects the overall model accuracy.

Additionally, we count the number of replicate GP models that include the feature. We combined these two measures to produce an overall importance measure for each feature. However, it is unknown whether this is optimal. We may be missing important parts of the GP models. This problem is somewhat alleviated for RF and LR models. Extensions to this study may include using machine learning techniques designed for easy identification of important features. Computational time may also be important to some researchers. The RF and LR analyses were relatively quick, completing in less than an hour on a single computer. The GP analysis, however, took several hours longer.

While we applied these classification techniques to two different datasets, these results are not comparable for a variety of reasons. Similar considerations will often apply to comparisons of techniques for classification-based diagnostics using multiple datasets. First, the types of samples collected in the two studies differed. The Srinivasan *et al.* study included women with and without a BV diagnosis. The Ravel *et al.* study included only asymptomatic participants. While both studies use Roche's 454 FLX sequencing platform, they amplify different regions of the 16S rRNA sequence. Srinivasan *et al.* use the V3-V4 region while Ravel *et al.* use the V1-V2 region. Additionally, the studies use different methods for classifying reads into taxonomic groups (the RDP classifier [20] in the Ravel *et al.* study and pplacer [21] in the Srinivasan *et al.* study).

In our study, we analyzed the results for each study individually, using the same read identification used in the original study. This allowed us to compare our results with the previous ones. However, this approach has the consequence of making it difficult to compare the results for the two datasets. This difficulty is shown clearly in the identification of different correlated groups. The correlated groups often include different microbial taxa. An additional difficulty is the comparison of a species level identification in the Srinivasan *et al.* study with genus level identification in the Ravel *et al.* study.

In spite of these dataset differences, a few patterns in the results may motivate future work. The ROC plots in Figure 2 show accuracies for the Nugent score BV classifiers that are remarkably similar between datasets. It seems possible that this similarity reflects a

consistent property of the dataset. Application of these classification techniques on different microbial community phenotypes (such as obesity or pH) may determine if these patterns are significant.

In this study we have shown that GP, RF, and LR generate models that classify samples by Amsel criteria BV with accuracies above 80%. These same techniques classify samples by Nugent score BV with accuracies above 90%. This study demonstrates the feasibility of using classification models to identify populations in a microbial community that are associated with BV. Determining the effect size of the important features may extend these results. Additionally, applying these techniques to different datasets and classifying on a variety of microbial community characteristics will determine how well these methods work for samples that may be very different from the vagina microbiome.

4.6 Materials and Methods

4.6.1 Dataset details

We use two different datasets drawn from studies published by Ravel *et al.* in 2011 [8] and Srinivasan *et al.* in 2012 [9]. The Ravel *et al.* study sampled the microbiome of 396 asymptomatic women. The study amplified and sequenced the V1-V2 variable regions of the 16S rRNA gene using Roche's 454 FLX sequencer. Reads were classified at the genus level using the RDP classifier [20]. The reads identified as *Lactobacillus* were further classified to the species level using a hidden Markov model based algorithm. The study identified a total of 282 microbial taxa across all samples. Out of 396 samples, 97 were BV+ using a Nugent score definition.

The Srinivasan *et al.* study sampled the microbiome of 220 women, 97 of whom were BV+ using Amsel criteria BV. Similarly, using Nugent score BV, 117 women were BV+. The study amplified and sequenced the V3-V4 variable regions of the 16S rRNA gene using Roche's 454 FLX sequencer. Reads were classified at the species or genus level using pplacer [21]. The study identified a total of 155 unique microbial taxa.

4.6.2 Classifier details

We implemented a GP classifier in C++. Table 4.2 shows many of the parameters used by the genetic program. We used tournament selection with the worst model in the tournament group replaced by the child of the best. The GP created the child by either mutating the best model or crossing over the best model with the second best model in the tournament group. Due to high variability in the results of GP, we repeated the analysis ten times. The model with the highest training fitness was then selected for evaluation with the testing dataset.

Table 4.2: This table lists the parameter values used by the GP classifier.

Parameter	**Value**
Population size	15000
Tournament group size	4
Cross-over probability	0.2
Total generations	300
Mutation probability	1
Available node functions	addition, subtraction, protected division, multiplication, if/then/else, sine, cosine, logical AND, logical OR, maximum, minimum, log

The fitness of each model was calculated using two steps. The first step calculated a cutoff value for the classifier results. To calculate this value the classifier results were averaged for BV+ and BV- training cases separately. The cutoff value is the average of these two numbers. Values from the model that fell on or above this cutoff were considered BV+ classifications and values below this cutoff were considered BV- classifications. In order to generate the ROC plots shown in Figure 2, the fitness value of the BV+ training cases was multiplied by a constant that varied between 0 and 20. This constant allowed us to vary the value of classifying BV+ vs. BV- samples. In the second step, the total number of incorrectly classified samples was added to the size of the

classification model multiplied by a small constant. This constant penalized larger models. The fitness was then minimized over the course of the program.

In order to identify features important to the GP models, we varied the values for each feature individually in every sample. We then determined whether varying the feature value changed the classification of the sample. This resulted in two summary values for each feature; the number of GP models in which varying the feature resulted in a different classification for at least one sample, and the number of samples in each model which changed classification due to changing the value of the feature. We rescaled these summary values to between 0 and 1 and added them together in order to obtain a single value describing the importance of each feature.

In order to implement the random forest classifiers we used the R package randomForest [22]. We used the *randomForest* function with default parameters to generate the classification model. To determine feature importance, we ranked the features by the increase in node purity. This is a measure of how much each feature increases the separation of the samples into BV+ and BV- categories for each classification tree. The increase in node purity was then averaged over all trees in the forest to obtain the total importance of each feature to the classification model.

To build a logistic model with linear regression, we used a maximum likelihood method implemented in the R package glmnet [18]. We ran the analysis using default parameters with a binomial response type. To determine feature importance, we ranked the features by the magnitude of the mean coefficient across the cross validation replicates divided by the standard deviation.

4.6.3 Microbial correlation reduction

In order to reduce the number of parameters and to increase the interpretability of our results, we collapsed highly correlated microbes into groups. We calculated pairwise correlations on microbial relative abundances using sparCC [23]. We converted the correlations into dissimilarities by subtracting the magnitude of the correlation from one.

We then used average hierarchical clustering and a dynamic tree-cutting algorithm to break the microbes into correlation groups. To do this we used the function *cutreeDynamic* from the R package dynamicTreeCut [24] with a 0.9 cut height and a three taxa minimum group size. This cut height was chosen to account for nearly all of the correlation present between microbes (Figure 4.3). Further analysis of the correlated groups is shown in Figure 4.4, which shows the mean cluster silhouette widths for varying cut heights. Uncorrelated microbes were left as individuals. A single feature in the dataset represented each correlated microbe group.

4.6.4 Cross-validation and accuracy determination

In order to avoid model over-fitting, we used ten-fold cross validation [25]. Cross validation detects over fitting and indicates how well the model is expected to perform with new data. We randomly broke the data into ten different parts. We used nine of these parts to train the model and the remaining part to test the performance of the model. We repeated this nine other times, using each of the ten parts as the testing data. We then averaged the accuracy of the model in classifying the testing samples over each of the 10 datasets to obtain a measure for the accuracy of each machine-learning technique.

4.7 Acknowledgments

We would like to thank Larry Forney, Terence Soule, and Mark McGuire for helpful discussions and two anonymous reviewers for their suggestions and comments.

