Diversity and dynamics of the vaginal microbiome: exploring the boundaries between health and disease

Presented in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy

with a Major in

Bioinformatics and Computational Biology

in the

College of Graduate Studies

University of Idaho

by

Roxana Jo Hickey

July 2015

Major Professor
Larry J. Forney, Ph.D.

Committee
James A. Foster, Ph.D.
Paul A. Hohenlohe, Ph.D.
Eva M. Top, Ph.D.

Department Administrator
Eva M. Top, Ph.D.

Authorization to Submit Dissertation

This dissertation of Roxana Jo Hickey, submitted for the degree of Doctor of Philosophy with a Major in Bioinformatics and Computational Biology and titled "Diversity and dynamics of the vaginal microbiome: exploring the boundaries between health and disease," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor:	Larry J. Forney, Ph.D.	Date
Committee Members:	James A. Foster, Ph.D.	Date
	Paul A. Hohenlohe, Ph.D.	Date
	Eva M. Top, Ph.D.	Date
Department Administrator:	Eva M. Top, Ph.D.	Date

Abstract

The vaginal microbiome has a rich history of research dating back more than a century. In recent years modern sequencing technologies have yielded significant new insights that challenge old paradigms and blur the boundaries between health and disease. We currently recognize that multiple kinds of vaginal communities exist in healthy women but are still very much in a natural history phase as we attempt to translate compositional diversity to meaningful differences in ecological function, community stability and, ultimately, women's health. The objective of this dissertation is to begin bridging this gap by incorporating functional characterization and longitudinal sampling into vaginal microbiome studies. In my opening chapter I provide an overview of vaginal microbiome research, recounting significant historical advances and highlighting remaining knowledge gaps. I advocate approaching the vaginal microbiome from an ecological perspective to facilitate our understanding of its role in health and disease.

Next, I describe the development and validation of a microarray designed for rapid functional screening of vaginal microbial communities. I demonstrate efficacy of the microarray on different types of samples and conclude that it could be used to develop informed hypotheses of community function and launch additional detailed studies.

In my third chapter, I characterize longitudinal changes in the vaginal communities of adolescent girls. To date most research has focused exclusively on reproductive age women, so this study addresses an important knowledge gap in understanding the transitions that occur in the microbiome during the formative years of puberty. Importantly I document trends in the rise of lactic acid bacteria, which occurs earlier in puberty than previously thought. I also report observations of *Gardnerella vaginalis* in several young adolescents, a notable finding considering this species is frequently considered a potential pathogen.

Finally, I explore the genomic diversity of *Gardnerella vaginalis*, which is increasingly recognized as a common inhabitant of healthy vaginal communities despite its strong association with bacterial vaginosis. Adopting the ecotype concept of intraspecies diversity, I show that *G. vaginalis* can be separated into multiple phylogenetic clades each in possession of a unique suite of functional traits that may be relevant to both its ecology and postulated virulence.

Acknowledgments

Wow, what a ride this has been. I have so many people to thank.

Above all, I thank my advisor Larry Forney for being an incredible mentor, role model and friend to me over the past nine—yes, *nine*—years. I first met Larry in 2006 when I enrolled in a freshman bioethics course that he co-taught with James Foster and Jason Johnstone-Yellin. Inspired by our lively class discussions, and on learning about Larry's pioneering work on the vaginal microbiome, I joined the lab in 2007 as an inexperienced yet eager undergraduate research assistant. I flourished under his mentorship and eventually became something of a poster child—quite literally, there is a life-sized poster of my image—for undergraduate research at the University of Idaho. By the end of college I remained deeply immersed in my research, so in 2010 I enrolled in the BCB program and pressed ahead toward new horizons. Over the years Larry and I developed a close working relationship built on mutual trust and respect. He came to know me uncannily well, and whether I was overjoyed or overwhelmed at the peaks and pitfalls of grad school, he could always 'read me like a book' and help me refocus on what was important. His keen intuition and unwavering commitment to the success of his students—not just as scientists, but as human beings living meaningful lives—set him leagues ahead of many advisors. I thank him, from the bottom of my heart, for believing in me and being there for me all these years. It has made all the difference.

I thank my committee members for their wisdom and encouragement. James Foster has known me from the beginning and was the first person to suggest that I look into undergraduate research (turned out to be a best idea ever). He introduced me to the fundamentals of computational biology and has been a fountain of great ideas. Eva Top has been an exemplary role model to me for many years, and I enjoyed interacting with her both in and out of our academic roles. Her analytical sensibility and attention to experimental details are worthy of emulation. Paul Hohenlohe has been a great addition to my committee, and I am grateful for the unique perspective he brings from his background in ecological genomics. He also happens to be a talented bluegrass musician, and I hope there are many more PEES sing-alongs down the road. Zaid Abdo is a past member of my committee, and I worked closely with him on a few projects while I was an undergraduate. I am grateful for his advice, patience and kindness.

I intersected with so many wonderful people during my time in the Forney lab. In chronological order (testing my memory): Maria Schneider, Xia Zhou, Jacob Pierson, Ursel Schütte, Hyo-Jin Ahn, Hyun-Joon La, Sanqing Yuan, Rachel Westman, Ana Cornea, Dora Cohen, Angie Spangler, Megan Lopez, Vandhana Krishnan, Shelby Thornton, Juliane Smith, Erika Bengtson, Jian Shen, Robin Baker, Renee Nuhn, Angie Buvel, Lygia Peralta, Dorah Mtui, Helena Mendes-Soares, Michael France, Hanna Kehlet, Leila Yazdani, Karol Gliniewicz and Kenetta Nunn. I know there are more whose names escape me now. Thank you all for making the lab such a fun place to work, and special thanks to Maria, Xia, Jacob and Ursel for showing me the ropes when I was a bumbling novice. I will miss you all dearly and think of you fondly whenever I watch Marcel the Shell with Shoes On ("compared to what?!").

I am grateful to many talented individuals with whom I was fortunate to collaborate outside of the lab. I especially thank Matthew Settles and Sam Hunter for their help and patience as I cultivated my computational skills and learned the nitty gritty details of genomic analysis. I thank Omar Cornejo for taking me on as a rotation student for a semester and showing his support and enthusiasm from day one. I also thank Dennis Fortenberry, Pawel Gajer, Bing Ma, Jacques Ravel and Haruo Suzuki for working closely with me on various projects and papers.

I am so fortunate to have been part of the BCB-Biology community of graduate students and postdocs: Daniel Beck, Maia Benner, Daniel Caetano, Lucius Caldwell, Erin Clancey, Mitch Day, Matthieu Delcourt, Simone Des Roches, Anahi Espindola, Travis Hagey, Kayla Hardwick, Tyler Hether, Sarah Jacobs, CJ Jenkins, Denim Jochimsen, Maribeth Latvis, Wesley Loftie-Eaton, Hannah Marx, Tim McGinn, Ailene McPherson, Genevieve Metzger, Diego Morales-Briones, Siavash Riazi, Pavitra Roychoudhury, Brice Sarver, Katie Shine, Matthew Singer, Thibault Stalder, Chloe Stenkamp-Strahm, ET Thornquist, Simon Uribe-Convers, Josef Uyeda and Janet Williams. Together we laughed, cried, danced, dressed up, brunched, flipped cups, celebrated marriages, welcomed babies into the world, cooked delicious dinners, cheered on the World Cup, crafted art, won trivia nights, drank questionable cocktails at the Garden and floated down the Snake River until the sun went down. Thank you for these wonderful memories.

IBEST provided the ideal setting to develop intellectually as a fledgling scientist. I learned a great deal from our outstanding faculty, and the camaraderie here is something to be cherished. In particular I acknowledge Holly Wichman and David Tank, whose exemplary demonstrations of scholarship and mentorship inspire me. I thank Lisha Abendroth, whose exceptional organiza-

tional skills and approachable personality made getting through graduate school immeasurably more pleasant. I also thank Benji Oswald of the IBEST Computational Resources Core for responding to endless inquiries about installing software and setting up analyses on the server.

I thank Matthew Pennell for more things than I can possibly enumerate here: for inspiring me with his insatiable curiosity and incisive intellect; for cheering me to the finish line; for debugging my R code and helping me choose pretty colors for graphs; for making me laugh; for providing a shoulder to cry on; for listening to me rant about the same frustrations over and over again; and for always motivating me to be a better version of myself. Finally, I thank my parents Joe and Remie Hickey for their enduring support (I am certain they often wondered when I would finally be done with school) and for raising me to believe I could accomplish anything I set my mind to. I know I have made them so proud.

To all of you, with all my heart, thank you.

Dedication

To my parents, with love and gratitude

Joe & Remie Hickey

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction: Understanding vaginal microbiome complexity from an ecological perspective[1]

1.1 Importance of the vaginal microbiome to health

Bacterial communities in the human vagina are thought to have a critical role in protecting the host against infectious disease. In reproductive-age women, it is thought they do so through the production of lactic acid resulting in a low pH environment that restricts the growth of pathogens and other opportunistic organisms (Linhares *et al.*, 2010b; O'Hanlon *et al.*, 2011). Thus, maintaining high numbers of lactic acid bacteria is a hallmark of healthy conditions. Although there are marked differences in the species composition and rank-abundances of populations in vaginal bacterial communities (Zhou *et al.*, 2010; Ravel *et al.*, 2011) among women, it appears that all are probably dominated by homofermentative lactic acid bacteria (Zhou *et al.*, 2007). This suggests the ecological function of various vaginal communities in reproductive-age women—creating a low pH environment through the production of organic acids—is conserved despite differences in the bacterial species present.

1.1.1 *The vaginal microbiome throughout a woman's lifespan*

The vaginal microbial ecosystem undergoes significant structural changes at various stages in a woman's life that are directly linked to the level of estrogen in the body (Farage and Maibach, 2006). Initial colonization occurs at birth, when the infant is first exposed to her mother's vaginal tract if delivered vaginally, or by the skin bacteria of persons handling the infant in the case of a Caesarian-section delivery (Dominguez-Bello *et al.*, 2010). While little is known about the importance of this initial colonization event, it is believed to establish the gut, skin and vaginal microbiota, and in the weeks and months following birth, these differentiate into communities distinct to each habitat (Palmer *et al.*, 2007; Dominguez-Bello *et al.*, 2010; Koenig *et al.*, 2011). During the first 2–4 weeks following birth, maternal estrogen mediates thickening of the vaginal

[1]This chapter has been published in a modified form as: Hickey R.J., Zhou X., Pierson J.D., Ravel J., and Forney L.J. 2012. Understanding vaginal microbiome complexity from an ecological perspective. Translational Research 160:267-282.

epithelium and the production of glycogen that is fermented by indigenous bacteria resulting in a lowering of the vaginal pH. This is transitory however, as subsequent metabolism of maternal estrogen is accompanied by thinning of the vaginal mucosa, a reduction of the level of glycogen, and a concomitant increase in vaginal pH (Farage and Maibach, 2006)

During childhood (Tanner stage 1; Marshall and Tanner, 1969), the pH of the vagina is nearly neutral and cultivation-dependent methods have shown it to be colonized by diverse assemblages of aerobic, strictly anaerobic, and enteric species of bacteria (Hammerschlag *et al.*, 1978a,b; Alvarez-Olmos *et al.*, 2004; Randjelović *et al.*, 2005). Between the ages of 8 and 13 years of age, pubertal changes in the vulva and vagina occur that are induced by adrenal and gonadal maturation. During the maturation process, follicular development causes estrogen production to rise, and once again this is accompanied by a thickening of the vaginal epithelium and intracellular production of glycogen. These new environmental conditions select for microorganisms capable of fermenting glycogen to lactic acid and the concomitant acidification of the vaginal environment that is characteristic of reproductive-age women (Farage and Maibach, 2006). Remarkably, the shifts in microbial community composition that occur during this transition have seldom been studied. Using cultivation-dependent methods, Alvarez-Olmos *et al.* (2004) found that the vaginal microbiota of many adolescent girls (14–18 y) resembled those of adult women with bacterial vaginosis. Yamamoto *et al.* (2009) assessed the vaginal microbiota of adolescent girls (13–18 y) using cultivation-independent methods and observed that the bacterial communities were comparable to those found in adults but remarked that this may not be the case for premenarcheal or perimenarcheal girls.

When vaginal epithelial cells are sloughed in reproductive-age women the glycogen present is then presumably metabolized by bacterial populations to produce organic acids; however, as widely cited as this mechanism is it is backed by very little evidence collected from *in vitro* analyses (Wylie and Henderson, 1969). The resulting low pH (4.0–4.5) of the vagina creates an environment that restricts or precludes the growth of many pathogenic organisms. However, with the onset of menopause, estrogen levels again decrease, and menstruation ceases. This is accompanied by atrophy of the vaginal epithelium and reduced cervico-vaginal secretions (Farage and Maibach, 2006). In most menopausal women the vaginal microbiota is thought to shift from populations of lactic acid producing bacteria to an assortment of species that include strictly anaerobic and enteric bacteria (Larsen *et al.*, 1982; Ginkel *et al.*, 1993). The dynamic nature

of this ecosystem underscores the importance of resolving its microbial constituents at different stages of human development and the prominent influence of estrogen levels in the host on the vaginal environment.

1.1.2 *Community performance in reproductive-age women: lactic acid production*

In general, the presence of high numbers of lactic acid bacteria in the vagina is often equated with 'healthy' and low numbers, or absence thereof, as being 'abnormal' (Priestley *et al.*, 1997; Pybus and Onderdonk, 1997; Donders, 2007). This has historically focused attention on members of the genus *Lactobacillus* as keystone species because of their well known ability to produce lactic acid through the fermentation of sugars. This view originates from the earliest studies of the vaginal microbiota over a century ago, when Professor Albert Döderlein first reported culturing the bacteria from vaginal secretions. He found they produced lactic acid, which in turn inhibited growth of pathogens both in vitro and in vivo. Döderlein's bacillus was later classified in 1928 as *Lactobacillus acidophilus* (Thomas, 1928). Several decades later in the 1980s, it was determined that *L. acidophilus* was not a single species, but rather a group of closely related, obligately homofermentative species collectively known as the *Lactobacillus acidophilus* complex (Lauer *et al.*, 1980). Because species within this complex are difficult to distinguish phenotypically or biochemically (Johnson *et al.*, 1980), they were differentiated on the basis of DNA homology (Schleifer and Ludwig, 1995; Du Plessis and Dicks, 1995). All of the *Lactobacillus* spp. found to be prevalent in the vagina today are members of this complex.

1.1.3 *Beyond Lactobacillus: findings from cultivation-dependent and -independent studies*

Following Döderlein's discovery of what later came to be known as *Lactobacillus*, cultivation-dependent studies eventually revealed that a diverse array of facultative and strictly anaerobic bacteria, and sometimes the yeast *Candida*, can be present in the healthy vagina but typically in much lower numbers (Redondo-Lopez *et al.*, 1990; Larsen and Monif, 2001; Marrazzo *et al.*, 2002). Furthermore several species of *Lactobacillus* belonging to the *Lactobacillus acidophilus* complex were identified, including *L. jensenii, L. casei, L. gasseri, L. crispatus, L. plantarum, L. fermentum, L. cellobiosus, L. brevis, L. minutus,* and *L. salivarius* (Rogosa and Sharpe, 1960; Levison *et al.*, 1977; Reid *et al.*, 1996; Antonio *et al.*, 1999). It is interesting to note that *L. minutus*

was reclassified in 1992 as a member of a new genus, *Atopobium*, and subsequently renamed *A. minutum* (Collins and Wallbanks, 1992). Given the uncertainty and controversy surrounding the potential role of *Atopobium* species in bacterial vaginosis (Verhelst *et al.*, 2004; Verstraelen *et al.*, 2004; Srinivasan and Fredricks, 2008), future studies might seek to better understand the traits that distinguish these genera and the species within them.

Efforts to characterize vaginal microbial communities using cultivation methods undoubtedly led to significant improvements in understanding the role of microbes in vaginal health, but they were limited due to the inherent biases in cultivation methods. Today it is well known that most host-associated and environmental microbes resist cultivation in the laboratory using traditional techniques (Bakken, 1985). Undoubtedly cultivation of microorganisms is fundamental to understanding their physiology and phenotypic characteristics, and it remains a very useful tool in studies of microbial ecology. Promising developments in cultivation of fastidious bacteria using state-of-the-art techniques (Connon and Giovannoni, 2002; Stingl *et al.*, 2007; Park *et al.*, 2011) are likely to enable the cultivation of many previously inaccessible microbes. However, studies aimed at assessing fine-scale variation in host-associated microbial communities within and among individuals or exploring ecological relationships within these communities require methods that provide detailed information about microbial diversity while also being cost-effective and scalable to high-throughput sample processing. In response to this need, cultivation-independent methods have in recent years become the standard approach to characterizing the diversity of microbes residing in and on the human body (Hugenholtz *et al.*, 1998; Dekio, 2005; Eckburg, 2005; Bik *et al.*, 2006; Turnbaugh *et al.*, 2007).

Major advances in DNA sequencing technology over the last decade have fundamentally changed the way we assess microbial community structure and composition. For investigations of bacterial diversity, these methods commonly utilize 16S rRNA gene sequences as a means to compare and classify taxa. This approach circumvents the need for cultivation by analyzing DNA sequences extracted directly from samples. Typically, partial 16S rRNA gene sequences are amplified using primers that anneal to highly conserved sequences in the gene, and the resulting amplicons are sequenced. Phylogenetic analyses of the sequences allows for classification of phylotypes and determination of the numerically dominant taxa in a community. Other methods that rely on other conserved genes—such as cpn60 (Schellenberg *et al.*, 2009), rpoC, uvrB or recA (van der Lelie *et al.*, 2006)—have also been developed but are not as widely used.

1.2 THE VAGINA CATALOGUES: WHAT IS NORMAL?

1.2.1 *Vaginal microbial community types*

Using these methods numerous studies have been done to characterize the vaginal microbial communities of healthy, asymptomatic, reproductive-age women (Burton *et al.*, 2003; Verhelst *et al.*, 2004; Zhou *et al.*, 2004; Hill *et al.*, 2005; Hyman *et al.*, 2005; Zhou *et al.*, 2007, 2010; Ravel *et al.*, 2011). Although these studies have relied on various analytical methodologies and study designs—sampling different regions of the vagina, women from various ethnic backgrounds, different geographical locations of populations, sampling times in relation to the menstrual cycle, and so on—their findings consistently demonstrate that vaginal bacterial community composition differs both within and between women, and several types of communities are known to exist. Together, these findings paint a much more complicated picture of the vaginal microbiota than had been considered in the past.

Previous studies performed in our laboratory have shown that several distinct kinds of vaginal communities with markedly different species composition occur in white, black, Hispanic and Asian women in North America (Zhou *et al.*, 2007; Ravel *et al.*, 2011) and Japanese women in Tokyo, Japan (Zhou *et al.*, 2010). Since vaginal bacterial communities differ in species composition they are likely to differ in how they respond to disturbances. Conceptually this is important since vaginal communities continually experience various kinds of chronic and acute disturbances caused by human behaviors such as the use of antibiotics, hormonal contraceptives and other methods of birth control, sexual intercourse, vaginal lubricants, douching and many others. In addition, the structure and composition of vaginal microbial communities are known to be influenced by natural changes in normal healthy women including aging (Larsen *et al.*, 1982; Cauci *et al.*, 2002), time in the menstrual cycle (Eschenbach *et al.*, 2000), menstruation (Smith *et al.*, 1982; Onderdonk *et al.*, 1987; Shiraishi *et al.*, 2011), pregnancy (Verstraelen *et al.*, 2009), and stress (Culhane, 2002; Nansel *et al.*, 2006).

In a previous study we analyzed 144 vaginal samples from healthy Caucasian and black women in North America (Zhou *et al.*, 2007). The results showed that 80% of the women had microorganisms phylogenetically related to *Lactobacillus iners, L. crispatus, L. jensenii,* or *L. gasseri* as a numerically dominant member of the vaginal microbiota. Overall, *L. iners* was the most common species of *Lactobacillus* in women of both ethnic groups having been recovered in 66% of the

women sampled, and other groups have reported this organism as being highly common in the vaginas of reproductive-age women as well (Tärnberg *et al*., 2002; Vásquez *et al*., 2002; Hill *et al*., 2005; Verhelst *et al*., 2005; Anukam *et al*., 2006; Tamrakar *et al*., 2007; De Backer *et al*., 2007; Ferris *et al*., 2007; Thies *et al*., 2007; Nam *et al*., 2007; Alqumber and Burton, 2008; Shi *et al*., 2009; Zozaya-Hinchliffe *et al*., 2010). Surprisingly *L. iners* was only first described in 1999 (Falsen *et al*., 1999) as it does not grow on the media typically used to isolate and enumerate *Lactobacillus*, so it was absent from earlier cultivation-dependent studies of the vaginal microbiota. The remainder of communities in our study contained a relatively low proportion of lactobacilli, exhibited greater species evenness and included high numbers of clones most closely related to *Atopobium* and genera of the order Clostridiales, including *Megasphaera, Dialister, Anaerococcus, Finegoldia, Peptostreptococcus,* and *Eubacterium*. Additionally, 20–30% of the clones from these communities were from novel clades in the phylum Firmicutes. Comparable results were obtained in a recent study of healthy, reproductive-age Japanese women (Zhou *et al*., 2010). The findings of these studies indicate there are a limited number of different kinds of vaginal microbial communities in asymptomatic, apparently healthy women. Moreover, from studies of adolescent women (Yamamoto *et al*., 2009), it appears that these communities are established in puberty and may reside in women until menopause.

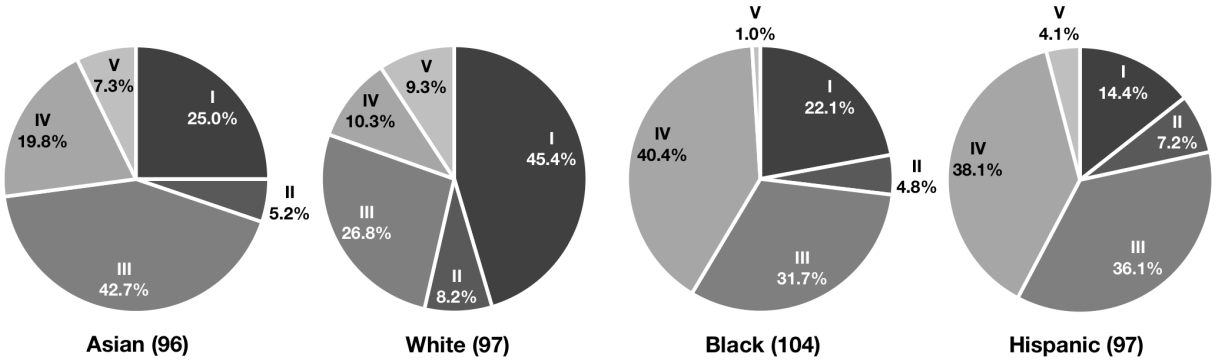

Figure 1.1: Representation of vaginal bacterial community groups within four ethnic groups of women. The number of women from each ethnic group is in parentheses. The roman numerals indicate five common vaginal bacterial community groups. Community groups I, II, III and V are predominated by *Lactobacillus crispatus, L. gasseri, L. iners* and *L. jensenii*, respectively, while community group IV contains a diverse assemblage of facultative and strictly anaerobic bacteria. Percent values are the percentages of women in each ethnic group whose vaginal bacterial community clustered with a particular community group (reproduced from data in Ravel *et al*., 2011).

Recently, we completed a more detailed and expansive study to characterize vaginal microbiota using high-throughput methods based on pyrosequencing of barcoded 16S rRNA genes (Ravel *et al.*, 2011). The subjects were a cohort of 394 North American asymptomatic, reproductive-age women equally representing four ethnic backgrounds (Asian, white, black, and Hispanic). We found a total 282 phylotypes among these women. Their vaginal bacterial communities were characterized into five groups, four of which were dominated by *Lactobacillus iners, L. crispatus, L. gasseri,* or *L. jensenii*, and the fifth which had lower proportions of lactic acid bacteria and higher proportions of strict and facultative anaerobes. The latter community type accounted for about 25% of the women sampled, a notable finding considering the prevailing view that high numbers of *Lactobacillus* are necessary for a healthy vaginal tract. Furthermore we observed high bacterial species diversity in all vaginal communities, even those in which the phylotype abundance distribution was highly skewed toward one or very few numerically dominant phylotypes.

An important finding from these studies is that the distribution of community types varies significantly among women from different ethnic backgrounds (Figure 1.1). For example, in the study by Ravel *et al.* (2011) vaginal bacterial communities dominated by *Lactobacillus* spp. were found in 80.2% and 89.7% of Asian and white women, respectively, but just 59.6% and 61.9% of black and Hispanic women, respectively. On the other hand, occurrence of communities with low proportions or no detectable *Lactobacillus* species community type were elevated in Hispanic (38.1%) and black (40.4%) women compared to Asian (19.8%) and white (10.3%) women. These findings are in accordance with results obtained by Zhou *et al.*, who assessed the vaginal bacterial communities of white, black and Japanese women (Zhou *et al.*, 2007, 2010). Moreover, vaginal pH was found to differ among ethnic groups as well, with the overall median vaginal pH of black (4.7 ± 1.04) and Hispanic (5.0 ± 0.74) women being slightly elevated over what is typically considered to be healthy (4.0–4.5). Vaginal pH was elevated (5.3 ± 0.6) among women of all racial groups in the 'diversity group', and it was above 4.5 for the community types dominated by *L. gasseri* (5.0 ± 0.7) and *L. jensenii* (4.7 ± 0.4). *L. crispatus* and *L. iners*-dominated community types had median pH values of 4.0 ± 0.3 and 4.4 ± 0.6, respectively (Ravel *et al.*, 2011).

1.2.2 *A rose by any other name*

There is compelling evidence to suggest *Lactobacillus* spp., the production of lactic acid, and the resulting low pH are important for preventing the proliferation of nonindigenous organisms in the vagina (Boskey *et al.*, 1999, 2001; Aroutcheva *et al.*, 2001a; Valore *et al.*, 2002; O'Hanlon *et al.*, 2011). However, these observations have been over-interpreted and through faulty logic have led to the assertion and common wisdom that *Lactobacillus* spp. must be present for health to be maintained. This has been extended and some claim that women whose vaginal communities are depauperate of *Lactobacillus* spp. are somehow abnormal. Unfortunately, this fallacy is the premise of the commonly used Nugent criteria used for the diagnosis of bacterial vaginosis wherein the degree of 'healthiness' is in part assessed by scoring the abundance of *Lactobacillus* morphotypes, ignoring the possibility that their ecological function could be supplanted by bacteria with other morphotypes. It might be more reasonable to postulate that a key ecological function of vaginal communities, namely the production of lactic acid, might be accomplished by a variety of taxa capable of homolactic and heterolactic fermentation of substrates. *Atopobium, Streptococcus, Staphylococcus* and *Leptotrichia* are among the genera found in the vagina that possess this capability in addition to the better known *Lactobacillus*. If the maintenance of a low environmental pH is indeed a key function of the vaginal microbial community then perhaps it may be more appropriate to consider 'lactic acid bacteria' *in toto* to be members of the same ecological guild because they use the same resource pool to accomplish the same ecological function (Jaksić, 1981). If this is the case, then the prevailing view that species of *Lactobacillus* are both necessary and sufficient for maintaining health may well be overly simplistic because functionally equivalent species may in fact 'substitute' for species of *Lactobacillus*.

1.3 MICROBIAL ECOLOGY OF THE VAGINAL MICROBIOME

Vaginal bacterial communities reside in an ecosystem that is strongly influenced by characteristics of the host, local environment, and constituent populations (Figure 1.2). The human microbiome is often referred to as a commensal relationship in which one member derives benefit from the other member without providing benefit or causing harm in return. This is almost certainly not the case for the vaginal microbiome wherein the bacteria are entirely dependent on the host for nutrients, and in turn the bacterial communities play a role in protecting against

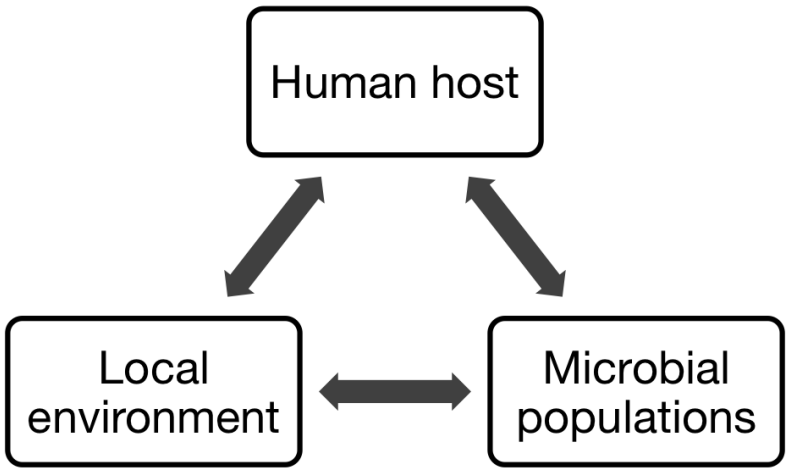

Figure 1.2: Interactions shaping the vaginal ecosystem. The vaginal ecosystem and bacterial communities therein are strongly influenced by characteristics of the host, local environment, and constituent populations.

disease-causing organisms. Consequently, it should be viewed as a mutualism in which an understanding of community composition, function, and dynamics requires that the vagina be viewed as an ecosystem and not simply the sum of its parts. For example, it is now clear that bacterial communities in the vaginas of reproductive-age women are reasonably complex and include diverse species from several bacterial lineages. Hence, although vaginal bacterial communities are dominated by lactic acid producing bacteria, they coexist and interact with a wide array of other bacterial species by competition for space and resources. Through metabolic activities these populations modify their environment in ways that either facilitate or preclude colonization by other species through resource competition, predation by bacteriophage and the production of antimicrobial substances. Likewise the host influences the composition of communities by determining the quantity and composition of vaginal transudates that constitute an important source of nutrients for resident bacterial populations. Moreover, it is likely though not proven that vaginal mucus and epithelial cell receptors play an important role in colonization by certain bacterial species. Further, innate and local immune systems may work to exclude or select community members (reviewed by Linhares *et al.*, 2010b).

The available evidence suggests vaginal communities are in a state of dynamic equilibrium, in which short term fluctuations (at least in reproductive-age women) occur in response to changes

driven by hormonal changes that are a part of normal menstrual cycles. It is unknown whether any stage in a menstrual cycle should be thought of as a 'disturbed' state and therefore more susceptible to invasion. However, it is known that many normal human activities are associated with destabilization of vaginal communities. For example, frequent sexual intercourse, having multiple sex partners, and frequent episodes of receptive oral sex (Schwebke *et al.*, 1999; Vallor *et al.*, 2001; Schwebke *et al.*, 2004; Beigi *et al.*, 2005) cause destabilization of vaginal microbial communities, as do douching (Brotman *et al.*, 2008) and use of spermicides (Klebanoff *et al.*, 2010; Rosenstein *et al.*, 1997; Gupta *et al.*, 2000). Each of these has the potential to destabilize vaginal bacterial communities and increase their invasibility. Here we discuss an ecological model as it may apply to the vaginal microbiota and community stability. First we must emphasize that a relevant and meaningful ecological theory for the vaginal microbiome is currently not feasible until additional research on community dynamics has been done. However, while the model presented here is conceptual in nature it does provide a useful framework for evaluating the possible importance and roles of community members and guiding future studies in this field.

