5-20-2011

Assessing Microbial Diversity Through Nucleotide Variation

Ahmet Eren
University of New Orleans

Assessing Microbial Diversity Through Nucleotide Variation

A Dissertation

Submitted to the Graduate Faculty of the
University of New Orleans
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Engineering and Applied Science
Computer Science

by

Ahmet Murat Eren

B.S., Canakkale Onsekiz Mart Universitesi, 2002
M.S., Canakkale Onsekiz Mart Universitesi, 2005

May, 2011

To Gülşan Eren, my little sister whom I never had a chance to spend enough time with.

ACKNOWLEDGEMENTS

I would like to thank my advisor Christopher M. Taylor for his help, guidance and his open mindedness. When I decided to change my research focus to microbial ecology in the middle of my Ph.D. education, he accepted the offer to be my advisor, facilitated an energetic collaboration with Michael J. Ferris and let me be a part of it.

I owe my deepest gratitude to Michael J. Ferris. It has been almost 2 years since I started working under his supervision, and I can clearly see that meeting with Michael J. Ferris was one of the most important milestones of my life. Among many other reasons, I am very thankful to him for tolerating my excitement, patiently teaching me and introducing me to the world of microbes.

I would also like to thank the invaluable members of my Ph.D. committee, Christopher M. Summa, Daniel Bilar and Huimin Chen. I wish I had been wise enough to benefit from their vision just a little more.

I honestly do not think that I would be able to finish this Ph.D. if it were not for the support of Seth Pincus, director of the Research Institute of Children's at Children's Hospital, and Mahdi Abdelguerfi, chair of the Department of Computer Science, University of New Orleans.

I am grateful to my dear colleagues Marcela Zozaya-Hinchliffe, Caroline E. Hennigan and Johana Norori for their support. Thanks to all faculty and staff of Children's Hospital and the Department of Computer Science in University of New Orleans.

My appreciation also extends to James Nachtwey, Richard Feynman, Kurt Vonnegut and Oğuz Atay for all the inspiration and strong influence.

I would also like to express my deep gratitude to those whom I met, and who had an impact on the course of my life: Nefin Huvaj, Mihail Guber, Mevzun Yüksel, Necdet Yücel, Doruk Fişek, Koray Löker, Oğuzhan Dinçer, Alp Öztarhan, Barış Metin, S. Çağlar Onur, Gürer Özen, Erkan Tekman, Murat Gündüz, Kevin Simpson and Aslı Şahin. Thank you very much.

There are people that I cannot thank enough. Therefore I will not dare to try. However, I would like to mention their names here: B. Duygu Özpolat, Arpat Özgül and Ashley Wright.

Finally, I frankly admit that I am here, writing this page in this library carrel, simply because I was lucky.

Thanks to serendipity.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Microbes are the most abundant and most diverse form of life on Earth, constituting the largest portion of the total biomass of the entire planet. They are present in every niche in nature, including very extreme environments, and they govern biogeochemical transformations in ecosystems. The human body is home to a diverse assemblage of microbial species as well. In fact, the number of microbial cells in the gastrointestinal tract, oral cavity, skin, airway passages and urogenital system is approximately an order of magnitude greater than the number of cells that make up the human body itself, and changes in the composition and relative abundance of these microbial communities are highly associated with intestinal and respiratory disorders and diseases of the skin and mucus membranes. In the early 1990's, cultivation-independent methods, especially those based on PCR-amplification and sequences of phylogenetically informative 16S rRNA genes, made it possible to assess the composition of microbial species in natural environments, advances in high-throughput sequencing technologies in recent years have increased sequencing capacity and microbial detection by orders of magnitude. However, the effectiveness of current computational methods available to analyze the vast amounts of sequence data is poor and investigating the diversity within microbial communities remains challenging. In addition to offering an easy-to-use visualization and statistical analysis framework for microbial community analyses, the study described herein aims to present a biologically relevant computational approach for assessing microbial diversity at finer scales of microbial communities through nucleotide variation in 16S rRNA genes.

INTRODUCTION

Life is one of the most interesting phenomena in the observable universe –and as far as we know, the most abundant, diverse, and adaptable form of life in the universe is "microbial life".

Biology, the study of life, is itself living its golden age thanks to the recent breakthroughs in genetics and advancements in sequencing technologies. It has been only 40 years since Walter Fiers and his team sequenced a complete gene for the first time (Min Jou, Haegeman, Ysebaert, & Fiers, 1972), and today high-throughput sequencing platforms can sequence more than 400,000,000 nucleotides in a matter of hours.

It is impossible to make sense of this tremendous amount of sequence data without state-of-the-art computer technology and bioinformatics; but in order to approximate what nature has to say, computational methodologies should be more biology-aware. This dissertation presents a novel computational approach in order to assess diversity, which is being missed by *de facto* approaches, at finer scales of microbial communities through tracing subtle nucleotide variations.

Aside from the Introduction, Conclusion and Epilogue, there are three main chapters in this study.

Chapter I provides background information on the subjects of microbial life, microbial ecology, state-of-the-art sequencing technologies, phylogenetically informative Ribosomal

RNA genes, the impact of microbial communities on human health, and finally, commonly used techniques to analyze vast amounts of genetic sequence data. In this section, notable publications and suggested readings from the literature are cited.

Chapter II describes a computational framework, developed by the author, which empowers biologists to analyze metagenomic sequence data. This framework was developed to provide a powerful, easy to use, computational tool to researchers in the field of microbial ecology. It is essentially a visualization and statistical analysis environment, which is currently being used by numerous research projects and was supported by funding from the Research Institute for Children in New Orleans and a grant from the NIH 5RO1AI79071-2.

Chapter III presents a discussion regarding the novel contributions of this study: assessing microbial diversity through nucleotide variation within 16S rRNA gene tags.

Finally, the Epilogue chapter will propose an overview about the caveats that are created or confronted by current computational approaches and emphasize the significance of the study presented.

CHAPTER I

Microbial Beings

In addition to their critical role in evolution, microbes are an essential component of life on Earth.

Microbes are ubiquitous in nature. They dominate life on the planet, not only in terms of total number of individuals and total biomass, but also in terms of their incredible metabolic diversity (Whitman, Coleman, & Wiebe, 1998). They are the driving force behind global biogeochemical cycles; they govern global carbon flux and carbon fixation, nitrogen flux and nitrogen fixation. Microbes globally control sulfur, iron and other essential molecules for biological processes. In addition, their vigorous metabolic adaptations have allowed them to colonize even the most extreme environmental niches our planet has to offer (Rothschild & Mancinelli, 2001).

Microbial beings make up the vast majority of the species in all three domains of life on earth, which were first proposed by Woese (Woese & Fox, 1977; Woese, Kandler, & Wheelis, 1990), the Archea, Bacteria and Eukarya. In fact, two of the domains, the Archea and Bacteria are solely comprised of prokaryotic (non-nucleated) microbes. Although these domains are composed of simpler single-celled organisms, and in contrast to eukaryotes they are missing nucleus and membrane-bound organelles, the metabolic diversity of these two domains is far greater than the Eukarya, and include, for example, all photosynthetic, nitrogen fixing and methane producing organisms. Even though today there are different

theories about whether bacteria, archea or eukaryotes, or other now extinct cells, were the first life forms to evolve (Zimmer, 2009; Brown, 2003), archea and bacteria constitute the most diverse and ecologically interesting clade on the tree of life and they are the primary focus of this study.

Microbial Ecology

Microbial ecology, the relatively young and flourishing juncture of ecology and microbiology, is the study of microbes and their interactions with their environment.

Identifying and classifying microbes, and studying them in their natural environments, has long been an obstacle to progress in microbial ecology. The advent of PCR, genetic sequencing and other cultivation-independent molecular methods led to remarkable advancements in microbiology and microbial ecology has become a well-established field. Advancements in microscopy made it possible to observe bacteria in environmental samples however the simple morphological characteristics of bacteria are not sufficient to identify individual species or to establish a phylogenitically meaningful system of taxonomy based on the principles of evolution. In the past, to identify bacteria, it was necessary to cultivate them in a laboratory setting and test them for certain physiological and biochemical traits. Bacterial culture studies done under controlled laboratory conditions were needed in order to produce enough cells to observe their responses to certain required tests for identification. Unfortunately, most of the bacteria that are observable in nature are resistant to standard cultivation techniques; actually, the bacteria that are *not* resistant to cultivation are account for only 0.001% to 0.3% of total cell counts in natural environments such as seawater, freshwater and soil (Amann, Ludwig, &

Schleifer, 1995). Recent developments in sequencing technologies, in addition to the introduction of 16S Ribosomal RNA (rRNA) gene as a phylogenetic marker for identification (Olsen, Lane, Giovannoni, Pace, & Stahl, 1986), ultimately made cultivation-independent studies possible and allowed microbial ecologists to develop a better understanding of the diversity within microbial communities.

Sanger's method for DNA sequencing (Sanger, Nicklen, & Coulson, 1977) was one of the critical breakthroughs in biological sciences and became one of the core components of basic biological research. Almost a decade after Sanger's achievement, Pål Nyrén developed the idea of bioluminometric DNA sequencing, which depends on measuring pyrophosphate release during DNA synthesis (Nyrén, 2007). Nyrén's idea eventually enabled 454 Life Sciences, a biotechnology company now a subsidiary of Roche, to introduce massively parallel high-throughput pyrosequencing method (Margulies et al., 2005) with the capacity of sequencing millions of bases in a matter of hours (as of today, $\sim 5 \times 10^5$ reads in less than 8 hours with 454 GS FLX). Since 2005, pyrosequencing has been utilized in numerous applications from whole genome sequencing to genotyping (King & Scott-Horton, 2007) and detecting single-nucleotide polymorphisms (Fakhrai-Rad, Pourmand, & Ronaghi, 2002).

Pyrosequencing has its own challenges such as a relatively high error rate (~1%), and shorter read lengths compared to Sanger capillary sequencing (Mashayekhi & Ronaghi, 2007). By contrast, pyrosequencing reduced the cost and time of the sequencing process and produced more reads while liberating researchers from laborious steps of cloning of DNA fragments and purification of individual templates. Combined with the power of 16S

Ribosomal RNA gene, pyrosequecinging today is a lieutenant of microbial ecology to obtain sequence data that would undergo an analysis to characterize microbial populations sampled from natural environments. Due to the vast amount of sequencing data being produced, microbial ecology research depends heavily on computer science and computational statistics.

16S Ribosomal RNA Gene

For a long time, investigating relationships between organisms depended on observable phenotypical patterns. Then molecular techniques such as DNA-DNA hybridization (Southern, 1975) let researchers measure whole genome similarities between two or more species (Socransky et al., 2004; Sibley & Ahlquist, 1984). Along with the emergence of sequencing technologies, measurement of evolutionary relationships and building phylogenies based on the genomic information gained a new direction (Zuckerkandl & Pauling, 1965).

Woese and Fox were the first to reveal the potential of Ribosomal RNA genes as efficient sources of information to build phylogenetic trees that would reflect evolutionary descent of species (Woese & Fox, 1977).

Ribosomal RNA genes, namely 5S, 16S and 23S in prokaryotes, are essential components of bacterial and archeal genomes as they are responsible for the formation of the ribosome: the vital organelle in all living cells that administers protein assembly. Even though 5S and 23S Ribosomal RNA genes also exhibit phylogenetic value, the 16S Ribosomal RNA (16S rRNA) gene is the most conserved one of all three (Woese, 1987) and therefore it is more suitable for phylogenetic inference.

The 16S rRNA gene, due to its structural function in the formation of the ribosome, has been under relentless evolutionary pressure. Through the journey of prokaryotes, this pressure created compartments within this 1542 nucleotide long curious and ancient gene into numerous highly *conserved* and *variable* regions.

DNA-DNA hybridization studies show a significant correlation between 16S rRNA gene similarity and whole genome similarities (Keswani & Whitman, 2001). This proves the 16S rRNA gene very useful for phylogenetic purposes on one hand. On the other hand, there are also purely biological concerns about relying on Ribosomal RNA sequence similarity as a measure of evolutionary relatedness, especially for closely related taxa. The majority of these concerns stem from poor ultrametric attributes of the 16S rRNA gene (Sneath, 1993), mostly due to low resolution and incongruities between Ribosomal RNA divergence and evolutionary divergence (Keswani et al., 1996). Despite these concerns, today, the 16S rRNA gene is still the most popular tool for studying microbial populations in depth.

Powerful Duo: Pyrosequencing and 16S Ribosomal RNA Gene

Although pyrosequencing gives relatively longer read lengths than other prominent high-throughput sequencing technologies (~150bp for Genome Analyzer IIx by Illumina Inc., and ~75bp for SOLiD™ by Applied Biosystems as of today), the full-length 16S rRNA gene is quite a bit longer than the maximum read length of pyrosequencing (~400-500bp), rendering it unviable to sequence the entire 16S rRNA gene without requiring assembly.

Meanwhile, highly conserved regions of the 16S rRNA gene make it possible to design universal primers to amplify desired hypervariable regions of the gene via polymerase chain reaction (PCR), which is a well-established molecular biology technique. Fortunately,

some of the hypervariable regions of the 16S rRNA gene are not only below the maximum read length that pyrosequencing can reach, they are also specific enough to be used as phylogenetic markers (Liu, Lozupone, Hamady, Bushman, & Knight, 2007) for taxonomical assignment purposes, and diversity analyses of the microbial communities in environmental samples (Pace, 1997). This method is known as hypervariable tag sequencing (Huse et al., 2008), and it allows researchers to utilize massively parallel high-throughput pyrosequencing to observe a very large number of individuals from a microbial sample by using only partial information from the 16S rRNA gene.

Another technique called DNA barcoding (Binladen et al., 2007; Hamady, Walker, Harris, Gold, & Knight, 2008), allows researchers to generate 16S rRNA gene tag sequences from multiple samples with one pyrosequencing run by incorporating unique tags into each PCR primer that are going to be used during the PCR amplification. As of today, one pyrosequencing run can generate hundreds of thousands of sequences from hundreds of samples simultaneously and provides great advantages over the traditional methods of working with microbial samples.

There are numerous other ways to study microbial populations depending on the questions to be addressed. Shotgun metagenomics (Petrosino, Highlander, Luna, Gibbs, & Versalovic, 2009), metatranscriptomics (Warnecke & Hess, 2009) and metaproteomics (Wilmes & Bond, 2006) are recently developed methods that allow researchers to investigate a great deal of functional diversity of microbial communities by analyzing samples collected from their natural environments. Nevertheless, in respect to the degree

of sample coverage, these methods are not comparable to the depth of unexplored biodiversity revealed with the pyrosequencing of 16S rRNA gene tags.

Microbial Communities and Human Health

Bacteria colonize the human body in the gastrointestinal tract, oral cavity, skin, airway passages and urogenital system (Group et al., 2009). One estimation indicates that there are 500 to 1000 different bacterial species harbored in different sites of the human body (Sears, 2005). Yet another striking discovery is that the number of bacterial cells that live on an average person's body is ten times more than the number of eukaryotic cells that make up the body itself (Savage, 1977; Berg, 1996). Considering the astronomical numbers and the diversity of bacterial cells that are harbored in different sites of the human body, it is not surprising that there are correlations between health and disease states and changes in the composition of microbial communities (Sandoval & Seeley, 2010).

From extracting nutrition trapped in the diet, to training our immune system, we do rely on bacteria throughout our lives (Dethlefsen, McFall-Ngai, & Relman, 2007). At the moment of birth all body sites that are eventually going to become home to an assemblage of microbes are essentially sterile. Heavily influenced by the type of delivery (Biasucci, Benenati, Morelli, Bessi, & Boehm, 2008), swift colonization starts with the bacteria from the mother as well as the surrounding environment (Mackie, Sghir, & Gaskins, 1999). Advancements in microbial ecology have shed light on the dynamics of early gastrointestinal tract colonization (Mackie et al., 1999) as well as the temporal and cross-sectional distributions of bacterial communities. For instance, as of today, we know that bacterial communities harbored in different sites of the human body have different

compositions; moreover, these communities show similarities among other people in respect to body site, more than other criteria such as sex or age (Costello et al., 2009). When pervasive effects of antibiotic treatments on gastrointestinal tract microbial communities have been characterized with deep sequencing, it has also revealed the sensitive and resilient nature of bacterial communities (Dethlefsen, Huse, Sogin, & Relman, 2008).

Struggling to understand the actual impact of microbial communities in the gastrointestinal tract is an active field of research. Using model organisms to investigate the relationships between microbial communities and their hosts is also being conducted. Recent studies with rats, mice and primates that try to address how microbial communities affect the organ development and organ function rendered previously unknown aspects of this relationship, and gave hints about the rest of the iceberg. While a recent study showed that microbial flora in the gastrointestinal tract changes the liver function of mice by modulating gene expression patterns (Björkholm et al., 2009), another one presented evidence indicating that early microbial colonization might affect brain development (Heijtz et al., 2011). Yet another study shed light on how commensal bacteria in the gastrointestinal tract plays a key role on adaptive immunity against respiratory virus infections and shows that after antibiotic treatment viral replication remains high in lungs while immune response decreases (Ichinohe et al., 2011). Moreover, links between microbial communities and kidney stone development (Sidhu, Allison, Chow, Clark, & Peck, 2001), maturation of the host immune system (Mazmanian, Liu, Tzianabos, & Kasper, 2005) and increasing risk of asthma as a result of antibiotic disturbance of early microbial communities (Kozyrskyj, Ernst, & Becker, 2007) are some of the physiological associations

that have been investigated. Besides the impact on physiological functions, recent preliminary findings suggested that commensal bacteria modulate serotonin levels (Desbonnet, Garrett, Clarke, Bienenstock, & Dinan, 2008) and anxiety-like behavioral patterns (Heijtz et al., 2011) in rats, which provokes us to question the extent of the bacterial influence on psychological traits.

Due to the intimate relationship between bacteria and humans (Sekirov & Finlay, 2006), consideration of a human being as a superorganism (Wilson & Sober, 1989) is not a new suggestion (Goodacre, 2007). Even though bacteria certainly have an interest in its hosts' well-being (Lederberg, 2000), there are tormenting diseases and medical conditions associated with the changes of the bacterial flora, such as obesity, Crohn's disease, necrotizing enterocolitis and bacterial vaginosis.

Obesity is one of the medical conditions that are associated with microbial communities. According to biochemical studies on microbial communities sampled from gastrointestinal tracts of obese and lean mice, the obese microbiome is more capable than the other in terms of the ability of extracting energy from diet (Turnbaugh et al., 2006), and this correlation has been explained with a slight shift in diversity and the relative abundance of two phyla; *Bacteroidetes* and *Firmicutes* (Ley et al., 2005), indicating that even subtle changes in the diversity and composition may affect the functional outcome of the community.

Crohn's disease is one of the two major types of inflammatory bowel disease and it is defined as a medical condition that causes the body's immune system to attack the healthy gastrointestinal tract (Baumgart & Sandborn, 2007). Even though there is evidence for a

genetic predisposition for developing Crohn's disease (Barrett et al., 2008), 16S rRNA gene tag-based analysis shows that the dominant members of Crohn's disease patients' microbial communities manifest a significant temporal instability compared to the healthy group (Scanlan, Shanahan, O'Mahony, & Marchesi, 2006).

Necrotizing enterocolitis is a frequent disease among premature infants with a very high fatality rate (Wang et al., 2009). Necrotizing enterocolitis causes portions of the intestines to undergo tissue death and fall off. A very low diversity of microbes occupying the gastrointestinal tract has been associated with necrotizing enterocolitis cases and it is believed that microbial communities play an important role in the pathogenesis of the disease (Emami et al., 2009).