4.8 Funding

This publication was made possible by the INBRE Program, NIH Grant Nos. P20 RR016454 (National Center for Research Resources) and P20 GM103408 (National Institute of General Medical Sciences). Computational support provided by NIH COBRE award P20GM16448. This material is based in part upon work supported by the National Science Foundation under Cooperative Agreement No. DBI-0939454. Any opinions,

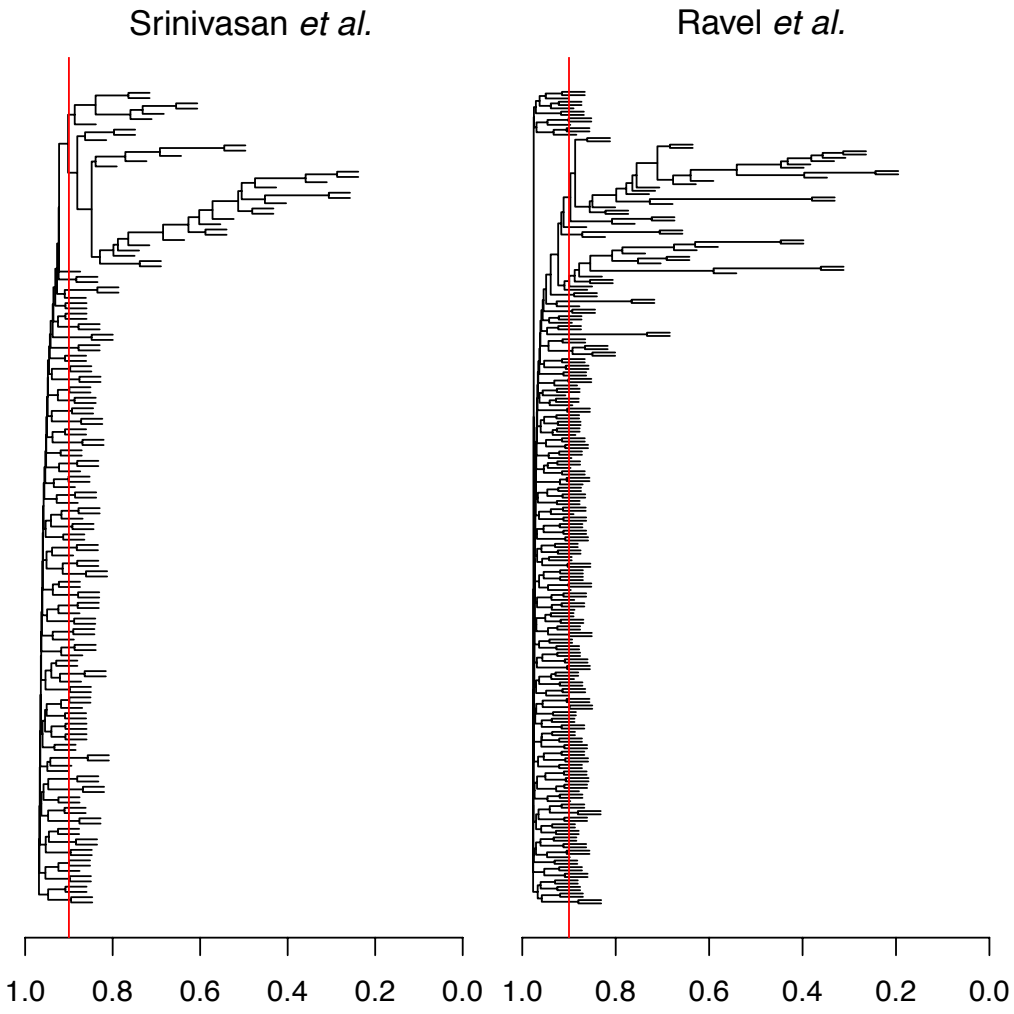

Figure 4.3: This figure shows the complete dendrogram resulting from average hierarchal clustering of microbial correlations. The vertical red line shows the 0.9 cutoff used to define correlated microbe groups. As can be seen, this cutoff accounts for most of the correlation between microbes.

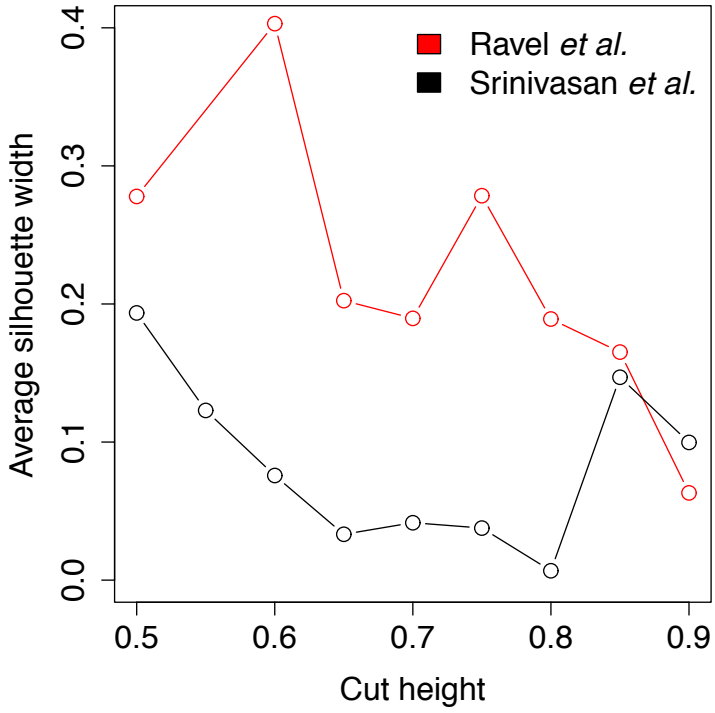

Figure 4.4: This figure shows the mean silhouette width for correlated microbe groups at varying cutoff levels. We converted the sparCC correlations between microbial taxa to distances by subtracting the absolute value of the correlation from one. We then clustered the taxa using average hierarchal clustering. For each cutoff level, we defined correlated groups using the cutreeDynamic tree-pruning algorithm.

findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

4.9 References

[1] R. E. Ley, P. J. Turnbaugh, S. Klein, and J. I. Gordon, "Microbial ecology: human gut microbes associated with obesity.," *Nature*, vol. 444, no. 7122, pp. 1022–1023, 2006.

[2] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon, "An obesity-associated gut microbiome with increased capacity for energy harvest," *Nature*, vol. 444, no. 7122, pp. 1027–131, 2006.

[3] C. D. Sibley, M. D. Parkins, H. R. Rabin, K. Duan, J. C. Norgaard, and M. G. Surette, "A polymicrobial perspective of pulmonary infections exposes an enigmatic pathogen in cystic fibrosis patients," *Proceedings of the National Academy of Sciences*, vol. 105, no. 39, pp. 15070–15075, 2008.

[4] Z. Ling, J. Kong, F. Liu, H. Zhu, X. Chen, Y. Wang, L. Li, K. E. Nelson, Y. Xia, and C. Xiang, "Molecular analysis of the diversity of vaginal microbiota associated with bacterial vaginosis," *BMC genomics*, vol. 11, no. 1, p. 488, 2010.

[5] S. Srinivasan and D. N. Fredricks, "The human vaginal bacterial biota and bacterial vaginosis," *Interdisciplinary perspectives on infectious diseases*, vol. 2008, 2009.

[6] R. M. Brotman, "Vaginal microbiome and sexually transmitted infections: an epidemiologic perspective," *The Journal of clinical investigation*, vol. 121, no. 12, p. 4610, 2011.