1.3.1 *Drivers and passengers*

The strong linkage between high numbers of lactic acid bacteria and a 'healthy' vaginal microbial community is consistent with Walker's Driver-Passenger model of community structure and function (Walker, 1995). Under this model, species of lactic acid bacteria would be considered 'drivers' that strongly influence the function or structure of the ecosystem by producing lactic acid and maintaining a low pH. The environment thus created would be a strong determinant of community species composition and activity because they would all have to flourish or at least tolerate an environmental pH of 4.0–4.5. The non-lactic acid bacteria would be considered 'passengers' that are typically present at lower numerical abundance, may have little influence on the ecology of the system, and might be lost from the community or change over time without markedly affecting community function. Vaginal communities seem to conform to this model at least from a numerical perspective since the rank abundance of species is highly skewed and lactic acid bacteria often outnumber others by two orders of magnitude, with diverse kinds of organisms present in lower numbers.

The occurrence of functional redundancy among 'drivers' (i.e., multiple species of lactic acid bacteria) may impact the stability and resilience of a community in the face of disturbances that

lead to changes in community structure or function. Such disturbances will differ in magnitude, frequency and source. Those disturbances that are small in scale or infrequent may not affect communities that have functional redundancy in driver species, because if one driver is disadvantaged or lost, another can compensate and thereby maintain community function. If, however, the disturbance is overwhelming or the community is driven by a single species, it may be more susceptible to change in function. This change in function could be harmful if it is necessary for health, or it can provide opportunities for colonization by pathogens.

1.3.2 *Resilience and stability*

Ecological theory and empirical data indicate that communities are not equally resilient or equally stable. The stability (or resistance) of a community reflects its capacity to resist change in structure or function in response to a disturbance event (McCann, 2000) as depicted in Figure 1.3, while resilience reflects the ability of a community to recover from a disturbance and return to a 'quasi-stable' equilibrium state (Pimm, 1984; Gunderson, 2000) as depicted in Figure 1.4. The resistance of a community is reflected in the magnitude of change that can occur without having an impact on community function, whereas resilience is a measure of disturbance frequency or intensity that alters community function. Both the resistance and resilience of communities are largely determined by the ability of ecological networks of indigenous species to tolerate stresses and disturbance events, the physical, chemical and metabolic interactions among the species present, and the degree of functional redundancy present. A disturbance event is an environmental change that causes shifts in population densities, the gain or loss of species, and concomitant changes to community function (White and Jentsch, 2001). The response of any community to one or more disturbance events (or to a disturbance regime characterized by a distribution of disturbance sizes, frequencies, intensities, and timing) is determined by the attributes of component species (Hobbs and Huenneke, 1992). Disturbed communities may or may not return to their previous state (White and Jentsch, 2001). The ability of an ecosystem to buffer against perturbations and resist species invasions is dependent on the redundancy of species that have important stabilizing roles as well as their ability to differentially respond to perturbations. This 'insurance hypothesis' posits that increasing diversity increases the odds that at least some species will respond differentially to variable conditions and perturbations, and that

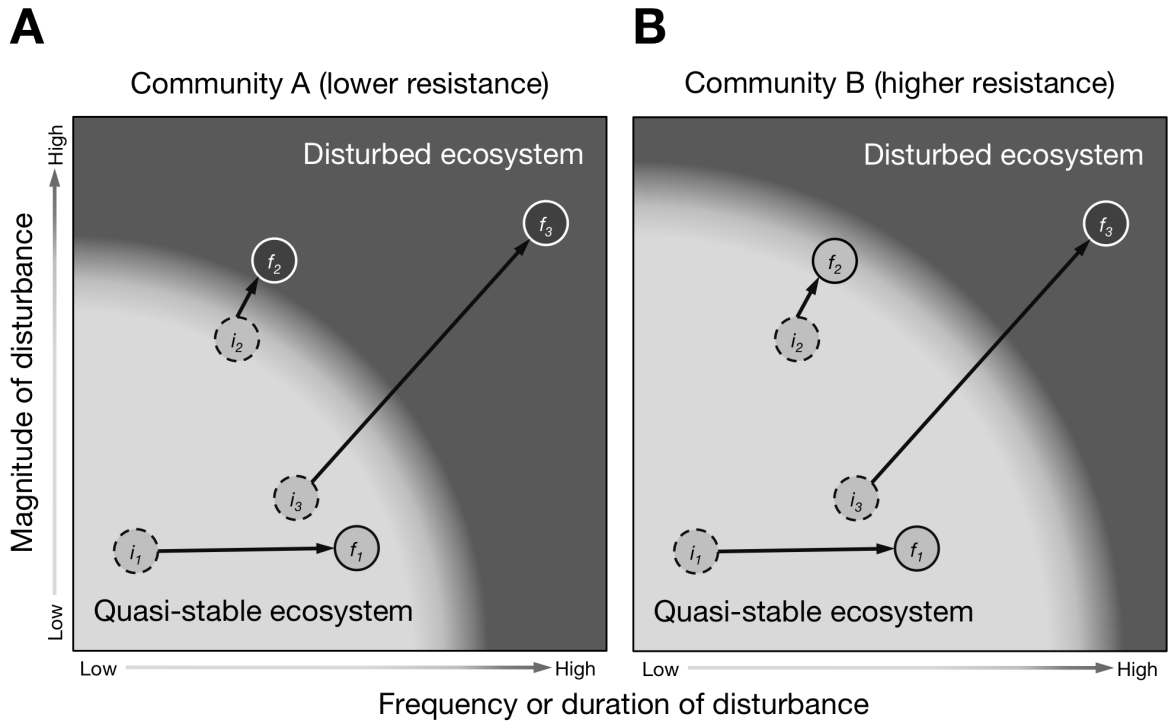

Figure 1.3: Conceptual schematic of community resistance. The resistance or 'quasi-stability' of a community reflects its capacity to resist change in structure in response to a disturbance event. Ecosystem disturbances can occur at varying intensities and frequencies or durations, indicated here on the y-axis (magnitude or intensity) and x-axis (frequency or duration), respectively. Panels (A) and (B) represent two communities with different levels of resistance. The lighter portion in the bottom left-hand portion of each space represents an ecosystem's quasi-stable state in which changes may occur to the community structure without pushing it into a 'disturbed' state. The darker portion in the upper right represents the disturbed ecosystem. Circles surrounded by dashed lines and marked *i* represent various initial states of an ecosystem, and circles with solid lines and marked *f* represent the final state following a disturbance event. Some disturbances may push the ecosystem to another point within its quasi-stable space (e.g., f_1 in [A]; f_1 and f_2 in [B]) whereas some disturbances may be great enough to push the community into a 'disturbed' state (e.g., f_2 and f_3 in [A]; f_3 in [B]). Communities that differ in species composition are likely to have different degrees of resistance. In our example, communities A and B experience the same disturbances, but in (A) disturbance events 2 and 3 push the community into a disturbed state whereas in (B) only disturbance event 3 is strong enough to disturb the ecosystem from its quasi-stable state.

greater diversity increases the odds that an ecosystem has functional redundancy by containing species that are capable of functionally replacing important species (McCann, 2000).

Events that destabilize microbial communities are not equal in terms of intensity or duration. Communities with low resistance and resilience may be disturbed by a single intense event of short duration, or multiple events of moderate to low intensity. This can result in transitory changes to the structure of these communities rendering them more susceptible to invasion by species that are not indigenous to the human vagina including transient species of fecal origin and opportunistic pathogens. In contrast, robust vaginal communities will exhibit stability in the face of more frequent events of low to moderate intensity and retain ecological functions that are characteristic of healthy communities. Given this, we postulate that stability and resilience of vaginal bacterial communities are likely to vary widely since the species composition and structure of these communities differs among women and this in turn may account for differences in the susceptibility of individuals to urogenital infectious diseases. Nonetheless, this hypothesis has yet to be empirically tested. To date, most studies of vaginal microbiology have relied on cross-sectional study designs in which individuals are sampled at one time point or over regular intervals of a few weeks or months (Wilks and Tabaqchali, 1987; Eschenbach *et al.*, 2000; Coolen *et al.*, 2005; Srinivasan *et al.*, 2010). As a result, very little is known about the dynamics of vaginal microbiota over short time scales and in response to various host-associated perturbations.

1.3.3 *Susceptibility to invasion by nonindigenous species*

As discussed above, acidification of the vaginal milieu via lactic acid production is probably important to preventing the invasion of vaginal communities by nonindigenous organisms. It has been suggested that hydrogen peroxide may also help prevent such invasions since some vaginal species of *Lactobacillus* produce H_2O_2 *in vitro* (Eschenbach *et al.*, 1989; Hawes *et al.*, 1996; Rosenstein *et al.*, 1997; Antonio *et al.*, 1999; Vallor *et al.*, 2001; Wilks *et al.*, 2004). However recent studies have shown this is not likely the case *in vivo* since dissolved oxygen levels in the vagina are exceptionally low (O'Hanlon *et al.*, 2010; Linhares *et al.*, 2010b; O'Hanlon *et al.*, 2011). Other mechanisms such as production of bacteriocins have also been suggested (Aroutcheva *et al.*, 2001a) but their importance has yet to be documented.

The establishment of a pathogen in a community closely mirrors the process of exotic species invasion in plant and animal communities. One concept from invasive species research that

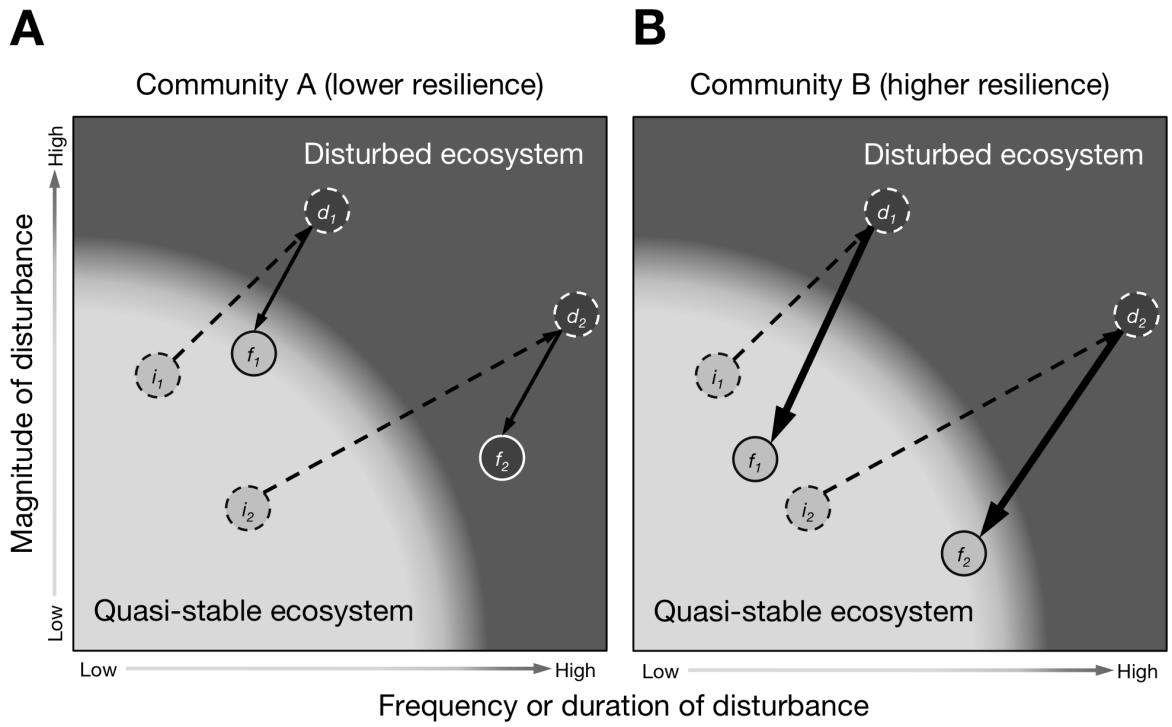

Figure 1.4: Conceptual schematic of community resilience. Resilience is the ability of a community to return to a quasi-stable state following a disturbance event. Different communities, particularly if they differ in species composition, are presumed to possess different degrees of resilience. The x-axis and y-axis are the same as in Figure 1.3. Panels (A) and (B) represent two communities with different levels of resilience. Circles surrounded by dashed lines and marked *i* represent various initial states of an ecosystem, dashed circles marked *d* represent intermediate disturbed states, and circles with solid lines and marked *f* represent the final state following the disturbance. Dashed arrows indicate disturbance events (these are the same location, direction and magnitude in [A] and [B]), and the solid arrows indicate the community rebounding toward its quasi-stable state. The thickness of the solid line represents the relative degree of resilience. In this case, the resilience of community A is sufficient to restore quasi-stability in disturbance event 1 but not 2, whereas the resilience of community B is sufficient to recover from both disturbance events.

applies most readily is the relationship between disturbance and invasion. A review by Didham *et al.* (2005) examined whether invasive species were drivers of change in community composition or passengers that took advantage of a disturbance that had changed the community's structure. The implication for medical microbiology is that without disturbance a pathogen maybe be able to invade some community types but not others. The second concept is the role of redundancy in driver species to maintain function and prevent invasion. Understanding the extent of functional redundancy and the contribution of specific community members to a healthy vaginal microbiota could help define risk factors for infection.

Ecologists have long known that the biological communities of disturbed ecosystems are more susceptible to invasion by non-indigenous, 'weedy' species (Hobbs and Huenneke, 1992). This is likely true for the bacterial communities of the human vagina too. If the proliferation of these invasive species proceeds unchecked, it could lead to clinically significant symptoms and disease. If the resilience of a vaginal community is low, then transitory changes to the structure of these communities may occur more readily in response to disturbances and these disturbed communities may be more susceptible to invasion by species that are not indigenous to the human vagina. These might include transient species of fecal origin and opportunistic pathogens. Moreover, we speculate that the disturbed state may itself constitute the clinical syndrome of BV where there is a reduction in the presence of lactic acid bacteria.

1.3.4 *The role of the host in shaping the vaginal microbiome*

Evidence that host specific characteristics influence the species composition and dynamics of microbial communities that colonize the vagina is accruing, though direct evidence is lacking. As described previously, recent studies have shown that the vaginal communities of women can be classified into several types based on similarities in bacterial community composition. This can be viewed in two ways. One is that vaginal communities can show marked differences in the composition and rank abundance of species present, and these differences may be potentially important in terms of ecosystem resilience and resistance to infectious agents. On the other hand, it also demonstrates that the differences among women are apparently not boundless, and therefore colonization of the host and vaginal community composition are probably not random events. Said plainly, there appear to be host factors that facilitate or select for bacterial species with particular characteristics. These might be linked to the presence of epithelial cell surface

receptors, variation in the composition or amount of vaginal secretions, the host immune system, or other factors (reviewed in Linhares *et al.*, 2010a and Wira *et al.*, 2005).

The notion that there is selection for particular bacterial species by host-determined characteristics is also supported by the observation that a rather limited number of different *Lactobacillus* species are found to dominate vaginal communities. Recent reports suggest only four species of *Lactobacillus*, namely, *L. crispatus, L. jensenii, L. gasseri* and *L. iners*, are found as dominant members of vaginal communities. This is surprising given the plethora of different species of lactobacilli that are recognized (Schleifer and Ludwig, 1995) and suggests that species of lactobacilli found in the vagina possess characteristics that allow them to compete and be successful under the environmental conditions of the vagina. If there is selection within an individual for species (or strains of species) that possess a suite of specific characteristics, then this could have important implications for efforts to develop prebiotics and probiotics for maintaining or re-establishing normal, healthy vaginal communities, because ideally they would be tailored to reflect differences in the species composition of an individual's normal community.

1.4 Defining disease: mysteries and myths of bacterial vaginosis

1.4.1 *Bacterial vaginosis: an enigma of women's health*

Bacterial vaginosis (BV) is the most frequently cited cause of vaginal discharge and malodor and the most common vaginal disorder of reproductive-age women, resulting in millions of health care visits annually in the United States alone (Sobel, 2000; Koumans *et al.*, 2007). Moreover, in non-pregnant women BV is associated with serious adverse sequelae including infertility (Sweet, 1995), endometritis (Haggerty *et al.*, 2004), and pelvic inflammatory disease (Wiesenfeld *et al.*, 2002), as well as an increased risk of acquiring HIV, *Neisseria gonorrhoeae*, and other sexually transmitted diseases (Martin *et al.*, 1999; Wiesenfeld *et al.*, 2003; Schmid *et al.*, 2000) During pregnancy, BV is associated with several adverse outcomes including preterm delivery of low birth weight infants (Hillier *et al.*, 1995), spontaneous abortion (Ralph *et al.*, 1999), and postpartum endometritis (Leitich *et al.*, 2003). The exact etiology of BV remains elusive, but different pathogenicity models have been proposed involving either the depletion or displacement of lactobacilli in the development of BV as depicted in Figure 1.5 (Srinivasan and Fredricks, 2008).

The prevalence of BV among women varies widely and depends on the subject population. It is present in 10–20% of white, non-Hispanic women and 30–50% of African-American women, and it may occur in up to 85% of sex workers in Africa (Sobel, 2000; Morris *et al.*, 2001; Newton *et al.*, 2001). It has a prevalence of 5–26% in pregnant women worldwide (Goldenberg *et al.*, 1996).

Despite over a century of work, attempts to find a single causative agent have failed. Since Gardner and Dukes (1955) first implicated *Gardnerella vaginalis* as a causative agent of BV, numerous efforts have been made to associate BV with the presence of certain bacteria in hopes of identifying an infectious agent. In recent years investigators have used cultivation-independent methods to continue the search for organisms that might cause BV. Fredricks *et al.* (2005) used broad-range PCR and sequencing of 16S rRNA genes to find that women with BV had a relatively high prevalence abundance of bacteria such as *Atopobium vaginae, Leptotrichia amnionii, Sneathia sanguinegens, Porphyromonas asaccharolytica, G. vaginalis*, and novel members of the Clostridiales referred to as BV-associated bacteria (BVAB). Moreover, in a study by Ferris *et al.* (2007) broad range PCR assays were also used to characterize the vaginal microbiota before and after metronidazole treatment to find that the diversity of anaerobic bacterial types of BV flora was shifted to a predominantly *L. iners* microbiota in cured patients and that unresponsive patients had the highest concentrations of *A. vaginae*. While these studies corroborate the co-occurrence of BV with so-called BV-associated bacteria, it is unclear if these bacterial species are causally related to the symptoms of BV or whether the association is contingent on the criteria used to diagnose BV. Studies in our lab and others have demonstrated these very same suspected agents and other closely related bacteria are found in asymptomatic women though perhaps at somewhat lower number (Figure 1.6). This argues that these organisms might not be 'infectious agents' in the strict sense, but when present in high number they might elicit some or all of the symptoms classically associated with bacterial vaginosis. That said, it is risky to deduce that certain organisms cause BV simply because they are abundant when symptoms are present.

Others have pursued the question of whether BV is a sexually transmitted disease. Evidence supporting this notion comes from the observed concordance of BV status in monogamous lesbian couples ranges up to 95% (Marrazzo *et al.*, 2002, 2008). Also, on average women with BV have more sex partners and an earlier age of sexual debut than women without BV (Schwebke *et al.*, 1999, 2004), and an association between receptive oral sex and BV has been suggested. However, BV has been diagnosed in asymptomatic virginal women (Yen *et al.*, 2003; Jones *et al.*,

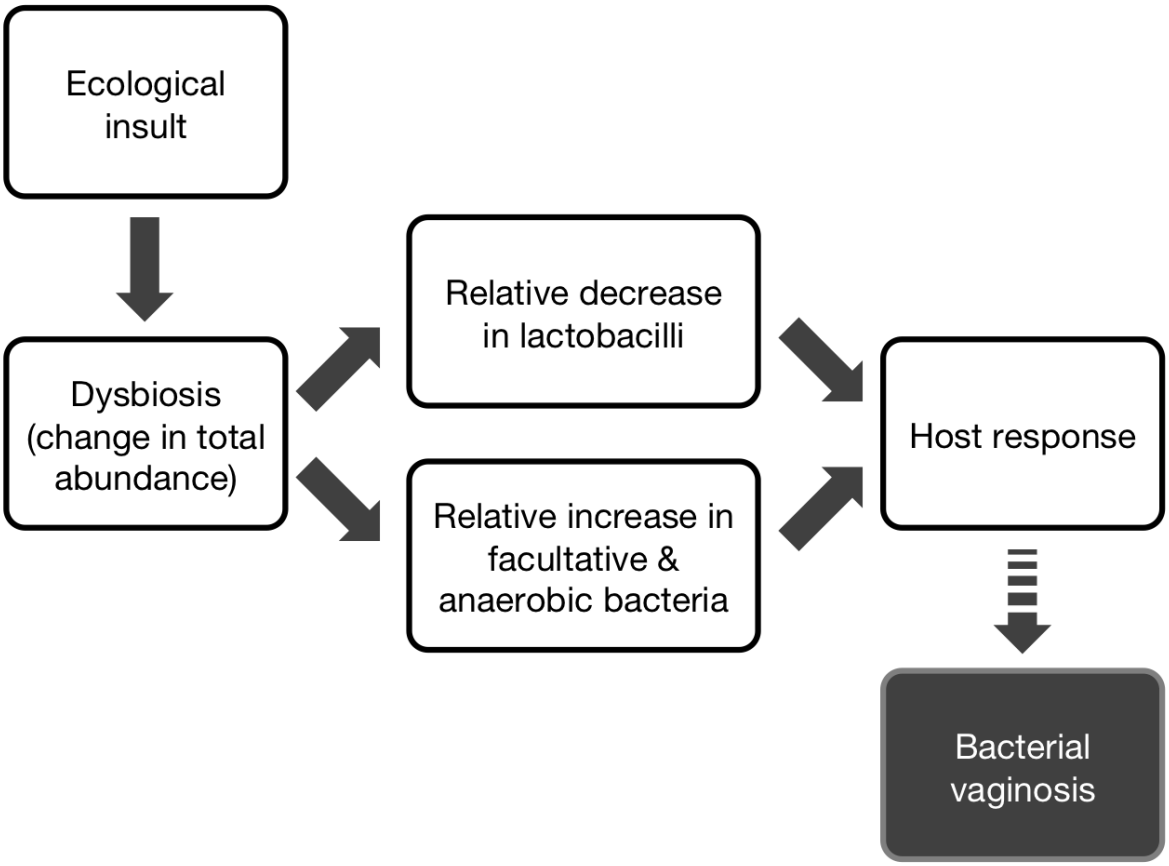

Figure 1.5: Conceptual models for the pathogenesis of bacterial vaginosis. Following an ecological insult or disturbance, dysbiosis may result when there is a change in the total abundance of microorganisms. This could result in a relative decrease in lactobacilli or a relative increase in facultative and anaerobic bacteria. Both scenarios may elicit a host response that eventually results in bacterial vaginosis (BV). This schematic is modified from one presented by Srinivasan and Fredricks (2008) wherein the two models were referred to as the '*Lactobacillus* depletion model' and 'primary pathogen model.'

2007; Vaca *et al.*, 2009), and this calls into question what, if any, role sexual behaviors may have in the acquisition of BV. On balance the data available support the hypothesis that BV is not an infectious disease.

1.4.2 *Is bacterial vaginosis being mischaracterized as a disease?*

Over the years, the definition of BV and the diagnostic criteria commonly used have been conflated, and they remain mired in controversy. The Amsel test, which is often used for the clinical diagnosis of BV, is based on four criteria: (a) a vaginal pH of >4.5, (b) the presence of clue cells, (c) a fishy odor upon addition of 10% KOH to vaginal discharge, and (d) a white, thin, homogenous vaginal discharge (Amsel *et al.*, 1983). The diagnosis of BV is made if at least three of these criteria are confirmed. The gold standard for the diagnosis of BV in research and laboratory settings has been the Nugent score (Nugent *et al.*, 1991). This diagnostic test is a scored scale based on (1) the presence of Gram-positive rods (*Lactobacillus* morphotypes) (2) the presence of Gram-variable rods and cocci (*Gardnerella vaginalis, Prevotella, Porphyromonas*, and peptostreptococci morphotypes) and (3) the presence of curved Gram-variable rods (*Mobiluncus* morphotypes). In a formal sense, an obvious potential problem is the logic of the Nugent score premise that high numbers of *Lactobacillus* spp. define 'health', and this imposes a bias against normal vaginal microbial communities that lack appreciable numbers of lactobacilli, yet maintain a low pH. The Amsel test, on the other hand, may lack sensitivity due to the subjectivity of the clinician's interpretation. Reports comparing the two diagnostic measures arrive at opposing conclusions (Chaijareenont *et al.*, 2004; Sha *et al.*, 2005), which has led many to suspect the accuracy of these tests.

Two fallacies permeate thinking about the diagnosis and treatment of BV. The first is that BV is an infectious disease. This seems to arise from classical thinking about infectious diseases in the framework of Koch's postulates, wherein a single species is both necessary and sufficient to cause infection. Although this is certainly true for a number of pathogens, it may be inadequate as an explanation for other diseases caused by mixtures of organisms. These so-called polymicrobial infections may not be infections in the strict sense of the word, wherein a nonindigenous organism invades a community. Instead they may be caused by indigenous populations that are typically rare but become abundant due to changes to ecologically important characteristics modulated by the host (e.g., nutrient levels) or disturbances that alter the competitive dynamics

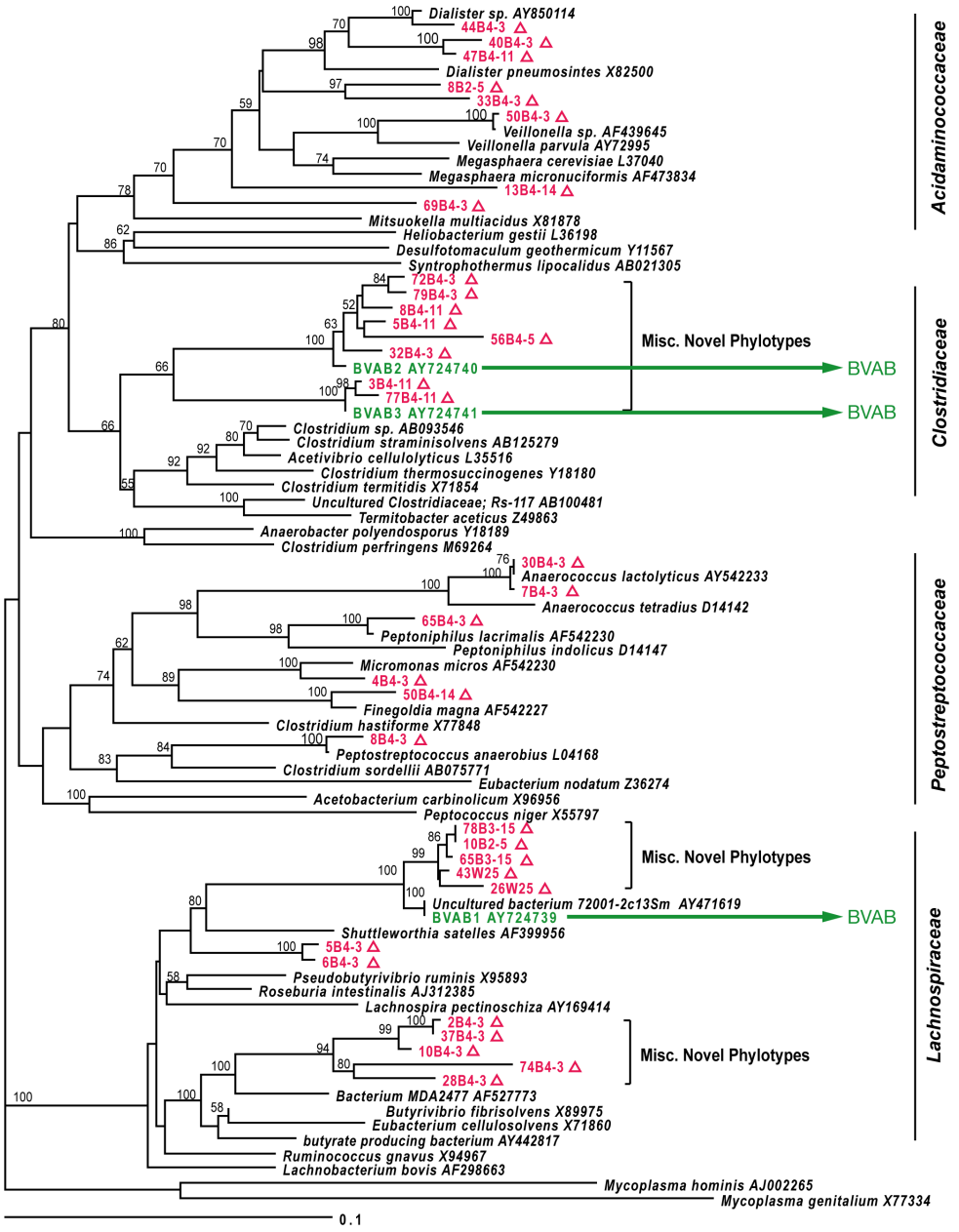

Figure 1.6: Diverse bacterial taxa found in the healthy vaginal microbiome. The phylogenetic tree shows the relationship of selected phylotypes from vaginal communities of healthy Caucasian and black women (marked by triangles; unpublished data from Zhou *et al.*, 2007), type strains from the RDP database (unmarked) and three BV-associated bacteria (BVAB) (marked with arrows; sequences deposited by Fredricks *et al.*, 2005). The tree was constructed using a neighbor-joining algorithm with *Mycoplasma* spp. serving as the outgroup. Bootstrap values (from 500 replicates) > 50% are shown at the branch points, and the bar indicates 10% sequence divergence.

of bacterial populations. In other words, the incidence of infectious disease may often depend not only on the competitiveness of an invasive species (i.e., the infectious agent) but also on the ecological dynamics of the habitat (i.e., anatomical site). To us it seems there is much to be gained from borrowing and testing ecological theory that has been developed over decades by plant and animal ecologists who have studied invasive species in a wide variety of circumstances.