Bacterial vaginosis is a very common vaginal disorder associated with preterm delivery (Hillier et al., 1995). Drastic shifts in vaginal flora with the absence of *Lactobacillus* species in Bacterial vaginosis diagnosed women causes a decrease in the pH of the vaginal flora, which has been suggested as a cause of weak immune responses to sexually transmitted viruses (Sha et al., 2005). 16S rRNA gene-tag based studies revealed novel bacterial species in the flora of patients that are diagnosed with Bacterial vaginosis that were previously unknown (Oakley, Fiedler, Marrazzo, & Fredricks, 2008). There will be a more comprehensive discussion of bacterial vaginosis in Chapter III.

Indeed, analyzing the differences between the flora of healthy patients and the flora of patients who are suffering from a disease is crucial to developing a better understanding of the dynamics of these medical conditions and eventually preventing and treating them.

Microbial Community Analysis from a Computational Perspective

Microbial communities have been explored in a wide range of environments and broadened our understanding about unforeseen diversity of prokaryotes (Sogin et al., 2006). There are various challenges attached to almost every step of working with microbial communities from sampling to sequence analysis, but nonetheless, 16S rRNA gene tag sequencing-based studies have provided invaluable information about the dynamics of microbial life in almost every natural environment.

Numerous steps and methods are required to reveal the structure of a microbial community of interest. Once pyrosequencing reads are gathered from an environment as raw sequence data, it is necessary to quantify distributions of organisms. Following this step statistical and computational inference methods are utilized to shed light on microbial diversity, compare different environments to each other, perform cluster analysis, and to emphasize similarities between environments based on various distance metrics.

Quality Control

Both PCR amplification and pyrosequencing introduce a considerable amount of noise to the sequencing results, such as chimeric sequences resulting from complications regarding the DNA templates during enzymatic amplification (Meyerhans, Vartanian, & Wain-Hobson, 1990) or ambiguous lengths of homopolymer regions due to the chemistry of pyrosequencing. Most downstream computational and statistical methods assume that the sequencing data is trustworthy and in some cases this assumption leads to a distorted view of the landscape spawned from skewed frequency count curves and diversity

estimations. Therefore, in order to improve the trustworthiness of the data to meet the assumption of conventional algorithms, it is a required step to improve the overall quality of pyrosequencing reads by eliminating the low-quality ones before any analysis. Both the formulation of the quality issues and their impacts on analysis results (Quince et al., 2009; Reeder & Knight, 2009; Kunin, Engelbrektson, Ochman, & Hugenholtz, 2010), and suggestions to overcome these (Huse, Welch, Morrison, & Sogin, 2010; Reeder & Knight, 2010; Malde, 2011; Quince, Lanzen, Davenport, & Turnbaugh, 2011) are an active field of research.

Taxonomic Classification

Once the quality expectations are satisfied, the next step for most of the studies working with a large number of sequences is to start revealing the content of their libraries by either (1) assigning taxonomy to their sequences by comparing sequences with the ones in public databases that contain sequences from known species, (2) clustering sequences into groups of operational taxonomic units (OTUs) based on an arbitrary sequence similarity threshold, or (3) a hybrid of these two to ease the time and computational power requirements: first clustering sequences into OTUs and then picking representative sequences from every group, then assigning taxonomy to representative sequences and propagating it back to the original OTU group.

This important step is where raw, scattered sequences become labels and counts. Once it is known for an environment how many different OTUs have been observed and how many reads fell into those OTU groups, the data is ready to be analyzed by the means of computational and statistical approaches.

The Ribosomal Database Project (Cole et al., 2009) maintains a curated database of full length, high quality 16S rRNA genes and provides a naïve Bayesian classifier (Wang, Garrity, Tiedje, & Cole, 2007) to query environmental sequences against the RDP database for rapid assignment of higher-order taxonomy up to the Genus level. The RDP Classifier is the most commonly used tool to assign taxonomy to 16S rRNA gene tag sequences; but it does not provide any genetic distance metric.

BLAST (Basic Local Alignment Search Tool) (Altschul, Gish, Miller, Myers, & Lipman, 1990) is also used to find the most similar sequence in a database by measuring the pairwise aligned distance of a given sequence and the sequences in a target database. GreenGenes (DeSantis et al., 2006) provides a database specifically designed for 16S rRNA gene tag sequence analyses along with a pre-computed phylogenetic tree, so identification numbers of highest BLAST hits can later be used to generate phylogenetic comparison-based distance matrices for various analyses.

Alpha Diversity Analyses

Alpha diversity analysis methods are among the most fundamental descriptive tools of ecology and have been used to describe species richness or diversity in an environment. Today, most alpha diversity measurements are being used for statistical examination of samples collected from microbial communities. These measurements may be broken up into the following categories: True Diversity, Species Diversity Assessment and Rarefaction Curves.

True Diversity Estimation

One of the major issues of microbial ecology is to address how well a sample represents a community's true diversity since it confronts researchers with a fundamental problem: sampling bias. Mostly due to the vast scaling differences involved with sampling, reliable and applicable solutions to measure how well a sample represents a community's true diversity is very hard to develop. However, microbial ecologists still have to rely on their samples to speculate about the diversity of their original communities and this requires heavy use of computational statistics (Hughes, Hellmann, Ricketts, & Bohannan, 2001).

There are several widely used non-parametric and computationally lightweight "*true diversity*" estimators that rely on abundance data, such as Chao1 (Chao, 1984) and ACE (Chao & Lee, 1992). But these are known to be prone to skewed results when working with very high diversity situations, which is expected to be the case with most microbial community analysis studies (Sogin et al., 2006), where rare members create a long tail in the frequency count distribution curve of a sample (Bunge & Barger, 2008). One of the ideas to overcome this problem as much as possible is to combine statistics with heuristics, rather than only using a single coverage-based nonparametric richness estimation method for approximation. CatchAll (Bunge, 2011), a recently developed method, is a precursor to this approach.

CatchAll aims to find the optimal finite-mixture of models with the best parameters in order to realistically explain the distribution of operational taxonomic units in a sample, so that the actual diversity of the parent population can be computed by extrapolating the final estimation curve. Result of the analysis with CatchAll is a list of estimation

recommendations along with confidence intervals, goodness-of-fit estimations and standard errors for researchers to investigate and select. It is the first application to carry out parametric species richness estimation this way and being used by The International Census of Marine Microbes (ICoMM, http://icomm.mbl.edu), and various software packages designed to study microbial communities such as MOTHUR (Schloss et al., 2009) and QIIME (Caporaso et al., 2010b).

Species Diversity Assessment

Diversity indices, generally, aim to provide a statistic about the distribution of different types of objects in a set. Shannon-Wiener Index and Simpson's diversity index (Simpson, 1949) are two widely used indices within ecology to measure species diversity in an environment, and serving microbial ecology for the same purpose, despite the tremendous difference in observation size between conventional ecology and microbial ecology. Shannon index is more sensitive to changes in abundance of rare OTUs (Hill, Walsh, Harris, & Moffett, 2003) compared to other diversity indices including Simpson's, and is considered a better general diversity index. Since neither of these indices considers the species richness of other samples and they are biased with the sample size, comparability of these measurements among different samples is very limited. But it is a very common practice to compare microbial communities to each other with respect to Shannon's diversity index.

Rarefaction Curves

Rarefaction is yet another statistical technique that has been used within various ecological contexts (Sanders, 1968). A rarefaction curve is a graphical representation of the

expected number of species versus the total number of individuals in a sample. One of the uses of rarefaction in conventional ecology was to predict whether two samples might be sampled from the same environment (Simberloff, 1978). A somewhat useful and concurrent utilization of rarefaction curves in microbial ecology is to estimate whether the minimum feasible sample has been met; in order to cover most of the diversity in a given environment by answering how many more OTUs would have been observed if sample size were larger. Rarefaction curves are expected to reach a horizontal asymptotic convergence, which would indicate that increasing sample size would not have revealed more OTUs, and therefore it could be argued that environment was sampled sufficiently.

Beta Diversity Analyses

Beta diversity analyses are useful to exhibit patterns in a data set that would otherwise be missed. Microbial ecologists struggle to make sense of the massive and complex data extracted from environmental samples using various multivariate analysis methods that have become standards of the field.

Here in this part, some of those methods are going to be explained briefly, namely: Clustering, Principle Coordinate Analysis and Correspondence Analysis. Ramette's survey (Ramette, 2007) would be a suitable suggestion for a more comprehensive review of available methods.

Clustering

In a general sense, clustering analysis refers to numerous computational methods that aim to partition data into one or more groups, by not only minimizing within-group

variation, but also maximizing inter-group variation so that every group would be composed of data that present similar characteristics.

Depending on the type of data, various types of clustering algorithms are popular in different fields. For microbial ecology, where OTUs are believed to branch out from common origins based on their phylogenetic distribution, hierarchical clustering algorithms seem to be more popular to explain sample similarities than partitional clustering algorithms.

Hierarchical clustering is a clustering approach where hierarchies of clusters are built based on distances between pairs of objects. Complete-linkage is one of the ways for mediating the order of clusters to be merged in an agglomerative way. This is achieved by calculating distances between clusters via distances between farthest objects in two clusters (Levcopoulos, 1998). It is the most common practice to perform hierarchical cluster analysis of microbial communities. Results of complete-linkage clustering are usually shown as tree diagrams (dendrograms), as it is a very standard way of visualizing and communicating relationships between groups of samples.

The anticipated outcome of a successful clustering is an accurate description of the underlying structure of samples based on the relationships of components that assemble them. Naturally, relationships between those components are subject to the chosen way of computing the distance between them.

One of the ways to compute distances between samples is to use percent abundance normalization of the reads in order to generate numerical representations of each sample,

which can be treated as an *n*-dimensional feature vector from the contingency table, where *n* is the number of different OTUs in all samples. A distance matrix can then be populated from Euclidian distances between these vectors and used for clustering analysis and visualization. One big problem with this method arises from the fact that phylogenetically very similar OTUs and phylogenetically very distinct OTUs contribute equally to the distance between samples. This may result in missing subtle differences during the cluster analysis. For instance, the distance between two communities that have different but phylogenetically very closely related OTUs may not be smaller than a third community that has different, and also phylogenetically very distinct OTUs than the first two communities, and this may result in an unrealistic portrayal of their relative overall association.

One other way to compute distances between samples relies on the phylogenetic inference. Phylogeny provides better resolution for asserting the resemblance or the magnitude of diversification between communities more accurately (Martin, 2002). UniFrac (Lozupone & Knight, 2005; Hamady, Lozupone, & Knight, 2010), a relatively new method that allows the computation of differences between microbial communities based on phylogenetic information, has become one of the standard metrics for comparing communities. For every pair of samples, the UniFrac algorithm measures the total length of all unique branches on the tree that lead to OTUs that belong to either of the samples. The ratio of the branches that are unique to either of the samples to all branches in the tree gives a metric that reflects the similarity between those two samples to each other, based on the given phylogeny. Comparison results populate a distance matrix, which can be used to explain similarities. Today, using the UniFrac distance matrix for clustering and

investigating the statistical significance of clustering results with Monte Carlo simulations is a common approach for comparing different environments.

Principal Coordinate Analysis

Principal coordinate analysis (PCoA) is a multi-dimensional scaling method that makes it possible to visualize a distance matrix in a two or three-dimensional Cartesian coordinate system by choosing the orthogonal axes ordered to explain maximum variation between objects.

PCoA is a popular way to exploit UniFrac results that are stored in a distance matrix in order to visualize distances between samples based on their overall phylogenetic profiles. Since distances between the objects on the projection are expected to be reflective of the distances within the matrix, PCoA is a handy method to understand and communicate comparison results.

Correspondence Analysis

Correspondence analysis (CA) has a long history in ecology. It operates on contingency tables where samples and OTUs are rows and columns and intersections of them are the numbers of occurrences or percent abundances of these two discrete varieties (Hill, 1974). The aim of the method is to compare the correspondence between rows and columns to represent underlying structures in a data set in a lower dimension by retaining most of the information about the distances of the samples. When there is an excessive number of OTUs in the analysis, CA could provide a quick overall look, but again, in most cases it would not be as accurate as a method where phylogeny among samples is considered.

CHAPTER II

As discussed in the previous chapter, advances in sequencing technologies that can utilize 16S rRNA gene made deep exploration of microbial communities in environmental samples easy and affordable.

Also the correlations between human microbial community composition and health conditions generated even a greater interest in analyzing microbial communities.

Due to the large amount of sequence information associated with PCR amplification and pyrosequencing of 16S rRNA genes from the microbial communities, a variety of statistical methods and extensive computational aid is mandatory for the analysis of the data. While sequencing becoming more and more affordable, most of the available software pipelines remained arduous to be utilized by people who are not computer experts and required researchers to invest a substantial amount of time to analyze their own sequences.

The primary goal of the work that is going to be presented in this chapter was to bring the analysis of large amounts of microbial community sequence data within the reach of scientists who have only basic computer skills and to create an extensible framework where new methods can be incorporated and applied to existing data easily.

A Framework for Analysis of Metagenomic Sequencing Data

There are several methods available to understand and compare microbial community structures in samples from different environments through 16S rRNA gene sequence data. However, most of these methods are not designed to manipulate large high-throughput pyrosequencing data. One solution that is being employed by researchers in the field is to prepare individual scripts in order to manipulate large sequencing files for each analysis, which requires extensive programming skills and experience to maintain. Another solution is to rely on more general approaches and to use online tools and/or pipelines to perform basic analyses and tests on data, which introduces another set of caveats that are going to be addressed in the Motivation subsection. The framework that is going to be presented in this section is designed to overcome many of the challenges of metagenomic sequencing data analysis and to provide researchers an easy way to analyze and interpret their data with a lot of visualization capabilities.

The framework was reported in a paper titled "A Framework for Analysis of Metagenomic Sequencing Data" (Eren, Ferris, & Taylor, 2011) in The Pacific Symposium on Biocomputing 2011, Hawai'i. It is an open source project and licenced with General Public Licence. The source code is available at a repository hosted by GitHub:

https://github.com/meren/viamics

Motivation

Software packages that are available to researchers to process 16S rRNA gene sequence data can be divided into two groups: those that are hosted on a server and used via web interfaces, and those that are downloaded and run locally. Both approaches have their

benefits and their limitations. Online ribosomal sequence analysis applications and pipelines, such as the Ribosomal Database Project (RDP) pipeline (Cole et al., 2009) or the online tool chain of GreenGenes (DeSantis et al., 2006), require researchers to upload their data over the Internet and work using web interfaces that are designed to be easy to operate. However, online analyses usually have stringent limitations on the number of sequences that can be analyzed (or number of runs, or number of permutations). This is primarily due to the fact that scarce resources, such as CPU time, memory size and network bandwidth, must be shared by many researchers in any centralized approach. Another limitation of this approach is that the software cannot be customized and enhanced for specialized analysis since it is running on another group's server.

On the other hand software that can be downloaded and run locally such as MOTHUR (Schloss et al., 2009) and QIIME (Caporaso et al., 2010b), permits researchers to use their own computational resources without requiring them to upload their data to another server. However, since most of these applications necessitate the use of command line interfaces to perform function calls, the learning curve for these tools is steep and a significant investment of time is required to learn and operate them.

Another aspect of available 16S rRNA analysis software that limits its utility is the "pipeline" approach. Pipeline approaches are a model of computing where a set of applications are connected to each other such that output from one application becomes input to one or more applications in the subsequent stage. A pipeline approach is not an efficient structure for an application that is designed to analyze sequencing data. Applications in a pipeline cannot use previous applications' resources; these resources may

need to be re-allocated or re-computed at every stage of the pipeline. This redundancy is not efficient use of computational resources and negatively impacts overall performance. In addition, the process of file upload, analysis and download, which may be repeated at different stages, is time consuming since the user must wait for output and must often upload results again for the next stage of analysis. Lastly, the preponderance of intermediate results from different stages of the pipeline that the user must manage is a large burden that can easily lead to mistakes due to human error.

It was necessary to design an extensible and easy-to-use software framework that is liberated from these issues as much as possible by offering a hybrid solution.

During its development, the framework has been tested and used by microbial community researchers studying the microbiota associated with various diseases such as bacterial vaginosis and necrotizing enterocolitis. Researchers using the framework were empowered to analyze their own samples, test hypotheses, and produce publication quality figures in order to communicate their results.

Technical Features

The framework is developed on the Pardus Linux distribution using the Python programming language and open source scientific computing tools and libraries such as SciPy (http://scipy.org) and matplotlib (http://matplotlib.sourceforge.net/). Among other numerous direct and indirect benefits, a reliance on open source development tools and libraries allows the framework to be extended effortlessly and makes it easier to port the framework to non-Linux-based environments.

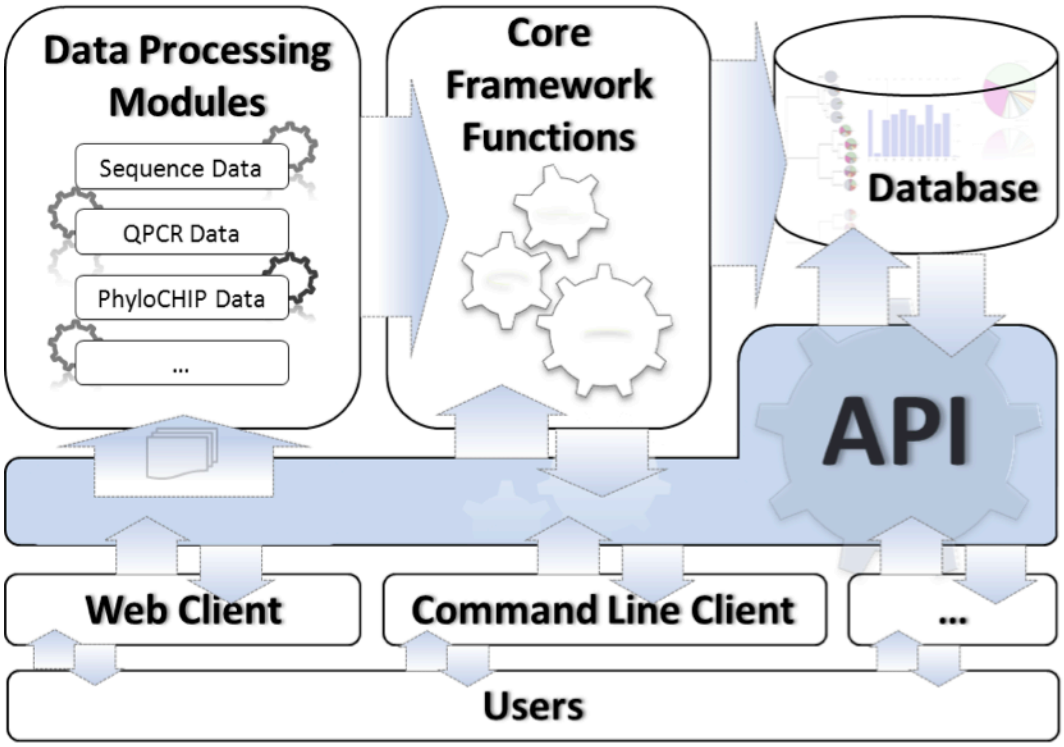

Figure 1. Architectural overview of the framework.

Figure 1 shows an architectural overview of the framework with two major components: A multi-threaded server application that runs in the background performing data processing and core framework functions and interfaces for users to interact with the server.

Server

The server performs all manner of computational tasks. The multi-threaded design of the server allows it to run multiple analyses concurrently and handle queries simultaneously. The server exposes its functions via an application programming interface (API). This makes it possible for different types of clients to be written and interact with the server seamlessly (Figure 1). This flexibility also allows the framework to be used in

both the graphical, user-friendly manner, or invoked by scripts for automated analysis of large numbers of data sets.