[7] R. I. Amann, W. Ludwig, and K.-H. Schleifer, "Phylogenetic identification and in situ detection of individual microbial cells without cultivation.," *Microbiological reviews*, vol. 59, no. 1, pp. 143–169, 1995.

[8] J. Ravel, P. Gajer, Z. Abdo, G. M. Schneider, S. S. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, *et al.*, "Vaginal microbiome of reproductive-age women," *Proceedings of the National Academy of Sciences*, vol. 108, no. Supplement 1, pp. 4680–4687, 2011.

[9] S. Srinivasan, N. G. Hoffman, M. T. Morgan, F. A. Matsen, T. L. Fiedler, R. W. Hall, F. J. Ross, C. O. McCoy, R. Bumgarner, J. M. Marrazzo, *et al.*, "Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria," *PloS one*, vol. 7, no. 6, p. e37818, 2012.

[10] X. Zhou, S. J. Bent, M. G. Schneider, C. C. Davis, M. R. Islam, and L. J. Forney, "Characterization of vaginal microbial communities in adult healthy women using cultivation-independent methods," *Microbiology*, vol. 150, no. 8, pp. 2565–2573, 2004.

[11] E. H. Koumans, M. Sternberg, C. Bruce, G. McQuillan, J. Kendrick, M. Sutton, and L. E. Markowitz, "The prevalence of bacterial vaginosis in the united states, 2001-2004; associations with symptoms, sexual behaviors, and reproductive health," *Sexually transmitted diseases*, vol. 34, no. 11, pp. 864–869, 2007.

[12] R. Amsel, P. A. Totten, C. A. Spiegel, K. Chen, D. Eschenbach, and K. K. Holmes, "Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations," *The American journal of medicine*, vol. 74, no. 1, pp. 14–22, 1983.

[13] R. P. Nugent, M. A. Krohn, and S. Hillier, "Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation.," *Journal of clinical microbiology*, vol. 29, no. 2, pp. 297–301, 1991.

[14] J. H. Moore, N. Barney, C.-T. Tsai, F.-T. Chiang, J. Gui, and B. C. White, "Symbolic modeling of epistasis," *Human heredity*, vol. 63, no. 2, pp. 120–133, 2007.

[15] A. E. Eiben and J. E. Smith, *Introduction to evolutionary computing*. springer, 2003.

[16] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1, pp. 278–282, IEEE, 1995.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[18] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[20] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy," *Applied and environmental microbiology*, vol. 73, no. 16, pp. 5261–5267, 2007.

[21] F. A. Matsen, R. B. Kodner, and E. V. Armbrust, "pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree," *BMC bioinformatics*, vol. 11, no. 1, p. 538, 2010.

[22] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[23] J. Friedman and E. J. Alm, "Inferring correlation networks from genomic survey data," *PLoS computational biology*, vol. 8, no. 9, p. e1002687, 2012.

[24] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r," *Bioinformatics*, vol. 24, no. 5, pp. 719–720, 2008.

[25] R. R. Picard and R. D. Cook, "Cross-validation of regression models," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 575–583, 1984.

CHAPTER 5
MACHINE LEARNING CLASSIFIERS PROVIDE INSIGHT INTO THE RELATIONSHIP BETWEEN MICROBIAL COMMUNITIES AND BACTERIAL VAGINOSIS

5.1 Notes

Chapter 4 describes using feature subsets to validate the important features of BV classification models generated using random forests and logistic regression. This study had two main goals. The first goal was to determine how many features were necessary for the models to obtain high classification accuracy. The second goal was to determine how much each feature contributed to the overall classification accuracy.

This study is currently being prepared for publication.

5.2 Abstract

Bacterial vaginosis (BV) is a disease associated with the vagina microbiome. It is highly prevalent and is characterized by symptoms including odor, discharge and irritation. No single microbe has been found to cause BV. In this paper we use random forests and logistic regression classifiers to model the relationship between the microbial community and BV. We use subsets of the microbial community features in order to determine which features are important to the classification models. We find that models generated using logistic regression and random forests perform nearly identically. Additionally, they identify largely similar important features. These results are in contrast to a previous study in which the important features identified by the classifiers were dissimilar. This difference appears to be the result of using different feature importance measures.

5.3 Introduction

Advances in sequencing technology allow researchers to study microbial communities in new ways. Researchers use 16S rRNA sequencing to identify the bacteria present in microbial communities. These studies have found highly complex communities composed of hundreds or thousands of different bacteria types. Some microbial communities are found in or on other organisms. Known as microbiomes, these communities have been shown to play important rolls in host health and disease. For example, in humans, gut microbiomes are important parts of digestion [1] and have been associated with obesity [2]. Microbial communities in the lungs may exacerbate cystic fibrosis [3].

The vagina microbiome is often composed of hundreds of different bacteria types, although only a few taxa may be at high abundance [4]. The composition of the vagina microbiome can be highly variable, both between women and through time [5]. Additionally the microbiome is associated with bacterial vaginosis (BV).

BV is a disease characterized by an overgrowth of certain microbe types in the vagina. It is highly prevalent, with estimates of affected women as high as almost 30% [6]. Symptoms of BV include odor, discharge, and irritation. It is also associated with increased rates of preterm birth [7] and increased susceptibility to some STDs [8]. While no single microbial cause of BV has been found, the microbial community as a whole is associated with BV [9].

Researchers often use two main BV diagnostics. The Nugent score is a measure based on cell morphology that can range from 0 to 10, with a score of 7 or greater indicating BV [10]. The Amsel criteria include a vaginal pH greater than 4.5, a positive whiff test, the presence of clue cells, and the presence of discharge. The presence of three of these four criteria indicates BV [11].

Determining which parts of the microbial community are associated with BV is difficult. This is partly due to the large number of taxa found in the community and the even larger number of potential interactions between taxa. Variation in the microbial community between women and over time adds to the difficulty of the problem.

Computational tools, however, may provide methods for studying these highly complex communities. Machine learning methods may allow us to model complex relationships in the microbial community related to BV.

Machine learning methods are able to generate complex models describing the relationship between the microbial community and BV. Every machine learning method has a different technique for generating a classification model. However, the end result for each method is a model that classifies samples into BV categories. Two model characteristics are interesting. First, the model accuracy describes how well the model fits the data. Second, the important features of the model are those features that the model uses to classify the samples. These features allow the researcher to generate hypotheses about the underlying biology.

Previous research has found that classification models generated using genetic programming, random forests, and logistic regression classify microbial communities into BV categories with between 80 and 90% accuracy [12]. This research has identified two challenges to using machine learning classifiers to study microbial communities. First, when the classification models are deconstructed to determine which features are important to the model accuracy, each machine learning technique results in different identified features. This makes it difficult to determine if the identified features are actually important, or if they are the result of technical artifacts. Additionally, it is difficult to distinguish between features that are critical to the accuracy of the classifier and features that are only marginally helpful. While an importance measure is calculated for each feature, this measure is often only effective in ranking features, rather than determining how much each feature adds to the overall accuracy.

In this study, we use subsets of the full feature set in order to address these problems. We add features sequentially to the classification models and observe how the accuracy changes. This allows us to determine how many features are necessary to obtain high classification accuracy. Additionally, we generate models using random feature subsets in order to obtain a feature importance measure that is consistent across machine learning

techniques. We find that random forests and logistic regression classifiers identify largely similar microbial community features. Nugent score BV appears to be closely associated with more features of the microbial community than Amsel BV.