A second fallacy is directly tied to the notion that the vaginas of normal healthy women are populated by high numbers of *Lactobacillus* spp. This statement is accurate so far as it goes. However, the converse statement—that women whose vaginal communities have few or no *Lactobacillus* spp.—are abnormal is unsupported by data. We postulate that because of this logical fallacy, BV is often over-diagnosed. This could partly account for the reported high incidence of so-called 'asymptomatic' BV in reproductive-age women (Sobel, 2000), and also explain a proportion of BV treatment failures and apparent recurrences of BV in women. Acknowledging that not all vaginal communities of healthy women are dominated by *Lactobacillus* spp. would also be in accordance with the observation that the vaginal communities of post-menopausal women (not taking hormone replacement therapy) often lack appreciable numbers of *Lactobacillus* spp., yet these individuals do not exhibit other untoward symptoms. We suspect that the causes and cures of BV will continue to be enigmatic until it is recognized that while 'normal and healthy' can be equated with high numbers of lactobacilli, the converse statement ('unhealthy' is equated with low numbers of or no lactobacilli) is not necessarily true. We must be vigilant and realize that for a significant proportion of women 'normal and healthy' can also occur in the absence of appreciable numbers of *Lactobacillus* spp.

1.5 CONCLUDING REMARKS

Multiple kinds of normal vaginal microbial communities are found in healthy women. The data provide strong evidence that more species than lactobacilli can dominate the vaginal microbial ecosystem of healthy women. The community function of maintaining low pH is highly conserved among women despite the differences in their vaginal microbial community composition and structure. These communities are postulated to provide different levels of protection against disease and infection, and their ability to offer protection may be lessened if the communities are disrupted. We propose that all vaginal microbial communities are not equally resilient and

their stabilities differ in the face of disturbances. Moreover, differences in the resilience of various vaginal microbial communities may account for the differential susceptibility of races to HIV, BV and other urogenital infectious diseases. Obviously our knowledge of the factors that affect and control the vaginal microbiota is still incomplete and increasingly we should view the vagina as a microbial ecosystem so that we can better understand the full range of factors that affect risk to disease.

CHAPTER 2

VChip, a microarray for functional analysis of the vaginal microbiome[2]

2.1 SUMMARY

We report on the development and validation of the VChip, a DNA microarray for exploratory analysis of the species and gene composition of vaginal microbial communities. The three-plex microarray consists of 1.4 million 60-mer probes derived from the coding DNA sequences of 313 strains of bacterial species commonly found in the vagina along with 716 human immunity genes. We performed a series of validation experiments with the VChip to demonstrate its efficacy with mock bacterial communities and clinical vaginal swabs. Through these experiments we confirmed that the VChip specifically detected expected bacterial species from DNA in mock communities and from both DNA and cDNA in vaginal swabs. Furthermore, it was sensitive to detection of bacterial and human immunity genes in a wide range of concentrations. VChip produced similar overall patterns of species and gene presence as high-throughput sequencing methods including 16S rRNA gene amplicon pyrosequencing and Illumina shotgun metatranscriptome sequencing of vaginal swabs. The species-specific probe sets can be extended to other applications for qualitative analysis of the vaginal microbiome, such as associating gene expression with health outcomes or rapidly screening a large number of samples for patterns of interest.

2.2 INTRODUCTION

Bacterial communities in the human vagina exhibit a mutualistic association with their host and play an important role in maintaining health and preventing disease. Several different kinds of vaginal communities occur in reproductive-age women (Zhou *et al.*, 2007; Ravel *et al.*, 2011), and the majority of these are dominated by lactic acid bacteria that provide a key ecosystem service of lactic acid production, which lowers the environmental pH and restricts the growth of non-indigenous bacteria (Boskey *et al.*, 1999, 2001; Linhares *et al.*, 2010b). *Lactobacillus* spp. are most

[2]This chapter is to be submitted for publication as: Hickey R.J., Hunter S.S., Settles M.L., Ma B., Myers G.S.A., Sun Y., Ravel J., and Forney L.J. VChip, a microarray for functional analysis of the vaginal microbiome.

commonly identified as the predominant members of healthy vaginal communities (Tärnberg *et al.*, 2002; Hyman *et al.*, 2005; Hill *et al.*, 2005), but other lactic acid bacteria such as *Atopobium* and *Streptococcus* probably serve a similar role in some communities (Zhou *et al.*, 2004). A diverse assortment of other bacterial taxa are also present in the vaginal communities of healthy women, and these may further modify the environment in ways that either facilitate or preclude colonization by other species (Redondo-Lopez *et al.*, 1990; Aroutcheva *et al.*, 2001a; Valore *et al.*, 2002). Several studies have reported differences in community composition among women of different ethnicities, suggesting that host genetics may contribute to shaping the microbiome (Zhou *et al.*, 2007, 2010; Ravel *et al.*, 2011). Furthermore, vaginal communities are dynamic and species composition can vary significantly over time (Srinivasan *et al.*, 2010; Gajer *et al.*, 2012; Ravel *et al.*, 2013). This is likely to be influenced by the metabolic activities and interactions among bacteria and with the host immune response. Currently, the relationship between species composition, ecological function and stability of vaginal microbial communities is not well understood. To better understand these factors, we need to take a closer look at how bacterial communities function in their natural environment and expand our thinking of ecosystem services beyond lactic acid production.

Detailed studies of the vaginal microbiome have relied primarily on phylogenetic analysis of partial 16S rRNA gene sequences (Fredricks *et al.*, 2005; Zhou *et al.*, 2007; Ravel *et al.*, 2011) or other conserved genes (Hill *et al.*, 2005) to survey the types and relative abundances of bacterial taxa present. While these studies yield significant insight to the diversity of bacterial phylotypes found in vaginal communities, the use of a single gene such as 16S rRNA provides only limited information about gene function and metabolic capabilities, either potential or expressed. This is due to the common observation of high levels of intraspecies genomic diversity in bacteria even in cases where 16S rRNA gene sequences are nearly identical (Lan and Reeves, 2000). Therefore, there is a pressing need for research tools that enable fast, targeted and simplified functional analysis of vaginal communities to facilitate comprehension of how community function differs in relation to species composition, metabolic activity and interspecies interactions. Analyses of this nature warrant a more comprehensive approach involving many genes that are present among many members of a community. Metagenomic approaches that enable characterization of the total genomic content of microbial communities have gained popularity in recent years (Qin *et al.*, 2010; Gilbert and Dupont, 2011) but so far have not been widely adopted for vaginal

microbiome research, perhaps owing to the significant computational demands of metagenomic analysis.

The ability to screen many samples for differences in species composition and gene content would be greatly beneficial for conducting large-scale studies and developing informed hypotheses about community function in the vaginal microbiome. DNA microarrays are ideal for this purpose because they are relatively simple to use, rely on established laboratory techniques and can be analyzed using statistical methods that are robust across many datasets (Irizarry *et al.*, 2003). Microarrays have been used in other areas of microbial ecology research for many years (Wagner *et al.*, 2007; Gentry *et al.*, 2006). These primarily consist of phylogenetic microarrays that probe group-specific marker genes (e.g., 16S rRNA) used to classify taxa in a community or estimate species richness (DeAngelis *et al.*, 2011; Tottey *et al.*, 2013; Ballarini *et al.*, 2013; Rajilić-Stojanović *et al.*, 2009) much the same way that 16S rRNA amplicon sequencing is used to characterize community composition. In addition, some microarrays probe specifically for genes of functional relevance such as those involved in biochemical pathways and nutrient cycling (He *et al.*, 2007; Zhou *et al.*, 2013; Lee *et al.*, 2013). Numerous microarrays developed for terrestrial, aquatic and human-associated bacterial communities have been reviewed previously (Gentry *et al.*, 2006; Paliy and Agans, 2012; Zhou, 2003). To our knowledge two vaginal microbiome-targeted microarrays have been developed to screen for bacteria associated with either healthy conditions or bacterial vaginosis (Dols *et al.*, 2011; Cruciani *et al.*, 2015); however, both arrays target only 16S rRNA gene sequences and therefore have limited utility for characterizing community function.

We developed a DNA microarray—termed the VChip—for exploratory analysis of community composition and gene expression in vaginal microbial communities. The motivations for this project were an incomplete understanding of the ecological functions of the vaginal microbiome and its role in women's health as well as a lack of high-throughput research tools to rapidly evaluate the composition of vaginal microbial metagenomes and metatranscriptomes. Using recent data on the taxonomic composition and rank-abundances of species in the vaginal communities of healthy women (Ravel *et al.*, 2011) and publicly available genome sequences of host-associated vaginal bacteria, we designed an extensive probe set (1.4 million 60-mer probes per sub-array of a three-plex glass slide array) targeting the genomes of 313 bacterial strains found in the vagina along with 716 human immunity genes. We performed a series of validation

experiments with the VChip to demonstrate its efficacy for analysis of species composition in mock bacterial communities and DNA and cDNA prepared from clinical vaginal swabs from healthy adult women. Our findings support the utility of VChip as a functional screening tool with a variety of potential applications, including identification of genes or taxa that contribute to community ecological function or health outcomes of the host, or rapidly screening a large number of samples to facilitate selection of a manageable subset for further investigations.

2.3 MATERIALS AND METHODS

2.3.1 *Probe design and array production*

Literature from previous studies of the vaginal microbiome (Ravel *et al.*, 2011; Gajer *et al.*, 2012) guided the selection of bacterial genomes to be used in designing probes for the microarray. The NimbleGen array format used was a three-plex 4.2 million-probe custom DNA array (1.4 million probes per sub-array). Probe sets were designed based on the coding DNA sequences (CDS) of 313 bacterial strains (184 species, Table A.1) associated with the vaginal microbiome and included species that are typically associated with healthy conditions (e.g., *Lactobacillus* spp.) as well as others that have been associated with bacterial vaginosis (e.g., *Gardnerella vaginalis*) or sexually transmitted diseases (e.g., *Chlamydia trachomatis*). We also included 716 human immunity genes in the array design to facilitate assessment of the human immune response. Detailed procedures for the design and production of the microarray are described in Appendix A. Briefly, we clustered sequences based on 80% identity and coverage in Cd-hit (Li and Godzik, 2006) and performed multiple sequence alignments in MUSCLE (Edgar, 2004) to generate a representative sequence for each cluster. We submitted representative sequences to Roche NimbleGen (Madison, WI, USA) to design unique 60-mer probes for each gene cluster (at least two and up to five probes per cluster). Final array design and production were completed in collaboration between Roche NimbleGen and the Institute for Bioinformatics and Evolutionary Studies (IBEST) Genomics Resources Core at the University of Idaho. The NimbleGen Design File (NDF) containing probe sequences and other information about the array is available at `http://github.com/roxanahickey/vchip`.

Table 2.1: Composition of mock communities tested on the VChip

Species	Mock community								
	MC-1	MC-2	MC-3	MC-4	MC-5	MC-6	MC-7	MC-8	MC-9
	Proportion of genomic DNA (prop. expected genome copy equivalents[a])								
Anaerococcus hydrogenalis (vaginal isolate)	0.200 (0.183)	0.100 (0.096)	0.010 (0.010)	0.010 (0.096)	0.001 (0.018)	-	-	-	-
Anaerococcus tetradius (vaginal isolate)	0.200 (0.165)	0.100 (0.087)	0.010 (0.009)	0.010 (0.087)	0.001 (0.016)	-	-	-	-
Atopobium vaginae ATCC BAA-55	0.200 (0.246)	0.100 (0.130)	0.010 (0.014)	0.010 (0.129)	0.001 (0.024)	-	-	-	-
Finegoldia magna (vaginal isolate)	0.200 (0.184)	0.100 (0.097)	0.010 (0.010)	0.010 (0.096)	0.001 (0.018)	-	-	-	-
Gardnerella vaginalis ATCC 14018	0.200 (0.222)	0.100 (0.117)	0.010 (0.012)	0.010 (0.116)	0.001 (0.022)	-	-	-	-
Lactobacillus crispatus ATCC 33820	-	0.500 (0.473)	0.950 (0.945)	0.050 (0.470)	0.050 (0.889)	0.050 (0.987)	0.010 (0.937)	1.000 (1.000)	-
Homo sapiens (female genomic DNA)	-	-	-	0.900 (0.006)	0.945 (0.011)	0.950 (0.013)	0.990 (0.063)	-	1.000 (1.000)

[a] Genomic DNA proportions were converted to expected genome copy equivalent proportions based on estimated genome sizes (see Table A.2).

2.3.2 *Bacterial strains and vaginal swabs*

We constructed nine mock communities containing different proportions of genomic DNA (2.5 μg total) from six bacterial strains and human female genomic DNA (Promega, Madison, WI, USA) as shown in Table 2.1. Human DNA was included to evaluate potential burden on bacterial detection and cross-hybridization with bacterial probes, as well as to test hybridization of the 716 human immunity genes on the array. Bacteria in the mock communities included *Lactobacillus crispatus* ATCC 33820, *Atopobium vaginae* ATCC BAA-55, *Gardnerella vaginalis* ATCC 14018, and three isolates from vaginal swabs classified as *Finegoldia magna, Anaerococcus tetradius*, and *Anaerococcus hydrogenalis* based BLAST (Altschul *et al.*, 1997) results with >97% similarity in V1–V5 16S rRNA gene sequences of species in the NCBI GenBank database (`http://www.ncbi.nlm.nih.giv/genbank/`).

We also included three vaginal swabs from a 10-week longitudinal study of the vaginal microbiome (Ravel *et al.*, unpublished) to evaluate hybridization of DNA and mRNA (converted to cDNA) from clinical samples. The study was approved by the Institutional Review Board of the University of Maryland School of Medicine. Sample VM-1 was collected from an individual (Subject 1) whose vaginal community was stably dominated by *Lactobacillus iners* over the course of the study (Figure 2.1A). Samples VM-2 and VM-3 were collected at two time points in a second individual (Subject 2) during a period in which the vaginal community experienced an abrupt

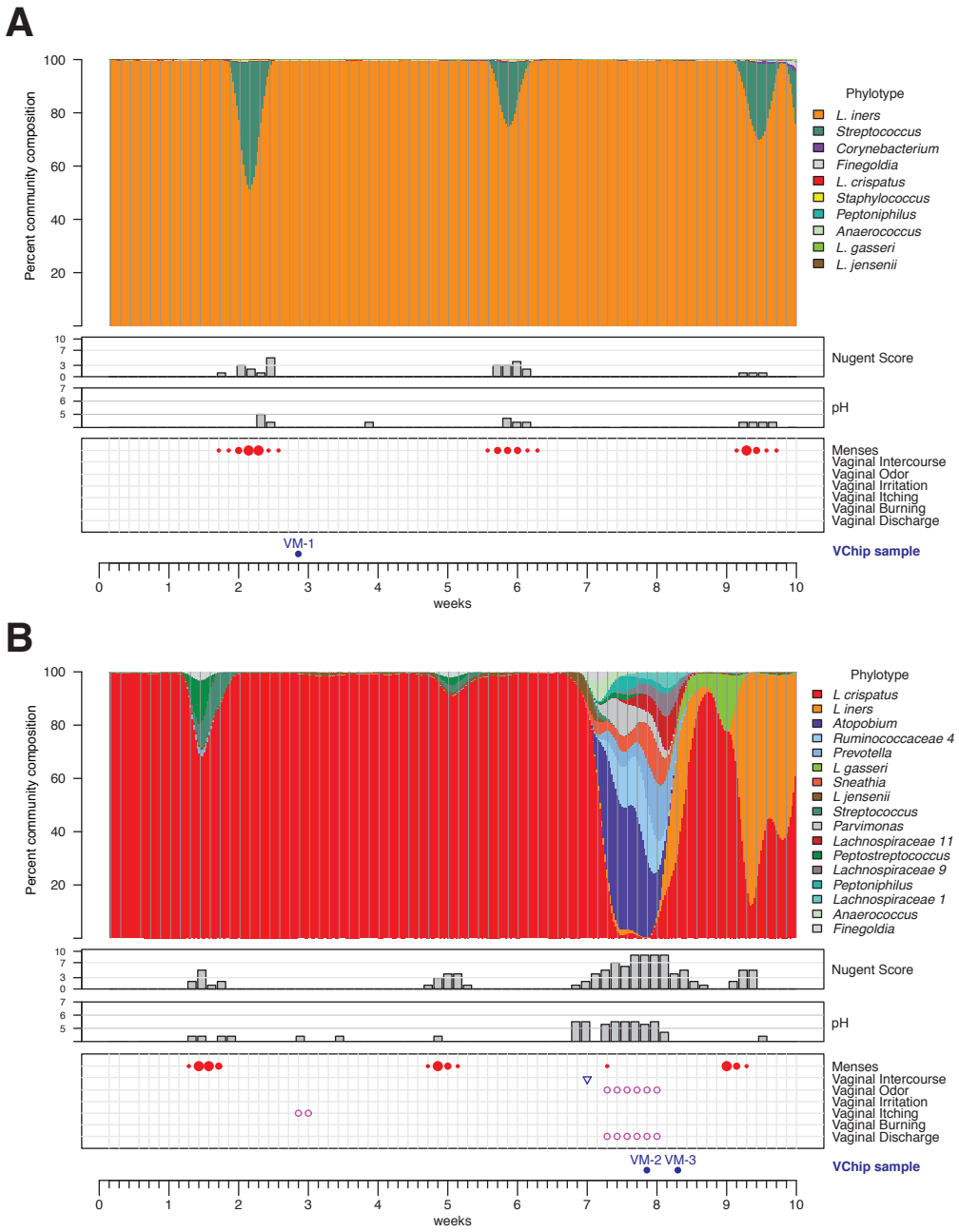

Figure 2.1: Daily temporal dynamics of vaginal bacterial communities in two women over 10 weeks. The vaginal microbiota of Subject 1 is shown in (A) and Subject 2 in (B). The relative abundances of phylotypes in each community are depicted as interpolated bar plots (top panel). Beneath these are profiles of Nugent scores (range 0–10) and vaginal pH (range 4–7). Occurrence of menses (red dots), vaginal intercourse (blue inverted triangles) and vaginal symptoms (pink open circles) are indicated in the bottom panel, with samples selected for VChip analysis indicated directly below.

but temporary shift in community composition before returning to a *Lactobacillus*-dominant state (Figure 2.1B).

2.3.3 *Preparation of genomic DNA*

Genomic DNA was extracted from vaginal swabs that had been stored in Amies transport medium (Copan Diagnostics, Murietta, CA, USA) at -80°C. We employed a validated procedure (Forney *et al.*, 2010; Yuan *et al.*, 2012) that includes steps for enzymatic and physical lysis of bacterial cells followed by purification of genomic DNA using a QIAsymphony robotic platform and Qiagen CellFree500 kits (Qiagen, Venlo, Limburg, Netherlands) according to the manufacturer's protocol. The same protocol was performed manually for extracting genomic DNA from bacterial pure cultures used to construct the mock communities.

2.3.4 *Sequencing of V1–V2 16S rRNA gene amplicons and cDNA derived from vaginal swabs*

We characterized bacterial composition of the three vaginal swabs using Roche 454 pyrosequencing of 16S rRNA gene V1–V2 hypervariable regions as previously reported (Gajer *et al.*, 2012). These data served as a comparison to the VChip hybridization of whole community DNA and cDNA to determine whether the species detected were reasonably consistent between methods. Metatranscriptome cDNA was prepared by extracting total RNA from vaginal swabs that had been stored in RNAlater and depleted of rRNA using a combination of Epicentre (Madison, WI, USA) Ribo-Zero rRNA removal kits for bacteria and human/mouse/rat kits. The remaining mRNA was reverse transcribed, and resulting double-stranded cDNA (~200 ng) was used to construct libraries for sequencing on an Illumina HiSeq 2000 instrument (Illumina, San Diego, CA, USA) at the University of Maryland Institute for Genome Sciences (IGS) using protocols recommended by the manufacturer and modified by the Genomic Resource Center at IGS. Relative abundances of individual transcripts were determined based on the depth of read coverage. Metatranscriptome data from this analysis are heretofore referred to as Illumina RNA-Seq data.

2.3.5 *Sample processing and hybridization*

Genomic DNA from mock communities and vaginal swabs and cDNA from vaginal swabs were processed in the IBEST Genomics Resources Core facility at the University of Idaho follow-

ing NimbleGen's protocols for comparative genomic hybridization (CGH) arrays (version 8.1) (Roche NimbleGen, 2011). Briefly, 0.5 μg of purified, unamplified, unfragmented genomic DNA or cDNA was labeled with high-efficiency Cy5 Random Nonamers, followed by hybridization and washing as described in the manual. Prepared samples were analyzed on a Roche NimbleGen MS200 scanner (Roche NimbleGen, Madison, WI, USA) along with standard quality controls.

Hybridization intensity signals were normalized using the Robust Multichip Average (RMA) method (Irizarry *et al.*, 2003). To facilitate comparisons between microarray hybridization results and expected mock community composition and vaginal swab sequencing data, we converted each species' RMA-normalized hybridization intensity value to a percent of the total signal across the entire array as follows: first, the 95th quantile of the hybridization values for the probes associated with a given species was calculated. Next, the median of the 95th quantile values was subtracted from the 95th quantile values, and negative values were set to zero. Finally, percentages of each signal were calculated from the values that were not log-transformed. The rationale for this approach is that due to stochastic processes not all probe sets within a given species will produce a hybridization signal. The 95th quantile represents probes with the highest signal within a group, while taking a median or mean results in values that are essentially indistinguishable from background signal. Hybridization values were adjusted accordingly according to this procedure.

2.3.6 *Data analysis*

Qualitative and statistical analyses were performed using custom R scripts and packages available from Bioconductor (`http://bioconductor.org`), including oligo (Carvalho and Irizarry, 2010), pdInfoBuilder (Falcon *et al.*, 2009), limma (Smyth, 2005), qvalue (Storey, 2002; Dabney *et al.*, 2004) and Biostrings (Pages *et al.*, 2013). Raw and summarized data files are available at `http://github.com/roxanahickey/vchip`.

2.3.7 *Analysis of mock communities*

To facilitate comparisons between mock community hybridization data with expected species proportions, we scaled the expected proportions of each species by their genome size. This was accomplished by dividing each species' genomic DNA proportion in a community by its

assembled genome size to estimate the expected proportion of genome copy equivalents (see Table 2.1). Because exact genome sizes were not known for many of the strains used in the mock communities, they were estimated based on the genome sizes of up to three strains with genomes available from NCBI (Table A.2). We excluded genomes that contained plasmids (in bacteria) or organelles (in human) and gave preference to contig and chromosome assemblies over scaffold assemblies. We compared the adjusted proportions of genome copy equivalents to the VChip species-specific normalized hybridization signal by calculation of Pearson correlation coefficients.

2.3.8 *Analysis of vaginal swab metagenomic DNA*

To compare hybridization of metagenomic DNA from vaginal swabs to community composition determined by V1–V2 16S rRNA gene sequencing, we considered only the subset of species targeted by VChip probe sets that could be matched directly to the 16S rRNA gene sequence data by genus or species name. This was necessary because many operational taxonomic units (OTUs) determined by bioinformatic analysis of the Roche 454 16S rRNA gene pyrosequencing data were not directly comparable due to nomenclature differences in the databases used for annotation. We calculated Pearson correlation coefficients for the species-level comparison of the two methods.

2.3.9 *Analysis of vaginal swab metatranscriptomic cDNA*

We performed *in silico* mapping of VChip probes against RNA-seq reads to compare hybridization of vaginal swab cDNA with Illumina RNA-Seq data. We mapped probe sequences (60 bp) against the Illumina RNA-Seq reads (100 bp) using Bowtie v0.12.9 ((Langmead *et al.*, 2009)) (parameters: "-v 3 -fullref -chunkmbs 512 -best -strata -m 20") and filtered the resulting alignments to include only those where the full length of each probe aligned to the read, allowing for two mismatches. The numbers of reads with mapped probes were summed and converted to a percent of total reads with mapped probes. These were compared to the metatranscriptome cDNA normalized hybridization values for vaginal swab samples VM-1 and VM-2 by calculation of Pearson correlation coefficients.

Finally, to demonstrate utility of VChip for comparative gene expression analysis, we compared the hybridization of cDNA from samples VM-2 and VM-3 (both from Subject 2). Following normalization of the data using procedures referenced above (Irizarry *et al.*, 2003), we calculated the $\log_2$-fold difference in expression values between the two samples. The number of gene clusters with >2 $\log_2$-fold difference (in magnitude) was determined for each species and averaged to evaluate whether overall change in gene expression for each species was overall up, down, or neutral in sample VM-3 relative to VM-2.

2.4 RESULTS

2.4.1 *Validation of VChip with mock communities*

We hybridized mock communities consisting of bacterial and human genomic DNA to assess VChip probe specificity to known species and sensitivity to different concentrations of DNA. Figure 2.2 compares the relative hybridization signals with expected proportions of genome equivalent copies for each species in the mock communities. The VChip specifically detected both the bacterial species and human DNA present, and Pearson correlation coefficients ranged from 0.34 to 0.95 (mean r=0.77, median r=0.85). Single-species mock communities had the highest correlations (MC-8, r=0.95; MC-9, r=0.92), while communities consisting of bacteria in uneven proportions had the lowest (MC-2, r=0.34; MC-3, r=0.60). Communities with both bacterial and human DNA also had relatively high correlation coefficients (MC-4 through MC-7; r=0.73–0.85), as did the mock community with balanced proportions of five bacterial species (MC-1, r=0.85). These results demonstrate that VChip probes were able to hybridize DNA from bacteria that constituted as little as 0.1% of the total (2.5 ng) genomic DNA.

Hybridization signal for probe sets assigned to species that were not present in the mock communities accounted for approximately 10% of the total signal overall, indicated by the 'Other' category in Figure 2.2. This was collectively made up of very low percentages of signal (typically <1%) across many species and was essentially indistinguishable from background noise (the RMA signal-to-percent conversion filters out most low-level signal, but some residual background noise remains). The species detected using the VChip were concordant with our expectations in the mock communities but also indicated an unexpected presence of *L. crispatus* in mock community MC-1 even though we had not knowingly added its DNA during sample prepara-

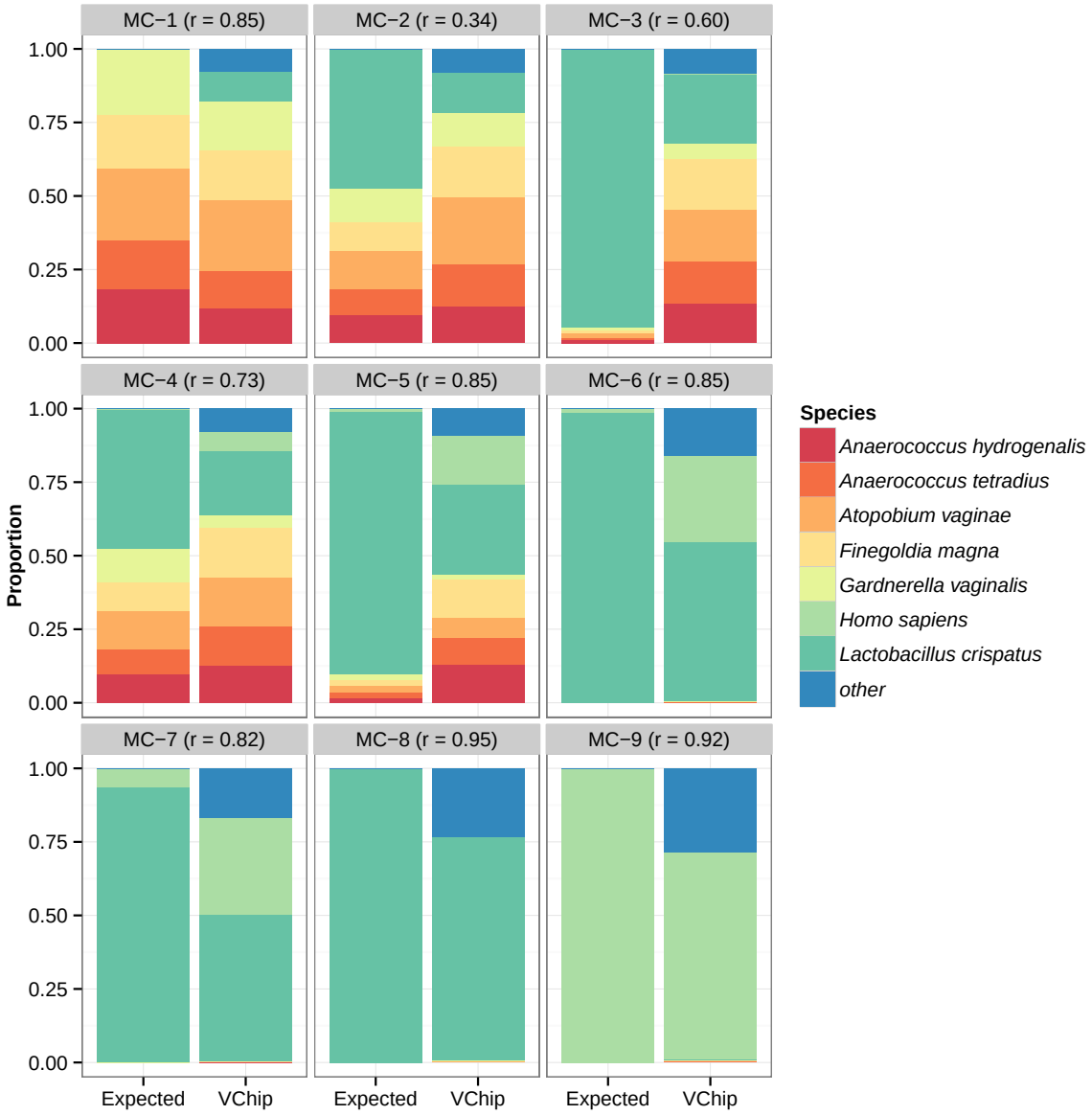

Figure 2.2: Comparison of VChip-derived species composition of mock communities to expected proportions of species' genome equivalent copies. Nine mock communities were constructed from genomic DNA as indicated in Table 2.1, and the proportions of DNA were converted to proportions of expected genome equivalent copies per species. The left bar in each plot indicates the expected proportions of genome equivalent copies per species in the mock community, and the right bar indicates the observed proportions of normalized hybridization signal attributed to each species on the VChip. The header of each subplot indicates the mock community label (MC-1 through MC-9) along with the Pearson correlation coefficient between observed and expected values for each sample. Species are indicated in the legend to the right of the plots. In addition to the six bacterial species and human, the 'other' category encompasses very low relative proportions (typically <1%) of signal across many species and is indistinguishable from levels of residual background noise.

tion. Comparisons with other arrays hybridized with samples containing *L. crispatus* suggested contamination of this sample and not a failure of the array (see Appendix A and Figure A.1).