The server has more than one data processing module, and a set of core functions that is separated from the data. This modularity allows server's core functions and analysis capabilities to be extended for different types of inputs, such as quantitative real-time PCR data.

Client

Any client that can communicate via UNIX domain socket or TCP/IP protocols can query and submit tasks to the server through the API. The default client of the framework is a set of Django (http://www.djangoproject.com) powered web interfaces. The web client allows users to connect to and use the framework via their web browser. Thus, they can interact with the default client of the framework using any operating system and Internet browser they choose. This separation also makes scaling easy: it is possible to host the framework on a computer in a local network for a group of researchers to use, it is also possible to use it on a personal computer.

Limitations

The framework is still under development. As of the day this dissertation was being written, 16S rRNA sequences were being analyzed by RDP's naïve Bayesian classifier (Wang et al., 2007) which limited researchers to perform analyses based on genus level taxonomy. However, the modular nature of the framework allows its capabilities to be easily extended. Implementing other data processing modules for phylogenetic analysis or

more intricate population comparisons based solely on sequence similarity may be possible.

When this dissertation was being written, several biological researchers were using the framework for sequence data analysis, statistical comparison and visualization purposes for a number of active research projects. The most demanding project that analyzed on the framework was consisted of 166 samples with more than 2 million sequences.

It is also important to note that the classification of sequence data is independent and orthogonal to the downstream analysis and visualization tools. In fact, any data set that contains names and associated abundance values can be slipped into the framework and processed through the downstream analysis and visualization. As a concrete example of this, a facility for quantitative PCR data was implemented. Quantitative PCR results can be loaded into the framework and analyzed in a similar manner to classified 16S rRNA sequencing data.

Workflow

Ease of use and extensibility were the key design concerns for the framework. Hence, most of the analysis tasks were performed on the framework without requiring any *a priori* knowledge to be provided by the researcher. The basic workflow of the framework for a sequence analysis is illustrated in Figure 2. Readers are also encouraged to visit http://meren.org/framework/ to view an example analysis performed with the framework.

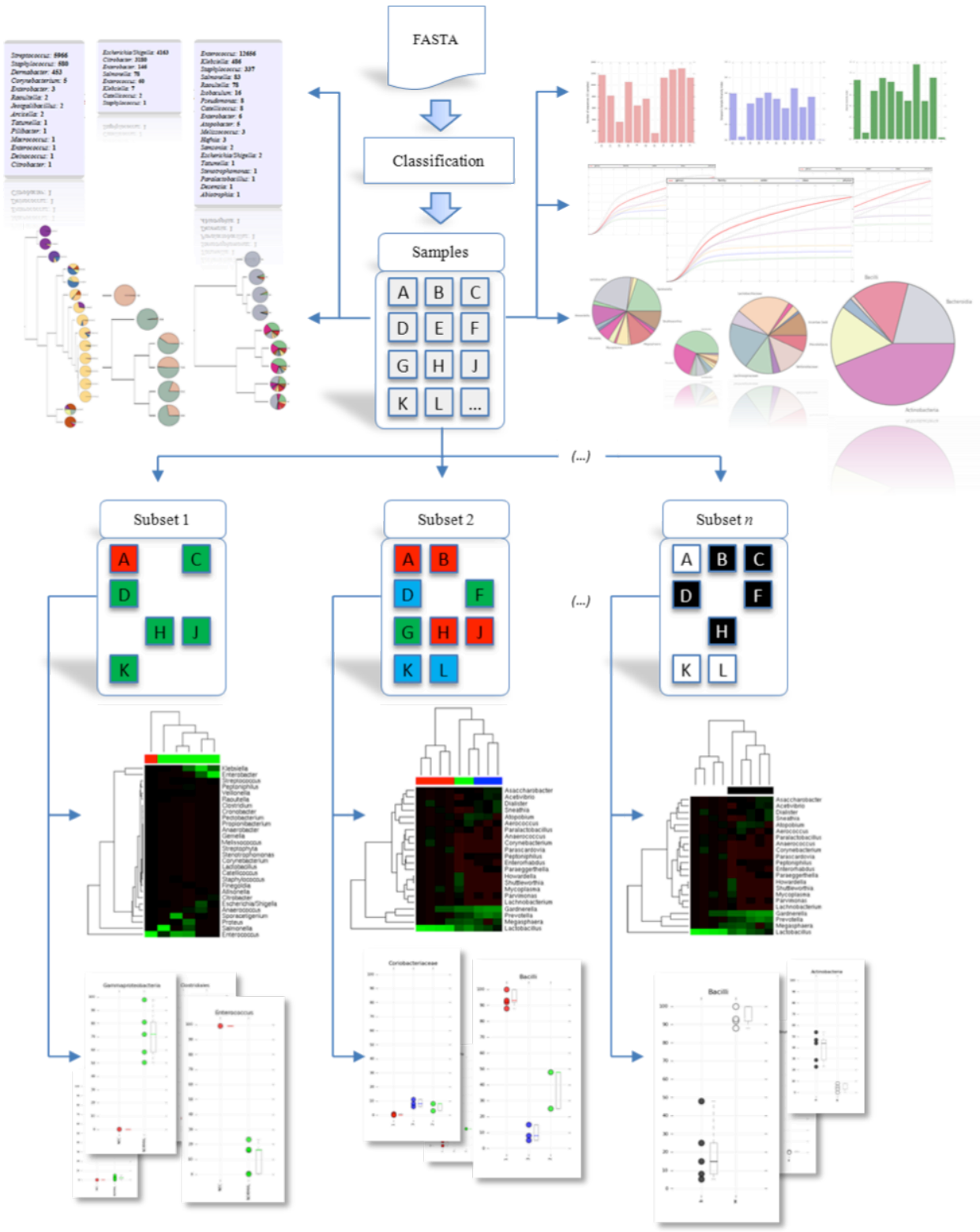

Figure 2. Basic workflow of the framework. Analysis begins with the submission of a FASTA formatted 16S rRNA sequence file.

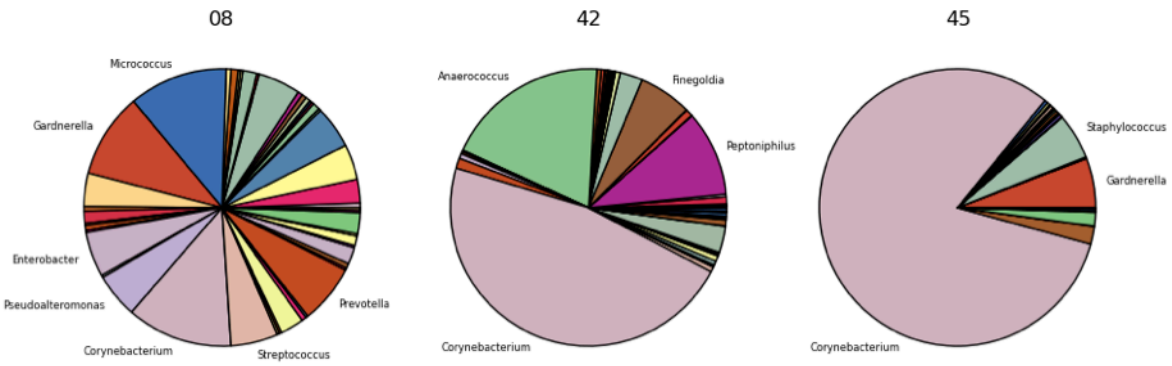

Figure 3. Example pie chart figures show bacterial composition at the genus level of three random samples from a bacterial vaginosis study analyzed using the framework.

An RDP-based sequence analysis begins by submitting a FASTA formatted file containing 16S rRNA gene sequences. The file can contain multiple FASTA files originating from multiple environmental or clinical specimens. The framework then employs RDP's naïve Bayesian classifier (Wang et al., 2007) for rapid assignment of sequences to the taxonomic groups at the phylum, class, order, family and genus levels. The framework proceeds to perform unsupervised preliminary analyses on the samples acquired from the RDP classifier results which include:

- Calculations of total and percent abundance of bacteria in every sample,
- Bar chart representation of the number of sequences acquired for each sample,
- Classification confidence parameters for each sample and operational taxonomic unit.
- Bar charts for Shannon and Simpson's diversity indices,

- Pie charts for samples based on their bacterial compositions at each taxonomic level ranging from phylum to genus (Figure 3),

- Rarefaction curves to illustrate the degree of diversity covered by each sample (Figure 4),

- Hierarchical clustering dendrograms that illustrate how samples clustered based on their bacterial composition at different taxonomic levels (Figure 7).

Once this set of unsupervised alpha-diversity analyses is completed, researchers can assign keys to desired samples and create subsets of samples for further investigation. The user defines subsets by assigning samples to groups, and then assigns a color to each of those groups for visualization. There is no limit on the number of subsets the user may define. The framework ignores samples that are present in the original library if they are not assigned into any groups in a defined subset.

When the newly defined subset of samples is submitted for analysis, dot plots of every operational taxonomic unit (OTU) at each taxonomic level ranging from phylum to genus are generated. Box plots are attached alongside the dot plots to illustrate the abundance of each individual OTU across subsets of samples (Figure 5). Complete linkage clustering analysis is performed to assess similarities between microbial communities based on the percent abundance of the taxa they contain. These clustering results are displayed as dendrograms along with heatmaps illustrating the abundance of taxa in each sample (Figure 6). Heatmaps can be refined further to eliminate very low abundance OTUs or to use logarithmic values.

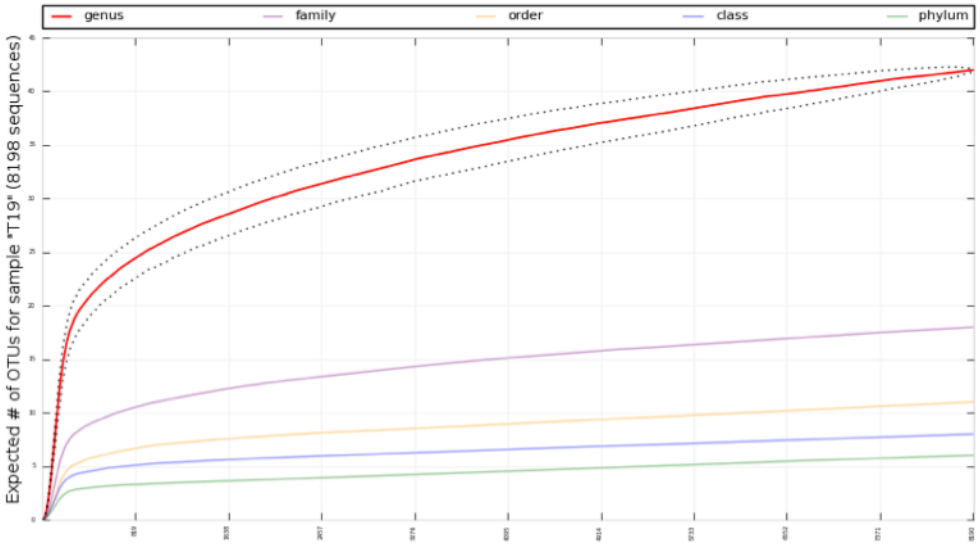

Figure 4. In this example set of rarefaction curves, species richness and expected number of OTUs are shown at different taxonomic levels of a sample that was analyzed using the framework.

Future Work

The framework was designed with the modularity in mind. Therefore enhancing the framework with a variety of additional components in the future is possible. Phylogeny-based beta diversity analysis methods; such as UniFrac (Hamady et al., 2010), as well as other experimental data analysis methods may be a part of it in order to give researchers a broader understanding of their data.

It is also worth noting that the longer read lengths being produced by the Illumina Genome Analyzer IIx have made deep sequencing of entire metagenomes feasible. This will allow researchers to go beyond simple classification based on 16S rRNA and on to analysis of complete metagenomes more frequently. The framework provides the infrastructure for further development of features to address assembly, classification and processing of

metagenomic sequencing data while maintaining the ease of use through web-based client interfaces.

Finally, this framework provides an important separation between data structure that is being generated, and the analysis and visualization of the data. The front-end that uses the RDP classifier to interpret pyrosequencing reads of 16S rRNA into their taxonomic categories was the first front-end that was implemented. Enhancing the utility of the framework by developing other front-end classifiers that may use the NCBI taxonomy or perform classification based solely on edit distance of sequences to further explore intra-genus and intra-species diversity is possible and would be invaluable.

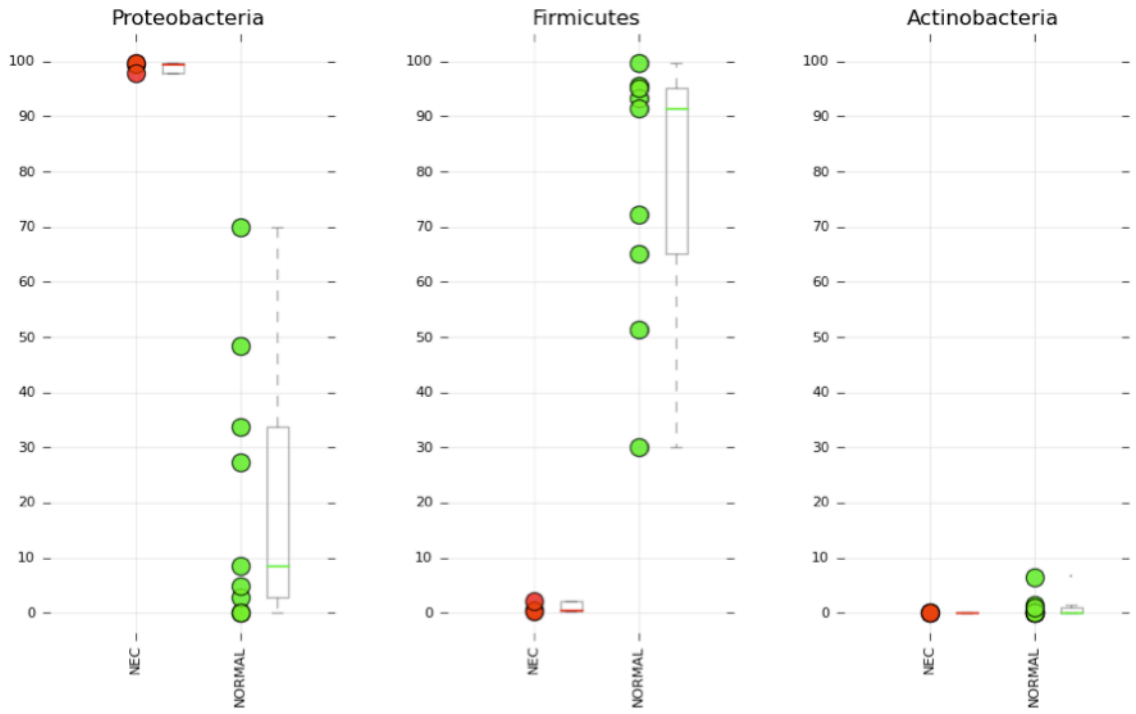

Figure 5. Example dot plot of a subset of samples assigned to two categories, NEC (green) or NORMAL (red) showing differences in the percent abundance of three different OTUs at the phylum level.

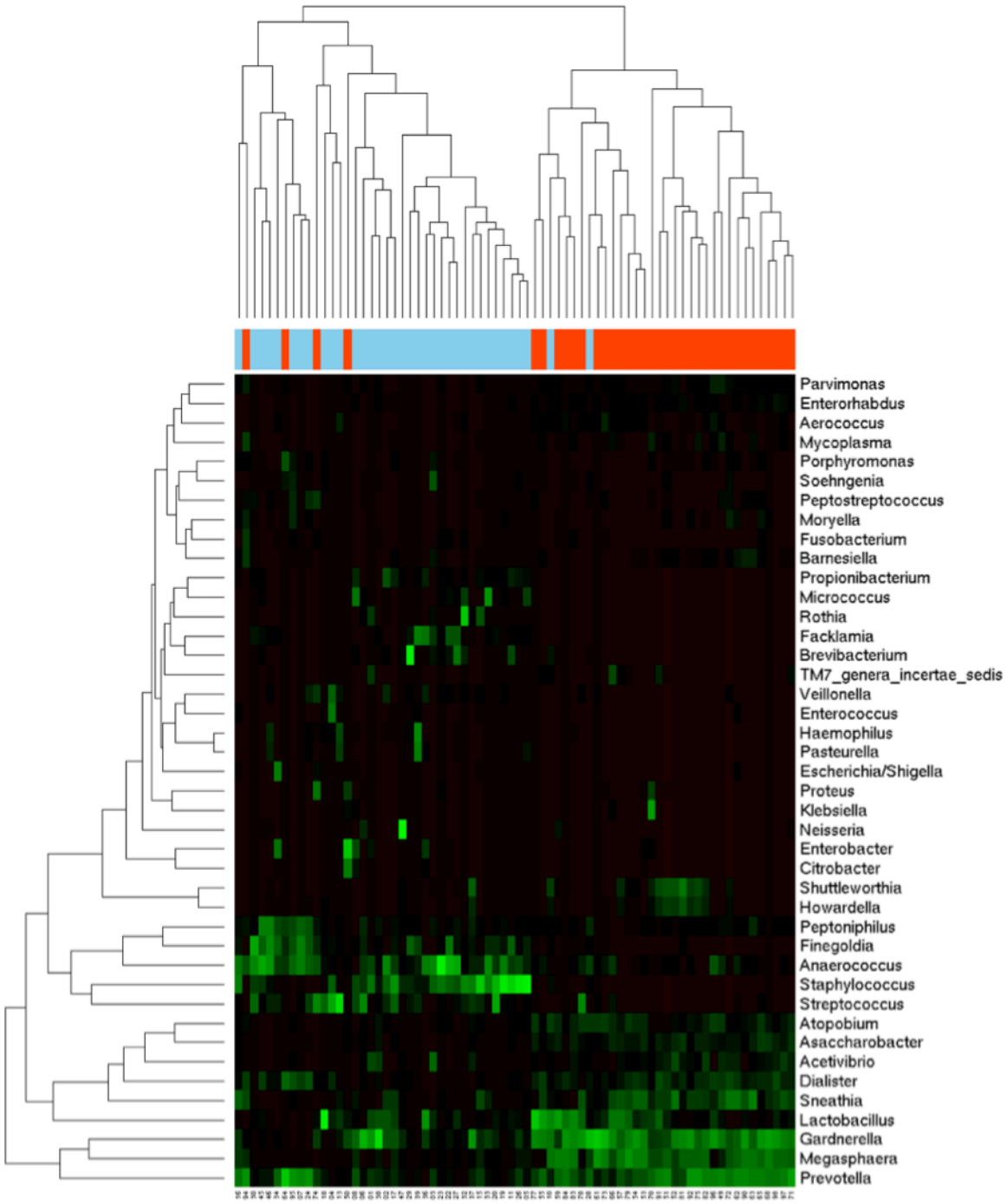

Figure 6. Example heatmap generated by the framework showing how a subset of samples clustered based on their microbial flora at the genus level. Within this particular subset of samples, the cyan color represents penile skin swab samples collected from male patients and the red color represents vaginal swab samples gathered from female patients. The vaginal swab samples largely cluster together on the left of the heatmap, while the penile skin swab samples cluster together on the right side of the heatmap.

Discussion

Although there are a variety of tools currently available for metagenomic sequence analysis, they impose unnatural paradigms or restrictive limitations on biological researchers who may have only rudimentary computer skills. Pipeline approaches force users to select analyses to perform on their data instead of performing a comprehensive analysis by default. They also place a burden on the user for maintenance and routing of intermediate results that can lead to errors. Web-based applications have advantages in terms of ease of use, but can be restrictive in the quantity of data they allow to be analyzed and the amount of user interaction required to perform an analysis. None of these approaches, by themselves, provide a viable alternative for microbial community researchers to analyze their data without scaling a significant learning curve.