5.4 Materials and Methods

We use datasets from studies published by Ravel *et al.* [4] and Srinivasan *et al.* [13]. The Srinivasan *et al.* dataset includes both Amsel BV and Nugent score BV, while the Ravel *et al.* dataset includes only Nugent score BV. The Ravel *et al.* dataset includes 396 asymptomatic women of whom 97 were BV+ using a Nugent score definition (Nugent score >=7). The Srinivasan *et al.* dataset includes 220 women, of whom 97 were BV+ using Amsel criteria and 117 were BV+ using Nugent score.

For this study, we processed the datasets using methods described in detail in our previous paper [12]. The key part of these methods is the reduction of the bacterial abundance data to correlated groups. Many bacteria in these datasets have correlated abundances. We used sparCC [14] correlations along with the R package dynamicTreeCut [15] in order to cluster the correlated taxa into groups. These groups were then reduced to single features in the dataset. The correlated groups are shown in Figure 5.1, which has been reprinted from [12].

We use two different machine learning algorithms to generate classification models, random forests (RF) and logistic regression (LR). The RF classifiers were implemented using the *randomForest* function in the R package randomForest [16]. We implemented LR classifiers using the *glmnet* function in the R package glmnet [17]. To identify important features of RF models, features were ranked according to their increase in node purity (INP). INP is a measure of how much each feature increases the classification accuracy of each decision tree, averaged across all trees in the ensemble. For LR, features were ranked by their mean coefficient magnitude in all cross-validation datasets divided by their standard deviation.

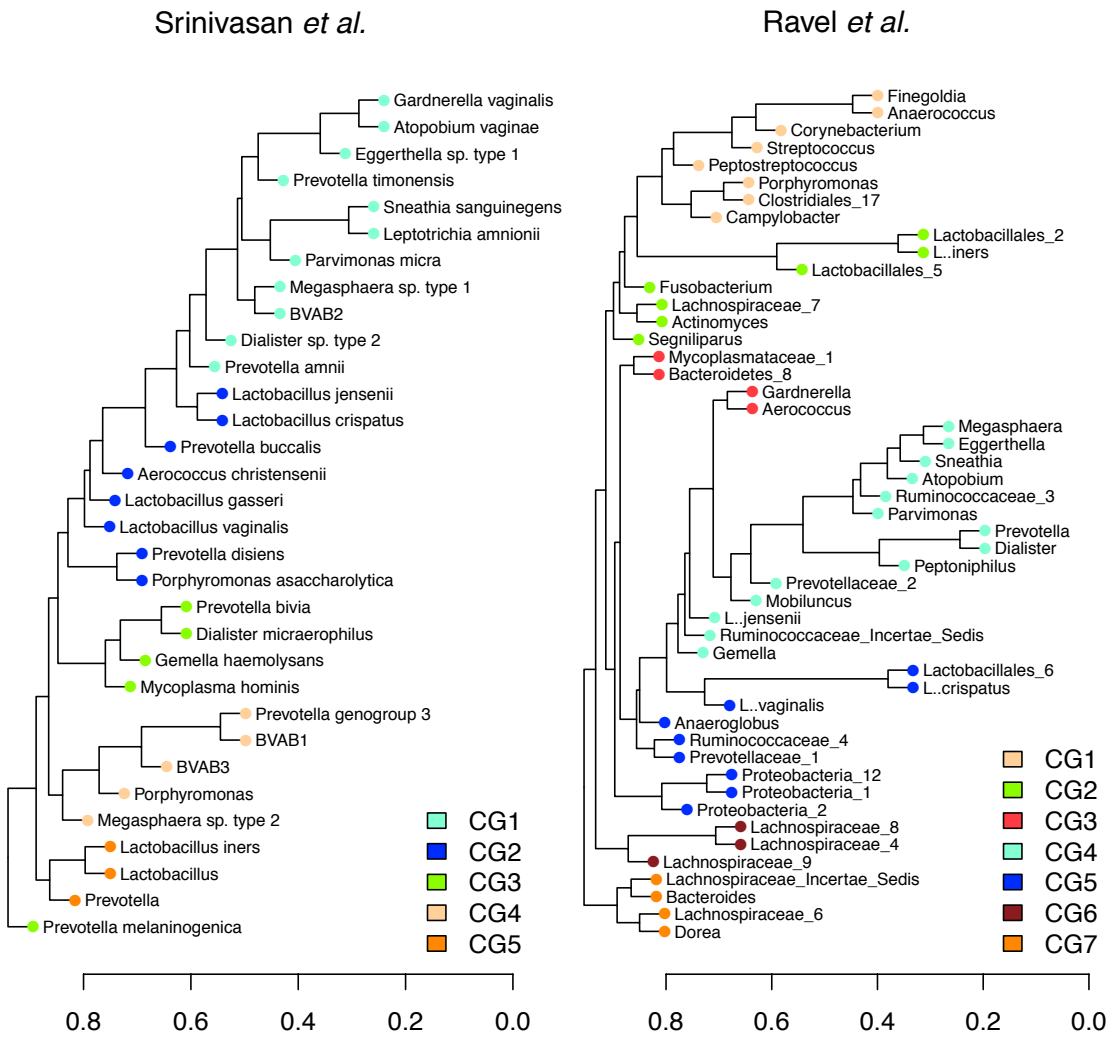

Figure 5.1: Reprinted from [12]. This figure shows the correlated microbe groups. Each of these groups is represented in the dataset as a single feature.

In addition to the RF and LR classifiers, we also calculated reliefF rankings and pairwise correlations for the features. ReliefF is a feature selection algorithm that estimates the relevance of each feature by how well it separates similar samples into classes [18]. To calculate the reliefF rankings, we used the *attrEval* function in the R package CORElearn [19]. The Pearson correlation between each feature and BV was calculated using R's *cor* function.

To prevent over fitting, we used ten fold cross validation. Each dataset was split randomly into ten parts. Nine of the parts were used to train the classification models. The remaining part was used to measure the model accuracy. This was repeated using each of the ten parts as the test dataset.

In the first step of the analysis, RF and LR models were generated using the full feature set of the training data. The importance of each feature to these models was then determined. ReliefF was used to generate a third feature ranking. These rankings were then used to select feature subsets in three different ways.

The first analysis selected the top N features from each of the feature rankings, where N ranged between 2 and 25. The second analysis used a five-feature sliding window across each of the rankings. The third analysis selected the top 25 features from each ranking and combined them into a single list. One thousand five-feature subsets were selected at random from this list of features. RF and LR classifiers were trained on each subset using the training data. The accuracy of each classifier was determined using the testing data.

The classification accuracy for each model was measured using the area under the receiver-operator curve (AUC). The receiver-operator curve (ROC) is a curve that describes the classifier accuracy in both positive and negative samples, thus representing both type 1 and type 2 error. The area under the ROC is often used as a summary of the model accuracy [20].

5.5 Results

Top N feature subsets help determine how accuracy improves with each feature addition. The features are added in order of perceived importance. If several features contribute additively and equally, a linear increase in accuracy would be expected. If only the top few features contribute substantially, the accuracy would reach its maximum quickly and then level off. More complex patterns may emerge if there are important interactions between features. Figure 5.2 shows the classification accuracy for RF and LR models as more features are added to the model. In every case, both RF and LR models classify samples with high accuracy after the inclusion of only a few features. Except for the LR feature rankings, high accuracy is obtained with five or fewer features.