2.4.2 *Validation of VChip with DNA from clinical samples*

As with the mock communities, the VChip hybridized DNA from bacteria present in a wide range of proportions in three vaginal swabs. Using Pearson correlation, we compared the VChip-derived bacterial species proportions with those based on V1–V2 16S rRNA gene sequencing for 42 bacterial species that had species name matches between the probe dataset and pyrosequencing dataset (Figure 2.3). Collectively those 42 species accounted for at least 80% of the hybridization signal from samples on the array, but only 15 species had relative abundances greater than 0% in the 16S rRNA gene pyrosequencing data. This is perhaps due to residual background noise on the microarray, or greater sensitivity of the VChip probes to low-abundance community members relative to 16S rRNA gene amplicon sequencing. In the case of the latter, the high sensitivity of VChip probes to low-abundance taxa could enable detection of important genes in species that might otherwise be overlooked. Pearson correlations between the hybridization signal (%) on the VChip and relative abundance based on 16S rRNA gene sequencing was high when the community was skewed toward a single dominant species. For instance, sample VM-1 was composed of >99% *Lactobacillus iners* based on 16S rRNA gene sequencing, and the Pearson correlation coefficient for comparison to VChip was 1.00. Samples VM-2 and VM-3, which had more evenly distributed species relative abundances, had lower correlation values of 0.53 and 0.84, respectively.

2.4.3 *Validation of VChip with cDNA from clinical samples*

We performed qualitative comparisons of cDNA hybridizations on the VChip with 16S rRNA gene pyrosequencing and Illumina RNA-Seq data to evaluate overall similarity in the species detected. One interesting outcome from this assessment was the unexpected abundance of transcripts detected from *Finegoldia magna* in sample VM-1. The relative abundance of this species was just 0.05% by 16S rRNA Roche 454 pyrosequencing, and *F. magna*-specific signal constituted less than 1% of the normalized signal in the metagenomic DNA hybridization. The cDNA hybridization from the same sample, on the other hand, constituted 17.5% of the normalized

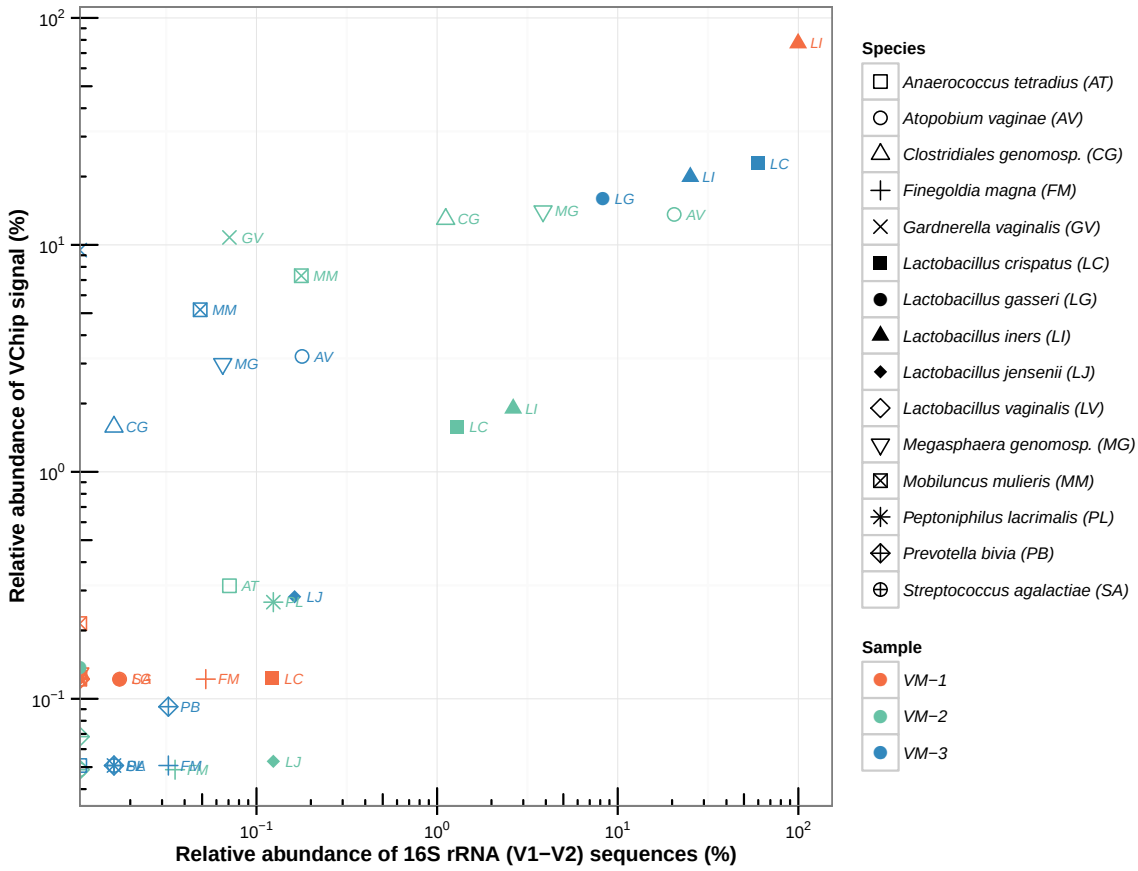

Figure 2.3: Comparison of species relative abundance in vaginal swabs detected by VChip and 16S rRNA V1–V2 pyrosequencing. Metagenomic DNA from three vaginal swab samples (sample VM-1, orange; VM-2, teal; VM-3, blue) were analyzed on the VChip. Hybridization signals were normalized and converted to relative abundances per species (y-axis) and compared to relative abundance data determined from V1–V2 16S rRNA pyrosequencing (x-axis). The data are plotted on a $\log_{10}$ scale to clearly separate out low relative abundance species. The top 15 most abundant of 42 species with exact name matches between the VChip taxa and pyrosequencing dataset are plotted as indicated in the legend to the right of the graph. Pearson correlation coefficients based on all 42 species are 1.00 for VM-1, 0.53 for VM-2 and 0.85 for VM-3.

Table 2.2: Pearson correlation of VChip vs. *in silico* mapping of Illumina RNA-Seq reads against probes[a]

Genus (above diagonal) / Species (below)	VM-1 cDNA	VM-1 reads	VM-2 cDNA	VM-2 reads
VM-1 cDNA	1.00	0.71	0.37	0.31
VM-1 reads	0.75	1.00	0.33	0.53
VM-2 cDNA	0.36	0.31	1.00	0.79
VM-2 reads	0.20	0.25	0.77	1.00

[a] RNA-Seq reads were mapped against VChip probe sequences *in silico* using Bowtie. Mapping results are compared to actual cDNA hybridizations on VChip at the genus and species level using Pearson correlation coefficients.

signal. Additionally, 3.41% of the Illumina RNA-Seq reads from this sample mapped to *F. magna*-specific VChip probe sets in the *in silico* analysis described below, providing further evidence of gene expression in *F. magna*. This possibility might have been overlooked based on 16S rRNA or metagenome sequencing due to the low relative abundance of this species. While observations like this could potentially spur further investigation, additional sample replicates would be necessary to draw statistically supported conclusions.

2.4.4 *In silico analysis of Illumina RNA-Seq data on VChip*

To assess how well the VChip hybridization of cDNA from vaginal swabs represented the 'true' species composition of the samples, we performed *in silico* hybridization by mapping VChip probe sequences against Illumina RNA-Seq reads and compared the results to those obtained directly from the VChip. Approximately 2.77% (n=2.26E6) of the total RNA-Seq reads from sample VM-1 and 2.23% (n=1.78E6) of the total reads from sample VM-2 were successfully mapped to VChip probe sequences (mapping was not performed for VM-3). Using the reads that mapped, Pearson correlation coefficients for species-specific relative hybridization abundances between the *in silico* mapping and the cDNA hybridization on the VChip were 0.71 and 0.79 at the genus level, and 0.75 and 0.77 at the species level, for VM-1 and VM-2, respectively (Table 2.2). Although only a small fraction of the reads mapped to the probe sequences, the high degree of correlation with the actual cDNA hybridized on the array indicates good agreement in the detection of cDNA fragments using two very different technologies.

There are several possible explanations for why most of the RNA-Seq reads did not map to the probe sequences. First, the percentage of the mapped reads is limited by the total number (n=1,443,693) and design of the probes. Each gene cluster represented on the array had at most

five 60-mer probes, so only up to 300 bp of any given gene cluster could potentially be mapped against the RNA-Seq reads. Assuming an average nucleotide length for a gene to be ~1000 bp, 70% of the DNA would be incompatible with the probes even if other segments of the genes were represented. Additionally, any human genes except for 716 immunity genes on the array would not have been detected, nor would eukaryotic or bacterial species not represented on the array. Further investigation indicated that ~20% of reads from VM-1 and ~2% of reads from VM-2 originated from human. Finally, although a ribosomal RNA depletion step was included prior to Illumina RNA-Seq sequencing, approximately 9% of the reads in each sample were identified as rRNA by bioinformatic assessment. Given these explanations, it is not surprising that a relatively low percentage of reads were mapped to VChip probes.

2.4.5 *Evaluation of gene expression changes between VM-2 and VM-3*

The cDNA hybridizations of samples VM-2 and VM-3 on the VChip were compared to observe differences in relative gene expression between two time points in an individual subject (Figure 2.4). A complete list of gene clusters with >2 $\log_2$-fold differences (in magnitude) between VM-2 and VM-3 are included in Supplementary File 1 (`http://github.com/roxanahickey/dissertation`). Many species, including several *Lactobacillus* spp., showed a large number of species-specific genes with increased average relative expression in the latter time point, which reflects the increase in the species' relative abundance over the timeframe observed (Figure 2.1B). These included genes such as a GNAT family acetyltransferase, initiator RepB protein and DNA methylase in *L. iners*, and a cell wall-associated hydrolase and endopeptidase O in *L. gasseri*, to name just a few examples. A large fraction of transcripts with the greatest number of differentially expressed genes with average positive change were attributed to less abundant taxa such as *Peptoniphilus lacrimalis, Dialister microaerophilus* and several others. *Atopobium vaginae* had the greatest number of genes with lower expression on average in the latter time point. Several other species including *Prevotella amnii, Anaerococcus tetradius, Clostridiales* genomosp. BVAB3 and *Mobiluncus curtisii* had large numbers of differentially expressed genes, but the average change across all species-specific genes was close to zero, indicating some genes were over-expressed in the earlier time point and others in the later. 190 human immunity genes were differentially expressed based on the >2 $\log_2$-fold threshold; of these, only two were lower (a myeloperoxidase, NCBI reference NM_000250; and NLRC4, NCBI reference NM_021209) while the rest were

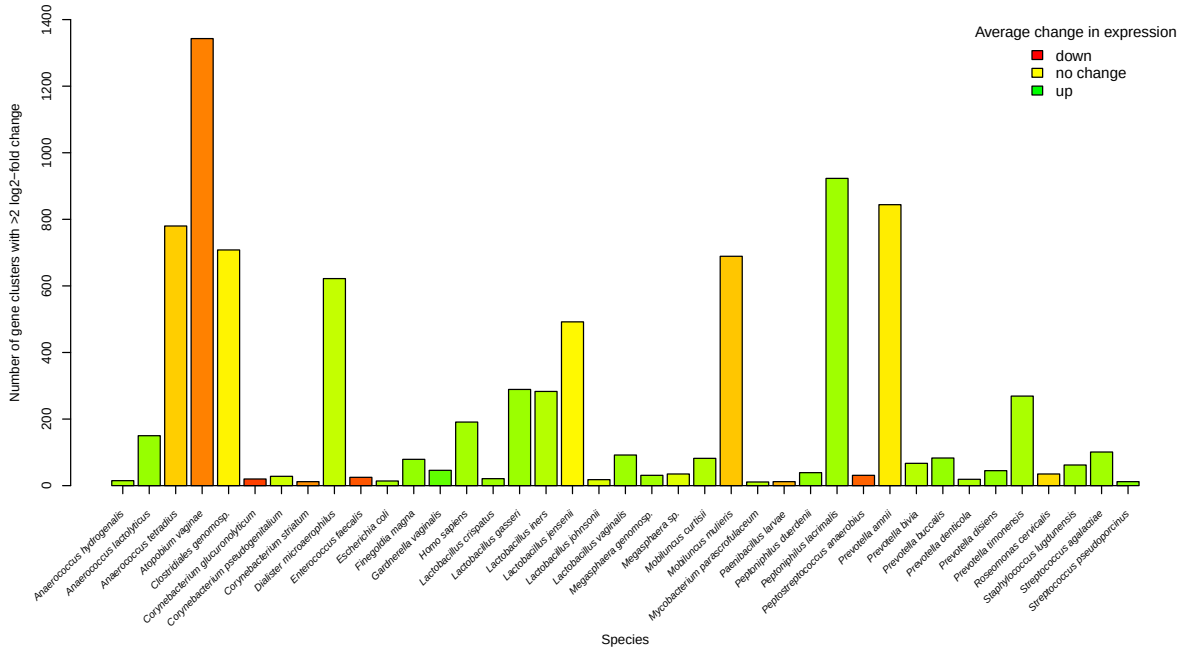

Figure 2.4: Species' gene expression changes in Subject 2. The number of species-specific gene clusters with >2 $\log_2$-fold differences (in magnitude) in normalized cDNA hybridization signal between samples VM-2 and VM-3 are indicated on the y-axis. The coloring of each bar represents the average $\log_2$ fold-change difference in hybridization of species-specific probe sets in VM-3 relative to VM-2. The bars are colored on a continuous scale ranging from red (greatest average negative change) to yellow (zero average change) to green (greatest average positive change).

higher in the later time point. Our results indicate that it is feasible to detect differences in patterns of gene expression between samples using the VChip. If similar comparisons were conducted on a larger scale, such changes in expression patterns could provide important insights into the mechanisms driving shifts in community composition and function.

2.5 DISCUSSION

Tools for streamlined functional analysis of vaginal microbial communities are needed to accelerate our understanding of ecosystem functioning in the vaginal microbiome and its relevance to vaginal community stability and urogenital health. Here we report on the development and validation of the VChip, a DNA microarray that surveys species composition and gene content of the vaginal microbiome more comprehensively than previous microarrays and is the first to our knowledge to probe human immunity genes along with bacteria. Because it is based on a standard microarray platform, data can be analyzed in a straightforward manner using established methods, and the probe sets can be tailored and reproduced on any oligonucleotide platform. The results of our validation experiments with both mock communities and clinical vaginal swabs support the utility of VChip as a tool for exploratory analysis of vaginal community gene content and expression. Although the NimbleGen array format we used has since been discontinued, future iterations of the VChip could be reproduced on alternative microarray platforms, and we have provided the necessary probe design files and additional information to enable such development (`http://github.com/roxanahickey/vchip`).

Other microarrays developed for human microbiome research have targeted bacteria residing in the human gastrointestinal tract (Tottey *et al.*, 2013; Rajilić-Stojanović *et al.*, 2009; Kang *et al.*, 2010), oral cavity (Crielaard *et al.*, 2011), vaginal tract (Dols *et al.*, 2011; Cruciani *et al.*, 2015) or multiple body habitats (Ballarini *et al.*, 2013). The majority of these are phylogenetic microarrays which are used to detect (and sometimes quantify the relative abundance of) microbes based on a single marker gene or small set of functional genes. We are aware of two published vaginal microarrays, both designed to detect vaginal bacteria using 16S rRNA gene amplicons. The first array described by Dols *et al.* (2011) probed a handful of species commonly associated with bacterial vaginosis (e.g. *Gardnerella vaginalis, Atopobium vaginae, Megasphaera*) but did not include any common *Lactobacillus* spp. associated with healthy conditions. The VaginArray

developed by Cruciani *et al.* (2015) included probe sets for 17 bacterial species, including the most common vaginal lactobacilli, *L. crispatus, L. iners, L. gasseri* and *L. jensenii*, but *Gardnerella vaginalis* was notably missing. Our VChip differs substantially from these because it probes both species-specific and conserved genes spanning whole genomes rather than a small set of marker genes, and it includes many representative of species that are associated with both healthy and unhealthy conditions. Moreover, because the VChip also includes human immunity genes, it could be used to evaluate the local host immune response along with vaginal microbial community gene expression.

A major advantage of the VChip's more comprehensive design is that it can be used in an exploratory manner to identify potential interactions between the host and vaginal microbial communities and relate those findings to health and disease. The process of gene selection for probe design of the VChip was agnostic toward metabolic or physiological function, so there are many genes represented on the array that have not yet been well characterized but could ostensibly be important for community ecology of the vaginal microbiota. Furthermore, probe sets are targeted at 184 bacterial species (313 strains) representing a wide spectrum of taxa found in the vagina at varying degrees of incidence and relative abundance. This enables detection of gene expression patterns even with so-called 'rare' or low-abundance bacteria, as was the case here with *F. magna* in the vaginal community of sample VM-1. Similarly, an investigator might want to determine which species contribute most to differences or changes in gene expression patterns across samples. Our comparison of two metatranscriptome samples from subject 2 (VM-2 and VM-3) revealed that several species displayed large changes in transcript abundance from one time point to another, which could prompt more targeted analysis in future studies. However, the hybridization results need not be partitioned by species-specific probes at all; it is also feasible to compare the entire hybridization signals to represent the 'total' community gene content or expression, with the caveat that the this is limited to genes represented on the array. This information could be leveraged to gain a better understanding of community function as well as generate hypotheses for further investigation.

In summary, we have demonstrated that the VChip produces similar overall patterns of species presence as 16S rRNA amplicon pyrosequencing and Illumina shotgun sequencing, and it may even be more sensitive to detection of genetic material from low-abundance bacteria that could be missed with shallow depth of sampling or sequencing. We have shown that the VChip is

suitable for exploratory functional analysis of vaginal communities, and we suggest it could be particularly useful as a screening tool to characterize and select samples of interest to study in greater detail with more comprehensive sequencing methods. Additionally, it can be used in the same way as traditional microarrays to evaluate differences in gene expression of vaginal microbial communities among samples. We conclude the VChip has potential to become a versatile research tool that could be adapted for a variety of applications.

CHAPTER 3

Dynamics of the vaginal microbiome during puberty[3]

3.1 SUMMARY

Puberty is an important developmental stage wherein hormonal shifts mediate the physical and physiological changes that lead to menarche, but up to now the bacterial composition of vaginal microbiota during this period have been poorly characterized. We performed a prospective longitudinal study of perimenarcheal girls to gain insight into the timing and sequence of changes that occur in the vaginal and vulvar microbiota during puberty. The study enrolled 31 healthy, premenarcheal girls between the ages of 10–12 years and collected vaginal and vulvar swabs quarterly for up to three years. Bacterial composition was characterized by Roche 454 pyrosequencing and classification of V1–V3 regions of 16S rRNA genes. Contrary to expectations, lactic acid bacteria, primarily *Lactobacillus* spp., were dominant in the microbiota of most girls well before the onset of menarche in the early to middle stages of puberty. *Gardnerella vaginalis* was detected in appreciable levels in approximately one-third of subjects, a notable finding considering this organism is commonly associated with bacterial vaginosis in adults. Vulvar microbiota closely resembled vaginal microbiota but often exhibited additional taxa typically associated with skin microbiota. Our findings suggest the vaginal microbiota of girls begin to resemble those of adults well before the onset of menarche.

3.2 INTRODUCTION

Understanding changes in vaginal bacterial communities over a woman's lifespan is essential to comprehending normal development, physiological function and health, and susceptibility to disease. Up to now, vaginal microbiota before puberty were thought to be relatively stable assemblages of aerobic, anaerobic and enteric bacterial populations (Hammerschlag *et al.*, 1978a; Gerstner *et al.*, 1982; Hill *et al.*, 1995; Myhre *et al.*, 2002). After menarche, the vaginal microbiota

[3] This chapter was previously published as: Hickey R.J., Zhou X., Settles M.L., Erb J., Malone K., Hansmann M.A., Shew M.L., Van Der Pol B., Fortenberry J.D., and Forney L.J. 2015. Vaginal microbiota of adolescent girls prior to the onset of menarche resemble those of reproductive-age women. mBio 6:e00097-15.

of healthy adults are typified by high numbers of homofermentative lactic acid bacteria that contribute to acidification of the vaginal microenvironment through the production of lactate and other organic acids (Boskey *et al.*, 1999; Linhares *et al.*, 2010b). Various species of *Lactobacillus* have been identified as the predominant lactic acid bacteria in most adult women, and the ecological function of lactate production is further conserved by genera such as *Streptococcus* and *Atopobium* that are found in a subset of women (Zhou *et al.*, 2007). Recent studies using cultivation-independent methods have revealed the substantial complexity and temporal variability of vaginal microbiota (Ravel *et al.*, 2011; Gajer *et al.*, 2012; Ravel *et al.*, 2013). Multiple community types distinguished by differences in the kinds and relative abundances of bacterial populations present are consistently found among healthy adult women (Zhou *et al.*, 2007; Verhelst *et al.*, 2005; Srinivasan *et al.*, 2010; Ravel *et al.*, 2011). These findings imply the absence of a 'core' healthy vaginal microbiota and underscore the importance of delineating differences in community composition among individuals as well as changes over time.

To date, most studies of vaginal microbiota have focused exclusively on reproductive-age women. As a result, little is known about when and how these communities are established during puberty. Changes in the composition and function of vaginal microbiota during puberty are thought to be mediated by estrogen-stimulated glycogen production in the vaginal epithelium (Linhares *et al.*, 2010b), but the timing and sequence of events during this period are not well studied. Menarche itself does not signal the completion of puberty, and pubertal hormonal influences on vaginal microbial communities may continue for months or even years following menarche. Using cultivation-dependent methods or microscopic examination, past studies of premenarcheal vaginal microbiota found bacterial communities with low numbers of strict and facultative anaerobes, with most species of apparently enteric origin (Gerstner *et al.*, 1982; Hill *et al.*, 1995; Myhre *et al.*, 2002). *Lactobacillus* species were rarely observed and, when found, constituted only a minor proportion of the total bacteria. Transition to adult-like vaginal microbial communities is not well documented but apparently occurs over a short time, as the vaginal microbiota of perimenarcheal and postmenarcheal 13–18-year-olds were found to resemble those of older women (Alvarez-Olmos *et al.*, 2004; Yamamoto *et al.*, 2009; Thoma *et al.*, 2011). However, most past studies are limited by inherent biases imposed by cultivation-dependent methods that fail to account for many bacterial taxa. Furthermore, we are unaware of studies that specifically characterized community composition in detail while evaluating subsequent physical

and physiological changes through menarche and thereafter. This lack of data highlights the need for longitudinal characterization of the vaginal microbial communities in perimenarcheal girls (i.e., before, during and following menarche).

There are several reasons to pursue a better understanding of the perimenarcheal vaginal microbiota. Clinically, vulvar and vaginal complaints such as vulvovaginitis are common among premenarcheal girls and are often ascribed to poor hygiene or physiologic leukorrhea (vaginal discharge due to estrogen stimulation) (Joishy *et al.*, 2005; Randjelović *et al.*, 2005; Yilmaz *et al.*, 2012). Numerous studies have reported bacterial vaginosis in adolescent girls, using diagnostic criteria developed for adult women (Bump *et al.*, 1986; Alvarez-Olmos *et al.*, 2004; Schwebke *et al.*, 2004; Brabin *et al.*, 2005; Brotman *et al.*, 2007; Schellenberg *et al.*, 2008; Vaca *et al.*, 2009). Without a frame of reference for 'normal' vaginal microbiota in healthy adolescents, the clinical relevance of microbiota resembling that associated with bacterial vaginosis is uncertain. Furthermore, as girls progress into menarche, menstrual hygiene behaviors including use of menstrual pads and tampons, bathing habits, and douching may alter existing vaginal microbiota (Merchant *et al.*, 1999; Blythe *et al.*, 2003; Schwebke *et al.*, 2004; Simpson *et al.*, 2004; Chase *et al.*, 2007; Brotman *et al.*, 2008; Ott *et al.*, 2009). Finally, changes in the early vaginal microbiota may have lasting influences on subsequent vaginal health, but our understanding of the complex interactions of immune tolerance of indigenous bacterial populations, immune surveillance for vaginal pathogens, variability in vaginal microbiota, and reproductive health outcomes remains primitive (Karlsson *et al.*, 2011; Mirmonsef *et al.*, 2011).

To better understand changes in both the vaginal and vulvar microbiota before, during, and after menarche, 31 healthy premenarcheal girls were enrolled between 10–12 years of age in a prospective longitudinal study in which girls were sampled quarterly for up to three years. The bacterial community composition of the vaginal and vulvar microbiota of girls, and vaginal microbiota of a subsample of their mothers, are described in this prospectively followed cohort. Our findings suggest the vaginal microbiota of adolescents resemble those of reproductive-age women, although not necessarily those of their mothers, well before the onset of menarche in early stages of pubertal development. Familiar bacterial species associated with the vaginal microbiota of adults were commonly found in girls, including *Lactobacillus crispatus, L. iners, L. gasseri, L. jensenii* and, notably, *Gardnerella vaginalis*. Following menarche, vaginal pH often remained above what is considered typical in healthy adult women even when lactobacilli were

present in high proportions, raising the possibility that total bacterial loads may not reach levels seen in adults until later in puberty. These analyses provide the first detailed investigation of the progression of changes that occur in vaginal microbiota over time during puberty.

3.3 RESULTS

3.3.1 *Clinical study and data collection*

A prospective longitudinal study was conducted in Indianapolis, Indiana, from June 2009 through June 2012 to assess changes in the vaginal and vulvar microbiota of perimenarcheal girls as they transitioned through menarche. Thirty-one healthy, asymptomatic girls representing black (n=21), white/non-Hispanic (n=7), white/Hispanic (n=1) and Native American (n=2) racial/ethnic groups were enrolled between June 2009 and August 2011. None of the girls had a history of sexual contact or recent antibiotic use. At enrollment, girls were between the ages of 10.0–12.9 years (mean 10.9 years) and premenarcheal. The age range was selected because it represents a developmentally meaningful interval during which pubertal development typically begins (Tanner, 1962; Marshall and Tanner, 1969), and we aimed to increase the chances that girls would reach menarche and experience subsequent menstrual cycling while participating in the study. Twenty-one girls (67.7%) reached menarche during the study. In addition, the mothers of 24 girls participated by providing vaginal swabs annually. Biological maternity was not recorded or required for participation. Although age, race and ethnicity of the mothers were captured, no analyses were performed on these data. Characteristics of the study participants are summarized in Table 3.1, with additional details included in Table B.1. Both parental and adolescent consent were obtained at the time of enrollment.

Girls returned for sample collection and clinical examination at three-month intervals for up to the duration of the three-year study (mean participation 1.6 years, range 1 day to 3.0 years). At each visit, up to three vaginal swabs and two vulvar swabs were collected by a female clinician as permitted by each girl. Breast and pubic development were assessed at each visit using Tanner's criteria (Tanner, 1962), and vaginal pH was determined using commercial pH paper. Mothers provided self-collected vaginal swabs on an annual basis. A total of 457 swabs were processed for Roche 454 pyrosequencing of the V1–V3 hypervariable regions of 16S rRNA genes: 198 vaginal swabs and 212 vulvar swabs from girls, and 47 vaginal swabs from mothers. A summary of

Table 3.1: Characteristics of adolescent study participants

Subject race and ethnicity	No. (row-wise %)			Mean age at enrollment (yr)	Mean duration of study participation (yr)
	Total	Who achieved menarche during study	Who had a participating mother		
All Subjects	31 (100)	21 (67.7)	24 (77.4)	10.9	1.6
Black, Non-Hispanic	21 (67.7)	14 (66.7)	17 (81.0)	10.9	1.7
White, Non-Hispanic	7 (22.6)	5 (71.4)	4 (57.1)	10.8	1.4
White, Hispanic	1 (3.2)	0 (0.0)	1 (100)	10.2	1.4
Native American, Non-Hispanic	2 (6.5)	2 (100)	2 (100)	11.5	1.3

samples collected at each visit (with Tanner stage and menarche status indicated) is shown in Figure B.1. In total there were 186 pairs of matched vagina-vulva samples from girls. Associated vaginal pH measurements were available for 65.7% of the sampling times from girls, while Nugent scores were obtained at only 24.8% of visits. We therefore elected not to analyze the Nugent score data.

3.3.2 *16S rRNA sequencing and taxonomic assignment of reads*

Bacterial composition was determined based on sequencing the V1–V3 hypervariable regions of 16S rRNA genes by Roche 454 pyrosequencing. High-quality reads were obtained from all but one vulvar swab sample, which was subsequently excluded from further analysis. The remaining 456 samples had a minimum of 462 and maximum of 31,569 reads after preprocessing (mean 4,723 reads, median 3,876 reads). Following taxonomic assignment of reads and summarization of the taxon abundance table as described in Materials and Methods, 78 taxa were identified to the species (n=9), genus (n=60), family (n=7) or order (n=2) level.

3.3.3 *Clustering analyses of the perimenarcheal vaginal microbiota*

The primary objectives of this study were to characterize the composition of vaginal microbiota in perimenarcheal girls, identify major community types, and assess similarities and differences in relation to menarche and pubertal development. Analyses were performed in R (`http://r-project.org`) using custom scripts available on GitHub.