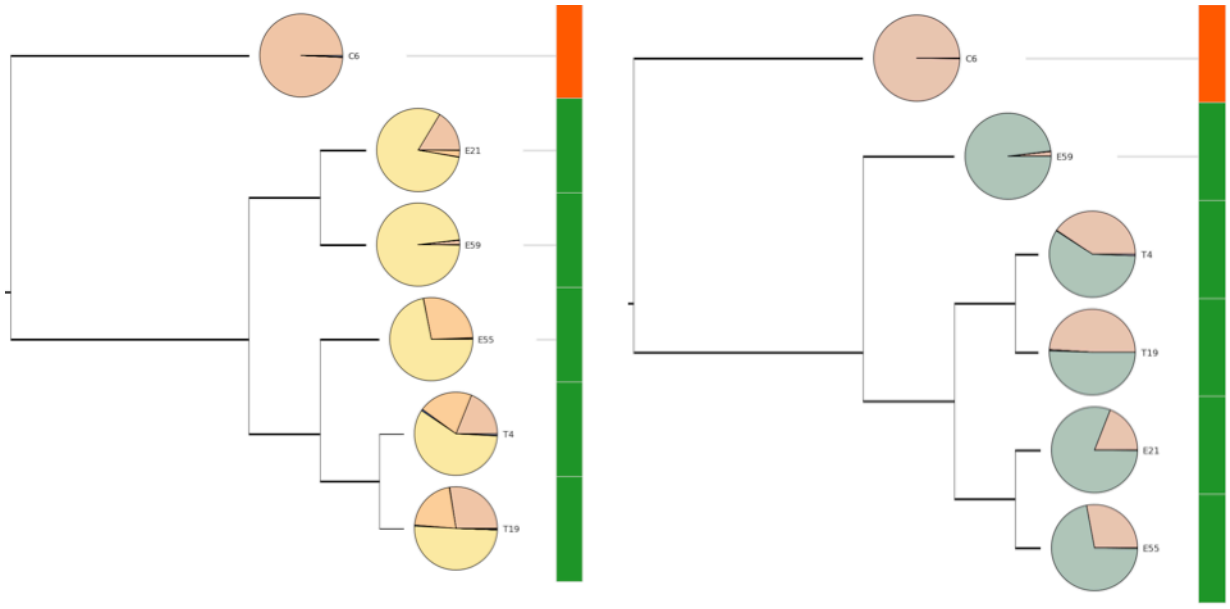

Figure 7. Example dendrograms generated by the framework showing samples from a necrotizing enterocolitis study clustered based on their microbial composition at the genus and family level. Smaller versions of the pie chart representations of samples attached to dendrograms to provide additional visual evidence for clustering results.

The framework presented here provides a scalable, hybrid approach to the problem of metagenomic sequence analysis. Researchers can run the framework on their own computational resources and are not faced with limitations on the quantity of sequences or number of analyses they can perform.

They are also able to use familiar web-based interfaces to access the server and do not need to shepherd analyses through a pipeline and manage intermediate results. All of the standard analysis methods are run at the push of a button and the user is presented with an intuitive interface to group samples for further directed analyses.

This flexibility and ease of use has allowed microbial community researchers to perform their own analyses and generate publication quality figures to communicate their results with relative ease.

CHAPTER III

Bacterial vaginosis (BV) is a vaginal infection that affects a high percentage of women and associated with serious sequelae, including adverse pregnancy outcomes (Hillier et al., 1995) and weaker immune responses to sexually transmitted viruses, including HIV (Sha et al., 2005). This enigmatic condition is characterized by a drastic change in the vaginal flora. While the number of *Lactobacillus* species in the vaginal environment decreases, the abundance of various non-*Lactobacillus* species, including *Gardnerella vaginalis*, a particular bacteria that is going to be the focus in this chapter, increase.

Although the etiology of BV is not yet fully understood (Larsson et al., 2005), several studies suggest that the disease can be sexually transmitted (Marrazzo et al., 2002) and that *Gardnerella vaginalis* may be the etiologic agent (Swidsinski et al., 2010). In contrast to the latter assertion, *G. vaginalis* is also commonly detected in vaginal specimens of women with clinically normal vaginal flora, albeit, at significantly lower concentrations than in BV (Numanović et al., 2008; Zozaya-Hinchliffe, Lillis, Martin, & Ferris, 2010).

Using a biotypization approach that incorporated three biochemical tests, (1) lipase, (2) hippuratehydrolysis, and (3) beta-galactosidase, Piot (Piot et al., 1984) defined eight biotypes of *G. vaginalis* almost three decades ago. Eight (2^3) biotypes is the maximum that can be defined using a set of three tests with two (+ or -) outcomes; which suggests that there may be more than eight *G. vaginalis* biotypes.

Biofilm formation has been shown to be a virulence trait in *G. vaginalis* (Swidsinski et al., 2005). Recently, a genomic study of two *G. vaginalis* strains, one of which was isolated from a BV patient and the other was isolated from a patient with normal vaginal flora, showed that the *G. vaginalis* strains differ in their capacity to form tightly adherent biofilms attached to vaginal epithelial cells (Harwich et al., 2010), indicating that there may be different types of *G. vaginalis* with different virulence capacities. Another genomic study of three *G. vaginalis* strains, two isolated from BV patients and one from a woman with normal vaginal flora, demonstrated that BV-associated strains produce proteins not found in the commensal strain (Yeoman et al., 2010).

Two 16S rRNA genes from the strains that were used in Harwich's genomic study differ from each other by only 6 nucleotides. This difference translates into 99.6% sequence similarity at the whole 16S rRNA gene level, suggesting that a very small degree of sequence variation in 16S rRNA genes in *G. vaginalis* strains may correspond to a significant phenotypic variation. This small genetic variation, though, is beyond the comprehension of available mainstream computational methods to detect and utilize to address different types of *G. vaginalis*.

One of the motivations triggered this study was the curiosity of whether these subtle genetic variations could be traceable among 16S rRNA gene tag sequences and whether a widely applicable analysis method could be suggested to further investigate similar instances.

Preliminary clustering analysis of *G. vaginalis* sequences obtained from vaginal swabs from patients diagnosed with BV and patients with normal vaginal flora showed that there

might be subtle differences between *G. vaginalis* sequences that are harvested from normal women and BV diagnosed women (see Figure 22 in Appendix section).

To gain further insight into the association between genetic variation in *G. vaginalis* and BV, we used sequences from 70,037 16S rRNA gene segments that were PCR-amplified from, 8 normal women, 35 women diagnosed with BV, 5 women with intermediate BV diagnosis, and penile skin and urethral specimens from their male sexual partners.

Using an entropy-based approach, nine highly variable nucleotide positions among 70,037 *G. vaginalis* 16S rRNA gene tags were identified, and nucleotides to define *G. vaginalis* "oligotypes". Also a phylogenetic correspondence between communities of *G. vaginalis* oligotypes and BV, and between sexual partners has been observed.

While Methods subsection explains the steps of analysis, Results subsection reveals the findings.

Methods

Sample Collection and Clinical Measurements

53 monogamous heterosexual couples were recruited for this study at the New Orleans STD clinic. All subjects were at least 18 years old with no history of antibiotic use in the past 28 days. Couples presented together for evaluation. A vaginal swab was collected from each woman for DNA extraction and pyrosequencing analysis of bacterial composition. A separate vaginal swab specimen was collected and characterized clinically as normal or BV using Nugent scores (Nugent, Krohn, & Hillier, 1991). The samples were designated “normal” (Nugent score= 0-3), “intermediate” (Nugent score= 4-6) or “BV” (Nugent score=

7-10). Two urethral swabs and two penile skin swabs were collected from males. For penile skin specimens, two sterile Dacron swabs were used; one was rolled with firm pressure around the circumference of the coronal sulcus and over the surface of the glans penis, and the second one was rolled with firm pressure all over the penile shaft. Urethral swabs were collected by inserting a sterile swab into the urethral meatus and rotating clockwise for approximately five seconds. The first urethral swab was rolled on a slide and stained with a modified methylene blue stain to evaluate for the presence of urethritis. The second swab specimen was immediately placed into a sterile tube containing 3 ml of DNA preservative (GeneLock™, Sierra Molecular Corp., Sonora, CA) for DNA extraction and pyrosequencing analysis.

Molecular Methods

Extraction of DNA from swab specimens was performed using commercial kits according to the manufacturer's instructions. An initial bacterial cell lysis step using lysozyme (20 mg/ml at 37 ^{o}C for 1 hour) was included (QIAamp DNA micro kit for male, QIAamp DNA mini kit for female specimens, Qiagen Inc., Valencia, CA). DNA obtained from the coronal sulcus and penile shaft swabs was combined for the analyses of bacterial composition. Bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP) was performed by the Research and Testing Laboratory (Lubbock, TX) using broad-range PCR-amplification of the V4 -V6 region of the 16S rRNA gene with primers 530F: GTGCCAGCMGCNGCGG and 1100R: GGGTTNCGNTCGTTG. The bTEFAP post-sequencing filtering algorithm is described elsewhere (Dowd et al., 2008).

Extracting Gardnerella vaginalis Sequences and Alignment

Pyrosequencing analysis generated a total of 1,245,347 reads from 157 specimens. The average nucleotide length of these sequences in the original pyrosequencing library was 482.72, with the standard deviation of 71.80. The average number of sequences per sample was 6,257.45 with the standard deviation of 3,518.42. In order to identify and segregate best *G. vaginalis* matching reads from the rest of the sequences in the pyrosequencing library, a local database created with 28 unique, full-length *G. vaginalis* 16S rRNA sequences that are obtained from the Ribosomal Database Project (GenBank accession numbers: AY958811; AY958823; AY958846; AY958890; AY958892; AY958951; AY958974; AY958991; AY958998; AY959004; AY959013; AY959031; AY959038; AY959153; AY959171; AY959189; DQ316100; MZH0805-03; EF194095; AM696904; EF653408; AY958925; GQ179718; GQ179719; GQ179720; HQ114564; CP002104; CP001849). A BLAST (version 2.2.24, with *e* value of 1e-30) search against this local database was performed for each sequence in the original pyrosequencing library. A total of 70,569 sequences that were ≥ 99% homologous to at least one of the *G. vaginalis* sequences in the RDP database were kept to be used on the subsequent analyses. The average nucleotide length of these sequences was 508.66, with the standard deviation of 19.80. Some of the samples did not yield any *G. vaginalis* sequences with given criteria; therefore those samples were excluded from the analysis. Table 2 under the *Summary of Specimens After BLAST Filtering* subsection in Appendix, shows the summary of BLAST results. Finally, in order to analyze the positional homology along the columns, resulting *G. vaginalis* sequences were aligned to GreenGenes (DeSantis et al., 2006) gold standard 16S rRNA gene sequence template for *G. vaginalis*, using PyNAST (Caporaso, J. G., et al. 2010).

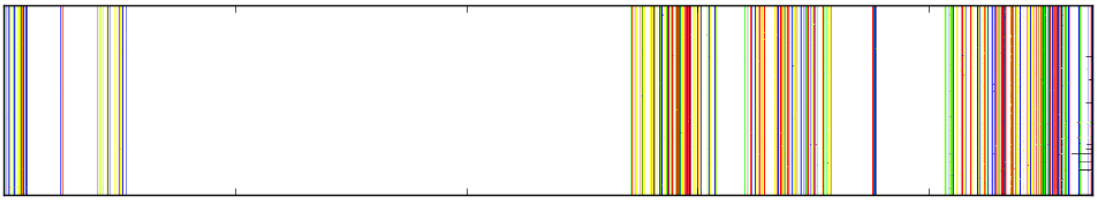

Figure 8. Visual representation of 500 aligned tag sequences that cover V4-V6 regions of 16S rRNA gene. Red, green, blue and yellow colors denote A, T, C and G bases respectively. White regions are gaps that were introduced by alignment process.

Figure 9. Shannon entropy analysis per column for 70,037 reads overlaid on the visual representation of aligned sequences. Peaks in entropy indicate nucleotide variation at given locations.

Identifying Variable Regions within Aligned Sequences and Generating Oligotypes

Since all sequences in this collection were similar to RDP sequences more than 99% identity, no or very little heterogeneity along the columns would be expected. Nevertheless, positional homology among aligned 16S rRNA gene tags that were classified as *G. vaginalis* showed consistent variation at certain locations that cannot be asserted to the caveats of pyrosequencing. Shannon entropy has been used to quantify the nucleotide variation along the columns of aligned *G. vaginalis* sequences and nine nucleotide positions that showed high variation in the V4-V6 region of *G. vaginalis* 16S rRNA gene were identified (figures 8, 9). Variable locations emerged from this analysis coincided with 2582nd, 3768th, 3788th, 3888th, 4408th, 4707th, 4709th, 4710th and 4714th nucleotide positions of the GreenGenes 16S rRNA gene alignment template. None of these positions were associated with

hompolymer regions, moreover, nucleotide variation at these locations also partially observed in full-length *G. vaginalis* 16S rRNA gene sequences obtained from the RDP database. 532 more sequences were eliminated because either they were not long enough to cover all variable locations or they had one or more ambiguous nucleotides at variable locations. 9-nucleotide long oligotypes generated by merging nucleotides from each variable location in tag sequences to represent *G. vaginalis*.

Creating Parsimony for Oligotypes and UniFrac Analysis of Samples

Total of 65 oligotypes, which were abundant more than 1% in any sample, analyzed with DNA parsimony algorithm (dnapars, version 3.67) from Phylogeny Inference Package (PHYLIP) (Felsenstein, 1989) to generate unrooted tree of parsimony from oligotypes extracted from *G. vaginalis* gene tag sequences (Figure 12). This tree has been used as a common phylogeny in UniFrac analysis (Lozupone & Knight, 2005) to calculate unique branch lengths of each sample laid on the tree to quantify similarities between samples based on their oligotype profiles. Hierarchical clustering analysis of vaginal, penile skin and urethra samples performed based on their UniFrac distance matrices (Figure 13). Both UniFrac clustering results, and the phylogenetic tree of oligotypes (Figure 12) were visualized via Interactive Tree of Life (Letunic & Bork, 2007).

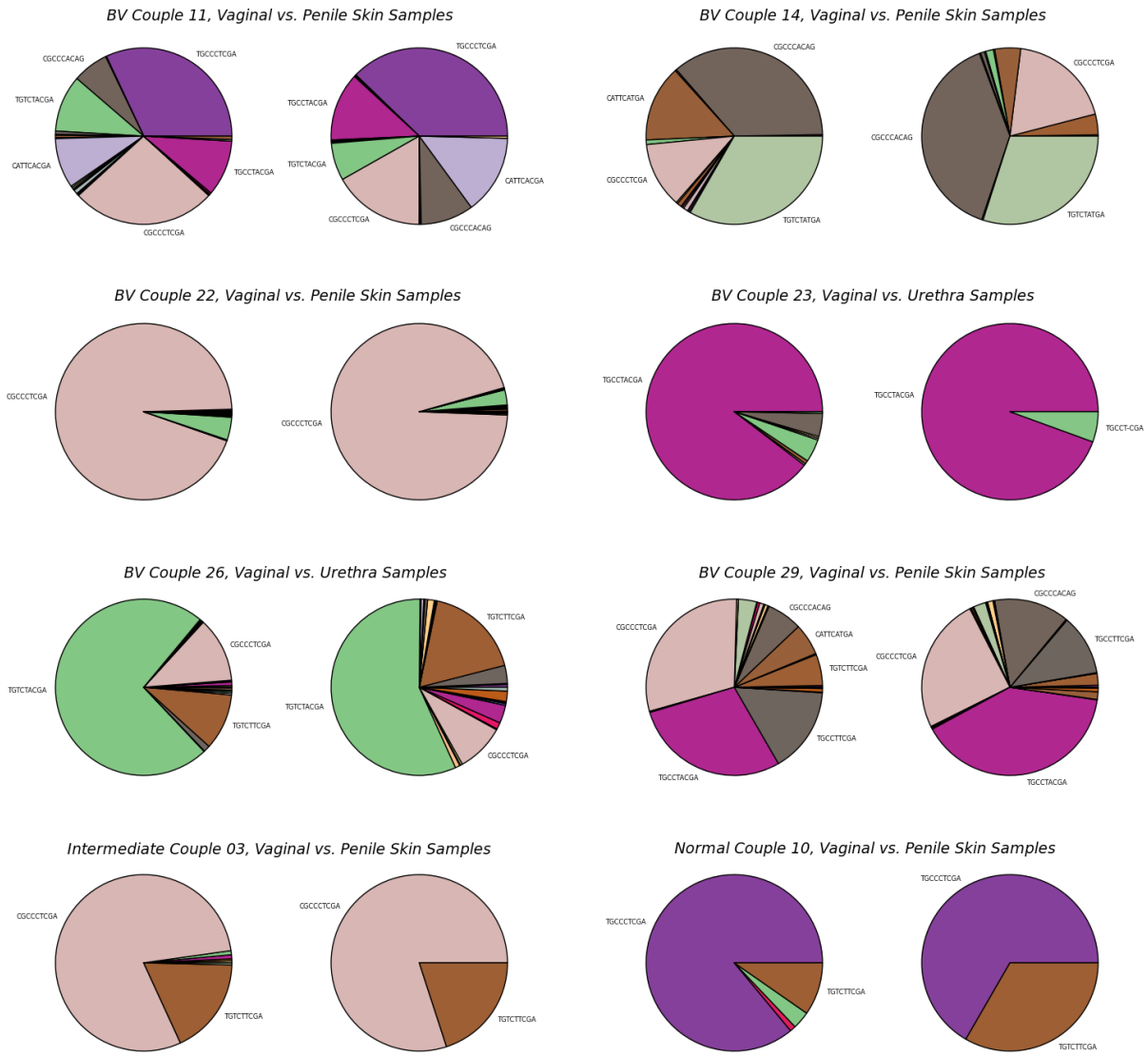

Figure 10. Oligotype profiles in various female patients and their sexual partners. Different colors in pie charts correspond to different oligotypes. Despite the number of different compositions among women, significant correlations between sexual partners supported the idea that *G. vaginalis* types are shared between sexual partners.

Results

By counting oligotypes in every sample it was possible to compare samples to each other based on the composition of *G. vaginalis* oligotypes they contained. The analyses revealed extensive diversity among *G. vaginalis* sequences from different samples, as well as significant correlations of oligotype profiles between some sexual partners.

Oligotype profiles of 24 of 44 total female patients, whose sexual partners yielded at least one *G. vaginalis* sequence, showed significant correlation ($r >= 0.9$, $p < 0.001$) with either the penile skin or urethral specimens of their partners. In an additional 8 couples there was a lesser, but still high degree of correlation ($r >= 0.5$, $p < 0.001$) between the vaginal and either the penile skin or the urethral sample of the partners (correlation values for every couple are listed in the *Pearson Correlation Table for G. vaginalis Oligotype Profiles Among Couples* subsection under Appendix).

Figure 10 shows 8 couples with oligotype profiles that are different from each other and similar between partners (more detailed information about the similarity of oligotype profiles available in in the *G. vaginalis Oligotype Profile Comparison Between Couples* subsection under Appendix).

Figure 11 shows the presence of major oligotypes among samples from different environments.