Sliding window subsets may show patterns that the top N features miss. For example, the first two features may individually be sufficient to obtain a high accuracy. The first feature in the top N subsets masks the relevance of the second feature. A sliding window makes it possible to determine how the features affect classification accuracy without the influence of the more important features of higher rank. Each successive window replaces the highest ranked feature in the previous window with the next lowest ranked feature. Figure 5.3 shows the accuracy of RF and LR models using a five-feature subset of the data.

The sliding window subsets for Nugent score BV and the reliefF and RF rankings show substantial stability in classification accuracy as lower ranked features replace high ranked ones. This pattern is reduced in both the LR rankings and for all Amsel BV classifiers. Additionally, the LR rankings appear to decrease with less consistency than reliefF and RF rankings. This may reflect inconsistency in the LR rankings.

Random subsets extend the sliding window analysis by removing its dependency on the initial feature ranking. This allows us to determine how each feature affects the model accuracy when combined with four other features. The size of the random group was chosen based on the top N analysis results. The inclusion of five features was sufficient to produce models with accuracy as good as the full model. Table 5.1 shows the top fifteen features for RF and LR. The features identified using the mean accuracy across random

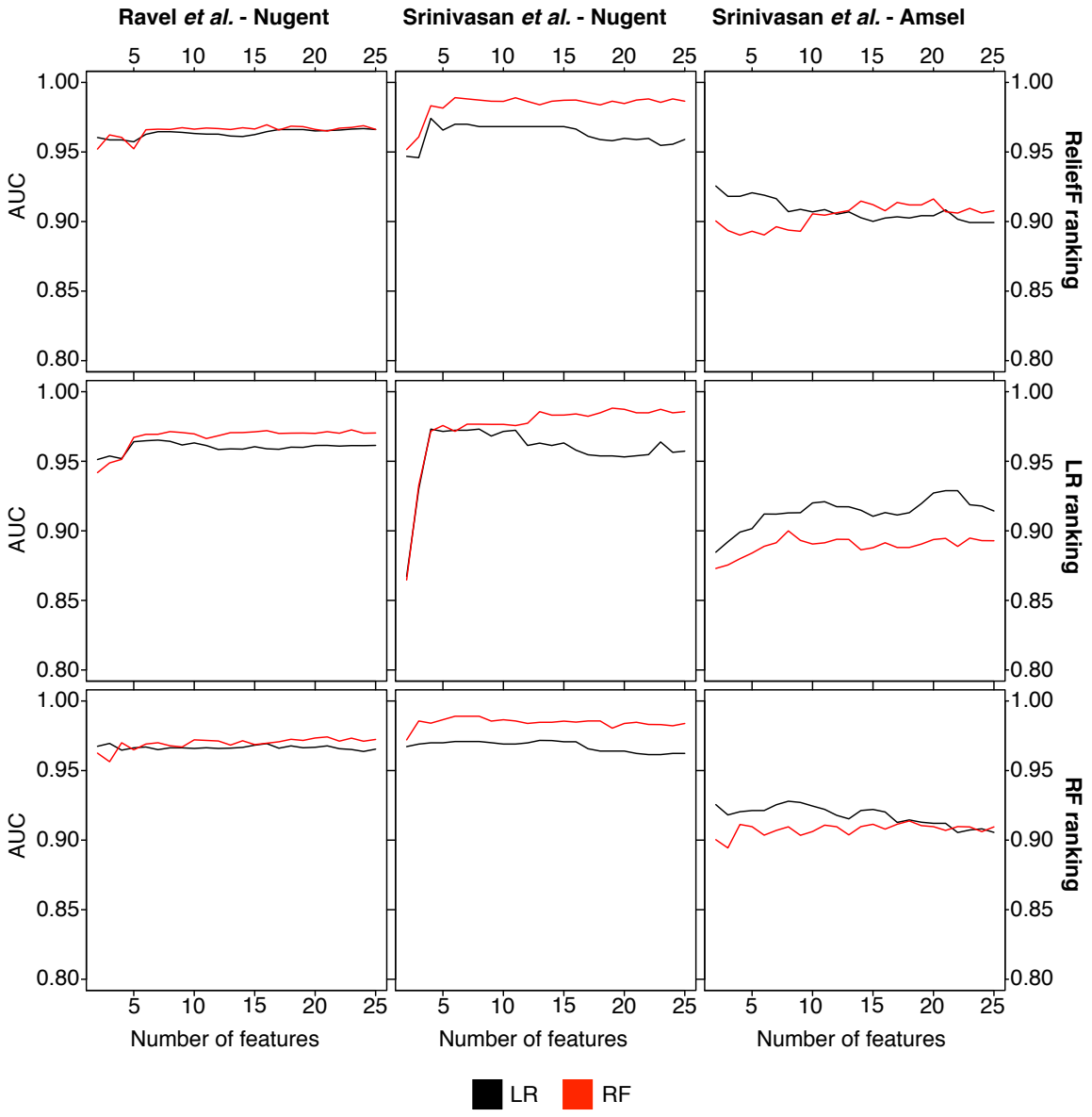

Figure 5.2: This figure shows how the classification models perform as the number of features available to the models increases. Features are added according to their performance ranking using reliefF (top row), logistic regression (middle row), and random forests (bottom row). The model performance is measured using the area under the ROC (AUC). As can be seen, the top five features are often sufficient to obtain high classification accuracy.