To define the vaginal community types present in both girls and mothers in the study, hierarchical clustering was performed on the Bray-Curtis dissimilarity matrix calculated from

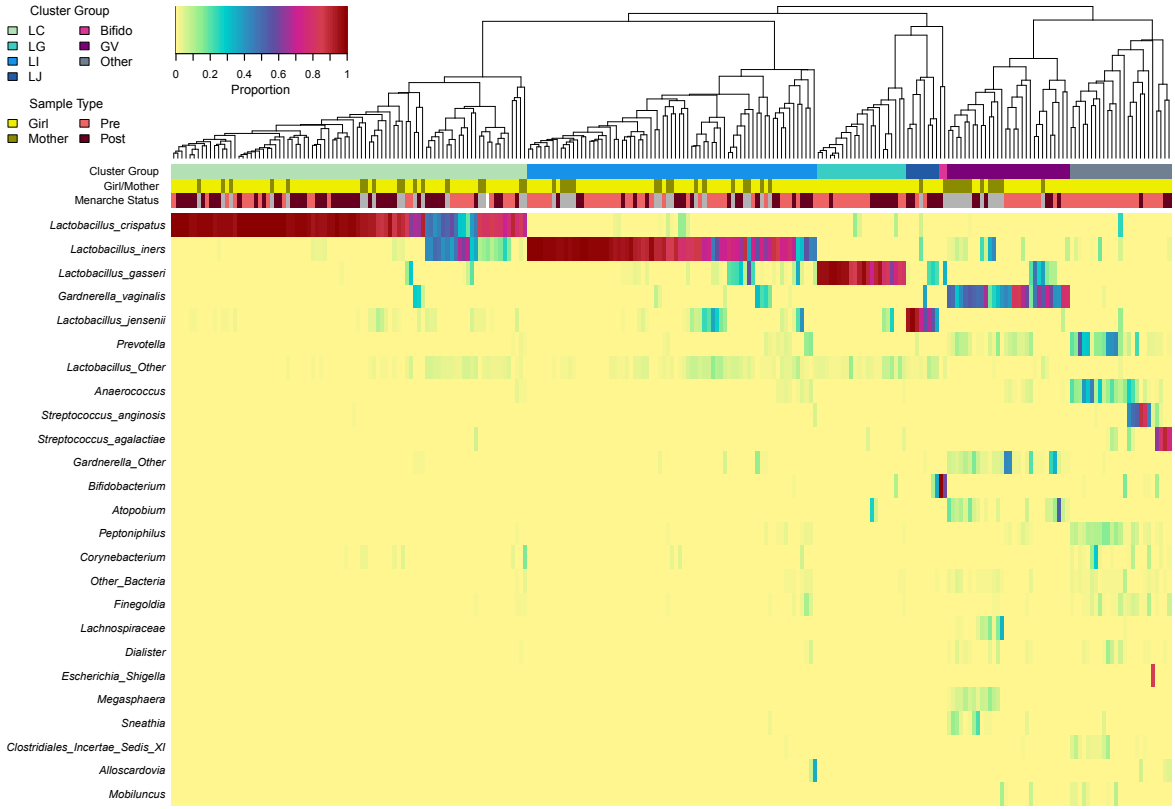

FIGURE 3.1: Composition of vaginal microbiota of girls and mothers sampled longitudinally. Each column in the dendrogram and heatmap represents the vaginal microbiota sampled from a single individual at a single time point. In total 198 samples from 31 girls and 47 samples from 24 mothers are represented. The dendrogram represents the average linkage hierarchical clustering of samples based on the Bray-Curtis dissimilarity matrix computed from Hellinger-standardized taxon abundance data. The colored bars below the dendrogram represent cluster group (top row) and sample type (second and third rows). Clusters are named to signify the most abundant taxon, when applicable: LC (*Lactobacillus crispatus* dominant, *n*=87), LI (*L. iners*, *n*=71), LG (*L. gasseri*, *n*=22), LJ (*L. jensenii*, *n*=8), 'Bifido' (*Bifidobacterium*, *n*=2), GV (*Gardnerella vaginalis*, *n*=30) and 'Other' (*n*=25). The heatmap represents proportions (before Hellinger standardization) of the 25 overall most abundant taxa within each community as indicated by the legend at top right. Sample type categories include girl/mother and premenarche/postmenarche (no menarche status is indicated for mother samples, colored gray).

Hellinger-standardized taxon abundance data. The average linkage method was identified as the best clustering method for the data using the minimum Gower distance (Gower, 1983), and seven clusters were selected as optimal using the silhouette method (Kaufman and Rousseeuw, 2009). Figure 3.1 shows the hierarchical clustering and community composition of all 198 vaginal swabs from girls and 47 vaginal swabs from mothers. Four clusters were characterized by high proportions of different *Lactobacillus* spp., including *L. crispatus* (cluster LC, *n*=87), *L. iners* (LI, *n*=71), *L. gasseri* (LG, *n*=22) and *L. jensenii* (LJ, *n*=8). A fifth cluster was characterized by high proportions of *Gardnerella vaginalis* (cluster GV, *n*=30). The sixth cluster ('Other', *n*=25) contained a mixture of various taxa including *Streptococcus agalactiae, Str. anginosis, Prevotella* and *Anaerococcus*, not all of which were detected in all samples within the cluster. Lastly, two samples characterized by high proportions of *Bifidobacterium* were grouped into a seventh cluster termed 'Bifido'; this included a postmenarcheal sample from subject 124 and a sample from her mother. Vaginal pH varied significantly among clusters, with samples in groups GV and Other each having significantly higher pH than the LC and LI clusters (Figure B.2).

Some clusters or community types appeared to be more commonly associated with either premenarcheal or postmenarcheal status in girls. Figure 3.2 compares the number and proportion of premenarcheal, postmenarcheal and mother vaginal samples assigned to each cluster. The six clusters besides 'Bifido' encompassed both premenarcheal and postmenarcheal vaginal samples from girls, but two clusters in particular were overrepresented in one group relative to the other. Among the premenarcheal samples (*n*=110), 20.0% (*n*=22) were assigned to the 'Other' cluster compared to only 3.4% (*n*=3) of the postmenarcheal samples (*n*=87), nearly a six-fold difference. Those three postmenarcheal samples were all from subject 104, two of which had high proportions of *Streptococcus agalactiae* and one with a high proportion of *Peptoniphilus*. On the other hand, 50.6% (*n*=44) of the postmenarcheal samples were assigned to cluster LC, compared to 22.7% (*n*=25) of the premenarcheal samples, more than a twofold difference. Cluster GV was more common among mothers (25.5% of samples from mothers) compared to girls (9.1% collectively of pre- and postmenarcheal samples), but even so, finding *Gardnerella vaginalis* in the vaginal microbiota of girls prior to onset of partnered sexual activity is notable.

The hierarchical clustering results indicate the vaginal microbiota of many premenarcheal girls were similar to those previously found in older adolescents and adults. To assess how long prior to menarche this was the case, we evaluated the cluster assignments over time within each

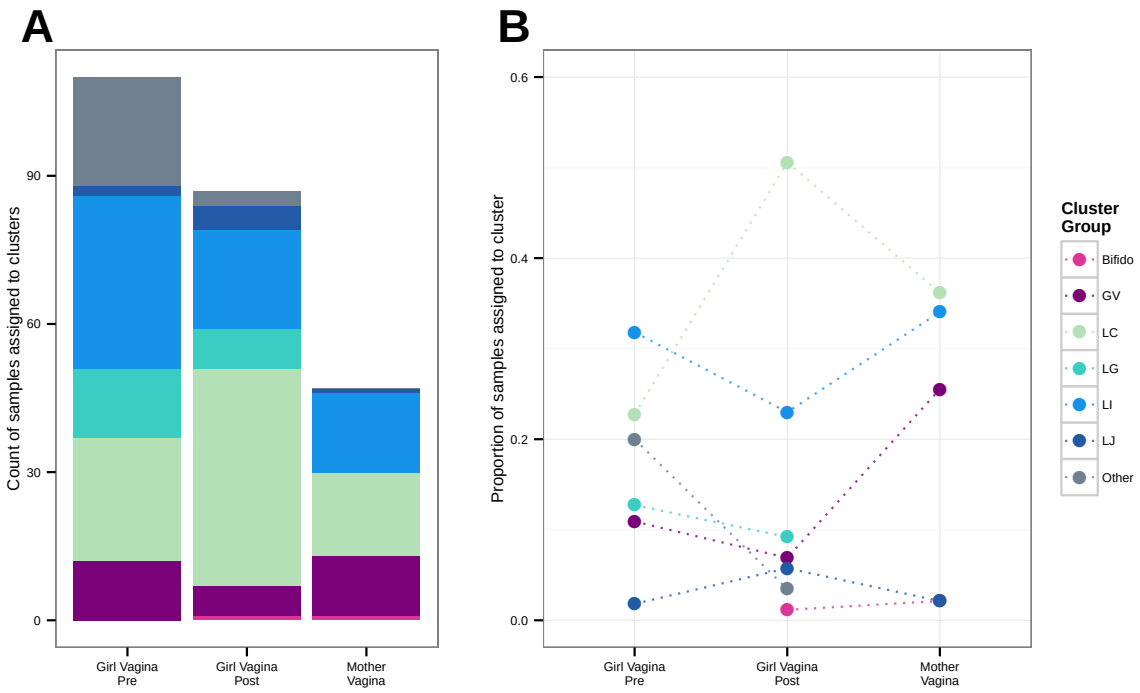

Figure 3.2: Hierarchical cluster assignment by sample type. 198 vaginal microbiota from 31 girls and 47 vaginal microbiota from 24 mothers were separated into seven groups by hierarchical clustering. (a) Count of girl premenarcheal (n=110), girl postmenarcheal (n=87), and mother vaginal microbiota (n=47) assigned to each cluster group (see Figure 3.1). (b) Proportion of samples in each group assigned to each cluster. The dotted lines serve to highlight differences between sample types and do not represent changes in cluster group prevalence over time.

individual, shown in Figure 3.3. Patterns of changes appeared to be highly individualized, but some common trends emerged. The vaginal microbiota of several girls were grouped into one of the *Lactobacillus*-dominant clusters (LC, LI, LG or LJ) more than a year before menarche (e.g., subjects 101, 108, 127, 128), many from the baseline visit. Some were initially in the 'Other' cluster before transitioning to a *Lactobacillus* cluster (e.g., subjects 103, 111, 123, 132). Except for subject 104, none transitioned back to the 'Other' cluster following menarche. While some girls' microbiota remained in the same cluster over long periods of time (e.g., subjects 101, 107), others transitioned between multiple groups (e.g., 103, 104, 112, 126). We note that other transitions may have occurred during the intervals between quarterly visits, since evidence from studies of reproductive-age women indicates that short-term changes are commonplace (Lopes dos Santos Santiago *et al.*, 2011, 2012; Gajer *et al.*, 2012; Lambert *et al.*, 2013; Ravel *et al.*, 2013). Nonetheless, communities typified by high proportions of *Lactobacillus* or other lactic acid bacteria (e.g., *Streptococcus*) were present well before the onset of menarche in most girls and were maintained after menarche.

Whereas hierarchical clustering was used to separate vaginal samples into major groups of community types, principal coordinates analysis (PCoA) was performed to obtain a more nuanced view of similarities and differences among samples in relation to other variables of interest. Figure B.3A shows that premenarcheal, postmenarcheal and mother vaginal samples are similarly distributed, suggesting differences in community composition are not strongly accounted for solely by age. Figure B.3B, the same PCoA plot color-coded by hierarchical cluster assignment, emphasizes the variability within clusters and reveals that some groups are more distinct while others are spread across a broader range of space. This serves as an important reminder that bacterial composition and rank-abundances within communities often lie along a continuum and need not be regarded as discrete types; rather, many communities may best be considered as 'intermediate' between groups characterized by different distributions of taxa.

3.3.4 *Longitudinal dynamics of the perimenarcheal vaginal microbiota*

To gain a detailed view of changes in community composition that occurred over time within each individual, we prepared summary plots of vaginal microbiota composition and associated metadata (Supplementary File 2, `http://github.com/roxanahickey/dissertation`). As anticipated based on the hierarchical clustering results, the vaginal microbiota of nearly all par-

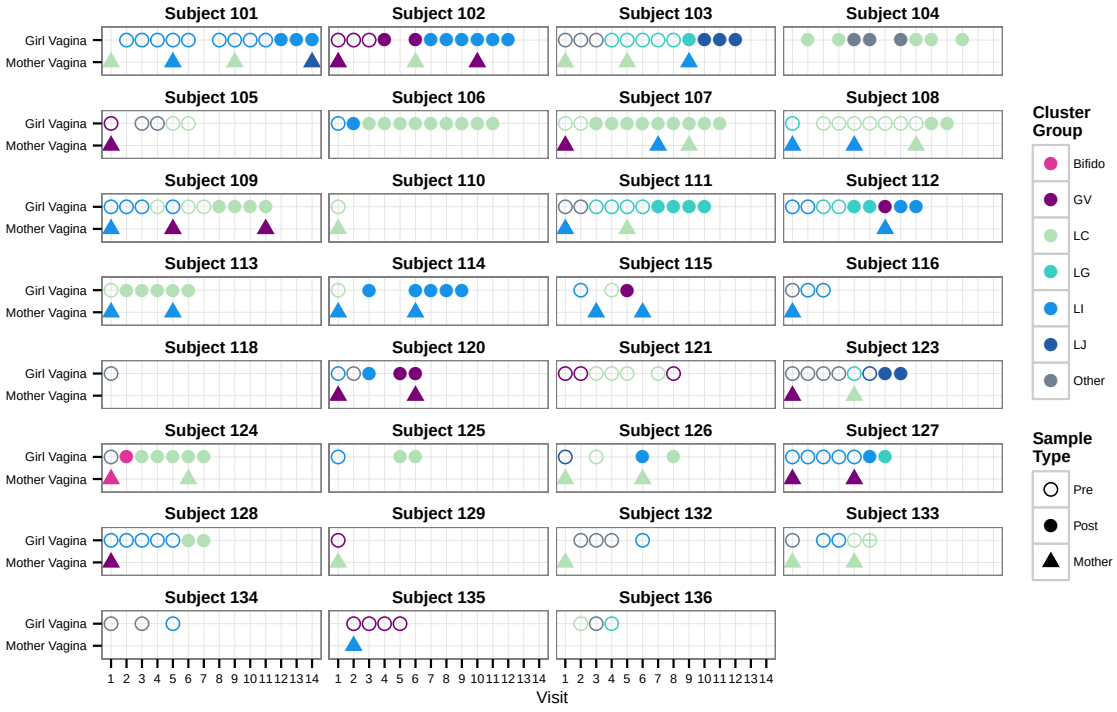

FIGURE 3.3: Hierarchical cluster assignment over time within individual participants. Each panel shows the hierarchical cluster assignment (see Figure 3.1) of vaginal microbiota samples from an individual girl (circles) and her mother (triangles), when applicable. The x-axis indicates the clinical visit at which each sample was collected (visits occurred approximately every three months). Open circles signify premenarcheal status, and filled circles signify postmenarcheal status in girls. The menarcheal status was not recorded for subject 133 at visit 6, indicated by an open circle with crosshatch.

ticipants who reached menarche (n=21) became characterized by a dominance of lactobacilli before or shortly after menarche. Interestingly, the microbiota of subjects 115 and 120 shifted from *Lactobacillus*-dominant in premenarche to *Gardnerella*-dominant in postmenarche. Even among premenarcheal vaginal microbiota, lactobacilli (primarily *Lactobacillus* spp., but in some cases *Streptococcus*) constituted at least 10% of the community in 27 girls (87.1%), and at least 50% in 25 girls (80.5%).

Figure 3.4 shows the progression of changes in the vaginal microbiota of four girls who had *Lactobacillus*-dominant communities before or at menarche, although the species composition and temporal dynamics differed considerably. For instance, while *L. crispatus* dominated the microbiota of subject 107 for more than two years, the microbiota of subject 109 gradually shifted from *L. iners*-dominant to *L. crispatus*-dominant. Subject 102 had *Gardnerella vaginalis* and *L. gasseri* in premenarche that later transitioned to *L. iners* following menarche. Subject 103 began the study two years before menarche with a diverse assortment of anaerobes that would usually be considered typical of prepubertal vaginal microbiota, but she eventually developed a microbiota dominated by *L. jensenii*, *L. gasseri* and *Bifidobacterium*. These examples demonstrate several trajectories to lactobacilli-dominant vaginal microbiota and emphasize the establishment of lactobacilli dominance does not necessarily result in static community composition. Moreover, multiple species of *Lactobacillus* may be numerically dominant at different times in the same individual, consistent with observations in adult women (Gajer *et al.*, 2012). Community profiles for all participants, complete with associated vulvar and mothers' vaginal microbiota where applicable, are shown in Supplementary File 2.

3.3.5 *Gardnerella vaginalis in the perimenarcheal vaginal microbiota*

Gardnerella vaginalis was commonly detected among adolescent participants, constituting 10% or more of the vaginal microbiota in at least one sampling of 11 girls (35.5%), seven of whom (22.6%) had communities dominated by *Gardnerella* at some point in time. The proportion of *Gardnerella* in these participants over time is shown in Figure B.4. Some girls had high proportions of *Gardnerella* at multiple consecutive visits (e.g., subjects 102, 121 and 135) while it was detected in others at only one time point (e.g., subjects 112, 116 and 126). Participants with *Gardnerella* represented both black (7/21, 33.3%) and white (4/8, 50.0%) racial groups. Although *G. vaginalis* is commonly associated with bacterial vaginosis (BV) and some have suggested it is a sexually

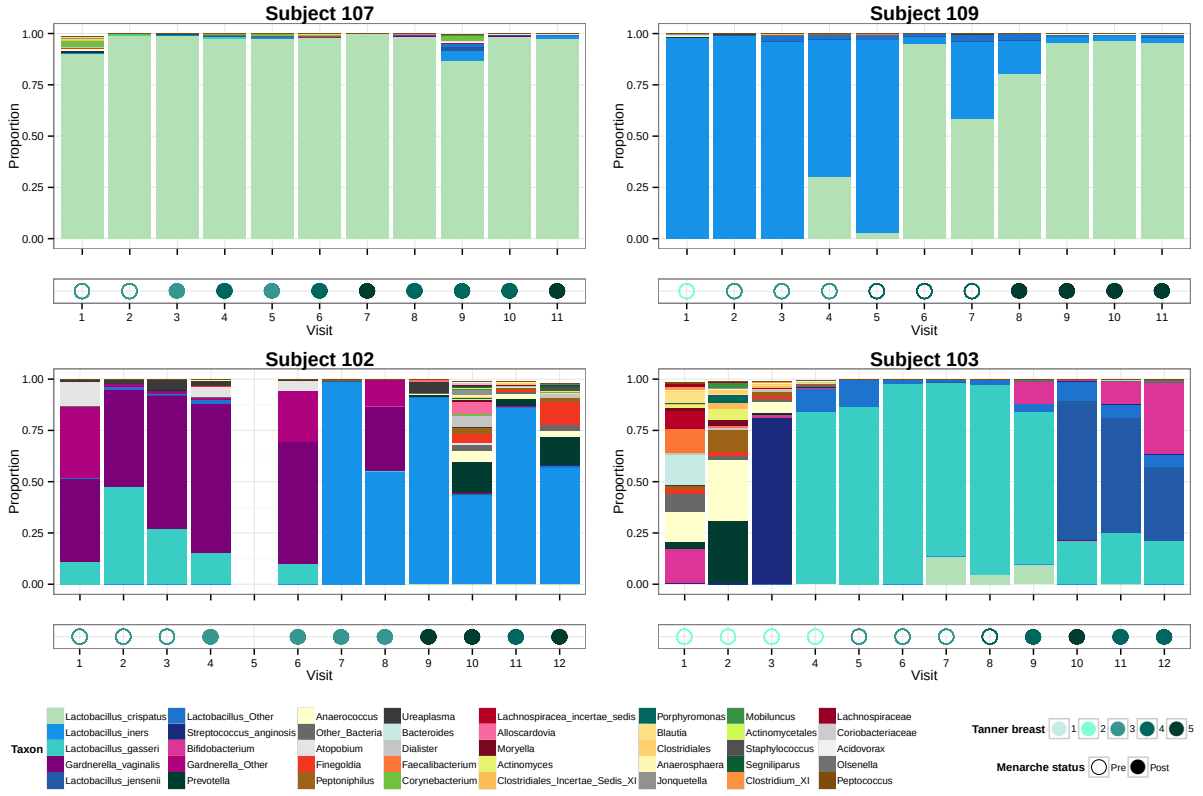

FIGURE 3.4: Transitions to *Lactobacillus*-dominant vaginal microbiota. Panels show the vaginal bacterial community profiles and associated pubertal development of four participants sampled longitudinally. Bar plots represent the proportions of bacterial taxa in the community (legend at bottom left). Below each bar plot the menarcheal status and clinician-assessed Tanner stage of breast development are indicated by point shape and color, respectively (legend at bottom right). Empty spaces in the plots indicate a skipped visit.

transmitted infectious agent (Brook, 2002; Schwebke *et al.*, 2014b), these findings present clear evidence that it can reside in the vaginal microbiota of healthy perimenarcheal girls without a history of partnered sexual behavior or symptoms characteristic of BV (both of which were exclusion criteria).

3.3.6 *Lactic acid bacteria and vaginal pH*

High numbers of lactobacilli, particularly species of *Lactobacillus*, are widely regarded as a hallmark of vaginal health in adult women. Their presence is typically associated with a low vaginal pH of around 4–4.5, although healthy women in black and Hispanic groups have been shown to have slightly higher pH on average (Ravel *et al.*, 2011). Since we observed high proportions of lactobacilli in the adolescent participants in this study, we next examined the relationship between the proportions of lactobacilli and vaginal pH with respect to pubertal development and menarche. In this case, we considered lactic acid bacteria (LAB) as the genera in our dataset contained within the order Lactobacillales, which included *Lactobacillus*, *Streptococcus, Aerococcus* and *Facklamia*. We note that other genera such as *Atopobium* and *Bifidobacterium* are also known producers of lactic acid, so our representation of LAB is therefore somewhat conservative.

Overall, the relative abundance of LAB tended to increase with pubertal development, while vaginal pH tended to decrease (Figure 3.5). We compared several linear mixed effects models, controlling for inter-subject variation, to determine how Tanner stage, age and menarcheal status affected LAB proportions and pH. Because Tanner breast and pubic stage were collinear and accounted for similar proportions of variance (analysis at `https://github.com/roxanahickey/adolescent/blob/master/03-community-dynamics.md`), we tested models that included either breast or pubic scores, but not both. We focus on models using Tanner breast scores here (Table 3.2), and models with Tanner pubic scores are reported in Table B.2. Following a stepwise model selection strategy, the optimal model accounting for changes in LAB proportions included only age and Tanner breast stage as fixed effects, since inclusion of menarche status added little explanatory value. Under this model, the transition from Tanner breast stage 2 to 3 was associated with a highly significant increase in logit-transformed LAB proportions (p=1.1e-5), while subsequent transitions to stage 4 and 5 had relatively small effect (p=0.26 and p=0.56, respectively). A similar pattern was detected using Tanner pubic scores instead (Table B.2). The best model accounting for changes in vaginal pH included Tanner breast stage, menarche status,

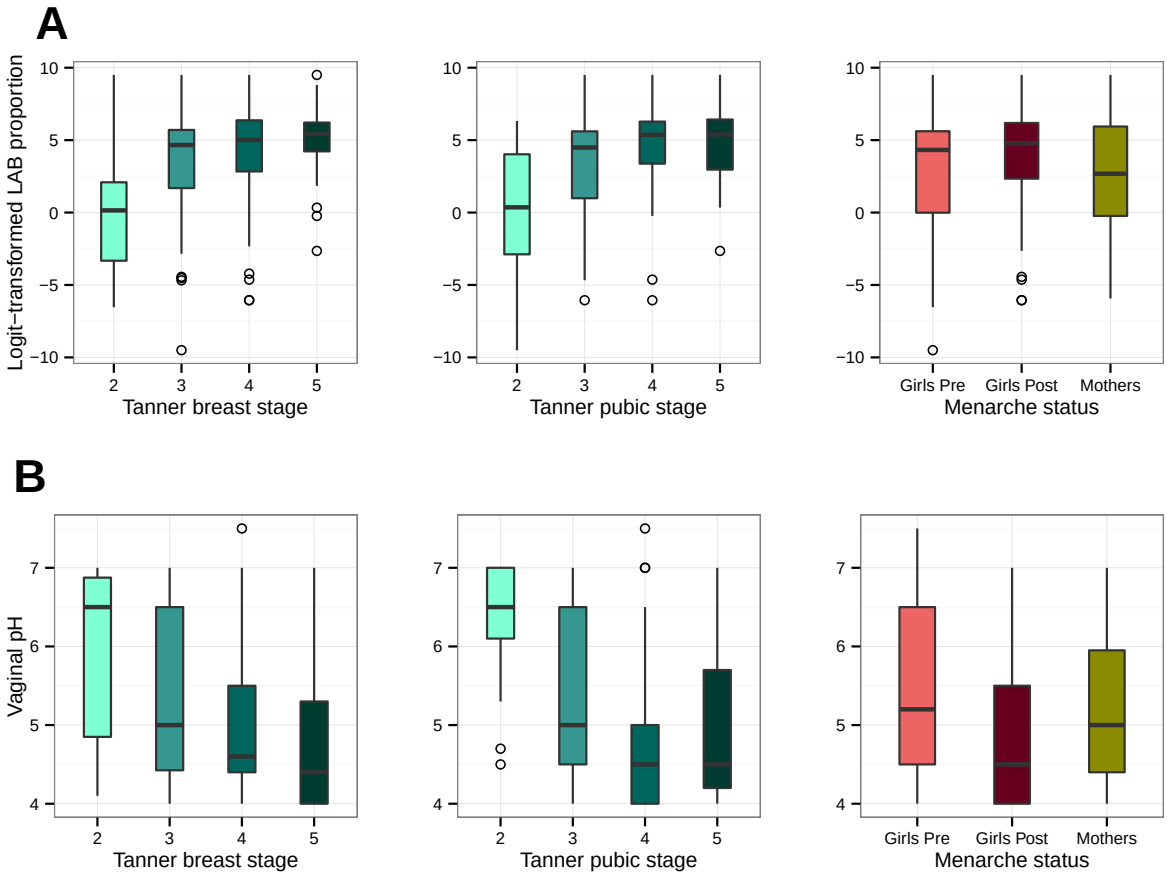

Figure 3.5: Trends in relative abundance of lactic acid bacteria and vaginal pH with pubertal development and menarche status. Upper and lower panels show box plots of (A) the logit-transformed proportion of lactic acid bacteria (LAB; includes *Lactobacillus, Streptococcus, Aerococcus* and *Facklamia*) and (B) vaginal pH of 31 perimenarcheal girls. In the left column, box plots show the relationship to Tanner breast stage; in the middle column, Tanner pubic stage; and in the right column, menarche status. The far right column also includes data for 24 mothers who participated in the study. In each plot the box represents the interquartile range, the whiskers represent the upper and lower quartiles, the horizontal line represents the median, and open circles represent outliers.

and age as fixed effects. Although the overall model accounted for a significant proportion of variance, only age had a significant marginal effect on pH (p=0.0071), and the transitions in Tanner stage and menarche status were not themselves significantly associated with changes in pH. However, a similar model using Tanner pubic scores indicated the transition from stage 2 to 3 was significantly associated with declining pH (p=0.0006), while subsequent transitions were not significant. Taken together, these results suggest the transition from Tanner breast or pubic stage 2 to 3 (i.e., early to mid-puberty) is associated with the most profound changes in LAB proportions and vaginal pH.

Although increased LAB proportions were often coupled with decreased vaginal pH over time, closer inspection revealed that many vaginal samples with high proportions of LAB still had a pH well above 4.5, including both premenarcheal and postmenarcheal samples at all Tanner stages. We hypothesized this may be due to lower bacterial loads in the perimenarcheal vaginal microbiota, leading to lower levels of lactic acid and elevated pH. To test this hypothesis, we performed qPCR to estimate the number of 16S rRNA gene copies in 24 randomly selected girl vaginal samples with high proportions LAB (> 0.75) and either 'low' (< 5.0, n=12) or 'high' vaginal pH (≥ 5.0, n=12). A detailed explanation of our approach and findings are summarized in Appendix B and Figure B.5. Although we did not find a statistically significant difference in the average estimated number of 16S rRNA gene copies in the 'low' versus 'high' pH groups (p=0.14), we suggest this phenomenon warrants consideration in future studies.

3.3.7 *The vulvar microbiota of perimenarcheal girls*

Although this study was primarily aimed at characterizing the vaginal microbiota of perimenarcheal girls, we also analyzed the composition of vulvar swabs and compared them to the vaginal microbiota. This included making qualitative comparisons, calculating correlations between paired vagina-vulva microbiota in taxonomic rank and relative abundances, performing hierarchical clustering as we did for the vaginal samples, and performing 'indicator species' analysis (De Cáceres and Legendre, 2009; De Cáceres *et al.*, 2010) to determine whether certain taxa were more highly associated with the vaginal or vulvar environment.

Composition of the vulvar microbiota typically mirrored that of contemporaneously sampled vaginal microbiota, although the vulva tended to have a greater variety of bacterial taxa present. Hierarchical clustering analysis of vaginal and vulvar samples together resulted in the

Table 3.2: Linear mixed effects modeling of lactic acid bacteria and vaginal pH

Model and parameters[a]	Result for model				
LAB model: logit(LAB) ~ TB + age + 1|subject + epsilon					
Random effects	Variance	SD	No. of observations	No. of groups	
Subject (intercept)	4.4	2.1	189	28	
Residual	7.3	2.7			
Fixed effects/contrasts[b]	Coefficient[c]	SE	df	*p*-value[d]	
Intercept	-6.1	4.4	150.6	1.70E-01	
TB 3 vs. 2	3.3	0.7	183.9	1.10E-05	***
TB 4 vs. 3	-0.6	0.5	176.2	2.60E-01	
TB 5 vs. 4	0.4	0.7	176.7	5.60E-01	
Age	0.7	0.4	151.7	4.10E-02	*
Vaginal pH model: pH ~ TB + menarche status + age + 1|subject + epsilon					
Random effects	Variance	SD	No. of observations	No. of groups	
Subject (intercept)	0.6	0.8	122	20	
Residual	0.4	0.6			
Fixed effects/contrasts	Coefficient	SE	df	*p*-value	
Intercept	9.3	1.4	112.5	2.10E-09	***
TB 3 vs. 2	-0.3	0.2	108.6	2.10E-01	
TB 4 vs. 3	0	0.2	105.5	8.40E-01	
TB 5 vs. 4	0	0.2	104.2	8.60E-01	
Postmenarche vs. premenarche	-0.2	0.2	115.5	3.70E-01	
Age	-0.3	0.1	113.4	7.10E-03	**

[a] LAB, lactic acid bacterium proportion; TB, Tanner breast stage; ε, random error; SD, standard deviation; SE, standard error; df, degrees of freedom.

[b] Contrasts between successive Tanner stages were made, excluding stage 1 (not represented).

[c] Marginal slope of the fixed effect on the response.

[d] Significance is indicated as follows: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

same groups overall (Figure B.6), although the distribution of samples across groups was slightly different (e.g., the *L. jensenii* cluster was much smaller, with many samples assigned instead to the 'Other' or *L. iners* clusters). Spearman's rank correlation coefficients were computed for the 186 matched vagina-vulva pairs based on the taxon proportion data. The mean correlation was 0.63 (median also 0.63), indicating a moderately high degree of concordance in the rank-order of taxa in paired vaginal and vulvar samples. Indicator species analysis, summarized in Table B.3, revealed that three taxa were more strongly associated with the vulva (i.e., *Segniliparus, Murdochiella, Fusobacterium*), and each of these were typically minor in relative abundance (on average, 1.0–3.5% in the vulva compared to 0.02–0.04% in the vagina). No taxa were more strongly associated with the vagina compared to the vulva, indicating most vaginal species are likely also found in the vulvar microbiota. Seventeen taxa were associated jointly with the pre-menarcheal vagina, premenarcheal vulva and postmenarcheal vulva, suggesting that, at least in this dataset, the premenarcheal vaginal microbiota shared more taxa in common with the vulvar microbiota than did the postmenarcheal vaginal microbiota. These findings suggest the vulvar microbiota of girls are generally similar to vaginal microbiota, perhaps even more so prior to menarche.

We qualitatively assessed changes in genus-level richness and Simpson's diversity index of vaginal and vulvar microbiota in relation to Tanner stage and menarche status (see Appendix B). As suggested by the indicator species analysis, the vulva tended to have greater richness and diversity than the vagina. Decreases in diversity with Tanner stage progression mirrored the increases in LAB proportions (Figures B.7). However, lower richness and diversity in the vaginal microbiota were not necessarily associated with lower vaginal pH (Figure B.8).