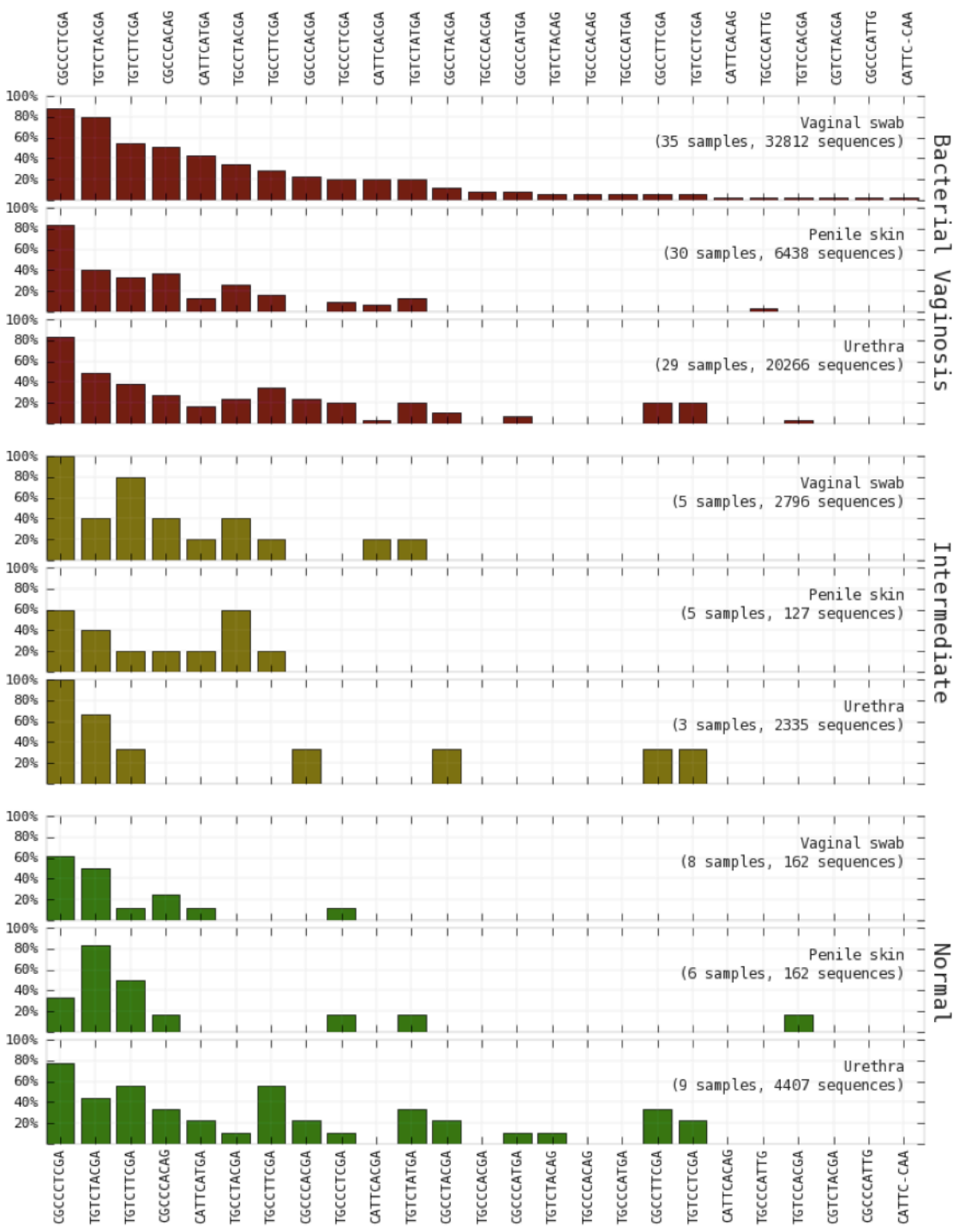

Figure 11. Oligotype distribution among samples. Bars show the percentage of samples in a group that has given oligotype at least once.

CGCCCTCGA was the most abundant oligotype and observed in most of the samples. It was the dominant oligotype of 28 vaginal samples out of 48 total. CGCCCACAG, TGTCTACGA and TGCCCTCGA were the dominant oligotypes in 10, 5 and 3 vaginal samples respectively. TGCCTTCGA was the dominant type in only 1 vaginal sample. TGCCTACGA was also the dominant oligotype in only one vaginal sample and the only urethra sample that had this type dominant, was her sexual partner (BV couple #23 in Figure 10). Correlation values for every couple are listed in Table 1, under the *Pearson Correlation Table for G. vaginalis Oligotype Profiles Among Couples* subsection in Appendix. Also, more detailed visualization of the oligotype profiles that are shared between sexual partners with high correlation is available under the *G. vaginalis Oligotype Profile Comparison Between Couples* subsection in Appendix.

Figure 12 shows the parsimony of 65 oligotypes that were present in any sample with more than 1% abundance. Presences of the oligotypes among sample groups are also listed in Table 3, under the *Oligotype Distribution Among Sample Groups* subsection in Appendix.

Hierarchical clustering analysis based on UniFrac distance matrix, which was populated by using the phylogenetic tree visualized in Figure 12, grouped oligotype profiles from normal women together and separated them from oligotype profiles of BV diagnosed women (Figure 13). Meanwhile neither oligotype profiles from urethra samples nor penile skin samples showed any grouping (Figure 13).

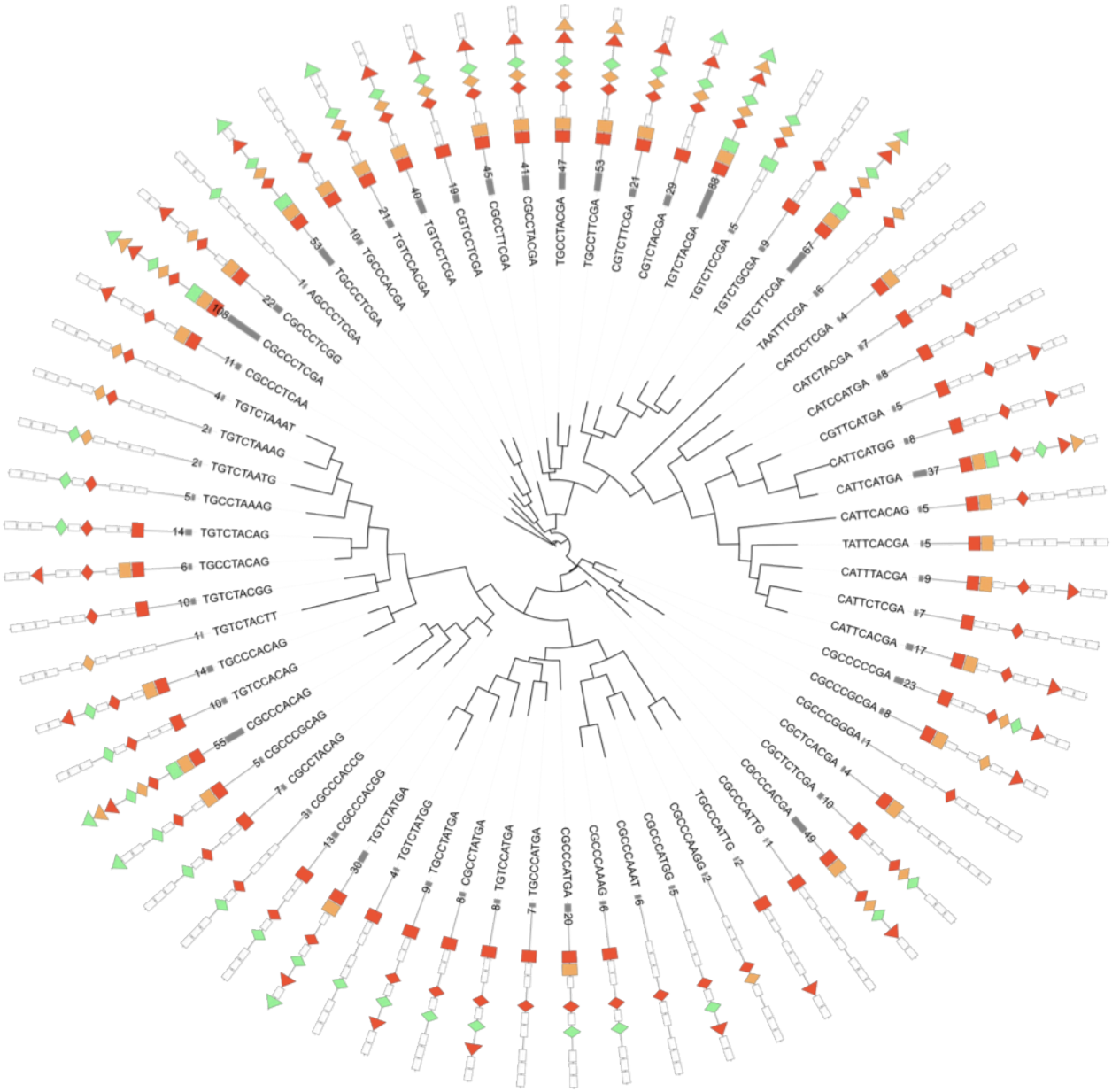

Figure 12. Parsimony of 65 oligotypes that were present in any sample with more than 1% abundance. Bars next to oligotypes indicate in how many samples they were present out of 130 total. Rectangles, diamonds and triangles denote the presence of a given oligotype in vaginal swab, urethra and penile skin samples, respectively. Red, orange and green colors indicate BV, intermediate and normal female patients and their sexual partners.

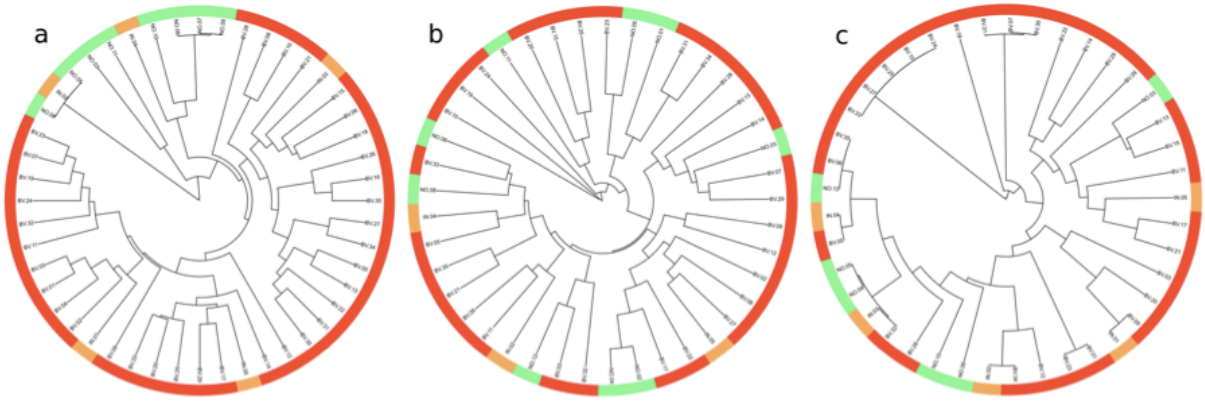

Figure 13. Hierarchical clustering of oligotype profiles from vaginal swab samples (a) (clustering significance: $p < 0.001$, UniFrac significance: $p = 0.016$), urethra samples (b) (clustering significance: $p < 0.001$, UniFrac significance: $p = 0.077$) and penile skin samples (c) (clustering significance: $p = 0.011$, UniFrac significance: $p = 0.001$) based on their UniFrac distances. Red, orange and green colors indicate samples from BV, intermediate and normal diagnoses, respectively.

Discussion

Different oligotypes inferred with Shannon entropy analysis from *G. vaginalis* sequences in this study actually correspond to 0.19% to 1.77% variation in whole length V4-V6 reads. Nevertheless, different *G. vaginalis* types that are consistent among couples were still traceable despite the incomplete information from V4-V6 region of 16S rRNA gene (see Figure 10).

Bacterial species may have more than one copy of 16S rRNA gene in their genomes, and different copies of 16S rRNA genes may differ from each other with more than 1% nucleotide similarity (Klappenbach, Saxman, Cole, & Schmidt, 2001). *G. vaginalis* have 2 copies of 16S rRNA gene in its genome (Nath, Chen, Ahn, & Chen, 2000), and some of the diversity that is being observed within *G. vaginalis* could be ascribable to the copy number. But the number of oligotypes revealed by this analysis is much more than the amount that could be explained with variation between different copies.

It has been shown that even phylogenetically very similar bacterial species may exhibit different ecological properties and species with identical 16S rRNA genes may represent different ecotypes and virulence properties (Jaspers & Overmann, 2004).

In this sense, sensitivity of 16S rRNA gene is limited. But it is specific: there is a relationship between 16S rRNA gene divergence and overall gene content (Konstantinidis & Tiedje, 2007). Also a recent study from Knight et al confirms the statistically significant correlation between 16S rRNA gene evolutionary distances and gene content conservation of species at the protein level where genomic differences could be traceable even with short pyrosequencing tag reads from V2 or V4-V6 regions of 16S rRNA gene (Zaneveld, Lozupone, Gordon, & Knight, 2010).

Ward has shown that even one nucleotide difference may indicate ecologically distinct strains, which would be overlooked by an arbitrary value, e.g. >97% 16S rRNA gene similarity to assign taxonomy (Ward, 1998).

Even though deep sequencing help collecting sequence data to cover a great deal of diversity within a microbial population, it does not provide enough resolution for mainstream tools to infer diversity at lower levels of taxonomy. This study presents a computational approach to explore diversity at an ecologically meaningful level that could be used to investigate potential ecotypes within a conventional group of species.

CONCLUSION

Microbial ecology is a multi-disciplinary research area in the juncture of microbiology, ecology, bioinformatics, and computational statistics.

Through the assistance of high-throughput sequencing technologies, analyses of 16S rRNA gene sequences from environmental samples provide new insights to the extraordinary richness of the bacterial world, and their relationships with the environment.

Handling vast amount of sequence data from bacterial communities is a nontrivial task and compels researchers to rely heavily on computational tools for analysis. Although there are a variety of tools currently available for metagenomic sequence analysis, they impose unnatural paradigms or restrictive limitations on biological researchers who may have only rudimentary computer skills. In this study, an easy-to-use and extensible software framework that was designed and developed to empower researchers to do their own analyses has been presented. The framework (Viamics) is currently is being used in multiple human microbiome studies.

Diversity assessment is also a major challenge in microbial ecology. Computationally feasible but ecologically calamitous approaches that are being widely used today in order to analyze 16S rRNA gene sequences from environmental samples lead to a considerable amount of diversity to be overlooked. The novel method that has been presented in this study revealed a remarkable amount of diversity within *Gardnerella vaginalis* by using

partial sequences from 16S rRNA gene. Results of this study suggested that biologically relevant computational approaches for microbial community analyses might offer more accurate ways to address missing microbial diversity.

EPILOGUE

The time I spent in microbial ecology as a bioinformatist helped me to develop an understanding of some of the fundamental challenges in this multi-disciplinary field. Some of these challenges could be addressed by computational biology and bioinformatics, while some of them are solely physical or biological obstacles. I would like to close by mentioning them for future references.

Diversity Assessment

The NIH Human Microbiome initiative (Group et al., 2009) has ignited a great deal of interest in microbial ecology, and in the era of metagenomic analyses made possible by the advent of next generation sequencing, knowledge of the diversity, ecology and symbiotic relationship between microbes and their human hosts will undoubtedly continue to accumulate at a rapid pace (Turnbaugh et al., 2007). However progress in microbial ecology may be limited by classic ecological assumptions and currently available computational and statistical approaches, almost all of which require sequence data to be organized and somewhat arbitrarily grouped into taxonomic units, such as OTUs, species or phyla prior to being analyzed.

A common convention is to group 16S rRNA gene sequences that are similar to each other by ≥ 97% into the same species (Stackebrandt & Goebel, 1994), and sequences that are ≥ 85% similar into the same genus, and ~60-70% similar into the same phylum (Schloss & Handelsman, 2006). Although a great deal of insight can be gleaned from the

study of groups of similar sequences, microbial ecologists realize that a significant amount of genetic diversity is being masked in the process (Keswani & Whitman, 2001). However, the use of arbitrary amounts of sequence differences to group bacteria into OTUs is still a common practice due to its practicality and convenience from the computational perspective. Ideally, genetic variability within a microbial community would be based on biologically relevant parameters, instead of arbitrary cutoffs. New computational and statistical methods, and new layers of numerical abstractions to analyze complex sequence data, are needed to gain a more realistic understanding of the ecology of microbial communities.

The chapter in this dissertation describing the analysis of *G. vaginalis* oligotypes in sexually active couples is a simple example of this process. The study demonstrates how modifying the computational perspective based on ecologically meaningful insights can reveal previously unrecognized patterns in diversity, even within a well-studied species.

Sampling

One major problem in microbial ecology emerges from the fact that operational scale of sampling media is not comparable with the scale where microbial communities reside. An ecologist, who uses conventional techniques for collecting data, has the luxury to select and observe a population and their phenotypic responses to changing environmental conditions in order to discuss broader ecological phenomena (Ozgul et al., 2010), meanwhile microbial ecologists have vast scale differences between them and the microbial communities they wish to observe.

Sampling of a microbial community, for instance with a swab, takes place at a several orders of magnitude larger scales while even microns change a lot when considering the bilateral interactions between microbial communities and their environment (Young, Crawford, Nunan, Otten, & Spiers, 2008). This is a step where all subtle gradients and natural patterns are being missed, and downstream analyses might be heavily affected.

Rare Species

High throughput pyrosequencing analyses of bacterial populations in environmental samples using PCR-amplified 16S rRNA gene sequences reveal an enormous diversity of microbes (Sogin et al., 2006). The maximum depth of sampling coverage of microbial species that can be attained via these technologies will likely to be increased in the future (Lazarevic et al., 2009; Caporaso et al., 2010c). However, there are still very rare members of microbial communities that create a long tail in the frequency count distribution curves of species. Even though they may have a strong impact on the ecological function of the community (Pester et al., 2010; Hol et al., 2010), unfortunately, it is still a cumbersome task to investigate the impact and function of these rare species, especially using whole genome and biochemical studies, and even using the 16S rRNA gene tag-based approach.

Because currently it is not practically feasible to gather enough sequence data from the rare members of communities to perform functional surveys, computational approaches may play a key role to infer rare members' impact on community's functional diversity. However, almost all available computational methods tend to ascribe the statistical impact of a member in proportion to its relative abundance in the environment. In some cases, this

may reflect little ecological relevance, and well skew our perceptions of the importance of rare microbial species in nature.

Species Concept

Perhaps the most fundamentally important barrier to progress in microbial ecology is the lack of a clear definition or rational concept of a bacterial species. As mentioned above, this is a fundamental problem that dramatically affects the efficiency of computational methods and statistical inferences. The modern biological species concept (de Queiroz, 2005) of macro-organisms does not apply to archea and bacteria, since they are essentially asexual (even though they can exchange small segment of DNA through conjugation) and the few observable phenotypic traits they offer, such as cell size, shape, motility, are insufficient for meaningful classification. The species concept in microbiology is still an open discussion and concerns about the necessity of a more functional definition have been voiced repeatedly in the scientific literature (Ward, 1998; Gevers et al., 2005; Cohan & Perry, 2007; Ward et al., 2008). However, how exactly microbial species contribute to ecosystem function, is not clear from a broader ecological perspective for given definitions.