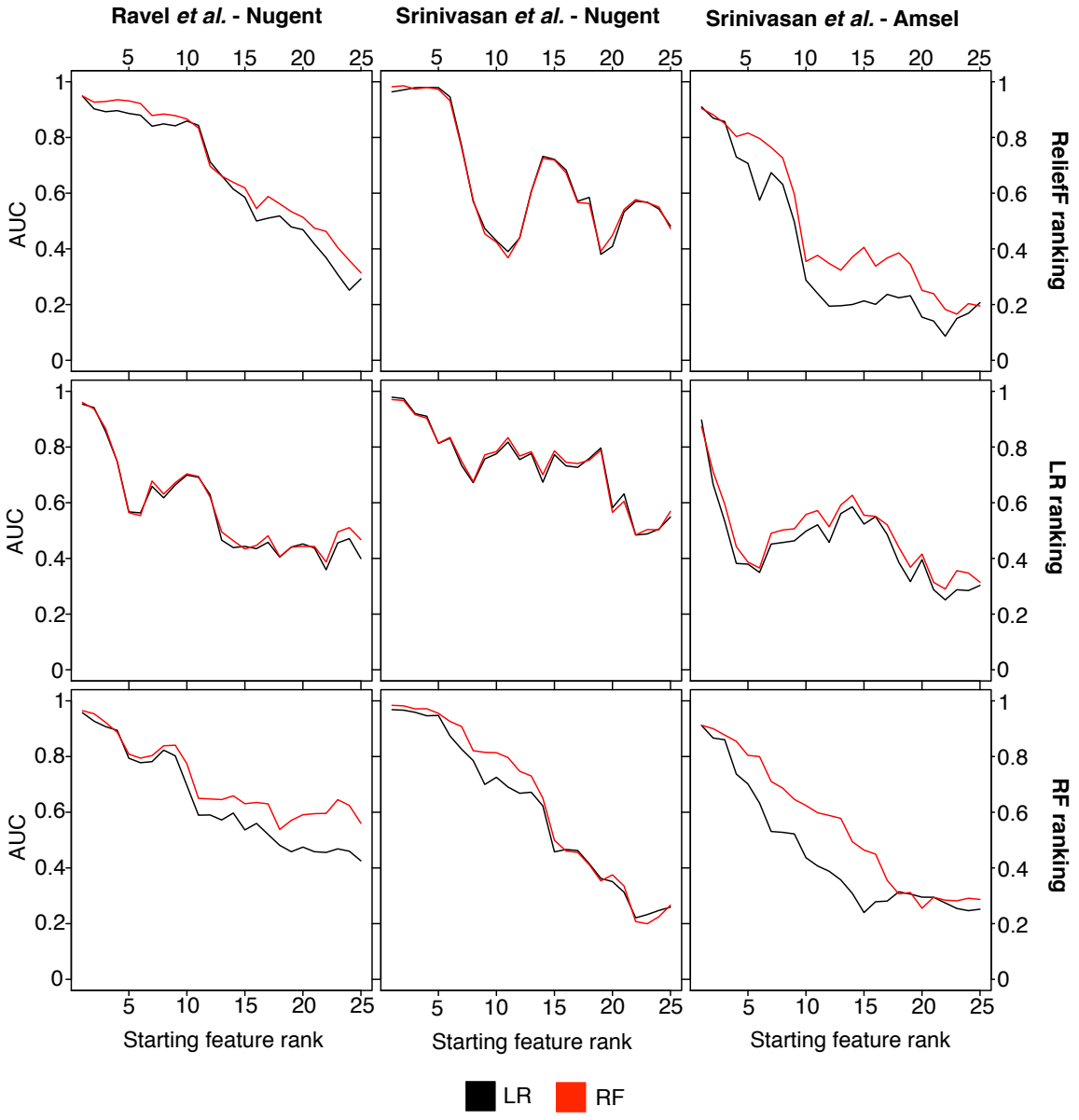

Figure 5.3: This figure shows the accuracy of models using a sliding window of five features chosen consecutively from the ranked feature lists. Features are added according to their performance ranking using reliefF (top row), logistic regression (middle row), and random forests (bottom row). The model performance is measured using the area under the ROC (AUC).

subsets appear to be very similar between classification methods.

Table 5.1: This table shows the top fifteen features for RF and LR classifiers. Features are ranked by their average accuracy in random five-feature subsets. There is substantial agreement in the features identified using RF and LR classifiers. The "CG" features are correlated groups of taxa. Taxa belonging to these groups may be seen in Figure 5.1

Ravel *et al.* Nugent		**Srinivasan *et al.* Nugent**		**Srinivasan *et al.* Amsel**	
RF	**LR**	**RF**	**LR**	**RF**	**LR**
CG4	CG4	CG1	CG1	nugent	CG1
CG3	CG3	CG2	CG2	CG1	nugent
CG5	CG5	pH	pH	CG2	CG2
pH	pH	clue	clue	CG3	CG5
CG2	CG2	vag_fluid	whiff	CG5	CG3
CG1	CG1	CG3	CG3	Uncorrelated taxa	CG4
Uncorrelated taxa	Total number reads	whiff	CG5	CG4	race
Anaerovorax	Clostridiales 1	CG5	vag_fluid	race	Anaerococcus prevotii tetradius
Community group	Anaerovorax	Uncorrelated taxa	Fusobacteriaceae	Peptostreptococcus	Candidate Division TM7 vaginal
Ethnic Group	Coriobacteriaceae 2	CG4	CG4	Anaerococcus prevotii tetradius	Peptoniphilus lacrimalis
Total number reads	Community group	race	race	Clostridiales	Bacteroides Porphyromonas
Coriobacteriaceae 2	Uncorrelated taxa	Fusobacteriaceae	Peptostreptococcus	Porphyromonas sp. type 1	Veillonella montpellierensis
L. gasseri	Lactobacillales 7	Peptostreptococcus	Anaerococcus prevotii tetradius	Fusobacterium nucleatum	Peptoniphilus harei
Clostridiales 1	L. gasseri	Anaerococcus prevotii tetradius	Atopobium minutum	Delftia	Fusobacteriaceae
Lactobacillales 7	Ethnic Group	Atopobium minutum	Bacteroides coagulans	Bacteroides Porphyromonas	Peptoniphilus

The features ranked highly by each importance measure appear to be very different. There is some agreement between the important measures, especially in the top five features. However, there are many features ranked highly by only a single measure. These results can be seen in Figure 5.4.

5.6 Discussion

The RF and LR classifiers identify very similar features using the random subset importance measure. This is in contrast to the previous results that found dissimilar rankings of important features [12]. Figure 5.3 may indicate problems with the LR importance measure. The sliding window subsets for reliefF and RF rankings generally show a consistent decrease in classification accuracy as the feature ranking decreases. The LR ranking, however, shows a more uneven decrease in accuracy with feature ranking. Similar patterns would be expected if the initial rankings were incorrect. While the reason for the poor performance of the LR rankings is unknown, it may be partially due to sensitivity of the importance measure to sparse data.

The important feature rankings appear highly dependent on the importance measure used. This can be seen in Figure 5.4. While there is some overlap in the top five to ten features identified by each importance measure, there are many features ranked highly by one importance measure and not others. It is not clear whether the ranking differences are due to noise or whether they may reflect some biological pattern.

A key advantage of using classification models to study microbial communities is the classification accuracy measure. The subset analysis allows for several interesting patterns and comparisons. It appears that the first few features are sufficient for classifying samples by BV. This can be seen in Figure 5.3. This is true for both Amsel BV and Nugent score BV.

Differences in Amsel BV and Nugent score BV are apparent from these results. The classification accuracy is higher for Nugent score BV, indicating a better model fit. This may result from a closer link between Nugent score BV and the microbial community. It