3.3.8 *Similarities between vaginal microbiota of girls and their mothers*

Lastly, we assessed similarities between the vaginal microbiota of girls and their mothers. This analysis was exploratory in nature, as we had no basis on which to expect more or less similarity within girl-mother pairs. We qualitatively assessed similarities by studying the hierarchical cluster assignments presented in Figure 3.3 and the paired community composition profiles depicted in Supplementary File 2. There was no general trend of resemblance between girls and mothers, but some interesting similarities were observed on a case-by-case basis. Some pairs often shared the same dominant taxon (e.g., *G. vaginalis* in subject pair 102/202, *L. iners* in 109/209 and 114/214, and *Bifidobacterium* or *L. crispatus* in 124/224), although even in those cases this similarity did not persist over repeated visits. In several cases, girls' microbiota bore virtually no resemblance to their mothers' (e.g., subject pairs 111/211, 113/213 and 132/232). The data are insufficient to determine whether any of these patterns are biologically meaningful, but the observation of similar vaginal microbiota in at least some girl-mother pairs is intriguing and warrants further studies aimed at understanding how the vaginal microbiota may be influenced by both genetic and environmental factors.

3.4 DISCUSSION

Vaginal microbiota play an important protective role in women's health, but our understanding of the microbiota in pre- and perimenarcheal girls has thus far been limited by a lack of detailed studies using modern techniques to assess community composition. In this prospective longitudinal study, we characterized changes in the vaginal and vulvar microbiota of 31 girls as they progressed through puberty to menarche. To our knowledge, it is the first study of perimenarcheal girls to employ high-throughput 16S rRNA gene sequencing, enabling finer resolution of the species

and genus membership of communities, along with measurements of pubertal development (Tanner stages) and vaginal pH. Our findings add significant new insight to the composition and dynamics of the vaginal microbiota of adolescents during puberty and serve as a foundation for future studies aimed at characterizing the vaginal microbiota over shorter time intervals and for longer spans of time, opening the door to the possibility of identifying drivers of community composition.

The vaginal microbiota of pre- and perimenarcheal girls in our study were similar to those previously described in older postmenarcheal adolescents (Alvarez-Olmos *et al.*, 2004; Yamamoto *et al.*, 2009) and adults (Hill *et al.*, 2005; Hyman *et al.*, 2005; Zhou *et al.*, 2007, 2010; Ravel *et al.*, 2011) in so far as lactic acid bacteria (LAB) were prominent members of most communities. For many girls this was true well before menarche, often a year or more in advance. This was unanticipated given the results of previous studies that employed cultivation-dependent methods and found lactobacilli to be infrequent or minor constituents of the microbiota in premenarcheal girls (Hammerschlag *et al.*, 1978b; Gerstner *et al.*, 1982; Myhre *et al.*, 2002; Yilmaz *et al.*, 2012). However, these studies did not focus on the appearance of adult-like microbiota during pubertal transitions. In our study we tracked these transitions using Tanner's criteria of breast and pubic development (Tanner, 1962), which are commonly used by clinicians to assess physical maturity. The five Tanner stages represent major changes in secondary sex characteristics that females undergo from pre-puberty (stage 1) to full maturity (stage 5) in response to increased estrogen levels and other physiological changes. Menarche is considered a late event in puberty, usually occurring in stage 3 or 4 (Marshall and Tanner, 1969). We observed the most substantial increases in proportions of LAB, along with declines in vaginal pH, in the transition from Tanner breast stage 2 to 3, followed by less pronounced changes thereafter. Most of the microbiota with low proportions or no LAB detected were sampled from girls earlier in puberty (breast stage 2 or 3 and premenarcheal). These communities were typically composed of an assortment of strict and facultative anaerobes, an observation more in line with what has been described previously in premenarcheal girls.

Interestingly, high proportions of LAB in girls were not always associated with low vaginal pH (i.e., pH $\leq$ 4.5, but see Ravel *et al.*, 2011) as commonly observed in reproductive-age women with similar microbiota composition. We postulated this could be due to lower bacterial loads (and therefore less lactic acid production) but were unable to detect a statistically significant difference

in 16S rRNA gene copy number as a proxy for bacterial cell number between a subset of samples with high proportions of lactobacilli and either low or high pH. A study by Brabin *et al.* (2005) found that higher vaginal pH occurred more often in adolescents with abnormal compared to normal menstrual cycles; this is another possible explanation for the patterns observed in our data. Our findings resonate with those reported by Thoma *et al.* (2011) that assessed longitudinal changes in the vaginal microbiota of 13–18-year-olds using Nugent Gram stain criteria. That study enrolled 49 virginal girls from a rural district of Uganda and sampled them weekly for two years. Among premenarcheal girls the authors observed a significant increase in large Gram-positive rods (presumably *Lactobacillus* spp.) and a concurrent decrease in small Gram-negative to variable rods (presumably *G. vaginalis* or *Bacteroides* spp.) over time. In contrast, changes in the abundance of those morphotypes among the perimenarcheal and postmenarcheal groups were attenuated over time. Furthermore, while declines in vaginal pH over time were observed for all three groups, the trend was statistically significant only in the perimenarcheal and postmenarcheal groups. The authors conclude vaginal microbiota transitioning occurs prior to menarche, but significant decreases in vaginal pH continue well after menarche. Similarly, we observed the steepest changes in relative abundance of lactobacilli in early to mid-puberty, with less pronounced increases in later stages of puberty. The observation of Thoma *et al.* of accelerated declines in vaginal pH after menarche is also consistent with our finding that pH often remained higher than expected even though lactobacilli were numerically dominant.

Although the emergence of lactobacilli as dominant members of the microbiota in early to mid-puberty was a consistent trend among our participants, community composition and dynamics varied considerably between and within individuals. Various lactobacilli were dominant in different community types, including *Lactobacillus crispatus, L. iners, L. gasseri, L. jensenii*, and in some cases, *Streptococcus* spp. The significance of communities dominated by different lactobacilli is not well understood, but recent studies have identified variation in the genetic and metabolic potential of vaginal lactobacilli that may influence the ecology of each species or strain in vivo (Witkin *et al.*, 2013; Mendes-Soares *et al.*, 2014). Furthermore, some girls had only one dominant species at consecutive three-month intervals, while others exhibited co-occurrence of two or more lactobacilli, or a succession of multiple species over time. As previously seen in longitudinal studies of reproductive-age women, transitioning through multiple community state types seems to be at least as common as having a community with stable or constant species

composition over time (Gajer *et al.*, 2012; Ravel *et al.*, 2013). Our findings suggest this pattern extends to earlier stages of life as well. This is an important reminder that the vaginal ecosystem is highly dynamic and individualized, and both temporal and inter-individual differences should be taken under consideration as we attempt to characterize and understand the 'normal' vaginal microbiota. Larger, longitudinal studies extending into late adolescence will be required to investigate the relevance of dynamic vaginal microbiota during puberty to long-term vaginal health.

The high prevalence of *Gardnerella vaginalis* among adolescent participants was a surprising finding in this study. It was the fourth most abundant species detected and constituted at least 10% of the vaginal microbiota at least once in approximately one-third of the participants. In some cases, *G. vaginalis* constituted the majority of the microbiota at multiple consecutive three-month intervals, suggesting long-term persistence. *G. vaginalis* is often associated with bacterial vaginosis (BV), a common yet poorly understood condition associated with increased risk of sexually transmitted infections and preterm birth (Koumans *et al.*, 2007; Workowski *et al.*, 2010). While *G. vaginalis* is undoubtedly correlated with BV, evidence for whether it is necessary or sufficient to elicit symptoms remains inconclusive (Srinivasan *et al.*, 2013; Yeoman *et al.*, 2013; Hickey and Forney, 2014; Schwebke *et al.*, 2014b). Postulated virulence mechanisms include biofilm formation (Swidsinski *et al.*, 2005), sialidase production (Santiago *et al.*, 2011), and synergism with other bacteria such as *Prevotella* (Pybus and Onderdonk, 1997). However, numerous studies have identified *G. vaginalis* in many reportedly healthy women, suggesting it may be a commensal in some individuals (Fredricks *et al.*, 2005; Hyman *et al.*, 2005; Zhou *et al.*, 2007; Ravel *et al.*, 2011). Considering evidence that identical *G. vaginalis* 16S rRNA sequences are detected among sexual partners (Eren *et al.*, 2011), some have suggested *G. vaginalis* is acquired exclusively through sexual contact (Schwebke *et al.*, 2014b). Our findings present evidence contrary to this hypothesis, as none of the girls in this study had a history of partnered sexual activity. Likewise, several studies report finding *G. vaginalis* in children and adolescents without sexual experience (Hammerschlag *et al.*, 1978b; Shafer *et al.*, 1985; Bump *et al.*, 1986; Myhre *et al.*, 2010; Fethers *et al.*, 2012; Yilmaz *et al.*, 2012). Whether microbiota containing *G. vaginalis* place the host at higher risk of infection remains to be seen, but recent investigations into the remarkable diversity of this species (Harwich *et al.*, 2010; Yeoman *et al.*, 2010; Ahmed *et al.*, 2012; Jayaprakash *et al.*, 2012) may

help to clarify whether commensal and pathogenic strains exist and can be distinguished from one another.

Study of the vaginal microbiota in adolescents may provide crucial insight into the complexity and variability of microbiota in adults, as well as whether the development of microbial communities during puberty is associated with vaginal health in adulthood. Our research demonstrates the feasibility of longitudinal study of girls to complement understanding of the vaginal microbiome of adult women. The 10–12-year-old participants reported feeling comfortable and safe with the study procedures, especially after the initial visit (Robbins *et al.*, 2012). The small, flexible swabs easily passed patent hymens and captured sufficient vaginal material for analysis. A trained, female provider and accompaniment by their mothers also allowed girls to become rapidly accustomed to sample acquisition and complete most scheduled study visits. Only six girls declined vaginal samples during the baseline visit, although our results, and those of others, suggest vulvar samples may serve as a rough proxy for vaginal samples (Elsner and Maibach, 1990; Farage and Maibach, 2006; Brown *et al.*, 2007) if vaginal samples are declined. Similar studies can and should be done to grow our understanding of both normal and abnormal vaginal microbiota, and ultimately improve strategies of gynecologic care for adolescent girls.

Many questions remain unanswered, including the relationship between estrogen, vaginal glycogen levels, lactobacilli levels and vaginal pH; factors that determine successional changes at earlier stages of puberty; and whether differences in community composition during adolescence can influence health outcomes later in life. Although concordance of girls and mothers was inconsistent, future studies of girl-mother or sister-sister pairs could elucidate the role of host genetics and interactions among individuals in shaping the microbiota. To gain a comprehensive understanding of the importance of vaginal microbiota composition and dynamics throughout puberty, it would be wise to include girls on the brink of puberty and even earlier in childhood, as well as older girls experiencing regular menstrual cycling. Ideally, similar studies should be performed on larger cohorts of girls with a balanced sampling of racial and ethnic groups to determine whether the patterns we observed are more broadly conserved.

3.5 MATERIALS AND METHODS

3.5.1 *Study design and enrollment criteria*

The study protocol was approved by the Institutional Review Board of Indiana University, and informed consent from girls and their mothers was documented before participation in the study. Participants and their mothers were recruited by referral from clinicians, by referral from participating mothers, or in response to advertisements placed in local newsletters and newspapers. Primary inclusion criteria included 10.0–12.9 years of age and premenarcheal status at enrollment, and initiation of pubertal development indicated by breast development of Tanner stage 2 or 3 (Tanner, 1962) as documented by self-reporting and corroboration by a clinician's examination. Eligible candidates were enrolled if they were in good health and willing to refrain from bubble baths and genital cleansing wipes for 48 hours before examination. Exclusion criteria included genitourinary symptoms (dysuria, vaginal discharge, genital ulcers); evidence of urinary tract infection; use of any systemic or topical antibiotic or antifungal treatment within the previous 60 days; prior genitourinary surgery, instrumentation or medical treatment for recurrent urinary tract infection, posterior urethral valves, or enuresis; any significant medical condition deemed cause for exclusion by the investigator (e.g., Type I diabetes, asthma, autoimmune diseases); prior evaluation for vulvovaginal symptoms; history of prepubertal bleeding; history of sexual abuse or sexual activity; or having already begun menarche.

Mothers of enrolled girls were invited to participate in annual sample collections. No age range was required, nor was biological maternity. Mothers were included if they reported themselves in good health and were willing to refrain from using bubble bath, douching substances, powders, perfumes, wet wipes or lotions to the genital area for 48 hours before sample collection, and were willing to refrain from sexual activity, bathing or swimming for 2 hours before sample collection. Lack of participation by the mother was not an exclusion factor for otherwise eligible girls. The study initially enrolled 32 girls and 25 mothers, but one participant (subject 117) was withdrawn at the baseline visit after the clinician determined she was at Tanner stage 1 for breast development, and no samples were collected. A sample collected from her mother (subject 217) was subsequently excluded from further analysis. Two participants (subjects 101 and 129) were also enrolled at breast stage 1 and permitted to continue at the investigator's discretion, but subject 129 was lost to observation after the baseline visit. One participant (subject 110) was enrolled

at breast stage 5 (her self-assessment was stage 3) but then lost to observation after two visits. Vaginal and vulvar samples from the latter three participants, along with any vaginal samples from their mothers, were retained for analysis.

3.5.2 *Collection of specimens and participant metadata*

At enrollment (baseline) and each quarterly visit, a female clinician completed a short physical examination to document breast, pubic hair, and genital development. Girls were then assisted to a sitting, 'frog-legged' position with their feet in stirrups. After additional inspection of the vulva, the clinician sequentially obtained two vulvar samples by rubbing both labia minora with a sterile, dry, flocked nasopharyngeal swab. The clinician (after confirmation of the adequacy of the hymenal opening) then sequentially inserted up to three swabs through the vaginal introitus to approximately 5 cm. Vaginal swabs were rotated twice. Lastly, vaginal pH was obtained by inserting a commercial pH paper into the vaginal opening and noting the pH in comparison to a provided color indicator. Clinicians omitted pH measure at their discretion to optimize sample integrity. Study-provided cell phones were used to identify the onset of menses, and visits were scheduled so as not to coincide with menses. Participants' mothers provided up to three self-obtained vaginal specimens for each annual collection. Vulvar swabs were not collected from the mothers. All swab samples were placed separately in labeled cryovials that were immediately placed on dry ice, then transferred to and stored in a -70°C freezer.

Of the swabs collected from girls and mothers, one was shipped to the University of Idaho for analysis of microbial community composition as described below, and a second was archived for use in subsequent studies. The third swab was used to assess vaginal health by Nugent criteria (Nugent *et al.*, 1991). Vaginal or vulvar sample collection was halted if the patient wished to discontinue sampling during the examination, or at the clinician's discretion to optimize sample integrity.

3.5.3 *16S rRNA sequencing and taxonomic assignment of reads*

Genomic DNA was extracted from vaginal and vulvar swabs with the use of a validated enzymatic lysis and bead-beating protocol (Yuan *et al.*, 2012), followed by purification with use of the QIAamp DNA Mini Kit (Qiagen, Venlo, Netherlands), which we have used in our previous

human microbiome studies (Ravel *et al.*, 2011; Gajer *et al.*, 2012; Ravel *et al.*, 2013; Hou *et al.*, 2013). Bacterial 16S rRNA genes were amplified by PCR using barcoded primers flanking hypervariable regions V1 and V3 (*Escherichia coli* positions 27F-534R), optimized by Frank *et al.* (2008) for improved detection of Bifidobacteriaceae (including *Gardnerella*), *Borrelia* and *Chlamydia*, as done previously (Ravel *et al.*, 2013). Amplicons were sequenced on a Roche 454 FLX pyrosequencer (Roche 454 Life Sciences, Branford, CT, USA) at the University of Idaho. Sequence reads were cleaned, filtered and taxonomically assigned using the Ribosomal Database Project (RDP) Naïve Bayesian Classifier to the first RDP level with a bootstrap score ≥50. Species of *Lactobacillus*, *Gardnerella*, and *Streptococcus* were further classified using a clustering approach with the R package WGCNA (Langfelder and Horvath, 2008) and 16S rRNA sequences from the PATRIC database (`http://patricbrc.org`). The above methods are described in detail in Appendix B. Following preprocessing of Roche 454 sequence data and taxonomic assignment of reads, data were processed using custom R scripts to calculate percentages of taxa within each sample. To simplify analysis of community composition, we retained named taxa that constituted either at least 1% of the community in two or more samples, or at least 5% of the community in at least one sample. Taxonomically assigned reads that did not meet this threshold were combined into an 'Other Bacteria' category, along with reads that could not be taxonomically assigned beyond the level of Bacteria.

3.5.4 *Availability of data and custom R scripts*

Sequences in standard flowgram format (SFF) for the 457 samples analyzed in this study are available to download from the NCBI Sequence Read Archive (BioProject PRJNA266340). Data and custom R scripts to reproduce the analyses, including walkthroughs of intermediate steps as well as some additional analyses, are available on GitHub at `https://github.com/roxanahickey/adolescent`. Analyses were conducted using R v3.1.0 (`http://r-project.org`).

3.5.5 *Hierarchical clustering and principal coordinates analysis of community composition*

Following approaches outlined by Legendre and Legendre (2012) and demonstrated by Borcard *et al.* (2011), we performed both hierarchical clustering and PCoA to obtain an overall picture of similarities and differences in bacterial community composition across vaginal samples from

girls and mothers. A similar analysis was performed for all vaginal and vulvar samples together. Prior to these analyses, the taxon abundance matrix was standardized using the Hellinger method with the R package vegan v2.0-10 (Oksanen *et al.*, 2013). This approach is recommended when applying clustering or ordination techniques to species abundance data with sparse representation of some taxa among samples (Legendre and Gallagher, 2001; Rao, 1995). The Bray-Curtis dissimilarity (Bray and Curtis, 1957) was then calculated from the Hellinger-transformed data and used to perform hierarchical clustering. Clustering was carried out using four different linkage methods: single, complete, average (unweighted pair group method with arithmetic mean, UPGMA), and Ward's minimum variance criterion. The best clustering method was identified by determining the cophenetic distance (Jain and Dubes, 1988) of each hierarchical clustering, followed by calculation of the Gower distance (Gower, 1983), the sum of squared differences between the original and cophenetic distances. The method with the smallest Gower distance was selected as the optimal clustering model for the distance matrix used. The optimum number of clusters was selected according to the maximum silhouette width (Kaufman and Rousseeuw, 2009), and resulting cluster assignments were used in subsequent analyses as a categorical representation of community composition. Combined heatmap and dendrogram plots were generated using custom code that used the R package gplots v2.14.1 (Warnes *et al.*, 2014). A Kruskal-Wallis rank sum test was performed to determine whether vaginal pH differed significantly across hierarchical cluster groups (excluding 'Bifido' due to having only two samples). This was followed by a pairwise multiple comparisons test using Tukey's method to identify significant differences between pairs of cluster groups.

As with hierarchical clustering, PCoA was performed with the Bray-Curtis dissimilarity matrix calculated from Hellinger-standardized taxon abundance data. To adjust for negative eigenvalues (the result of using a non-Euclidean distance matrix), a Cailliez correction was applied to the eigenvalues (Cailliez *et al.*, 1976) before calculating R^2-like ratios (essentially variance accounted for by each PCoA axis). PCoA plots were generated using the first two PCoA axes. These plots were used to make qualitative assessments of patterns in the participant metadata (e.g., sample type, Tanner stage, menarcheal status) associated with differences in community composition.

3.5.6 *Analysis of longitudinal trends in community composition and vaginal pH*

Qualitative assessments of changes in community composition associated with pubertal development and menarcheal status were complemented by linear mixed effects modeling to identify variables significantly associated with observed patterns. We used the R package lme4 v1.1-7 (Bates *et al.*, 2014) to perform separate analyses of the relationships between lactic acid bacteria proportions or vaginal pH and select participant metadata. LAB proportions were normalized by logit transformation prior to analysis (Warton and Hui, 2011). Subject was specified as a random effect in all models to control for inter-individual variability, and we elected to exclude subjects with less than three observations of the response variable of interest to minimize inflation of error estimates. Fixed (i.e., explanatory) effects included Tanner breast or pubic stage (ordered factor with levels 1 through 5), menarche status (pre- or post-), and age at sampling. When Tanner stage was included in the model, we performed contrasts between progressive stages (e.g., stage 3 vs. stage 2) to determine whether the response variable was significantly different between them. Starting with a simple model including Tanner stage as the sole fixed effect (e.g., LAB.logit ~ Tanner + 1|Subject + ε, where '1|Subject' specifies subject as a random effect and ε represents random error), we conducted a stepwise model comparison approach adding fixed effects and interactions one by one, using analysis of variance (ANOVA) to compare models at each step. If the models were not significantly different at α=0.05, the simpler model was favored over the more complex model. *P*-values for individual mixed effects models were obtained using the R package lmerTest v2.0-11 (Kuznetsova *et al.*, 2014), which employs Type III and Type I F-tests for fixed effects and likelihood ratio tests for random effects. Upon selecting the best model, residual and quantile-quantile plots were visually inspected to identify any obvious deviations from homoscedasticity or normality.

3.5.7 *Comparisons of the vaginal and vulvar microbiota of girls*

In addition to the hierarchical clustering and PCoA described above, indicator species analysis (De Cáceres *et al.*, 2010) was performed to identify bacterial taxa most strongly associated with groups of vaginal and vulvar samples. This was done by calculating indicator values with the 'IndVal' function (Dufrêne and Legendre, 1997) in the R package indicspecies (De Cáceres and Legendre, 2009). This statistic represents the association of each taxon with one or more groups

of samples. Groups were evaluated based on sample type (girl vagina, girl vulva, mother vagina) and menarche status. Similarity between paired vaginal and vulvar samples from girls (i.e., samples collected at the same visit from the same individual) were assessed by calculating Spearman's rank correlation coefficient from taxon relative abundances.

3.5.8 *Community richness and diversity analyses*

We made qualitative assessments of changes in genus-level community richness and diversity (Simpson's index) in relation to pubertal development (i.e. Tanner stage, menarche status) and vaginal pH, detailed in Appendix B. Rarefaction curves were used to determine a sampling depth of 2,000 observations per sample. Any taxa that could not be classified to the genus level were characterized as 'Other'; therefore the diversity estimates are somewhat conservative. Because of this, we elected not to perform additional quantitative analyses on these data.

CHAPTER 4

Focusing the diversity of *Gardnerella vaginalis* through the lens of ecotypes[4]

4.1 SUMMARY

Gardnerella vaginalis has long been associated with bacterial vaginosis (BV), but its prevalence among both healthy and BV-positive individuals and unanticipated within-species diversity have confounded efforts to clearly define its role. To disentangle the diversity of *G. vaginalis* we invoked the concept of ecotypes—lineages of genetically and ecologically distinct strains within a named species—to better understand their evolutionary history and identify functional characteristics that are likely to be adaptive. Using a variety of comparative genomics and phylogenetic techniques, we found that five clades of *G. vaginalis* were supported by phylogenetic concordance among the core genes. Genome size and GC content differed among clades, and each clade was enriched for a variety of functional genes. We also compared the genomes of *G. vaginalis* to several species in the closely related *Bifidobacterium* genus and found evidence suggesting that while *Gardnerella* appears to have undergone substantial genome amelioration over its evolutionary history, it possesses a large accessory genome relative to *Bifidobacterium*, including many genes that are unique to *Gardnerella*. Our findings present compelling evidence that ecotypes of *G. vaginalis* exist that may represent ecologically meaningful and clinically useful entities.

4.2 INTRODUCTION

Bacteria in the human vagina are known to play a significant role in urogenital health, but it is not always clear whether particular species or strains are beneficial or detrimental to health. *Gardnerella vaginalis*, perhaps more than any other species in the vaginal ecosystem, has long been scrutinized for its close association with bacterial vaginosis (BV). Commonly typified by vaginal itching, discharge and odor, BV has been linked with elevated risks to sexually transmitted infections and preterm birth (Koumans *et al.*, 2007). The connection between *G. vaginalis*

[4] This chapter is to be submitted for publication as: Hickey R.J., Cornejo O.E., Suzuki H., and Forney L.J. Focusing the diversity of *Gardnerella vaginalis* through the lens of ecotypes.

and BV dates back to 1955, when Gardner and Dukes first classified the small, Gram-positive (though variable staining), pleomorphic rods as *Haemophilus vaginalis* (Gardner and Dukes, 1955). It was later reclassified as *Corynebacterium vaginale* (Zinnemann and Turner, 1963) before eventually being renamed after its discoverer as *Gardnerella vaginalis* (Greenwood and Pickett, 1980). Today *G. vaginalis* remains the only recognized species in its genus, with its closest relatives found in the genus *Bifidobacterium*. Early studies reported a strong association of *G. vaginalis* with 'nonspecific vaginitis' (Gardner and Dukes, 1955), what we know today as BV. Since then, considerable research has been devoted to discovering a causal relationship between *G. vaginalis* and BV. Postulated virulence mechanisms include biofilm formation (Swidsinski *et al.*, 2005; Verstraelen and Swidsinski, 2013; Patterson *et al.*, 2010), secretion of exotoxins that facilitate attachment to vaginal epithelial cells (Cauci *et al.*, 1993; Gelber *et al.*, 2008), and the production of enzymes that enable degradation of vaginal mucus components (Wiggins *et al.*, 2001; Gilbert *et al.*, 2013). The prevalence of *G. vaginalis* in BV approaches 100% (Zariffard *et al.*, 2002; Verhelst *et al.*, 2004; Bradshaw *et al.*, 2006; Srinivasan *et al.*, 2012), and although other bacteria have been implicated in the etiology of BV, *G. vaginalis* is often considered a primary indicator of the disease (Muzny and Schwebke, 2013).

The strong correlation between BV and *Gardnerella* has sometimes been taken as direct evidence of causation (Schwebke *et al.*, 2014b). While it is certainly true that *G. vaginalis* is highly correlated with the occurrence of BV, there are many instances in which *G. vaginalis* is present but the symptoms of BV are not. Indeed, *G. vaginalis* is often a major constituent of the vaginal microbiota of healthy, asymptomatic women of all ages (Ravel *et al.*, 2011; Fredricks *et al.*, 2005; Schwebke *et al.*, 2014a) and even young girls (Hickey *et al.*, 2015). Detection of *G. vaginalis* has improved in recent years with changes to the 16S rRNA gene primers often used in procedures to classify bacteria (Frank *et al.*, 2008). Studies have shown that *G. vaginalis* can be a prominent member of vaginal communities in upwards of 40% of healthy individuals (Aroutcheva *et al.*, 2001b; Tabrizi *et al.*, 2006) and Balashov *et al.* (2014) found *G. vaginalis* in 97% of asymptomatic subjects using qPCR. Many earlier cultivation-based studies also reported finding *G. vaginalis* in healthy individuals (McCormack *et al.*, 1977; Sautter and Brown, 1980; Totten *et al.*, 1982), but these findings have been largely overlooked. This oversight has trickled into the very diagnosis of BV in clinical research, where the Nugent score (Nugent *et al.*, 1991) is commonly used to determine BV status even in the absence of clinical symptoms. Using this Gram stain based tech-

nique, the presence of *Gardnerella* or *Bacteroides* morphotypes (i.e., small Gram-variable rods and coccobacilli) in a vaginal smear results in a higher Nugent score and a BV-positive diagnosis. The Nugent score is widely used in research to diagnose BV despite its subjective nature and low sensitivity compared to Amsel's criteria (Chaijareenont *et al.*, 2004; Sha *et al.*, 2005), which are often considered more clinically relevant because they evaluate vaginal discharge, pH, odor and the presence of 'clue cells' (squamous epithelial cells heavily coated in bacteria resembling *G. vaginalis*) (Amsel *et al.*, 1983). The use of Nugent scores has perhaps led to an overestimation of BV in the literature, resulting in the strange notion of asymptomatic BV (Priestley *et al.*, 1997; Schwebke, 2000) and probably inflating the connection between *Gardnerella* and BV. As noted above, numerous studies suggest *G. vaginalis* might be another normal member of the healthy microbiota, at least for many women. This suggests that perhaps some strains are commensal organisms while others may be virulent. If this is the case, separating 'good actors' from 'bad actors' could vastly improve the diagnosis and targeted treatment of BV.

Many researchers have grappled with delineating 'good' and 'bad' strains of *Gardnerella* using a variety of techniques that characterize within-species diversity either by phenotype (e.g., 'biotyping' with biochemical tests; Piot *et al.*, 1984) or genotype (e.g., ARDRA profiling of 16S rRNA genes; Ingianni *et al.*, 1997). More recently, comparative genomics studies have revealed substantial differences in genomic composition that surpass even some of the most diverse species known (Yeoman *et al.*, 2010; Harwich *et al.*, 2010; Ahmed *et al.*, 2012). In the study by Ahmed *et al.* various phylogenetic approaches were applied to characterize the core genes of 17 *G. vaginalis* strains and four phylogenetic clades were identified. Some studies suggest particular biotypes or genotypes display a greater association with BV (Benito *et al.*, 1986; Numanović *et al.*, 2008), but results are inconsistent (Piot *et al.*, 1984; Aroutcheva *et al.*, 2001b)and may be confounded by erroneous biotype identification (Moncla and Pryke, 2009) or the presence of multiple types of *G. vaginalis* within a single individual (Briselden and Hillier, 1990; Santiago *et al.*, 2011; Balashov *et al.*, 2014)]. What we are left with, then, is both a strong sense that variants of *G. vaginalis* with different virulence potential exist and the dilemma of what to do about it in practice. The phenomenon of extensive intraspecies diversity is widespread in the bacterial world (Lan and Reeves, 2000), owing in large part to the peculiarities of bacterial evolution, such as lateral gene transfer and homologous recombination among asexual populations (Dykhuizen and Green, 1991; Riley and Lizotte-Waniewski, 2009). Despite decades of protracted philosophical debate over bacterial

species concepts, it is generally agreed upon that some system of naming organisms is essential to facilitating research and furthering our understanding of microbial diversity (Gevers *et al.*, 2005; Konstantinidis *et al.*, 2006; Doolittle and Papke, 2006).

We propose that the concept of bacterial ecotypes (Cohan, 2001, 2006) could be a useful framework in which to disentangle the diversity of *G. vaginalis* into ecologically meaningful and clinically useful entities. An ecotype is defined as a set of strains representing a particular evolutionary lineage of a named species that are both ecologically distinct and genetically similar to one another (Cohan, 2001). 'Ecological distinctness' can be inferred by determining sets of shared genes or similarities in gene expression patterns under the same environmental conditions. Genetic similarity (or 'cohesion') is characterized using a phylogenetic approach to identify sequence clusters that reflect shared evolutionary history. Ecotypes thus represent lineages within a species that possess unique adaptations and ecological capabilities. Most other taxonomic schemes for bacteria group strains into species based on defined similarity thresholds in phenotypic characteristics, whole genome hybridization, or 16S rRNA gene sequence similarity (Stackebrandt *et al.*, 2002). Instead of this, the ecotype concept relies on ecological and evolutionary theory to demarcate taxonomic units that are ecologically similar (though not necessarily identical) as a result of their common ancestry and shared evolutionary dynamics. Incorporating both types of information (ecological and evolutionary) provides a means to elucidate the specifics of adaptation within lineages and infer important functional characteristics including those that may contribute to pathogenicity (or lack thereof).