Hubbell's controversial work "A unified neutral theory of biodiversity and biogeography" (Hubbell, 2001) presents an interesting explanation for species assembly and distribution in ecological communities. Hubbell hypothesizes that the interactions among species are assumed to be equivalent on an individual *per capita* basis, implying that biodiversity is regulated by random events, every species in a saturated ecological community perform a *random walk* and phenotypical traits have no impact on their overall abundance and therefore their survival. Unlike the classical niche assembly theory that

contends that species live together and form a community only if they differ from each other in terms of resource uses, Hubbell diminishes the importance of species differences. Hubbell's hypothesis is based on two basic principles exploited from two biological observations: (1) "neutral speciation", which is supported by the observations that indicate that different individuals from different species belonging to the same trophic level appear to be controlled by similar birth, death and dispersal rates, (2) "zero-sum dynamics", which is supported by the observation that indicates that ecological systems are saturated (Alonso, Etienne, & McKane, 2006). The literature offers comprehensive studies that emphasize the merits (Alonso et al., 2006) and failures (McGill, Maurer, & Weiser, 2006; Dornelas, Connolly, & Hughes, 2006) of the theory at explaining biodiversity and species distribution. Discussing Hubbell's theory in detail is outside of the context of this dissertation, however, I believe it is important to consider that, even if a more natural way of defining species would emerge from microbial ecology, it might not be enough to fully explain ecological dynamics of microbial communities, especially if microbial ecology follows the dynamics of conventional ecology, which is by itself another area of research.

Studies suggest that neutral theories can be applicable to microbial communities to a degree. It has been demonstrated that stochastic neutral community models could explain the patterns in bacterial community structure using the relative abundances of 16S rRNA gene sequences (Sloan et al., 2006). On the other hand, two recently published studies that attempt to explain microbial community assembly processes, suggest that multi-trophic level bacterial community assembly is likely to be influenced by both neutral, and niche assembly rules in a stochastic manner (Langenheder & Székely, 2011; Caruso et al., 2011).

From a different perspective, several studies show that bacterial communities composed of different species may occupy similar environmental niches and exhibit similar metabolic functions. This implies that ecosystem functioning of a community might be related even though bacterial species composition is quite different (Langenheder, Lindström, & Tranvik, 2005). This idea is supported to an extent by a metagenomic study of three different microbial communities that developed in deep ocean environments around the nutrient-rich carcasses of whales that fall to the ocean floor (Tringe et al., 2005) which demonstrated that the microbial communities from similar environments with similar metabolic demands are essentially the same in terms of protein function, even though taxa diversity and abundance are markedly different.

Seeking for a meaningful definition of species to group prokaryotes is being challenged from more than one front and it is not certain that finding the optimum species definition is going to decrypt and fully explain seemingly stochastic patterns of community diversity and assembly.

Taxonomy and all other categorizations are the result of innate human intuition to group things that are similar; and human intuition, as Richard Feynman beautifully put in his 1986 book, *QED: The Strange Theory of Light and Matter,* may not be the best guide to understand nature, especially when observing little things:

> *"The theory of quantum electrodynamics describes Nature as absurd from the point of view of common sense. And it fully agrees with experiment. So I hope you can accept Nature as She is—Absurd"*

Today, almost all microbial ecology studies adhere to the hierarchical classification system of grouping bacteria at the species, genus, family, order and phylum levels, or alternatively, defining OTUs based on some arbitrary degree of genetic sequence similarity.

Computers are better at working with gradients and data that does not exhibit solid borders. Perhaps it is time for computational biology to take a lead role in developing a new approach to define and assess microbial diversity using a taxon-free inference that accounts for the entirety of genetic variation.

REFERENCES

Alonso, D., Etienne, R. S., & McKane, A. J. (2006). The merits of neutral theory. *Trends Ecol Evol (Amst)*, *21*(8), 451-457. doi:10.1016/j.tree.2006.03.019

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, *215*(3), 403-410. doi:10.1006/jmbi.1990.9999

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev*, *59*(1), 143-169.

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D. et al. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet*, *40*(8), 955-962. doi:10.1038/ng.175

Baumgart, D. C., & Sandborn, W. J. (2007). Inflammatory bowel disease: clinical aspects and established and evolving therapies. *Lancet*, *369*(9573), 1641-1657. doi:10.1016/S0140-6736(07)60751-X

Berg, R. D. (1996). The indigenous gastrointestinal microflora. *Trends Microbiol*, *4*(11), 430-435.

Biasucci, G., Benenati, B., Morelli, L., Bessi, E., & Boehm, G. (2008). Cesarean delivery may affect the early biodiversity of intestinal bacteria. *J Nutr*, *138*(9), 1796S-1800S.

Binladen, J., Gilbert, M. T., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R. et al. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, *2*(2), e197. doi:10.1371/journal.pone.0000197

Björkholm, B., Bok, C. M., Lundin, A., Rafter, J., Hibberd, M. L., & Pettersson, S. (2009). Intestinal microbiota regulate xenobiotic metabolism in the liver. *PLoS ONE*, *4*(9), e6958. doi:10.1371/journal.pone.0006958

Brown, J. R. (2003). Ancient horizontal gene transfer. *Nat Rev Genet*, *4*(2), 121-132. doi:10.1038/nrg1000

Bunge, J. (2011). Estimating the number of species with catchall. *Pac Symp Biocomput,* 121-130.

Bunge, J., & Barger, K. (2008). Parametric models for estimating the number of classes. *Biom J*, *50*(6), 971-982. doi:10.1002/bimj.200810452

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., & Knight, R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, *26*(2), 266-267. doi:10.1093/bioinformatics/btp636

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K. et al. (2010b). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*, *7*(5), 335-336. doi:10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J. et al. (2010c). Microbes and Health Sackler Colloquium: Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA*. doi:10.1073/pnas.1000080107

Caruso, T., Chan, Y., Lacap, D. C., Lau, M. C., McKay, C. P., & Pointing, S. B. (2011). Stochastic and deterministic processes interact in the assembly of desert microbial communities on a global scale. *ISME J*. doi:10.1038/ismej.2011.21

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*.

Chao, A., & Lee, S. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*.

Cohan, F. M., & Perry, E. B. (2007). A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol*, *17*(10), R373-86. doi:10.1016/j.cub.2007.03.032

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J. et al. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res*, *37*(Database issue), D141-5. doi:10.1093/nar/gkn879

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., & Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science*, *326*(5960), 1694-1697. doi:10.1126/science.1177486

de Queiroz, K. (2005). Ernst Mayr and the modern concept of species. *Proc Natl Acad Sci USA*, *102 Suppl 1*, 6600-6607. doi:10.1073/pnas.0502030102

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K. et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*, *72*(7), 5069-5072. doi:10.1128/AEM.03006-05

Desbonnet, L., Garrett, L., Clarke, G., Bienenstock, J., & Dinan, T. G. (2008). The probiotic Bifidobacteria infantis: An assessment of potential antidepressant properties in the rat. *J Psychiatr Res*, *43*(2), 164-174. doi:10.1016/j.jpsychires.2008.03.009

Dethlefsen, L., Huse, S., Sogin, M. L., & Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol*, *6*(11), e280. doi:10.1371/journal.pbio.0060280

Dethlefsen, L., McFall-Ngai, M., & Relman, D. A. (2007). An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature*, *449*(7164), 811-818. doi:10.1038/nature06245

Dornelas, M., Connolly, S. R., & Hughes, T. P. (2006). Coral reef diversity refutes the neutral theory of biodiversity. *Nature*, *440*(7080), 80-82. doi:10.1038/nature04534

Dowd, S. E., Callaway, T. R., Wolcott, R. D., Sun, Y., McKeehan, T., Hagevoort, R. G. et al. (2008). Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP). *BMC Microbiol*, *8*, 125. doi:10.1186/1471-2180-8-125

Emami, C. N., Petrosyan, M., Giuliani, S., Williams, M., Hunter, C., Prasadarao, N. V. et al. (2009). Role of the host defense system and intestinal microbial flora in the pathogenesis of necrotizing enterocolitis. *Surg Infect (Larchmt)*, *10*(5), 407-417. doi:10.1089/sur.2009.054

Eren, A. M., Ferris, M. J., & Taylor, C. M. (2011). A framework for analysis of metagenomic sequencing data. *Pac Symp Biocomput,* 131-141. doi:10.1142/9789814335058_0015

Fakhrai-Rad, H., Pourmand, N., & Ronaghi, M. (2002). Pyrosequencing: an accurate detection platform for single nucleotide polymorphisms. *Hum Mutat*, *19*(5), 479-485. doi:10.1002/humu.10078

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, *5*, 164-166.

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J. et al. (2005). Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol*, *3*(9), 733-739. doi:10.1038/nrmicro1236

Goodacre, R. (2007). Metabolomics of a superorganism. *J Nutr*, *137*(1 Suppl), 259S-266S.

Group, N. I. H. H. M. P. W., Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L. et al. (2009). The NIH Human Microbiome Project. *Genome Res*, *19*(12), 2317-2323. doi:10.1101/gr.096651.109

Hamady, M., Lozupone, C., & Knight, R. (2010). Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J*, *4*(1), 17-27. doi:10.1038/ismej.2009.97

Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J., & Knight, R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods*, *5*(3), 235-237. doi:10.1038/nmeth.1184

Harwich, M. D., Alves, J. M., Buck, G. A., Strauss, J. F., Patterson, J. L., Oki, A. T. et al. (2010). Drawing the line between commensal and pathogenic Gardnerella vaginalis through genome analysis and virulence studies. *BMC Genomics*, *11*, 375. doi:10.1186/1471-2164-11-375

Heijtz, R. D., Wang, S., Anuar, F., Qian, Y., Björkholm, B., Samuelsson, A. et al. (2011). Normal gut microbiota modulates brain development and behavior. *Proc Natl Acad Sci USA*, *108*(7), 3047-3052. doi:10.1073/pnas.1010529108

Hill, M. O. (1974). Correspondence analysis: a neglected multivariate method. *Applied Statistics*.

Hill, T. C., Walsh, K. A., Harris, J. A., & Moffett, B. F. (2003). Using ecological diversity measures with bacterial communities. *FEMS Microbiol Ecol*, *43*(1), 1-11. doi:10.1111/j.1574-6941.2003.tb01040.x

Hillier, S. L., Nugent, R. P., Eschenbach, D. A., Krohn, M. A., Gibbs, R. S., Martin, D. H. et al. (1995). Association between bacterial vaginosis and preterm delivery of a low-birth-weight infant. The Vaginal Infections and Prematurity Study Group. *N Engl J Med*, *333*(26), 1737-1742. doi:10.1056/NEJM199512283332604

Hol, W. H., de Boer, W., Termorshuizen, A. J., Meyer, K. M., Schneider, J. H., van Dam, N. M. et al. (2010). Reduction of rare soil microbes modifies plant-herbivore interactions. *Ecol Lett*, *13*(3), 292-301. doi:10.1111/j.1461-0248.2009.01424.x

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography.* Princeton University Press.

Hughes, J. B., Hellmann, J. J., Ricketts, T. H., & Bohannan, B. J. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol*, *67*(10), 4399-4406.

Huse, S. M., Dethlefsen, L., Huber, J. A., Mark Welch, D., Welch, D. M., Relman, D. A. et al. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet*, *4*(11), e1000255. doi:10.1371/journal.pgen.1000255

Huse, S. M., Welch, D. M., Morrison, H. G., & Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol*, *12*(7), 1889-1898. doi:10.1111/j.1462-2920.2010.02193.x

Ichinohe, T., Pang, I. K., Kumamoto, Y., Peaper, D. R., Ho, J. H., Murray, T. S. et al. (2011). Microbiota regulates immune defense against respiratory tract influenza A virus infection. *Proc Natl Acad Sci USA*. doi:10.1073/pnas.1019378108

Jaspers, E., & Overmann, J. (2004). Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Appl Environ Microbiol*, *70*(8), 4831-4839. doi:10.1128/AEM.70.8.4831-4839.2004

Keswani, J., Orkand, S., Premachandran, U., Mandelco, L., Franklin, M. J., & Whitman, W. B. (1996). Phylogeny and taxonomy of mesophilic Methanococcus spp. and comparison of rRNA, DNA hybridization, and phenotypic methods. *Int J Syst Bacteriol*, *46*(3), 727-735.

Keswani, J., & Whitman, W. B. (2001). Relationship of 16S rRNA sequence similarity to DNA hybridization in prokaryotes. *Int J Syst Evol Microbiol*, *51*(Pt 2), 667-678.

King, C. R., & Scott-Horton, T. (2007). Pyrosequencing: a simple method for accurate genotyping. *Methods Mol Biol*, *373*, 39-56.

Klappenbach, J. A., Saxman, P. R., Cole, J. R., & Schmidt, T. M. (2001). rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Res*, *29*(1), 181-184.

Konstantinidis, K. T., & Tiedje, J. M. (2007). Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol*, *10*(5), 504-509. doi:10.1016/j.mib.2007.08.006

Kozyrskyj, A. L., Ernst, P., & Becker, A. B. (2007). Increased risk of childhood asthma from antibiotic use in early life. *Chest*, *131*(6), 1753-1759. doi:10.1378/chest.06-3008

Kunin, V., Engelbrektson, A., Ochman, H., & Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol*, *12*(1), 118-123. doi:10.1111/j.1462-2920.2009.02051.x

Langenheder, S., Lindström, E. S., & Tranvik, L. J. (2005). Weak Coupling between Community Composition and Functioning of Aquatic Bacteria. *Limnology and Oceanography*, *50:3*, 957-967.

Langenheder, S., & Székely, A. J. (2011). Species sorting and neutral processes are both important during the initial assembly of bacterial communities. *ISME J*. doi:10.1038/ismej.2010.207

Larsson, P. G., Bergström, M., Forsum, U., Jacobsson, B., Strand, A., & Wölner-Hanssen, P. (2005). Bacterial vaginosis. Transmission, role in genital tract infection and pregnancy outcome: an enigma. *APMIS*, *113*(4), 233-245. doi:10.1111/j.1600-0463.2005.apm_01.x

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Osterås, M. et al. (2009). Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods*, *79*(3), 266-271. doi:10.1016/j.mimet.2009.09.012

Lederberg, J. (2000). Infectious history. *Science*, *288*(5464), 287-293.

Letunic, I., & Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, *23*(1), 127-128. doi:10.1093/bioinformatics/btl529

Levcopoulos, C. (1998). Fast algorithms for complete linkage clustering. *Discrete and Computational Geometry*.

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., & Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proc Natl Acad Sci USA*, *102*(31), 11070-11075. doi:10.1073/pnas.0504978102

Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D., & Knight, R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res*, *35*(18), e120. doi:10.1093/nar/gkm541

Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*, *71*(12), 8228-8235. doi:10.1128/AEM.71.12.8228-8235.2005

Mackie, R. I., Sghir, A., & Gaskins, H. R. (1999). Developmental microbial ecology of the neonatal gastrointestinal tract. *Am J Clin Nutr*, *69*(5), 1035S-1045S.

Malde, K. (2011). Flower: extracting information from pyrosequencing data. *Bioinformatics*. doi:10.1093/bioinformatics/btr063

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A. et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376-380. doi:10.1038/nature03959

Marrazzo, J. M., Koutsky, L. A., Eschenbach, D. A., Agnew, K., Stine, K., & Hillier, S. L. (2002). Characterization of vaginal flora and bacterial vaginosis in women who have sex with women. *J Infect Dis*, *185*(9), 1307-1313. doi:10.1086/339884

Martin, A. P. (2002). Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol*, *68*(8), 3673-3682.

Mashayekhi, F., & Ronaghi, M. (2007). Analysis of read length limiting factors in Pyrosequencing chemistry. *Anal Biochem*, *363*(2), 275-287. doi:10.1016/j.ab.2007.02.002

Mazmanian, S. K., Liu, C. H., Tzianabos, A. O., & Kasper, D. L. (2005). An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell*, *122*(1), 107-118. doi:10.1016/j.cell.2005.05.007

McGill, B. J., Maurer, B. A., & Weiser, M. D. (2006). Empirical evaluation of neutral theory. *Ecology*, *87*(6), 1411-1423.

Meyerhans, A., Vartanian, J. P., & Wain-Hobson, S. (1990). DNA recombination during PCR. *Nucleic Acids Res*, *18*(7), 1687-1691.

Min Jou, W., Haegeman, G., Ysebaert, M., & Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, *237*(5350), 82-88.

Nath, K., Chen, X., Ahn, K. S., & Chen, S. (2000). Characterization of the 16S rRNA gene V2 region and the rrn operons of Gardnerella vaginalis. *Res Microbiol*, *151*(9), 747-754.

Nugent, R. P., Krohn, M. A., & Hillier, S. L. (1991). Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *J Clin Microbiol*, *29*(2), 297-301.

Numanović, F., Hukić, M., Nurkić, M., Gegić, M., Delibegović, Z., Imamović, A. et al. (2008). Importance of isolation and biotypization of Gardnerella vaginalis in diagnosis of bacterial vaginosis. *Bosn J Basic Med Sci*, *8*(3), 270-276.

Nyrén, P. (2007). The history of pyrosequencing. *Methods Mol Biol*, *373*, 1-14.

Oakley, B. B., Fiedler, T. L., Marrazzo, J. M., & Fredricks, D. N. (2008). Diversity of human vaginal bacterial communities and associations with clinically defined bacterial vaginosis. *Appl Environ Microbiol*, *74*(15), 4898-4909. doi:10.1128/AEM.02884-07

Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R., & Stahl, D. A. (1986). Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol*, *40*, 337-365. doi:10.1146/annurev.mi.40.100186.002005

Olsen, G. J., & Woese, C. R. (1993). Ribosomal RNA: a key to phylogeny. *FASEB J*, *7*(1), 113-123.

Ozgul, A., Childs, D. Z., Oli, M. K., Armitage, K. B., Blumstein, D. T., Olson, L. E. et al. (2010). Coupled dynamics of body mass and population growth in response to environmental change. *Nature*, *466*(7305), 482-485. doi:10.1038/nature09210

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, *276*(5313), 734-740.

Pester, M., Bittner, N., Deevong, P., Wagner, M., & Loy, A. (2010). A 'rare biosphere' microorganism contributes to sulfate reduction in a peatland. *ISME J*, *4*(12), 1591-1602. doi:10.1038/ismej.2010.75

Petrosino, J. F., Highlander, S., Luna, R. A., Gibbs, R. A., & Versalovic, J. (2009). Metagenomic pyrosequencing and microbial identification. *Clin Chem*, *55*(5), 856-866. doi:10.1373/clinchem.2008.107565

Piot, P., Van Dyck, E., Peeters, M., Hale, J., Totten, P. A., & Holmes, K. K. (1984). Biotypes of Gardnerella vaginalis. *J Clin Microbiol*, *20*(4), 677-679.

Quince, C., Lanzén, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M. et al. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods*, *6*(9), 639-641. doi:10.1038/nmeth.1361

Quince, C., Lanzen, A., Davenport, R. J., & Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, *12*(1), 38. doi:10.1186/1471-2105-12-38

Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol*, *62*(2), 142-160. doi:10.1111/j.1574-6941.2007.00375.x

Reeder, J., & Knight, R. (2009). The 'rare biosphere': a reality check. *Nat Methods*, *6*(9), 636-637. doi:10.1038/nmeth0909-636

Reeder, J., & Knight, R. (2010). Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat Methods*, *7*(9), 668-669. doi:10.1038/nmeth0910-668b

Rothschild, L. J., & Mancinelli, R. L. (2001). Life in extreme environments. *Nature*, *409*(6823), 1092-1101. doi:10.1038/35059215

Sanders, H. L. (1968). Marine benthic diversity: a comparative study. *American Naturalist*, *102*, 243-282.

Sandoval, D. A., & Seeley, R. J. (2010). Medicine. The microbes made me eat it. *Science*, *328*(5975), 179-180. doi:10.1126/science.1188876

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, *74*(12), 5463-5467.