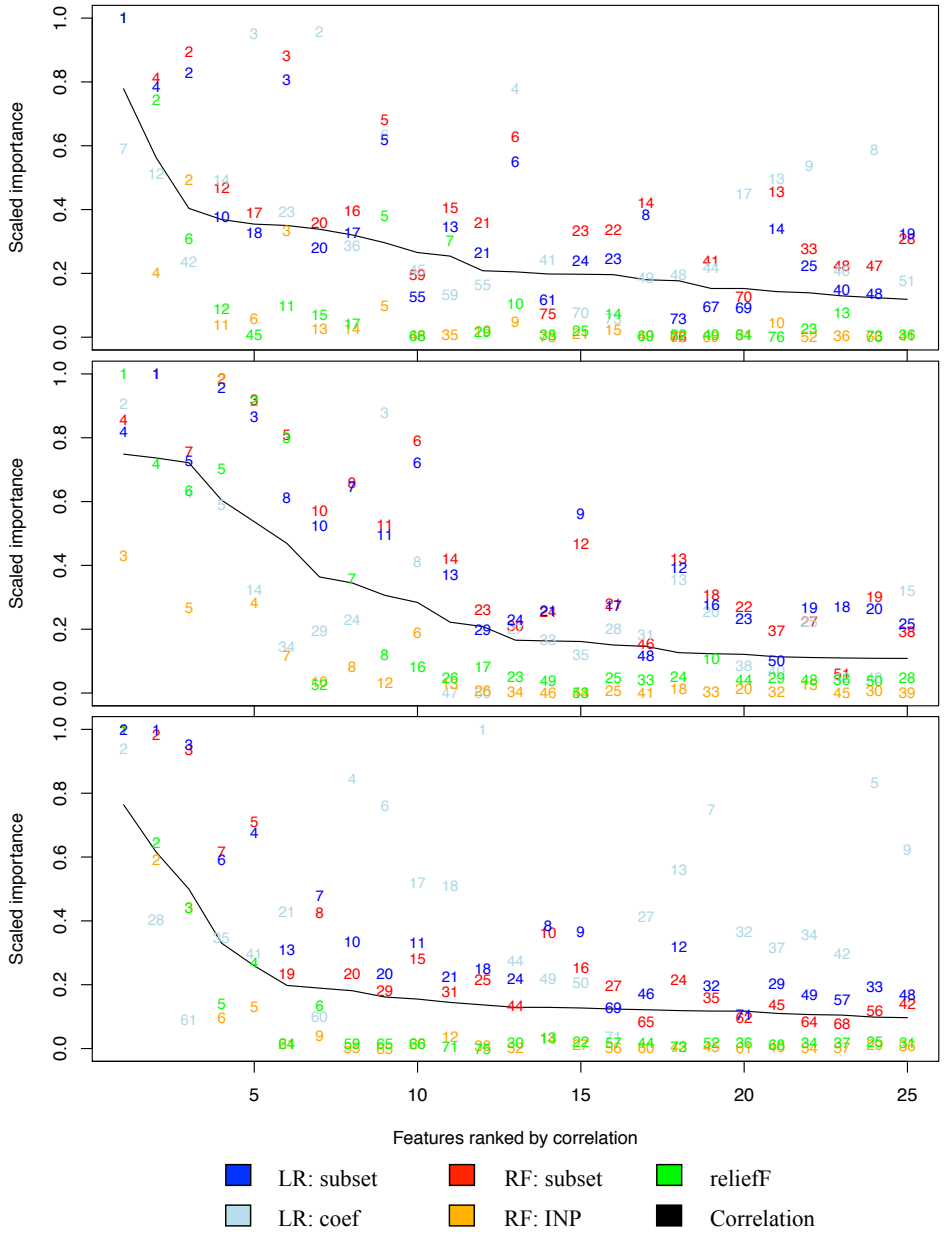

Figure 5.4: This figure compares the feature importance measures. The black line is the Pearson correlation between the feature and BV. Two importance measures are shown for LR; the mean classification accuracy of random five-feature subsets and the mean coefficient magnitude across validation datasets divided by the standard deviation. Two importance measures are also shown for RF; the mean classification accuracy of random five-feature subsets and the increase in node purity (INP). All measure have been scaled to between 0 and 1 for comparison purposes except for the Pearson correlations. The datasets from the top are Ravel *et al.* - Nugent BV, Srinivasan *et al.* - Nugent BV, and Srinivasan *et al.* - Amsel BV.

may also indicate that the relationship between Nugent score BV and the microbial community is more easily captured by the classification models. In other words, there may be a strong link between the microbial community and Amsel BV, but that link is complex and not fully exploited by the models. Alternatively, the Amsel BV classification may simply include more noise or error.

In addition to the overall accuracy magnitude, patterns in the accuracy decline shown in Figure 5.3 may reflect important differences in Amsel BV and Nugent score BV. In the case of Nugent score BV, it appears that the first few features are highly redundant. In many cases, the accuracy of the sliding window feature subsets only begins to decline substantially after five to ten top features have been excluded from the model. This is in contrast to the Amsel BV results. The decline in accuracy for the Amsel BV classifiers is nearly immediate, or only after the removal of a few high-ranking features. Many parts of the microbial community may be linked to Nugent score BV, while only a few are important indicators of Amsel BV.

The important features identified by the subset analysis are largely unsurprising. The correlated microbe groups that contribute substantially to the classification accuracy are composed of taxa that have been linked to BV in previous studies [21, 13, 22]. These taxa include *Gardnerella*, *Atopobium*, and *Eggerthella*. Correlated groups including various *Lactobacillus* species also rank highly. However, the grouping of many taxa into single features limits the interpretation of these results.

It is not clear if these classifiers find patterns that are any different from simple correlations. However, machine learning methods provide important accuracy measures that may help determine the number of features that are important. They may also indicate whether interaction terms are necessary to describe the system. Feature subset analysis illuminates many patterns and characteristics of the relationships between the microbial community and community characteristics such as BV. These methods may be generally useful for studying a wide range of microbial community related diseases and phenotypes.

5.7 Acknowledgments

We would like to thank Larry Forney, Terence Soule, and Mark McGuire for helpful discussions.

5.8 Funding

Funding for this project was provided by the NIH INBRE award P20GM016454 and by the NSF STC award DBI0939454. Computational support provided by NIH COBRE award P20GM16448.

5.9 References

[1] F. Bäckhed, R. E. Ley, J. L. Sonnenburg, D. A. Peterson, and J. I. Gordon, "Host-bacterial mutualism in the human intestine," *science*, vol. 307, no. 5717, pp. 1915–1920, 2005.

[2] P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, *et al.*, "A core gut microbiome in obese and lean twins," *Nature*, vol. 457, no. 7228, pp. 480–484, 2009.

[3] D. Willner, M. R. Haynes, M. Furlan, R. Schmieder, Y. W. Lim, P. B. Rainey, F. Rohwer, and D. Conrad, "Spatial distribution of microbial communities in the cystic fibrosis lung," *The ISME journal*, vol. 6, no. 2, pp. 471–474, 2012.

[4] J. Ravel, P. Gajer, Z. Abdo, G. M. Schneider, S. S. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, *et al.*, "Vaginal microbiome of reproductive-age women," *Proceedings of the National Academy of Sciences*, vol. 108, no. Supplement 1, pp. 4680–4687, 2011.

[5] P. Gajer, R. M. Brotman, G. Bai, J. Sakamoto, U. M. Schütte, X. Zhong, S. S. Koenig, L. Fu, Z. S. Ma, X. Zhou, *et al.*, "Temporal dynamics of the human vaginal microbiota," *Science translational medicine*, vol. 4, no. 132, pp. 132ra52–132ra52, 2012.

[6] E. H. Koumans, M. Sternberg, C. Bruce, G. McQuillan, J. Kendrick, M. Sutton, and L. E. Markowitz, "The prevalence of bacterial vaginosis in the united states, 2001-2004; associations with symptoms, sexual behaviors, and reproductive health," *Sexually transmitted diseases*, vol. 34, no. 11, pp. 864–869, 2007.

[7] S. L. Hillier, R. P. Nugent, D. A. Eschenbach, M. A. Krohn, R. S. Gibbs, D. H. Martin, M. F. Cotch, R. Edelman, J. G. Pastorek, A. V. Rao, *et al.*, "Association between bacterial vaginosis and preterm delivery of a low-birth-weight infant," *New England Journal of Medicine*, vol. 333, no. 26, pp. 1737–1742, 1995.

[8] H. C. Wiesenfeld, S. L. Hillier, M. A. Krohn, D. V. Landers, and R. L. Sweet, "Bacterial vaginosis is a strong predictor of neisseria gonorrhoeae and chlamydia trachomatis infection," *Clinical Infectious Diseases*, vol. 36, no. 5, pp. 663–668, 2003.

[9] B. B. Oakley, T. L. Fiedler, J. M. Marrazzo, and D. N. Fredricks, "Diversity of human vaginal bacterial communities and associations with clinically defined bacterial vaginosis," *Applied and environmental microbiology*, vol. 74, no. 15, pp. 4898–4909, 2008.

[10] R. P. Nugent, M. A. Krohn, and S. Hillier, "Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation.," *Journal of clinical microbiology*, vol. 29, no. 2, pp. 297–301, 1991.

[11] R. Amsel, P. A. Totten, C. A. Spiegel, K. Chen, D. Eschenbach, and K. K. Holmes, "Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations," *The American journal of medicine*, vol. 74, no. 1, pp. 14–22, 1983.