In the present study, we sought to evaluate the genomic diversity of *Gardnerella vaginalis* from the perspective of the bacterial ecotype concept. First, we characterized the pangenome of 35 strains of *G. vaginalis* to determine the sets of shared and unique genes as a rough estimate of ecological similarity and distinctiveness among strains. Second, we performed phylogenetic concordance analysis on the core genes of *G. vaginalis* to determine whether groups of strains clustered in a cohesive manner, suggesting distinct lineages. We then tested for differences in the representation of biochemical pathways and protein families among putative clades of *G. vaginalis* as additional evidence of their functional potential. Finally, we compared the genomes of *G. vaginalis* to strains of *Bifidobacterium* spp. to identify the functional characteristics that most readily distinguish *Gardnerella* from its closest evolutionary relatives. Although detailed population studies are needed to verify the existence and clinical relevance of *G. vaginalis* eco-

types, our findings suggest that approaching the within-species diversity of *G. vaginalis* in this manner is likely to yield new insight to the adaptive ecological features of a bacterium that has long perplexed researchers and clinical practitioners alike.

4.3 RESULTS

4.3.1 *General characteristics of Gardnerella vaginalis and Bifidobacterium spp. strains*

The overall goal of this study was to determine the phylogeny of *G. vaginalis* strains and to describe functional differences amongst members of genome clusters or clades as a means to identify putative ecotypes. For this purpose we selected 35 *G. vaginalis* strains with whole genome sequences available in the PATRIC database (`http://patricbrc.org`) that were reportedly isolated from the vagina or endometrium of (presumably) adult women. Clinical and phenotypic characteristics were available for several strains (Table 4.1). Although this study is primarily focused on the within-species diversity of *G. vaginalis*, we also included some of its close relatives in the *Bifidobacterium* genus to better understand how *Gardnerella* has diverged from *Bifidobacterium* over evolutionary time. *Bifidobacterium* is a well-studied genus with at least 34 named species and a large number of available genome sequences (Biavati *et al.*, 2000; Bottacini *et al.*, 2010). We chose 20 strains of six species (Table C.1) that were reportedly isolated from the human gastrointestinal tract (n=15), urogenital tract (n=2) or mammary gland (n=1); the body site of origin was not reported for two *B. animalis* strains. The phylogeny of 18 *G. vaginalis* and 20 *Bifidobacterium* strains based on 16S rRNA gene sequences is depicted in Figure 4.1. *G. vaginalis* formed a strongly supported monophyletic group nested within the *Bifidobacterium* genus, and its closest relatives appear to be *B. bifidum* and *B. thermophilum*. *G. vaginalis* strains were highly similar to each other in 16S rRNA sequence, with pairwise similarity ranging from 98.7–100.0%. The pairwise similarity of *G. vaginalis* strains to the 20 *Bifidobacterium* strains ranged from 91.5–94.8%. Members of the two genera were strikingly different in both genome size and GC content. Genomes of *G. vaginalis* varied from 1.47 to 1.73 Mb (average 1.59 Mb) and the GC content ranged from 41.0–43.4% (average 41.9%). In contrast, the genomes of *Bifidobacterium* species were much larger and ranged from 1.94–2.42 Mb (average 2.27 Mb) with a GC content of 58.6–62.6% (average 59.6%). This was consistent with many other species in the genus studied previously (Bottacini *et al.*, 2010; Lukjancenko *et al.*, 2011). These data indicate there has been

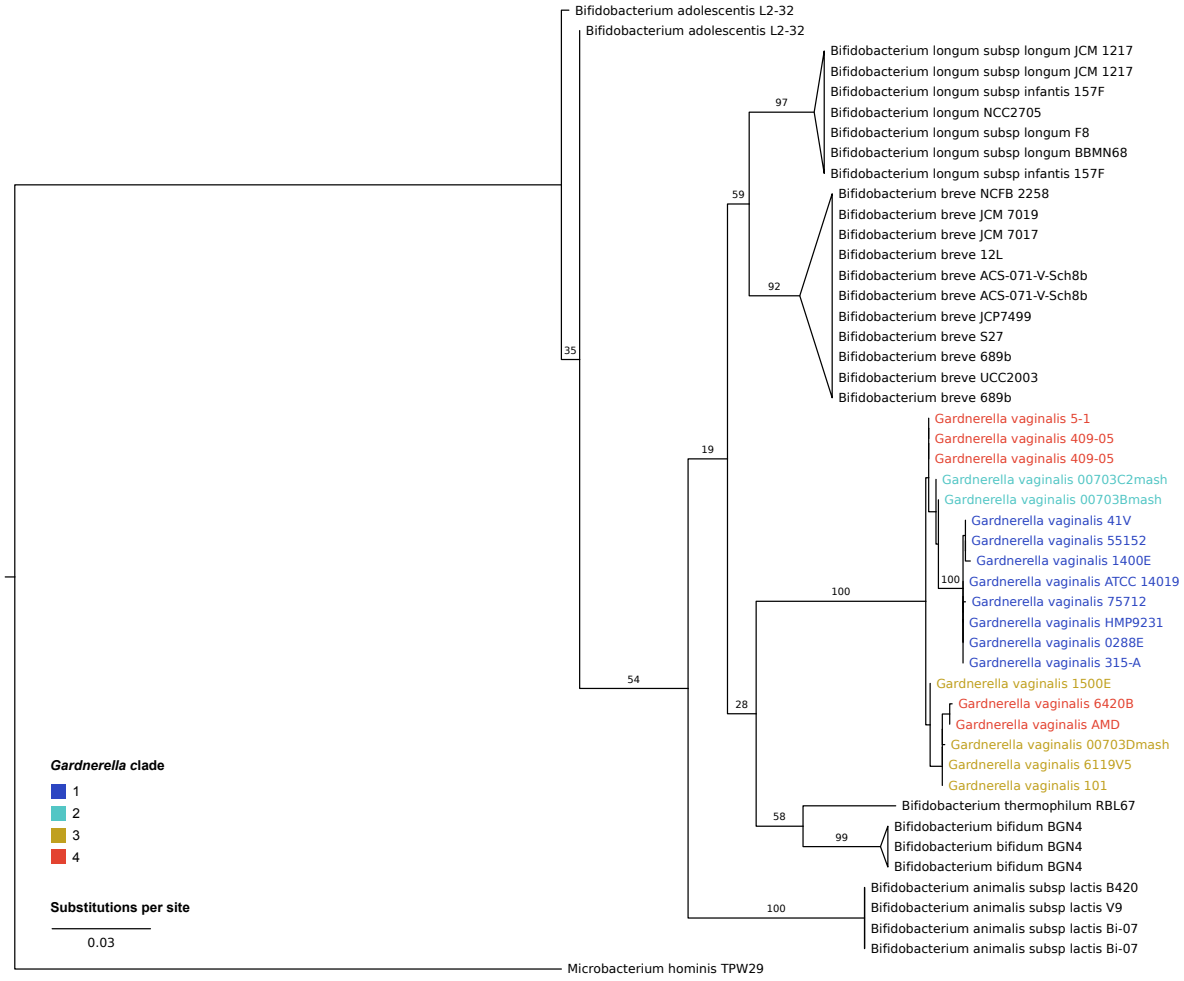

Figure 4.1: 16S rRNA maximum-likelihood phylogeny of *G. vaginalis* and *Bifidobacterium* spp. The maximum-likelihood phylogeny was computed on aligned DNA sequences of 16S rRNA genes (>1400 bp) under the GTR + gamma model of sequence evolution with 1,000 bootstrap replicates. Bootstrap support values are indicated on the branches. Terminal branches were collapsed by species except for *G. vaginalis*. Some strains possess multiple distinct gene copies and are represented more than once on the tree. Coloring of tip labels reflects the clade assignment of *G. vaginalis* strains as shown in Figure 4.4).

extensive genome amelioration in *Gardnerella* throughout its evolutionary history compared to its *Bifidobacterium* relatives. This is striking given that their 16S rRNA gene sequences are >91% identical, a level that would typically place bacterial taxa in the same genus (Janda and Abbott, 2007).

4.3.2 *Pangenome of Gardnerella vaginalis*

To define the pangenome of *G. vaginalis*, we clustered 44,505 coding DNA sequences (CDS) into protein families using an algorithm for grouping orthologous sequences with >70% amino acid sequence identity. Sequences that could not be grouped with any others at this threshold were deemed singletons. The pangenome of *Gardnerella* consisted of 2,392 protein families that were present in two or more strains, 7 protein families present in only one strain, and 1,495 singletons. Of the 2,399 protein families found, 49.4% were annotated as hypothetical proteins, so the functional attributes of a very large portion of *G. vaginalis* genomes are unknowable at this time. An accumulation curve of protein families (Figure 4.2) shows an ever increasing number of observed protein families with the inclusion of more genomes, which indicates there is an open pangenome with significant differences in gene content among strains of this species.

Partitioning the pangenome into the core genome (protein families conserved among all strains) and accessory genome (protein families shared by only some strains) revealed a small core genome and an expansive accessory genome (Figure C.1). The core genome of all 35 strains included 694 protein families, which represents just 29.0% of the 2,392 protein families found in two or more strains of the species. The remaining protein families constituted a large accessory genome in which many genes were present in only one or a few strains. Furthermore, each strain had a number of singletons that were not homologous to any others at a 70% identity threshold.

4.3.3 *Partitioning G. vaginalis strains into genome clusters based on protein family repertoire*

To identify ecologically distinct groups of strains they were clustered based on the similarity of their protein family repertoire. This produced four genome clusters of *G. vaginalis* strains (Figure 4.3). Clusters I and II contained a majority of the strains analyzed (14 and 11 genomes, respectively) and grouped more closely to each other than to clusters III and IV, which had 8 and 2 genomes, respectively. Grouping of genomes in clusters I, II and III were consistent with

Table 4.1: Genomic, clinical and phenotypic characteristics of *Gardnerella vaginalis* strains

Strain	GenBank accession	Genomic characteristics[a]					Clinical and phenotypic characteristics[b]				
		Size (Mb)	Contigs	Plasmids	GC%	CDS	Source	STI	Diagnosis/symptoms	Biotype[c]	Nugent score
00703Bmash	ADET00000000	1.566	16	0	42.3	1,273	Vagina	HSV-2	BV	2 or 5	7
00703C2mash	ADEU00000000	1.547	22	0	42.3	1,237	Vagina	HSV-2	BV	2 or 5	10
00703Dmash	ADEV00000000	1.491	11	0	43.4	1,172	Vagina	HSV-2	BV	3 or 7	3
0288E	ADEN00000000	1.709	17	0	41.2	1,364	Endometrium	Negative	Abnormal discharge, odor	1	8
101	AEJD00000000	1.527	43	0	43.4	1,190	NR	NR	NR	NR	NR
1400E	ADER00000000	1.716	28	0	41.2	1,370	Vagina/endometrium[d]	Negative	NR	4	9
1500E	ADES00000000	1.548	27	0	43	1,195	Vagina/endometrium[d]	Negative	NR	5	7
284V	ADEL00000000	1.651	16	0	41.2	1,304	Endometrium	Chlamydia trachomatis	Abnormal discharge, odor	1	7
315-A	AFDI00000000	1.653	13	0	41.4	1,320	Vagina	NR	NR	NR	NR
409-05	CP001849	1.618	1	0	42	1,190	Vagina	NR	Symptomatic	NR	9
41V	AEJE00000000	1.659	76	0	41.3	1,336	Vagina	NR	Healthy	NR	NR
5-1	ADAN00000000	1.673	94	0	42	1,294	Vagina	NR	Healthy	NR	NR
55152	ADEQ00000000	1.643	25	0	41.3	1,322	Vagina/endometrium[d]	Negative	Asymptomatic	3	8
6119V5	ADEW00000000	1.5	12	0	43.3	1,187	Vagina	Negative	Asymptomatic	7	5
6420B	ADEP00000000	1.494	14	0	42.2	1,162	Vagina/endometrium[d]	Negative	Asymptomatic	2	3
75712	ADEM00000000	1.673	3	0	41.3	1,314	Vagina	Negative	BV/asymptomatic	1	7
AMD	ADAM00000000	1.607	117	0	42.1	1,217	Vagina	NR	BV	NR	NR
ATCC 14018	ADNB00000000	1.604	145	0	41.2	1,313	NR	NR	Symptomatic	NR	NR
ATCC 14019	CP002104	1.667	1	0	41.4	1,345	Vagina	NR	Symptomatic	NR	NR
HMP9231	CP002725	1.727	1	0	41.2	1,376	Endometrium	NR	NR	NR	NR
JCP7275	ATJS00000000	1.56	202	0	41	1,230	Vagina	NR	NR	NR	10
JCP7276	ATJR00000000	1.656	179	0	41	1,315	Vagina	NR	NR	NR	5
JCP7659	ATJQ00000000	1.533	214	0	41.9	1,251	Vagina	NR	NR	NR	8
JCP7672	ATJP00000000	1.601	169	0	41.2	1,251	Vagina	NR	NR	NR	3
JCP7719	ATJO00000000	1.559	185	0	42	1,302	Vagina	NR	NR	NR	8
JCP8017A	ATJN00000000	1.606	187	0	42.1	1,343	Vagina	NR	NR	NR	8
JCP8017B	ATJM00000000	1.599	187	0	42	1,335	Vagina	NR	NR	NR	8
JCP8066	ATJL00000000	1.515	197	0	42.2	1,209	Vagina	NR	NR	NR	0
JCP8070	ATJK00000000	1.476	173	0	42.2	1,208	Vagina	NR	NR	NR	8
JCP8108	ATJJ00000000	1.663	176	0	41.1	1,351	Vagina	NR	NR	NR	8
JCP8151A	ATJI00000000	1.556	189	0	42	1,259	Vagina	NR	NR	NR	10
JCP8151B	ATJH00000000	1.551	185	0	42.2	1,276	Vagina	NR	NR	NR	10
JCP8481A	ATJG00000000	1.567	204	0	42.9	1,263	Vagina	NR	NR	NR	NR
JCP8481B	ATJF00000000	1.57	180	0	42.9	1,251	Vagina	NR	NR	NR	10
JCP8522	ATJE00000000	1.47	191	0	42.2	1,180	Vagina	NR	NR	NR	8

[a] Genomes were downloaded from the PATRIC database in February 2015 (ftp://ftp.patricbrc.org/patric2/). CDS = coding DNA sequence.

[b] From information available in PATRIC, BEI (`http://beiresources.org`), Ahmed *et al.* (2012), Harwich *et al.* (2010), and Yeoman *et al.* (2010). STI = sexually transmitted infection, NR = not reported.

[c] Based on biotype scheme proposed by Piot *et al.* (1984).

[d] PATRIC metadata differs from report by Ahmed *et al.* (2012).

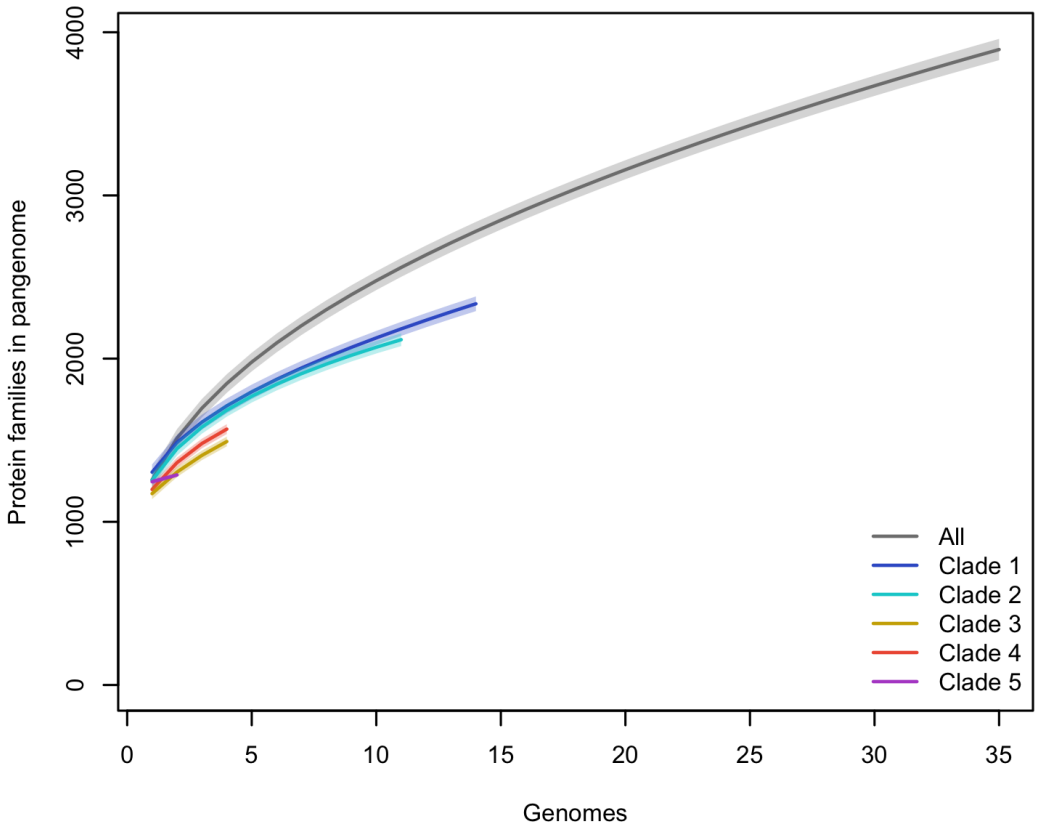

Figure 4.2: Number of protein families in the pangenome of *G. vaginalis*. Accumulation rarefaction curves across 35 *G. vaginalis* strains were calculated based on the presence or absence of protein families and singleton coding genes using the 'specaccum' function in the vegan R package and were estimated by bootstrapping 100 permutations of randomized sample order. The curves for all the genomes analyzed in this study (gray) and for the genomes of each putative clade (see Figure 4.4) are represented.

those determined by Ahmed *et al.* on a smaller set of 17 genomes. The two genomes in cluster IV have not been analyzed previously and may represent a genomic subgroup of *G. vaginalis* that has been under-sampled. Figure 4.3 also shows the Nugent scores and health status (if reported) of the women *G. vaginalis* strains were isolated from. These data reveal a potential sampling bias toward individuals with high Nugent scores (7–10). Nonetheless, isolates from women who had low Nugent scores (0–3) or were reported as asymptomatic are present within the three largest genome clusters.

4.3.4 *Prevalence of putative virulence factors in G. vaginalis genome clusters*

A closer look at the genes encoded by strains within these clusters showed differences in the presence or absence of putative virulence factors (Figure 4.3), notably vaginolysin, sialidase, galactosidases, glucosidases and hexosaminidases (Briselden *et al.*, 1992; Santiago *et al.*, 2011; Pleckaityte *et al.*, 2012; Moncla *et al.*, 2015). Thiol-activated cytolysin (commonly called vaginolysin) was found in most strains (30 of 35) and in all clusters. Surprisingly, three protein families and one singleton were all annotated as sialidase (enzyme EC 3.2.1.18) and there were striking differences in their distribution among strains in the four clusters. One sialidase family was present in 28 of 35 strains, spanning three genome clusters. The remaining two families of sialidase and one singleton were almost entirely found in cluster II. All strains in this cluster possessed at least two distinct sialidase families. While it is known that not all strains of *G. vaginalis* produce sialidase (Santiago *et al.*, 2011), allelic diversity and gene copy numbers of sialidase have not been well characterized. There were seven genomes that appear to lack sialidase, including strain ATCC 14018, the type strain that was originally isolated by Gardner and Dukes.

Recent work has shown that β-galactosidase, α-galactosidase and α-glucosidase activity are positively correlated with BV diagnosed based on Nugent scores (Moncla *et al.*, 2015). Like sialidase, these enzymes may enhance virulence by degrading components of vaginal mucus, thus enabling direct contact with epithelial cell surfaces (Wiggins *et al.*, 2001). α-fucosidase has also been suggested as a virulence factor, although a significant relationship between α-fucosidase activity and BV could not be demonstrated by Moncla *et al.* (2015). We found all of these enzymes in *G. vaginalis*, but only among genomes in cluster I. Notably, β-galactosidase and α-L-fucosidase were completely conserved among all 14 genomes in this cluster.

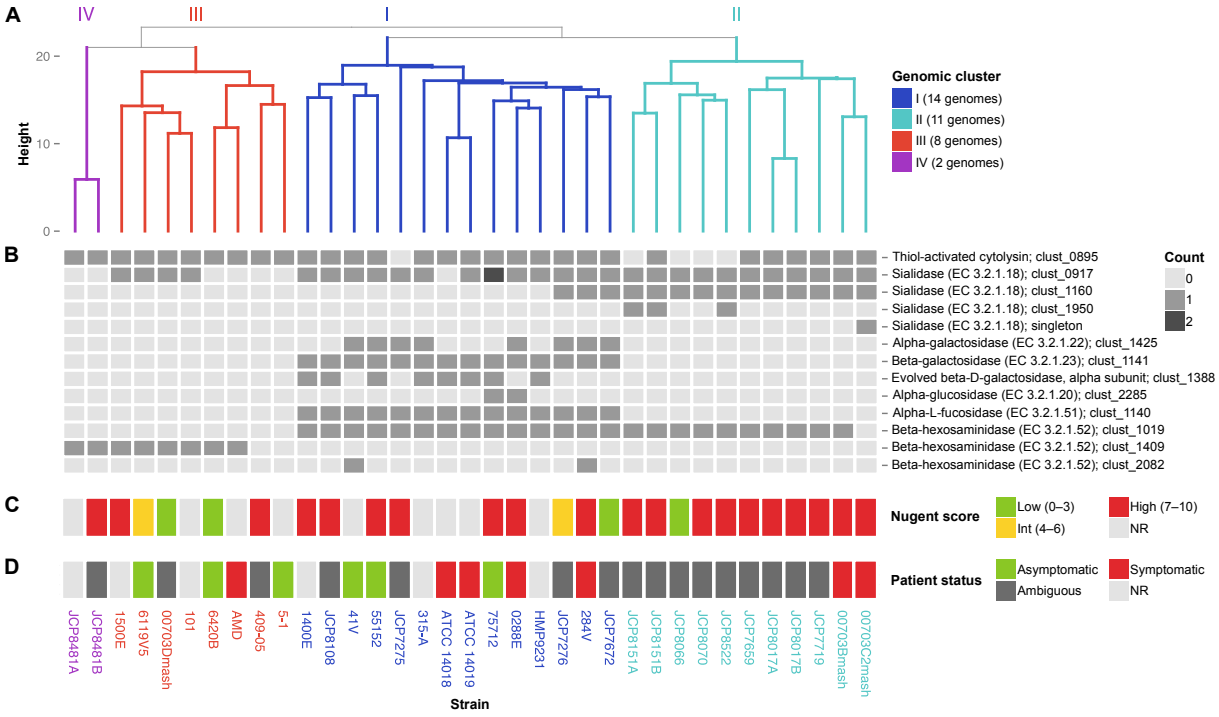

Figure 4.3: Genome clusters of *G. vaginalis* based on protein family repertoire. (A) Dendrogram constructed from the average linkage hierarchical clustering of genomes based on the Bray-Curtis dissimilarity matrix of count data for 2,399 protein families. The maximum silhouette method identified four optimal genome clusters indicated by different colors. (B) Counts of select protein families with putative virulence potential. The most prevalent annotation among each protein family is listed to the right of the graph along with the OrthoMCL cluster identifier. (C) Nugent scores of women *G. vaginalis* isolates were collected from. NR=not reported. (D) Health status of women *G. vaginalis* isolates were collected from. Samples were designated asymptomatic if the terms 'healthy' or 'asymptomatic' were explicitly stated; symptomatic if records indicated abnormal discharge, odor, or antibiotic treatment; or ambiguous if records did not describe clinical data. NR=not reported.

Table 4.2: Summary of core and accessory protein families and singletons among *G. vaginalis* clades

Clade	No. genomes	Core		Accessory		No. protein families	Unique core + accessory (%)	Singletons
		Count	% Total	Count	% Total			
All	35	694	28.9	1,705	71.1	2,399	-	1,495
Clade 1	14	901	51.5	847	48.5	1,748	16.5	588
Clade 2	11	886	52.5	802	47.5	1,688	15.6	428
Clade 3	4	998	76.9	299	23.1	1,297	4.1	195
Clade 4	4	969	72.5	368	27.5	1,337	4.4	231
Clade 5	2	1,204	97.6	30	2.4	1,234	9.6	53

4.3.5 *Partitioning G. vaginalis strains into phylogenetic clades based on the core genome*

In addition to possessing distinguishing ecological characteristics, ecotypes should be genetically cohesive as evident from clustering of gene sequences among strains. This provides evidence that ecologically similar strains are derived from a common ancestor and did not arise coincidentally through the convergent evolution of adaptive traits (Cohan, 2002). The core genome is ideal for this purpose because it likely to reflect genes that have been evolutionarily conserved, whereas accessory genes are probably genes that have been lost, acquired or retained by only some lineages. Therefore, we sought to determine whether the shared core genome of *G. vaginalis* strains formed distinct clades and whether these were consistent with the genome clusters based on protein families as determined above. We performed Bayesian phylogenetic concordance analysis (Figure 4.4) to determine the mostly likely relationships among strains based on the DNA sequences of 664 single-copy core protein families. The primary concordance tree suggests the existence of five clades with consistent support among the individual core gene trees. Clades 1, 2 and 5 were consistent with genome clusters I, II and IV. The strains in genome cluster III were better supported as two phylogenetic clades (clades 3 and 4). Each clade possessed a number of unique protein families constituting up to 16.5% (including both core and accessory) of the clade's protein repertoire (Table 4.2, Figure 4.5). Variability in genome size and GC content shrank considerably when strains were binned into their respective putative clades (Figure 4.6), bringing them more in line with other intraspecies ranges observed (Ahmed *et al.*, 2012). Whereas the range of genome size among all 35 strains was 257 Kbp, the largest range within a single clade was 179 Kbp (clade 4), which represents roughly a 30% reduction in variability. GC content variability was reduced even more drastically; compared to a species-wide range of 2.4%, the greatest within-clade range was just 0.4% (for clades 1, 2 and 3). Taken together with the phylogenetic concordance analysis,

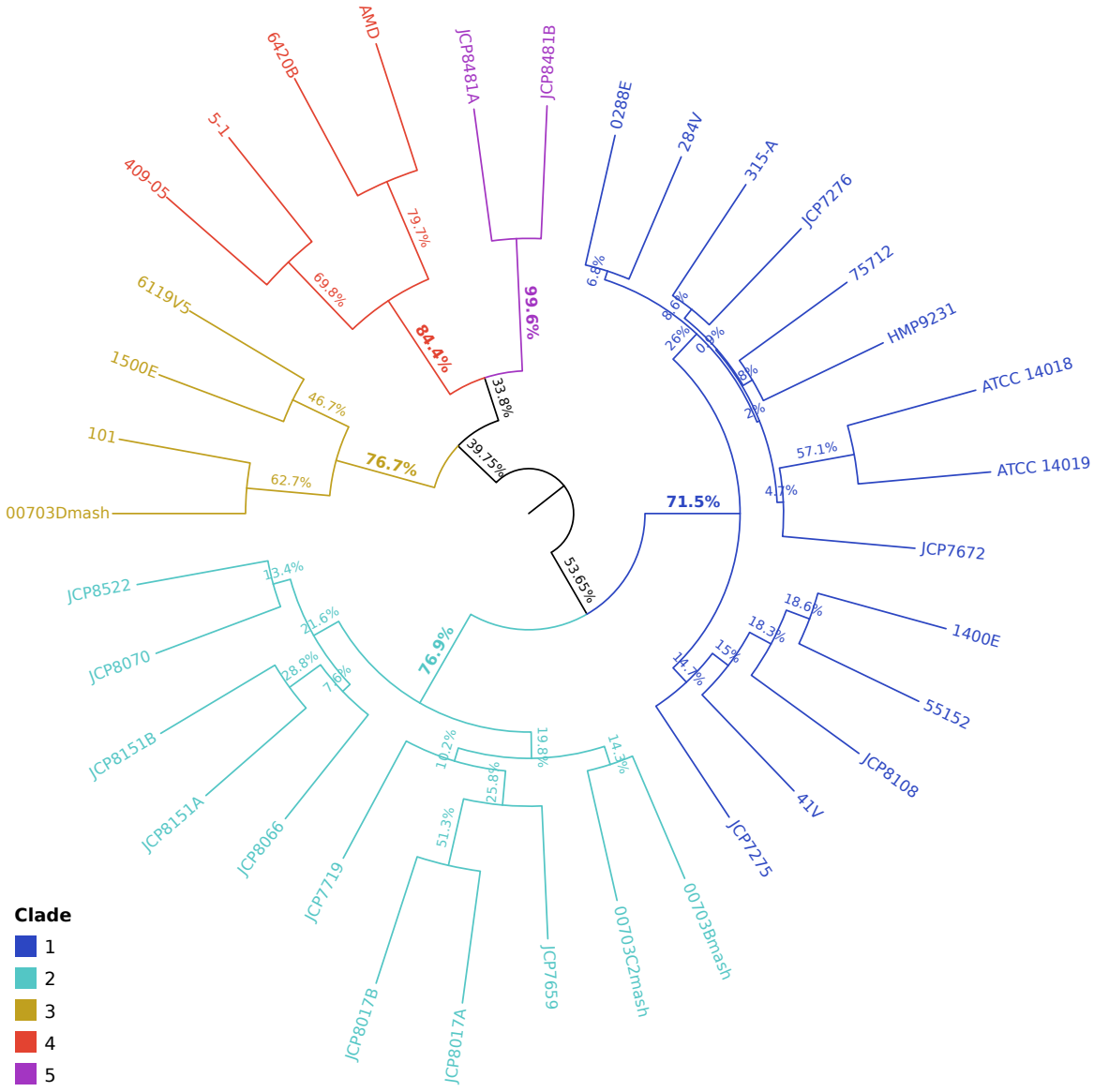

Figure 4.4: Primary concordance tree based on the core genome of *G. vaginalis*. Bayesian concordance analysis was performed with BUCKy software on 664 gene trees computed from the DNA sequences of core protein families for which there was a single representative in each of 35 genomes. The resulting primary concordance tree is shown here, rooted at the midpoint. Branch labels indicate concordance factors as the percentage of gene trees that contain a given split. Tips are labeled with the *G. vaginalis* strain identifiers and colored according to putative taxonomic clades, which are largely consistent with the groupings derived from hierarchical clustering of all protein families (see Figure 4.2). Clade concordance factors are emphasized at the branches leading to the parent node of each clade.

these data support the hypothesis that genetically distinct clusters of *G. vaginalis* exist and differ not only in their protein repertoire, but also in their evolutionary history.