Savage, D. C. (1977). Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol*, *31*, 107-133. doi:10.1146/annurev.mi.31.100177.000543

Scanlan, P. D., Shanahan, F., O'Mahony, C., & Marchesi, J. R. (2006). Culture-independent analyses of temporal variation of the dominant fecal microbiota and targeted bacterial subgroups in Crohn's disease. *J Clin Microbiol*, *44*(11), 3980-3988. doi:10.1128/JCM.00312-06

Schloss, P. D., & Handelsman, J. (2006). Toward a census of bacteria in soil. *PLoS Comput Biol*, *2*(7), e92. doi:10.1371/journal.pcbi.0020092

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B. et al. (2009). Introducing mothur: open-source, platform-independent, community-supported

software for describing and comparing microbial communities. *Appl Environ Microbiol*, *75*(23), 7537-7541. doi:10.1128/AEM.01541-09

Sears, C. L. (2005). A dynamic partnership: celebrating our gut flora. *Anaerobe*, *11*(5), 247-251. doi:10.1016/j.anaerobe.2005.05.001

Sekirov, I., & Finlay, B. B. (2006). Human and microbe: united we stand. *Nat Med*, *12*(7), 736-737. doi:10.1038/nm0706-736

Sha, B. E., Zariffard, M. R., Wang, Q. J., Chen, H. Y., Bremer, J., Cohen, M. H. et al. (2005). Female genital-tract HIV load correlates inversely with Lactobacillus species but positively with bacterial vaginosis and Mycoplasma hominis. *J Infect Dis*, *191*(1), 25-32. doi:10.1086/426394

Sibley, C. G., & Ahlquist, J. E. (1984). The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *J Mol Evol*, *20*(1), 2-15.

Sidhu, H., Allison, M. J., Chow, J. M., Clark, A., & Peck, A. B. (2001). Rapid reversal of hyperoxaluria in a rat model after probiotic administration of Oxalobacter formigenes. *J Urol*, *166*(4), 1487-1491.

Simberloff, D. (1978). Use of rarefaction and related methods in ecology. In C. J. Dickson K.L., Livingston R.J. (Ed.), *Biological Data in Water Pollution Assessment: Quantitative and Statistical Analyses* (pp. 150–165).

Simpson, E. H. (1949). Measurement of diversity. *Nature*.

Sloan, W. T., Lunn, M., Woodcock, S., Head, I. M., Nee, S., & Curtis, T. P. (2006). Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ Microbiol*, *8*(4), 732-740. doi:10.1111/j.1462-2920.2005.00956.x

Sneath, P. H. (1993). Evidence from Aeromonas for genetic crossing-over in ribosomal sequences. *Int J Syst Bacteriol*, *43*(3), 626-629.

Socransky, S. S., Haffajee, A. D., Smith, C., Martin, L., Haffajee, J. A., Uzel, N. G. et al. (2004). Use of checkerboard DNA-DNA hybridization to study complex microbial ecosystems. *Oral Microbiol Immunol*, *19*(6), 352-362. doi:10.1111/j.1399-302x.2004.00168.x

Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R. et al. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci USA*, *103*(32), 12115-12120. doi:10.1073/pnas.0605127103

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol*, *98*(3), 503-517.

Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int J Syst Bacteriol*, *44*, 846-849.

Swidsinski, A., Mendling, W., Loening-Baucke, V., Ladhoff, A., Swidsinski, S., Hale, L. P. et al. (2005). Adherent biofilms in bacterial vaginosis. *Obstet Gynecol*, *106*(5 Pt 1), 1013-1023. doi:10.1097/01.AOG.0000183594.45524.d2

Swidsinski, A., Doerffel, Y., Loening-Baucke, V., Swidsinski, S., Verstraelen, H., Vaneechoutte, M. et al. (2010). Gardnerella biofilm involves females and males and is transmitted sexually. *Gynecol Obstet Invest*, *70*(4), 256-263. doi:10.1159/000314015

Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W. et al. (2005). Comparative metagenomics of microbial communities. *Science*, *308*(5721), 554-557. doi:10.1126/science.1107851

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, *449*(7164), 804-810. doi:10.1038/nature06244

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., & Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, *444*(7122), 1027-1031. doi:10.1038/nature05414

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*, *73*(16), 5261-5267. doi:10.1128/AEM.00062-07

Wang, Y., Hoenig, J. D., Malin, K. J., Qamar, S., Petrof, E. O., Sun, J. et al. (2009). 16S rRNA gene-based analysis of fecal microbiota from preterm infants with and without necrotizing enterocolitis. *ISME J*, *3*(8), 944-954. doi:10.1038/ismej.2009.37

Ward, D. M. (1998). A natural species concept for prokaryotes. *Curr Opin Microbiol*, *1*(3), 271-277.

Ward, D. M., Cohan, F. M., Bhaya, D., Heidelberg, J. F., Kühl, M., & Grossman, A. (2008). Genomics, environmental genomics and the issue of microbial species. *Heredity*, *100*(2), 207-219. doi:10.1038/sj.hdy.6801011

Warnecke, F., & Hess, M. (2009). A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. *J Biotechnol*, *142*(1), 91-95. doi:10.1016/j.jbiotec.2009.03.022

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA*, *95*(12), 6578-6583.

Wilmes, P., & Bond, P. L. (2006). Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol*, *14*(2), 92-97. doi:10.1016/j.tim.2005.12.006

Wilson, D. S., & Sober, E. (1989). Reviving the superorganism. *J Theor Biol*, *136*(3), 337-356.

Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev*, *51*(2), 221-271.

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA*, *74*(11), 5088-5090.

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA*, *87*(12), 4576-4579.

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA*, *87*(12), 4576-4579.

Yeoman, C. J., Yildirim, S., Thomas, S. M., Durkin, A. S., Torralba, M., Sutton, G. et al. (2010). Comparative genomics of Gardnerella vaginalis strains reveals substantial differences in metabolic and virulence potential. *PLoS ONE*, *5*(8), e12411. doi:10.1371/journal.pone.0012411

Zaneveld, J. R., Lozupone, C., Gordon, J. I., & Knight, R. (2010). Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res*, *38*(12), 3869-3879. doi:10.1093/nar/gkq066

Zimmer, C. (2009). Origins. On the origin of eukaryotes. *Science*, *325*(5941), 666-668. doi:10.1126/science.325_666

Zozaya-Hinchliffe, M., Lillis, R., Martin, D. H., & Ferris, M. J. (2010). Quantitative PCR assessments of bacterial species in women with and without bacterial vaginosis. *J Clin Microbiol*, *48*(5), 1812-1819. doi:10.1128/JCM.00851-09

Zuckerkandl, E., & Pauling, L. (1965). Molecules as documents of evolutionary history. *J Theor Biol*, *8*(2), 357-366.

APPENDIX

G. vaginalis Oligotype Profile Comparison Between Couples

Following 8 figures are expanded version of the panel (Figure 10), which was utilized to expose similarities between the *G. vaginalis* oligotype profiles in various female patients and their sexual partners. These figures show the distinct compositions of *G. vaginalis* oligotypes among different women and how variable locations in the gene behave.

In each figure, while x-axis corresponds to nucleotides, y-axis corresponds to the variable locations in the gene. Red, green, blue and yellow colors used in circles denote A, T, C and G bases, respectively. Size of each circle is proportional to the number of *G. vaginalis* sequences that present given nucleotide at given location for a patient, divided by the number of all sequences from the same patient. Thicknesses of the lines in figures are proportional to the relative abundance of the oligotypes the represent.

BV Couple 11, Vaginal vs. Penile Skin Samples

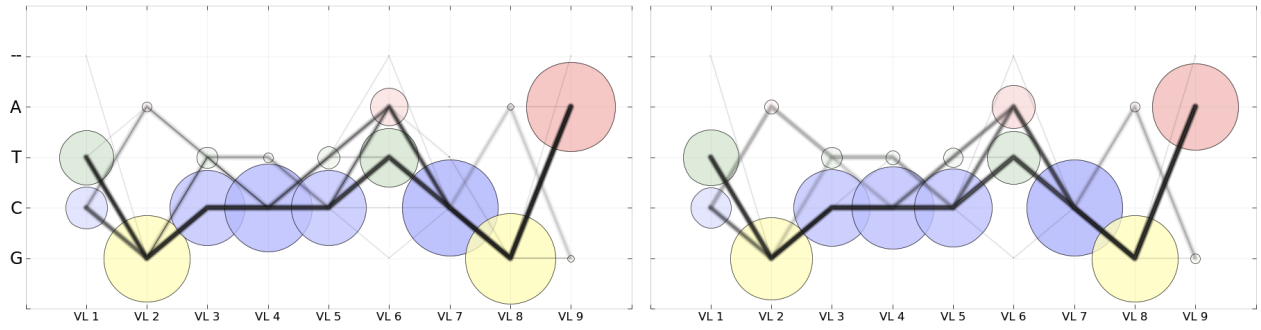

Figure 14. Comparison of *G. vaginalis* oligotype profiles from 654 vaginal swab sequences versus 495 penile skin sequences of BV couple 11 ($r = 0.956$, $p < 0.001$).

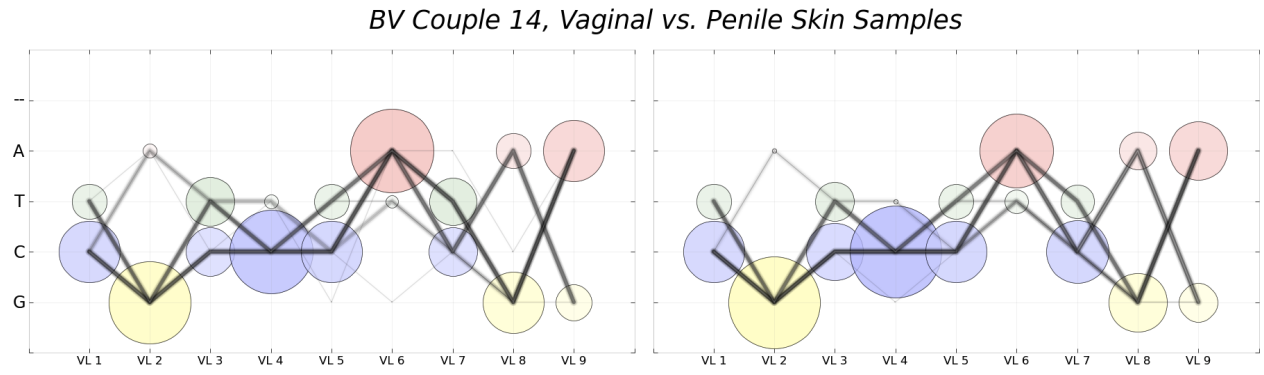

Figure 15. Comparison of *G. vaginalis* oligotype profiles from 357 vaginal swab sequences versus 421 penile skin sequences of BV couple 14 ($r = 0.970$, $p < 0.001$).

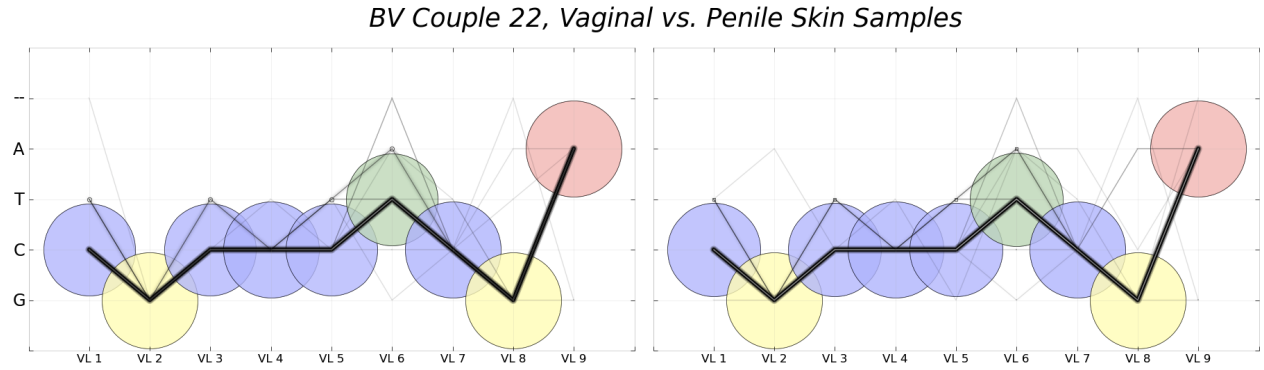

Figure 16. Comparison of *G. vaginalis* oligotype profiles extracted from 2686 vaginal swab sequences versus 3653 penile skin sequences of BV couple 22 ($r = 0.999$, $p < 0.001$).

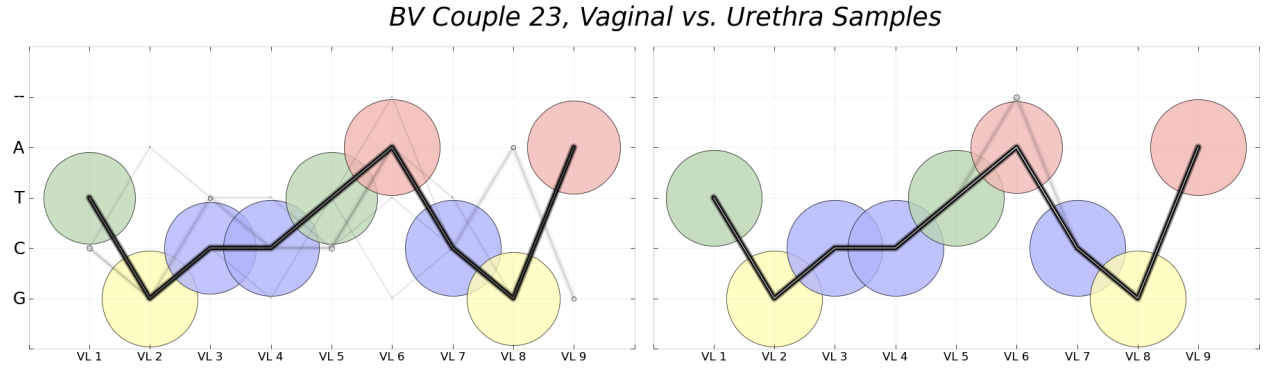

Figure 17. Comparison of *G. vaginalis* oligotype profiles extracted from 308 vaginal swab and 18 urethra sample sequences of BV Couple 23 ($r = 0.996$, $p < 0.001$).

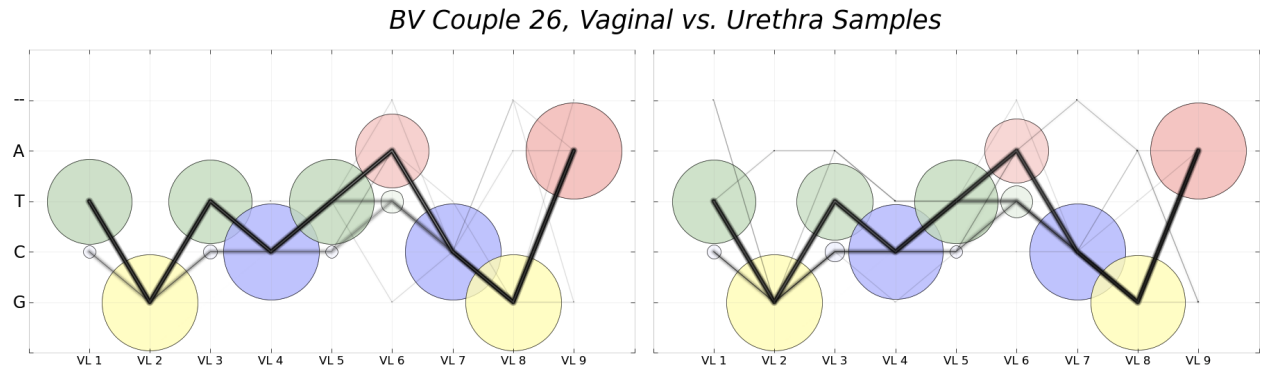

Figure 18. Comparison of *G. vaginalis* oligotype profiles extracted from 932 vaginal swab and 575 urethra sample sequences of BV Couple 26 (r = 0.984, p < 0.001).

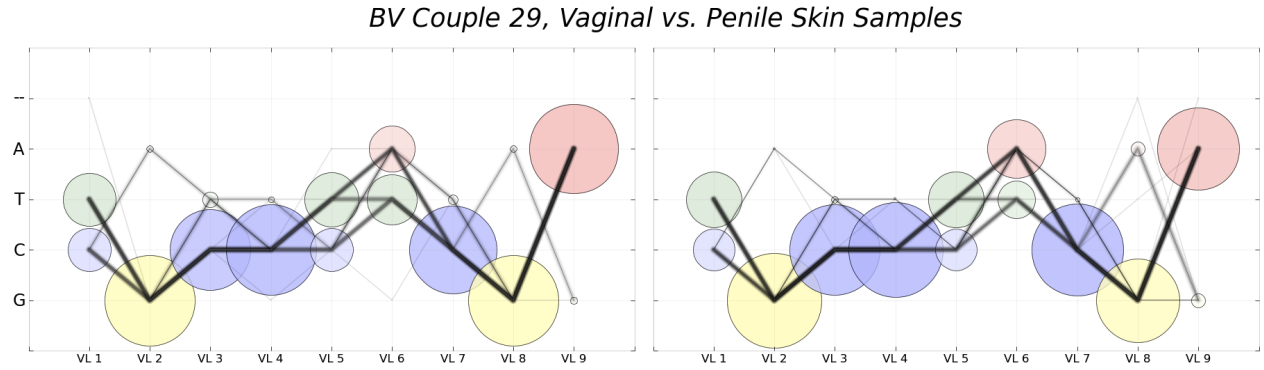

Figure 19. Comparison of *G. vaginalis* oligotype profiles extracted from 895 vaginal swab and 1313 penile skin sample sequences of BV Couple 29 (r = 0.948, p < 0.001).

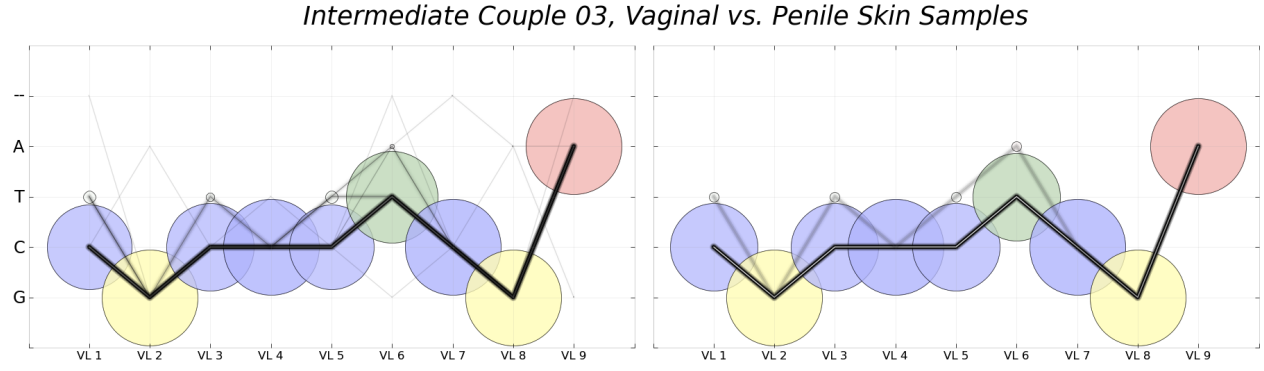

Figure 20. Comparison of *G. vaginalis* oligotype profiles extracted from 1193 vaginal swab and 10 penile skin sample sequences of Intermediate Couple 03 (r = 0.995, p < 0.001).