[12] D. Beck and J. A. Foster, "Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics," *PloS one*, vol. 9, no. 2, p. e87830, 2014.

[13] S. Srinivasan, N. G. Hoffman, M. T. Morgan, F. A. Matsen, T. L. Fiedler, R. W. Hall, F. J. Ross, C. O. McCoy, R. Bumgarner, J. M. Marrazzo, *et al.*, "Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria," *PloS one*, vol. 7, no. 6, p. e37818, 2012.

[14] J. Friedman and E. J. Alm, "Inferring correlation networks from genomic survey data," *PLoS computational biology*, vol. 8, no. 9, p. e1002687, 2012.

[15] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r," *Bioinformatics*, vol. 24, no. 5, pp. 719–720, 2008.

[16] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[17] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[18] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[19] M. Robnik-Sikonja and P. Savicky, *CORElearn: CORElearn - classification, regression, feature evaluation and ordinal evaluation*, 2013. R package version 0.9.42.

[20] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[21] D. N. Fredricks, T. L. Fiedler, and J. M. Marrazzo, "Molecular identification of bacteria associated with bacterial vaginosis," *New England Journal of Medicine*, vol. 353, no. 18, pp. 1899–1911, 2005.

[22] S. Srinivasan, M. T. Morgan, C. Liu, F. A. Matsen, N. G. Hoffman, T. L. Fiedler, K. J. Agnew, J. M. Marrazzo, and D. N. Fredricks, "More than meets the eye: Associations of vaginal bacteria with gram stain morphotypes using molecular phylogenetic analysis," *PloS one*, vol. 8, no. 10, p. e78633, 2013.

CHAPTER 6
CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Software Development

In the first two chapters of this dissertation, I presented software tools for analyzing microbial community datasets. These tools include OTUbase, which provides data structures and basic functions for analyzing OTU data in R, and Seed, which provides a simple visual interface for exploring microbial community data. Tools such as these are important for making complex analytical methods available to a broader audience. They increase research efficiency by eliminating the need for every lab to design and develop functionally similar software.

In general, software tools present many challenges over the course of their development and use. Previously unknown coding errors may become apparent and input file formats may change, requiring corresponding modifications to underlying functions. Similarly, new analysis techniques may reduce the usefulness of parts of the original tool, and require updates and modifications to maintain the software's relevance.

OTUbase has not been actively maintained. This has resulted in functions becoming obsolete, largely due to changing input file format and structure. Additionally, a lack of further development means key strengths of OTUbase are still unexploited. Improvements to OTUbase, such as expanding data input options, streamlining figure generation, and incorporating phylogenetic information remain unaccomplished.

OTUbase is also awkwardly positioned in the microbial community data processing pipeline. It requires either expert knowledge of R or external programs such as mothur to obtain relevant input files. After the data is in OTUbase, other R functions may be used to perform analyses and generate visualizations. This, however, requires that users are familiar with many other R packages in addition to OTUbase.

Seed appears somewhat better positioned than OTUbase. Seed uses a visual interface, which requires no programming skill or knowledge of R. The input files are simpler and

less likely to change in the near future. Overall, more users should be able to use Seed. However, it remains to be seen whether substantial numbers of users will find it helpful. Improvements, including user guidance on appropriate analyses or visualizations, may increase Seed's potential user base.

Future challenges facing Seed include maintenance and support. New analysis techniques and visualizations will be developed to study microbial communities. These techniques must be incorporated into Seed. Additionally, as microbial datasets increase in size, computational constraints may become important. Seed may require substantial increases in efficiency in order to continue to be an effective tool. As with any software, Seed will require an active group of developers in order to remain relevant.

6.2 Machine Learning Classification Models

The last two chapters of this dissertation demonstrate the potential of using machine learning classifiers to study microbial communities. Models generated using genetic programming, random forests, and logistic regression all classified samples by BV with high accuracy. Additionally, the subset analysis highlighted features of the microbial community that associate with BV.

There are many unanswered questions and future directions. In the case of BV and the vagina microbiome, while the classification models highlight potential associations, the specific relationship between the important features and BV remains to be determined. In addition, while the classification accuracy of the models was high, roughly 10% of the samples were misclassified. It is unclear whether more data or different classification models can account for these samples.

Un-sampled information about the microbial community may be important. This information may include fine-scaled microbial taxonomy, the presence or absence of specific microbial genes, or aspects of the woman's genetics. Additionally, the presence of bacteriophage or other viruses may be important parts of the microbiome. The incorporation of this information may help increase the classification accuracy.

The applicability of classification models to different microbial communities and phenotypes remains to be explored. These methods may be applied to other diseases such as inflammatory bowel disease or to other community characteristics such as pH. The performance of the classification models in these different situations may help determine the usefulness of machine learning approaches to microbial communities in general.

There are many limitations to using machine learning classifiers. In particular, complex models may capture intricate relationships in the microbial community; however, interpreting these models may be difficult. Mirroring the complexity of the interactions in the microbial community with computational models is only useful if the models can be used to generate hypotheses. The results shown in Chapter 4 seem to indicate that interactions between components of the microbial community are relatively unimportant to the classification accuracy. Alternatively, the interactions present in the community may be masked by the reduction of correlated taxa into groups. Small subsets of the total feature set result in classification accuracies as high as those using all features. If the classification models are capturing microbial interactions, the classification accuracy may not be maintained when only a few features are used. Simulated datasets may help determine how well the classifiers account for interactions and what the signatures of these interactions are in the subset analysis results.

Extensions to these methods may include using new data types and looking at microbial community characteristics more broadly. Datasets that include metagenomic information, or the abundances of viruses, fungi and other organisms, may help describe a more complete picture of the microbial community. Training classification models on several microbial community characteristics will yield multiple lists of important features. These important feature lists may show patterns reflecting the underlying microbial community structure.

With decreasing costs for genetic sequencing and computational processing, microbial community datasets are increasing in resolution and in scale. More samples can be collected and those samples can be analyzed at greater depth. Researchers now have the

opportunity to determine how microbial communities work, their structure and temporal dynamics. Ecological theories and models originally developed for macroscopic organisms can be tested and refined. The study of microbial communities may lead to greater understanding of evolutionary patterns. Using machine learning methods to generate classification models may contribute to the study of these complex ecosystems.

APPENDIX A

This appendix includes a license permitting the reprinting of the article "OTUbase: an R infrastructure package for operational taxonomic unit data". This article was originally published in the Journal Bioinformatics. It is included as Chapter 2 of this dissertation. The full citation for this article is shown below.

Daniel Beck, Matt Settles, and James A. Foster (2011) OTUbase: an R infrastructure package for operational taxonomic unit data. *Bioinformatics* 27(12): 1700-1701. PMC3106189

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Feb 17, 2014

This is a License Agreement between Daniel Beck ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3331650997838
License date	Feb 17, 2014
Licensed content publisher	Oxford University Press
Licensed content publication	Bioinformatics
Licensed content title	OTUbase: an R infrastructure package for operational taxonomic unit data:
Licensed content author	Daniel Beck, Matt Settles, James A. Foster
Licensed content date	06/15/2011
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Investigating the use of classification models to study microbial community associations with bacterial vaginosis.
Publisher of your work	n/a
Expected publication date	May 2014
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD
Terms and Conditions	

Figure 6.1: OTUbase License Agreement - Page 1

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

Figure 6.2: OTUbase License Agreement - Page 2

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:

v1.4

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK501228874.

Figure 6.3: OTUbase License Agreement - Page 3