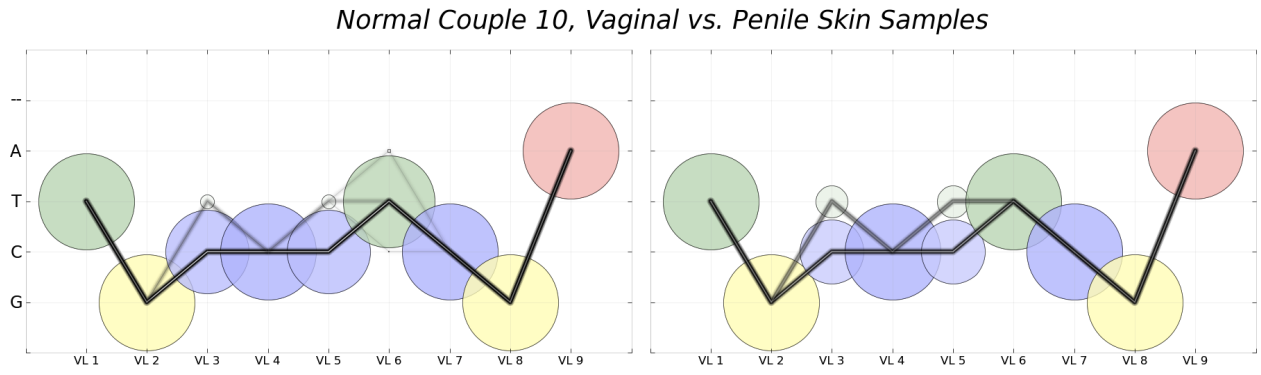

Figure 21. Comparison of *G. vaginalis* oligotype profiles extracted from 93 vaginal swab and 3 penile skin sample sequences of Normal Couple 23 (r = 0.937, p < 0.001).

Pearson Correlation Table for G. vaginalis Oligotype Profiles Among Couples

Table 1. Pearson correlation (r) between sexual partners based on their *G. vagnialis* oligotype profile. For every couple, oligotype profile of female patient's vaginal swab was compared to her sex partner's oligotype profile drawn from his urethra sample, and penile skin sample in order to quantify correlation. 2 Couples, whose male partners haven't yielded any *Gardnerella vaginalis* sequences, are not included.

Couple	Female Patient		Sex Partner's Penile Skin Sample			Sex Partner's Urethra Sample		
No	# seqs	Status	# seqs	r	p	# seqs	r	p
01	1562	BV	3	0.370	< 0.001	0	-	-
02	1841	BV	0	-	-	8	< 0.1	0.237
03	1269	BV	4	< 0.1	0.159	229	0.241	< 0.001
05	455	BV	11	**0.978**	< 0.001	2462	**0.963**	< 0.001
06	273	BV	42	**0.919**	< 0.001	0	-	-
07	495	BV	2	**0.978**	< 0.001	554	0.661	< 0.001
08	1316	BV	183	< 0.1	0.504	0	-	-
09	2044	BV	52	0.464	< 0.001	221	0.110	0.055
10	1087	BV	2	< 0.1	0.940	4	< 0.1	0.940
11	654	BV	495	**0.956**	< 0.001	891	0.224	< 0.001
12	179	BV	53	0.704	< 0.001	467	0.661	< 0.001
13	1939	BV	69	0.878	< 0.001	233	0.845	< 0.001
14	357	BV	421	**0.970**	< 0.001	1098	0.845	< 0.001
15	1057	BV	66	**0.968**	< 0.001	392	**0.923**	< 0.001
17	1031	BV	32	0.837	< 0.001	90	**0.908**	< 0.001
18	1361	BV	16	**0.900**	< 0.001	0	-	-
19	184	BV	0	-	-	15	< 0.1	0.404
20	594	BV	34	**0.998**	< 0.001	385	**0.984**	< 0.001
21	477	BV	8	< 0.1	0.218	123	**0.998**	< 0.001
22	2686	BV	3653	**0.999**	< 0.001	212	0.846	< 0.001
23	308	BV	4	0.706	< 0.001	18	**0.996**	< 0.001

(Table 1 cont.)

24	730	BV	1	0.666	< 0.001	55	0.666	< 0.001
25	615	BV	24	0.501	< 0.001	1424	0.501	< 0.001
26	932	BV	26	0.261	< 0.001	575	**0.984**	< 0.001
27	1340	BV	6	**0.989**	< 0.001	1320	**0.994**	< 0.001
28	355	BV	68	**0.955**	< 0.001	1591	< 0.1	0.144
29	895	BV	1313	**0.948**	< 0.001	4740	0.545	< 0.001
30	1449	BV	2	< 0.1	0.094	0	-	-
31	842	BV	1	< 0.1	0.939	17	0.141	0.013
32	211	BV	4	0.720	< 0.001	67	0.828	< 0.001
33	300	BV	1	0.805	< 0.001	2353	0.742	< 0.001
34	1117	BV	18	**0.997**	< 0.001	9	0.250	< 0.001
35	683	BV	17	**0.995**	< 0.001	662	**0.980**	< 0.001
01	784	IN	55	0.301	< 0.001	0	-	-
02	2	IN	51	< 0.1	0.953	189	< 0.1	0.648
03	1193	IN	10	**0.995**	< 0.001	0	-	-
04	270	IN	5	**0.943**	< 0.001	662	**0.980**	< 0.001
05	547	IN	6	**0.943**	< 0.001	1684	0.230	< 0.001
03	2	N	132	0.602	< 0.001	2129	0.551	< 0.001
05	10	N	3	< 0.1	0.953	1	< 0.1	0.953
06	35	N	6	**0.955**	< 0.001	1051	**0.988**	< 0.001
08	6	N	0	-	-	760	**0.999**	< 0.001
10	93	N	3	**0.937**	< 0.001	0	-	-
11	10	N	0	-	-	6	0.886	< 0.001

Summary of Specimens After BLAST Filtering

Table 2. The number of specimens in the original pyrosequencing library versus the number of specimens per environment that possessed at least one high quality *G. vaginalis* 16S rRNA gene tag sequence.

Specimen	Gram stain classification	# specimens in the original pyrosequencing library	# specimens after BLAST search for *G. vaginalis*	Average # of *G. vaginalis* sequences per category
Vagina	BV	36	33	928
Vagina	Intermediate	5	5	559
Vagina	Normal	12	6	26
Penile skin	BV	36	31	213
Penile skin	Intermediate	5	5	25
Penile skin	Normal	12	4	24
Urethra	BV	36	28	722
Urethra	Intermediate	3	3	507
Urethra	Normal	12	5	658

Oligotype Distribution Among Sample Groups

Table 3. Distribution of oligotypes that were present at minimum of 1% relative abundance in at least one sample in the library. Ratios in this table indicate the number of samples in a given group that exhibited the given oligotype.

	Gram stain BV			Gram stain Intermediate			Gram stain Normal		
Oligotype	Vaginal Swab	Penile Skin	Urethra	Vaginal Swab	Penile Skin	Urethra	Vaginal Swab	Penile Skin	Urethra
TGTCTACGA	33 / 35	13 / 30	19 / 29	4 / 5	2 / 5	3 / 3	4 / 8	5 / 6	5 / 9
CGCCCTCGA	33 / 35	25 / 30	25 / 29	5 / 5	3 / 5	3 / 3	5 / 8	2 / 6	7 / 9
TGTCTTCGA	27 / 35	11 / 30	13 / 29	4 / 5	1 / 5	1 / 3	1 / 8	3 / 6	6 / 9
CGCCCACGA	25 / 35	3 / 30	12 / 29	3 / 5	0 / 5	2 / 3	0 / 8	0 / 6	4 / 9
TGCCTTCGA	22 / 35	8 / 30	12 / 29	2 / 5	1 / 5	1 / 3	0 / 8	0 / 6	7 / 9
CGCCCACAG	22 / 35	11 / 30	12 / 29	2 / 5	1 / 5	1 / 3	2 / 8	1 / 6	3 / 9
CGCCTACGA	21 / 35	2 / 30	9 / 29	3 / 5	0 / 5	2 / 3	0 / 8	0 / 6	4 / 9
TGTCCTCGA	20 / 35	2 / 30	11 / 29	1 / 5	0 / 5	2 / 3	0 / 8	0 / 6	4 / 9
TGCCCTCGA	20 / 35	5 / 30	16 / 29	3 / 5	0 / 5	3 / 3	1 / 8	1 / 6	4 / 9
TGCCTACGA	19 / 35	9 / 30	9 / 29	4 / 5	3 / 5	2 / 3	0 / 8	0 / 6	1 / 9
CATTCATGA	19 / 35	4 / 30	8 / 29	2 / 5	1 / 5	0 / 3	1 / 8	0 / 6	2 / 9
CGCCTTCGA	18 / 35	2 / 30	16 / 29	3 / 5	0 / 5	2 / 3	0 / 8	0 / 6	4 / 9
CGTCTACGA	17 / 35	1 / 30	5 / 29	0 / 5	0 / 5	2 / 3	0 / 8	1 / 6	3 / 9
TGTCTATGA	13 / 35	5 / 30	7 / 29	1 / 5	0 / 5	0 / 3	0 / 8	1 / 6	3 / 9
CGCCCATGA	12 / 35	0 / 30	6 / 29	1 / 5	0 / 5	0 / 3	0 / 8	0 / 6	1 / 9
CGCCCTCGG	11 / 35	3 / 30	6 / 29	1 / 5	0 / 5	1 / 3	0 / 8	0 / 6	0 / 9
CATTCACGA	10 / 35	3 / 30	3 / 29	1 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CGTCCTCGA	9 / 35	1 / 30	6 / 29	0 / 5	0 / 5	1 / 3	0 / 8	0 / 6	2 / 9
TGTCTACAG	8 / 35	0 / 30	5 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	1 / 9
TGTCCACGA	8 / 35	0 / 30	7 / 29	1 / 5	0 / 5	2 / 3	0 / 8	1 / 6	2 / 9
TGCCCACAG	8 / 35	1 / 30	2 / 29	2 / 5	0 / 5	0 / 3	0 / 8	0 / 6	1 / 9
CGTCTTCGA	8 / 35	1 / 30	8 / 29	1 / 5	0 / 5	1 / 3	0 / 8	0 / 6	2 / 9
CGCCCCCGA	8 / 35	2 / 30	9 / 29	0 / 5	0 / 5	2 / 3	0 / 8	0 / 6	2 / 9
TGTCTGCGA	7 / 35	0 / 30	2 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
TGCCCACGA	6 / 35	0 / 30	2 / 29	2 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CATCTACGA	6 / 35	0 / 30	1 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CATCCATGA	6 / 35	0 / 30	2 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
TGTCTACGG	5 / 35	0 / 30	5 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
TGTCCACAG	5 / 35	0 / 30	4 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	1 / 9
TGCCTATGA	5 / 35	1 / 30	2 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	1 / 9
TGCCCATGA	5 / 35	0 / 30	2 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CGCCTATGA	5 / 35	0 / 30	2 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	1 / 9
CGCCCTCAA	5 / 35	1 / 30	4 / 29	1 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CGCCCGCGA	5 / 35	1 / 30	0 / 29	1 / 5	0 / 5	1 / 3	0 / 8	0 / 6	0 / 9
CGCCCACGG	5 / 35	0 / 30	6 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	2 / 9

(Table 3 cont.)

CATTTACGA	5 / 35	2 / 30	1 / 29	1 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CATTCTCGA	5 / 35	0 / 30	2 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CATTCATGG	5 / 35	2 / 30	1 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
TGTCCATGA	4 / 35	1 / 30	2 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	1 / 9
TATTCACGA	4 / 35	0 / 30	0 / 29	1 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CGCTCTCGA	4 / 35	0 / 30	4 / 29	0 / 5	0 / 5	1 / 3	0 / 8	0 / 6	1 / 9
TGCCTACAG	3 / 35	1 / 30	1 / 29	1 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CGTTCATGA	3 / 35	1 / 30	1 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CGCTCACGA	3 / 35	0 / 30	0 / 29	1 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CATTCACAG	3 / 35	0 / 30	1 / 29	1 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CATCCTCGA	3 / 35	0 / 30	0 / 29	1 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
TGTCTATGG	2 / 35	0 / 30	0 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	2 / 9
CGCCTACAG	2 / 35	0 / 30	4 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	1 / 9
TGCCCATTG	1 / 35	1 / 30	0 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CGCCCGCAG	1 / 35	0 / 30	1 / 29	1 / 5	0 / 5	0 / 3	0 / 8	1 / 6	1 / 9
CGCCCATTG	1 / 35	0 / 30	0 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CGCCCAAAG	1 / 35	0 / 30	4 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	1 / 9
TGTCTCCGA	0 / 35	0 / 30	2 / 29	0 / 5	0 / 5	1 / 3	1 / 8	0 / 6	1 / 9
TGTCTACTT	0 / 35	0 / 30	0 / 29	0 / 5	0 / 5	1 / 3	0 / 8	0 / 6	0 / 9
TGTCTAATG	0 / 35	0 / 30	0 / 29	0 / 5	0 / 5	1 / 3	0 / 8	0 / 6	1 / 9
TGTCTAAAT	0 / 35	0 / 30	3 / 29	0 / 5	0 / 5	1 / 3	0 / 8	0 / 6	0 / 9
TGTCTAAAG	0 / 35	0 / 30	1 / 29	0 / 5	0 / 5	1 / 3	0 / 8	0 / 6	0 / 9
TGCCTAAAG	0 / 35	0 / 30	3 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	2 / 9
TAATTTCGA	0 / 35	0 / 30	5 / 29	0 / 5	0 / 5	1 / 3	0 / 8	0 / 6	0 / 9
CGCCCGGGA	0 / 35	0 / 30	1 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
CGCCCATGG	0 / 35	1 / 30	2 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	2 / 9
CGCCCACCG	0 / 35	0 / 30	2 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	1 / 9
CGCCCAAGG	0 / 35	0 / 30	1 / 29	0 / 5	0 / 5	1 / 3	0 / 8	0 / 6	0 / 9
CGCCCAAAT	0 / 35	0 / 30	6 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	0 / 9
AGCCCTCGA	0 / 35	0 / 30	0 / 29	0 / 5	0 / 5	0 / 3	0 / 8	0 / 6	1 / 9

Cascading clustering of G. vaginalis Sequences from Normal and BV Diagnosed Women

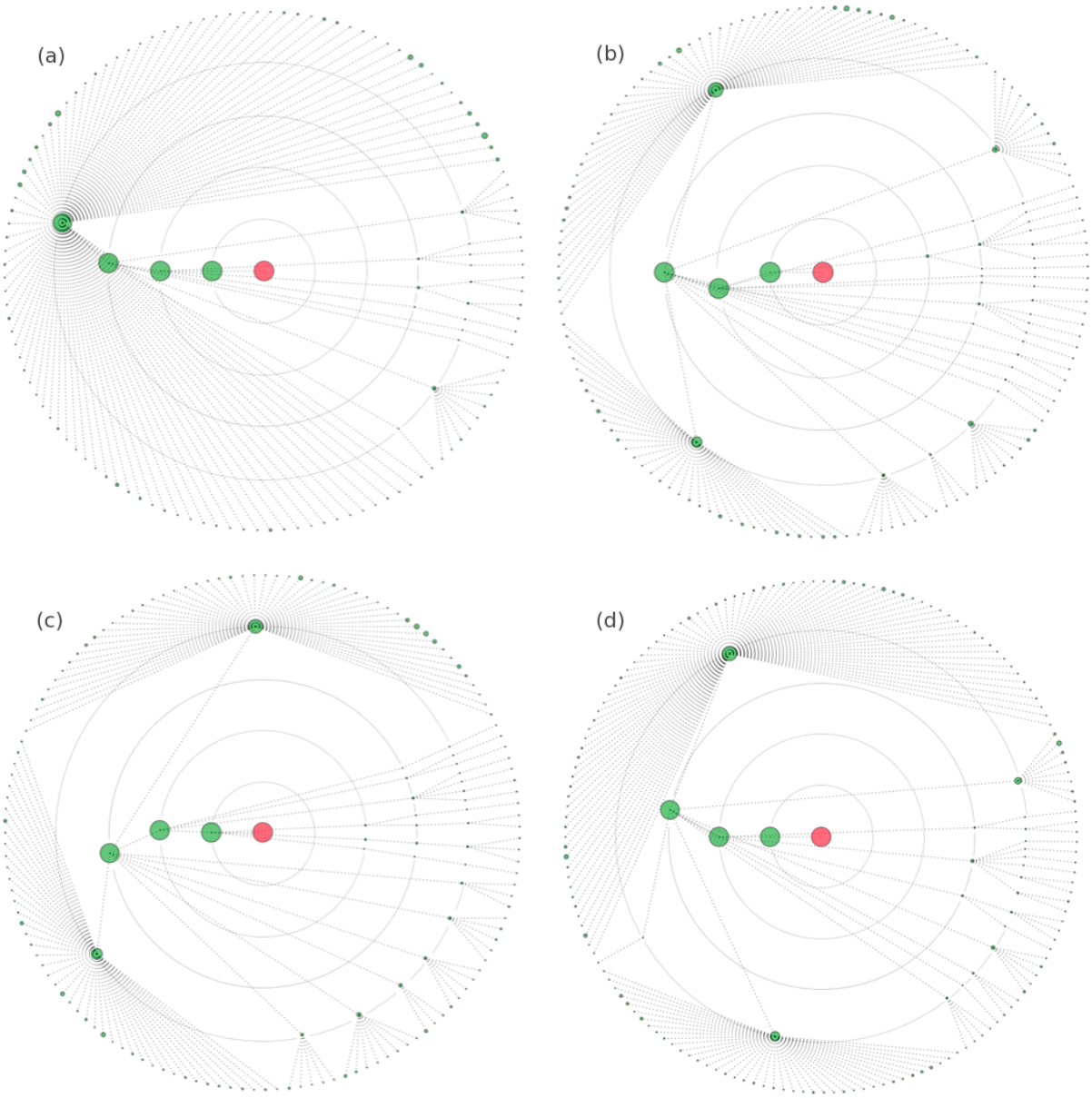

Figure 22. Clustering of 373 16S rRNA V4-V6 *G. vaginalis* sequences collected from vaginal swab samples of 13 normal women (a) and randomly selected 373 sequences from 39 women who were diagnosed with bacterial vaginosis (BV) (b; c; d). Every tier represents clusters at a certain sequence similarity level. Starting from the innermost tier (92% similarity) to the outermost (100%), clusters are being re-clustered with an increased similarity threshold by 2%. Unlike sequences from normal women, sequences from BV women splits into two or more clusters at 98% similarity level, which presents preliminary evidence that there may be more than one dominant type of *G. vaginalis* in BV women and it may be traceable from 16S rRNA.

VITA

A. Murat Eren was born in Turkey. During his undergraduate education he developed an interest in operating systems and cryptography and became proficient in GNU/Linux based operating systems and network security.

After graduating from university he became an advocate of the philosophy behind free/open-source software movement and joined forces of Turkish Linux Users Association (LKD) to create awareness towards open source as an economically and philosophically better alternative for universities and governmental agencies in Turkey.

In 2003, he started working at The National Research Institute of Electronics and Cryptology (UEKAE), an institute of The Scientific and Technological Research Council of Turkey (TÜBİTAK), as a part of a small developer team gathered to implement a Linux-based operating system, Pardus, to be used by the Turkish government and the public.

Following the first stable release of Pardus he moved to New Orleans and began his Ph.D. education at the Department of Computer Science in University of New Orleans in 2007.

After two years of Ph.D. education he developed a sudden interest in biological sciences, and decided to lend his expertise in computation to the exciting field of microbial ecology.

He likes photography, and has serious issues with any kind of authority.