4.3.6 *Enrichment of protein families, functional categories and pathways within clades*

Next, having confirmed multiple clades of *G. vaginalis* with similarities in their overall protein repertoires, we sought to explore the specific pathways and functions that distinguish these clades from one another. This information is essential to the development of hypotheses about potential physiological and ecological differences among clades of *G. vaginalis* and could help to elucidate not only differences in virulence potential but also nutritional requirements. We performed gene set enrichment analysis and examined the data from multiple perspectives by comparing the representation of protein families and corresponding functional annotations, Gene Ontology (GO) categories and KEGG biochemical pathways. Assessment of the GO categories and KEGG pathways is useful because it provides a broader view of protein function and metabolic capability by grouping proteins with similar or related functionality together. Furthermore, each protein may be assigned to multiple categories or pathways to maximize interpretation of the results. The limitation of this approach, however, is that not all of the 44,505 protein CDS are currently represented in these data (14,896 CDS assigned to 375 GO categories; 10,086 CDS assigned to 121 KEGG pathways). Assessment of the orthologous protein families, on the other hand, allows for inclusion of all predicted proteins with functions that have not yet been categorized in a hierarchical annotation system. Using these three datasets, we compared representation (i.e., enrichment) of protein classes between each clade vs. all others and calculated under- or overrepresentation as odds ratios (OR). We assessed significance using Fisher's exact test and adjusted p-values for multiple comparisons according to the false discovery rate (FDR), resulting in q-values. Comprehensive results are included in Supplementary File 3 (`http://github.com/roxanahickey/dissertation`), and we highlight some particularly intriguing results below. The GO categories and KEGG pathways significantly under- or overrepresented in each clade are listed in Table 4.3 and significant protein families in Table 4.4.

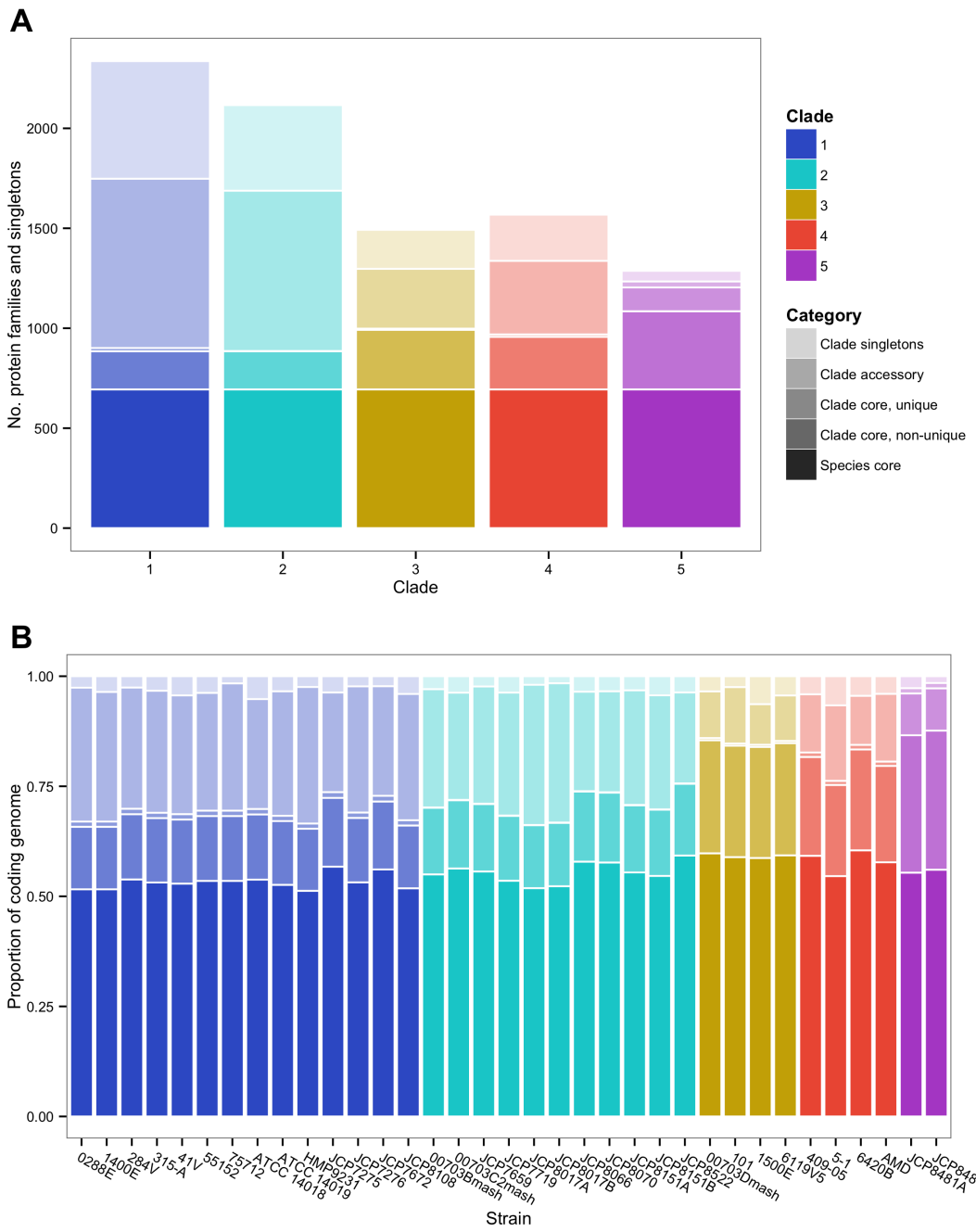

Figure 4.5: Core, accessory and unique protein families within each clade and strain of *G. vaginalis*. A) Genomes grouped and colored by phylogenetic clade. Protein family counts are plotted as stacked bars, with the density of shading corresponding to, in order from darkest to lightest: core protein families among the entire species; protein families core but not unique to each clade; protein families core and unique to each clade; protein families present in two or more genomes within each clade; and singletons present in a single genome. B) Proportions of core and accessory protein families and singletons in each strain's protein-coding genome. Coloring and shading of stacked bars are consistent with (A).

4.3.7 *Enrichment of galactose metabolism, degradation enzymes and other protein families in clade 1*

Clade 1, which had the most genomes (n=14) and the largest average genome size (1.66 Mb), had the greatest number of significantly enriched protein families, GO categories and biochemical pathways. The pathway for galactose metabolism was especially prominent in this clade, being enriched nearly two-fold compared to all other clades (OR=1.80, q=7.96E-08). Out of 13 enzymes assigned to this pathway in KEGG, 12 were unique to genomes in clade 1, including α- and β-galactosidase, α-glucosidase, and maltodextrin glucosidase. Several of the enzymes involved in galactose metabolism were also assigned to pathways for sphingolipid metabolism, glycerolipid metabolism, glycosaminoglycan degradation and glycosphingolipid biosynthesis. The pathway for "pentose and glucuronate interconversions" was also overrepresented in clade 1 (OR=2.85, q=1.30E-05). Most of the enzymes assigned to this pathway (7/8) were uniquely present in genomes of clade 1 and were primarily involved in xylose, ribose and arabinose metabolism. These findings suggest an enhanced ability among strains of clade 1 to utilize galactose and 5-carbon sugars as part of the cell's central metabolism.

In addition to the biochemical pathways just mentioned, clade 1 was also enriched for a number of protein families, including several ABC transporters and permeases (Table 4.4) that could serve a variety of functions involving transport of molecules across cell membranes. One of the most intriguing findings was the presence of a single protein family (clust_1151) annotated as zeta toxin, present in all strains of clade 1 and absent from all others (q-value=2.63E-03). Zeta toxins are highly homologous to PezT, the toxin component of the PezAT toxin-antitoxin (TA) system, which has been shown to kill bacteria by inhibiting peptidoglycan biosynthesis (Mutschler *et al.*, 2011). Mutschler *et al.* described it as a "potent Achilles' heel for microbes" and showed that partial autolysis caused by PezT in subpopulations of Streptococcus pneumoniae could favor biofilm formation. Zeta toxin might perform a similar function in strains of *G. vaginalis*, although this remains to be demonstrated experimentally. We also explored the prevalence of other proteins that may be involved in cell attachment and biofilm formation and found numerous protein families annotated as glycosyltransferases, LPxTG-specific sortase A and other LPxTG-motif recognition enzymes among strains of all clades (Figure C.2).

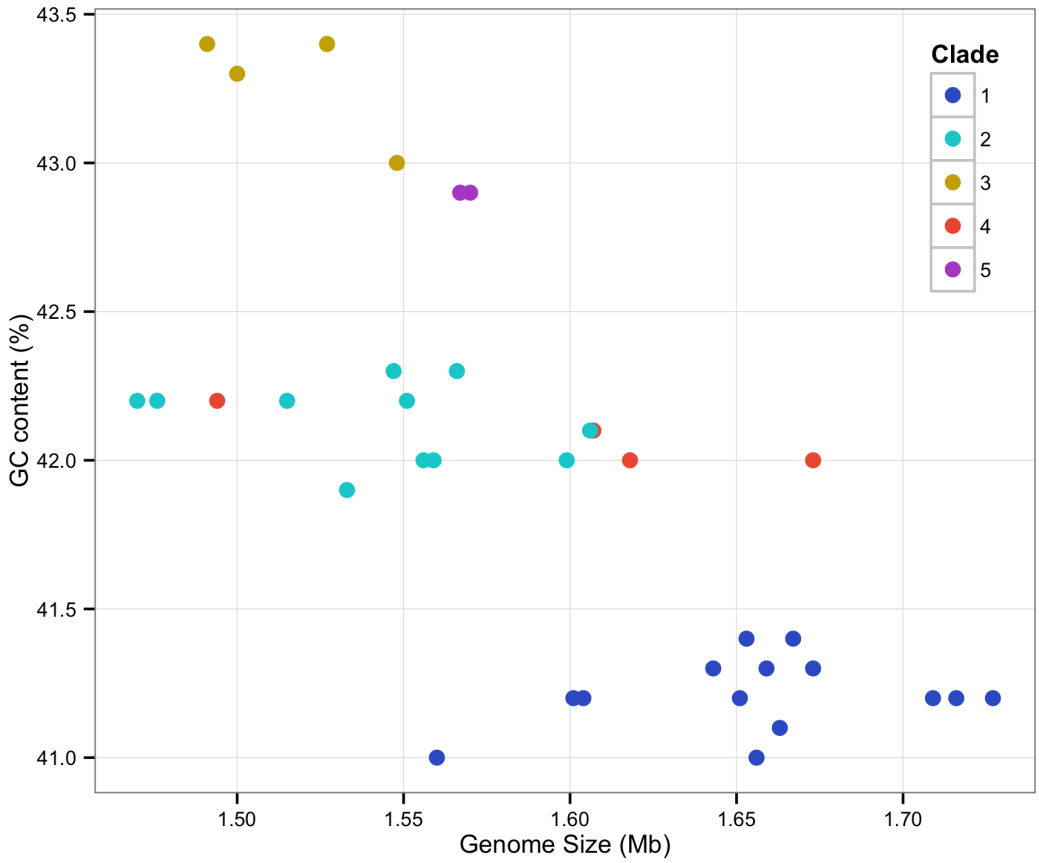

Figure 4.6: Variation in genome size and GC composition of *G. vaginalis* genomes. Genome size (Mb) and GC content (%) of 35 *G. vaginalis* genomes are plotted on the x-axis and y-axis, respectively. Points are colored to indicate the clade each strain was assigned to using phylogenetic concordance analysis of the core genes.

4.3.8 *Enrichment of sialidase and other protein families in clade 2*

Clade 2 had several enriched and unique protein families and GO categories. Perhaps the most striking overrepresented GO category was that of exo-α-sialidase activity (OR=2.868, q=3.30E-02), confirming our earlier observation of a greater number of sialidase protein families in genome cluster II. One sialidase family in particular (clust_1160) was highly enriched (OR=8.07, q=3.69E-04) and conserved among all 11 genomes (see also Figure 4.3). Members of clade 2 lacked enzymes assigned to the GO category for thiamine biosynthesis; however, it had three unique protein families annotated as the three components of an energy-coupling factor (ECF) transporter. ECF transporters are required for uptake of some micronutrients, with known substrates including riboflavin, thiamine, biotin and tryptophan (Rodionov *et al.*, 2009; Eitinger *et al.*, 2011), suggesting members of this clade may obtain thiamine or other essential micronutrients from external sources.

4.3.9 *Enrichment of pathways and protein families in clades 3, 4, and 5*

Only a small number of protein families and pathways were significantly enriched among clades 3 and 4 after adjustment for multiple comparisons, and none were significant in clade 5. This is likely due to the small number of genomes in these clades. The only pathway with a statistically significant result in clade 3 was the complete absence of polyketide sugar unit biosynthesis (q=8.12E-03) while clade 4 was enriched for three enzymes associated with degradation of allantoin, a metabolic byproduct of purine catabolism. Additionally, clade 4 was enriched with several proteins involved in thiamine biosynthesis. Notably, all four strains of clade 4 lacked protein families annotated as sialidase enzymes.

4.3.10 *Pangenome of Gardnerella and Bifidobacterium spp.*

To understand *Gardnerella's* functional potential at the genus level, we compared the protein-coding genomes of 35 *G. vaginalis* and 20 *Bifidobacterium* strains to estimate their collective pangenome. Using a relaxed threshold of 50% amino acid identity to allow for greater divergence in protein families between genera we identified 4,633 homologous protein families among the 82,781 CDS present in the two genera. *G. vaginalis* strains shared 703 core protein families, while *Bifidobacterium* strains shared 906 core protein families. Thus, even though *G. vaginalis* is

Table 4.3: KEGG biochemical pathways and Gene Ontology categories with significantly different representation among *G. vaginalis* clades

Clade	Database[a]	Category identifier and description	No. unique proteins	Odds ratio[b]	*q*-value[c]	
1	KEGG	52 Galactose metabolism	13	1.799	7.96E-05	***
1	KEGG	40 Pentose and glucuronate interconversions	8	2.85	1.30E-05	***
1	KEGG	120 Primary bile acid biosynthesis	1	Inf	1.02E-02	*
1	KEGG	121 Secondary bile acid biosynthesis	1	Inf	1.02E-02	*
1	GO	GO:0008728 GTP diphosphokinase activity	1	0	4.74E-02	*
1	GO	GO:0005694 chromosome	4	1.739	3.30E-02	*
1	GO	GO:0051276 chromosome organization and biogenesis	2	2.671	1.07E-03	**
1	GO	GO:0017103 UTP:galactose-1-phosphate uridylyltransferase activity	1	3.539	2.39E-02	*
1	GO	GO:0008982 protein-N(PI)-phosphohistidine-sugar phosphotransferase activity	2	16.66	2.38E-02	*
1	GO	GO:0004565 β-galactosidase activity	2	Inf	1.30E-05	***
1	GO	GO:0004335 galactokinase activity	1	Inf	1.72E-03	**
1	GO	GO:0004560 α-L-fucosidase activity	1	Inf	1.72E-03	**
1	GO	GO:0004856 xylulokinase activity	1	Inf	1.72E-03	**
1	GO	GO:0008733 L-arabinose isomerase activity	1	Inf	2.33E-03	**
1	GO	GO:0008741 ribulokinase activity	1	Inf	2.33E-03	**
1	GO	GO:0008742 L-ribulose-phosphate4-epimerase activity	1	Inf	2.33E-03	**
1	GO	GO:0004316 3-oxoacyl-[acyl-carrier-protein] reductase activity	1	Inf	4.84E-03	**
1	GO	GO:0004139 deoxyribose-phosphate aldolase activity	1	Inf	3.46E-02	*
2	KEGG	40 Pentose and glucuronate interconversions	8	0.472	4.77E-02	*
2	KEGG	363 Bisphenol A degradation	1	23.016	8.01E-03	**
2	KEGG	140 C21-Steroid hormone metabolism	1	Inf	1.84E-03	**
2	KEGG	591 Linoleic acid metabolism	1	Inf	1.84E-03	**
2	GO	GO:0009228 thiamin biosynthetic process	1	0	1.17E-02	*
2	GO	GO:0004565 β-galactosidase activity	2	0	4.89E-02	*
2	GO	GO:0004558 α-glucosidase activity	2	0.128	2.73E-02	*
2	GO	GO:0004308 exo-α-sialidase activity	1	2.868	3.30E-02	*
2	GO	GO:0005506 iron ion binding	2	23.125	7.85E-03	**
2	GO	GO:0016491 oxidoreductase activity	2	23.125	7.85E-03	**
2	GO	GO:0004640 phosphoribosylanthranilate isomerase activity	1	Inf	1.84E-03	**
2	GO	GO:0008865 fructokinase activity	1	Inf	1.84E-03	**
2	GO	GO:0009073 aromatic amino acid family biosynthetic process	1	Inf	1.84E-03	**
3	KEGG	523 Polyketide sugar unit biosynthesis	5	0	8.12E-03	**
4	GO	GO:0004038 allantoinase activity	1	Inf	1.90E-02	*
4	GO	GO:0006790 sulfur metabolic process	1	Inf	1.90E-02	*
4	GO	GO:0047652 allantoate deiminase activity	1	Inf	1.90E-02	*

[a] KEGG pathway and Gene Ontology (GO) category annotations were downloaded from the PATRIC database in February 2015 (ftp://ftp.patricbrc.org/patric2).

[b] Odds ratio (OR) >1, overrepresentation; OR < 1, underrepresentation; OR = 0, category absent from clade; OR = Inf (Infinite), category unique to clade.

[c] False discovery rate adjusted *p*-value (obtained by Fisher's exact test). Significance < 0.05 *, < 0.01 **, < 0.001 ***

considered a single species, it has fewer protein families common to all of its members than do six species of *Bifidobacterium* collectively. The core genomes of *Gardnerella* and *Bifidobacterium* had 553 protein families in common, representing 78.6% of *Gardnerella's* core genome and 61.0% of *Bifidobacterium's* core genome. This overlap reflects a common (though likely distant) ancestry between *Gardnerella* and *Bifidobacterium* and hints that *Gardnerella* has probably lost many core genes retained in *Bifidobacterium*. We found a relatively large accessory genome in *G. vaginalis* of 1,720 protein families, which was comparable in number to the accessory genome of all strains from six species of *Bifidobacterium* (n=1,757 protein families). These data reveal that the within-species differences in genomic content among *G. vaginalis* strains are on par or even greater

than those observed among multiple species of *Bifidobacterium*, suggesting extensive divergent evolution between the two genera as well as within *Gardnerella*.

4.3.11 *Comparison of G. vaginalis and Bifidobacterium spp. protein family repertoires*

To gain additional insight to the functional differences between *Gardnerella* and *Bifidobacterium*, we performed another gene set enrichment analysis on the protein families, GO categories and biochemical pathways that were represented in the genomes of the two genera (see Supplementary File 4 at `http://github.com/roxanahickey/dissertation`). After adjusting for multiple comparisons, we identified 785 protein families that were found in *Bifidobacterium* genomes but absent from *Gardnerella*; 30 that were underrepresented in *Gardnerella*; 88 that were overrepresented; and 282 families that were unique to *Gardnerella*. Among the protein families unique to *Gardnerella* were several putative virulence factors including zeta toxin, YafQ toxin, other toxin-antitoxin proteins and thiol-activated cytolysin. Additionally, several protein families were annotated as transporter proteins, which may be used to acquire nutrients that *Gardnerella* is unable to synthesize. Given their smaller genomes it is not surprising that *G. vaginalis* also had fewer GO categories and biochemical pathways than species of *Bifidobacterium*. The results show that β-galactosidase, α-galactosidase and α-glucosidase were significantly less prevalent in *G. vaginalis* relative to species of *Bifidobacterium* probably because only strains in clade 1 possessed such genes (Table C.2). However, the exo-α-sialidase GO category was significantly overrepresented in *G. vaginalis* (OR=3.59, q=7.88E-04) and only 10 of 20 *Bifidobacterium* strains possessed genes encoding sialidase (all were B. bifidum or B. breve). Furthermore, while 26 biochemical pathways were missing or significantly underrepresented in *Gardnerella* relative to *Bifidobacterium*, a few pathways were significantly overrepresented in *Gardnerella* by a factor of two or more, including "drug metabolism - cytochrome P450" and "metabolism of xenobiotics by cytochrome P450" (Table C.3). Collectively, these findings suggest that while *G. vaginalis* lacks many of the metabolic capabilities present in *Bifidobacterium* spp., it appears to encode many proteins that may confer greater resistance to antibiotics, kill other bacteria via toxin-antitoxin systems, or enhance degradation of vaginal mucus by sialidase.

Table 4.4: Protein families with significantly different representation among *G. vaginalis* clades

Clade	Cluster[a]	Representative protein product[b]	Odds ratio[c]	*q*-value[d]	
1	clust_0012	Galactose-1-phosphate uridylyltransferase (EC2.7.7.10)	3.567	2.57E-02	*
1	clust_1053	L-arabinose transport ATP-binding protein AraG (TC3.A.1.2.2)	Inf	1.24E-04	***
1	clust_1143	ABC transporter, solute-binding protein	Inf	2.63E-03	**
1	clust_1142	ABC-type sugar transport system, permease component	Inf	2.63E-03	**
1	clust_1140	α-L-fucosidase (EC3.2.1.51)	Inf	2.63E-03	**
1	clust_1141	β-galactosidase (EC3.2.1.23)	Inf	2.63E-03	**
1	clust_1148	Galactokinase (EC2.7.1.6)	Inf	2.63E-03	**
1	clust_1155	N-acylglucosamine2-epimerase (EC5.1.3.8)	Inf	2.63E-03	**
1	clust_1150	transcription regulator, probable	Inf	2.63E-03	**
1	clust_1146	Xylose-responsive transcription regulator, ROK family	Inf	2.63E-03	**
1	clust_1147	Xylulose kinase (EC2.7.1.17)	Inf	2.63E-03	**
1	clust_1151	Zeta toxin	Inf	2.63E-03	**
1	clust_1191	ABC transporter, ATP-binding protein	Inf	2.63E-03	**
1	clust_1170	ABC-type sugar transport systems, permease components	Inf	2.63E-03	**
1	clust_1181	Inositol transport system permease protein	Inf	2.63E-03	**
1	clust_1182	Inositol transport system permease protein	Inf	2.63E-03	**
1	clust_1184	L-arabinose isomerase (EC5.3.1.4)	Inf	2.63E-03	**
1	clust_1188	L-arabinose transport system permease protein (TC3.A.1.2.2)	Inf	2.63E-03	**
1	clust_1189	L-arabinose-binding periplasmic protein precursor AraF (TC3.A.1.2.2)	Inf	2.63E-03	**
1	clust_1185	L-ribulose-5-phosphate4-epimerase (EC5.1.3.4)	Inf	2.63E-03	**
1	clust_1179	Maltodextrin glucosidase (EC3.2.1.20)	Inf	2.63E-03	**
1	clust_1180	N-Acetyl-D-glucosamine ABC transport system, sugar-binding protein	Inf	2.63E-03	**
1	clust_1183	oxidoreductase of aldo/keto reductase family, subgroup1	Inf	2.63E-03	**
1	clust_1186	Ribulokinase (EC2.7.1.16)	Inf	2.63E-03	**
1	clust_1178	Transcriptional regulator	Inf	2.63E-03	**
1	clust_1190	Transcriptional regulator/sugar kinase	Inf	2.63E-03	**
1	clust_1176	Translation initiation factor2	Inf	2.63E-03	**
1	clust_1217	Crp-family transcriptional regulator	Inf	5.56E-03	**
1	clust_1213	PTS system, cellobiose-specific IIC component (EC2.7.1.69)	Inf	5.56E-03	**
1	clust_1216	sugar permease of ABC transporter system	Inf	5.56E-03	**
1	clust_1214	sugar transporter sugar binding protein	Inf	5.56E-03	**
1	clust_1256	2-keto-3-deoxygluconate permease (KDG permease)	Inf	1.13E-02	*
1	clust_1257	3-oxoacyl-[acyl-carrier protein] reductase (EC1.1.1.100)	Inf	1.13E-02	*
1	clust_1251	Choloylglycine hydrolase (EC3.5.1.24)	Inf	1.13E-02	*
1	clust_1248	periplasmic binding protein/LacI transcriptional regulator	Inf	1.13E-02	*
1	clust_1249	Putative cytoplasmic protein	Inf	1.13E-02	*
1	clust_1250	Putative inner membrane protein	Inf	1.13E-02	*
1	clust_1295	Choline binding protein A	Inf	1.87E-02	*
1	clust_1343	Cytoplasmic axial filament protein CafA and Ribonuclease G (EC3.1.4.-)	Inf	3.66E-02	*
1	clust_1352	Deoxyribonucleoside regulator DeoR (transcriptional repressor)	Inf	3.66E-02	*
1	clust_1353	Deoxyribose-phosphate aldolase (EC4.1.2.4)	Inf	3.66E-02	*
1	clust_1354	Homolog of fucose/glucose/galactose permeases	Inf	3.66E-02	*
1	clust_1355	Putative uncharacterized protein STY3991	Inf	3.66E-02	*
1	clust_1356	Ribokinase (EC2.7.1.15)	Inf	3.66E-02	*
2	clust_1002	Glycine oxidase ThiO (EC1.4.3.19)	0	2.24E-02	*
2	clust_1005	Hydroxymethylpyrimidine phosphate synthase ThiC / Thiamin-phosphate pyrophosphorylase (EC2.5.1.3)	0	2.24E-02	*
2	clust_1004	prophage ps3 protein 01, putative	0	2.24E-02	*
2	clust_0999	Sulfur carrier protein adenylyltransferase ThiF	0	2.24E-02	*
2	clust_1000	Thiazole biosynthesis protein ThiG	0	2.24E-02	*
2	clust_1160	Sialidase (EC3.2.1.18)	8.068	3.66E-02	*
2	clust_1316	ATPase component of general energizing module of ECF transporters	Inf	2.63E-03	**
2	clust_1315	ATPase component STY3233 of energizing module of queuosine-regulated ECF transporter	Inf	2.63E-03	**
2	clust_1323	Endoglucanase E precursor (EgE) (Endo-1,4-β- glucanase E) (Cellulase E)	Inf	2.63E-03	**
2	clust_1313	Fructokinase (EC2.7.1.4)	Inf	2.63E-03	**
2	clust_1318	Hypothetical sugar kinase in cluster with indigoidine synthase indA , PfkB family of kinases	Inf	2.63E-03	**
2	clust_1312	Inosine-uridine preferring nucleoside hydrolase (EC3.2.2.1)	Inf	2.63E-03	**
2	clust_1310	NADH-dependent butanol dehydrogenase A (EC1.1.1.-)	Inf	2.63E-03	**
2	clust_1311	Phosphoglycolate phosphatase	Inf	2.63E-03	**
2	clust_1314	Phosphoribosylanthranilate isomerase (EC5.3.1.24)	Inf	2.63E-03	**
2	clust_1309	Similar to tetracycline resistance protein	Inf	2.63E-03	**
2	clust_1317	Transmembrane component STY3231 of energizing module of queuosine-regulated ECF transporter	Inf	2.63E-03	**
4	clust_1669	Allantoate amidohydrolase (EC3.5.3.9)	Inf	1.65E-02	*
4	clust_1670	Allantoin permease	Inf	1.65E-02	*
4	clust_1671	Allantoinase (EC3.5.2.5)	Inf	1.65E-02	*
4	clust_1681	Cytoplasmic axial filament protein CafA and Ribonuclease G (EC3.1.4.-)	Inf	1.65E-02	*
4	clust_1673	Sulfur carrier protein adenylyltransferase ThiF	Inf	1.65E-02	*
4	clust_1672	Sulfur carrier protein ThiS	Inf	1.65E-02	*
4	clust_1674	Thiazole biosynthesis protein ThiG	Inf	1.65E-02	*
4	clust_1753	Translation initiation factor2	Inf	1.65E-02	*

[a] Orthologous clusters of coding DNA sequences (CDS) determined using OrthoMCL with 70% identity.

[b] Most common PATRIC protein feature annotation among protein family. 57 hypothetical proteins are excluded from this table.

[c] Odds ratio (OR) >1, overrepresentation; OR < 1, underrepresentation; OR = 0, pathway absent from clade; OR = Inf (Infinite), pathway unique to clade.

[d] False discovery rate adjusted *p*-value (obtained by Fisher's exact test). Significance: <0.05*, <0.01**, <0.001***.

4.4 DISCUSSION

In this study we sought to characterize the diversity of *Gardnerella vaginalis* within the framework of bacterial ecotypes, which are genetically cohesive lineages of a named species that have distinct ecological characteristics. We anticipated that this approach would constitute a significant step toward separating strains into ecologically meaningful units that could be further investigated for clinical relevance. We showed that 35 strains of *G. vaginalis* could be grouped into five clades based on congruence in the phylogenetic structure of their core genes, overlap in shared protein families, and similarity in overall genomic composition. Our findings provide evidence of the existence of multiple *G. vaginalis* ecotypes that probably evolved over time in response to distinctive selective pressures that reflect differences among microbial communities and hosts.

The results of our comparative genomic and phylogenetic analyses confirm and extend the findings of Ahmed *et al.* (2012) that identified four clades ('genovars') of *G. vaginalis* based on the analysis of 17 strains. The authors remarked that the genomic diversity among these strains was great enough to warrant designation as four separate species. We recapitulated the same four clades using a phylogenetic concordance analysis of the core genes and identified a new fifth clade consisting of two non-independent isolates. Clades 1 and 2 are presently the two most well-represented clades and include 25 of the 35 genomes analyzed here. It should be noted that most genome sequences were from strains isolated from women with high Nugent scores (7–10), suggesting that genomes of *G. vaginalis* strains from women with low (0–3) or intermediate (4–6) Nugent scores have probably been under-sampled. Specifically, there are few members of clades 3, 4 and 5, and this might reflect bias in the subject populations studied or the cultivation efficiency of selected isolates. This notion is supported by the work of Balashov *et al.* (2014) who developed clade-specific qPCR primers to assess the prevalence of each *G. vaginalis* clade among the vaginal microbiota of 60 women with (n=24) or without BV (n=36) as determined by both Nugent and Amsel criteria. The authors detected *G. vaginalis* in 59 of 60 specimens by qPCR targeting a species-specific tuf gene and reported that >70% of women carried two or more clades simultaneously. Clade 4 was most commonly identified (50/60 specimens), followed by clade 1 (32/60), clade 3 (19/60) and clade 2 (15/60). Notably, women who were co-colonized by two or more clades more often had both high Nugent (7–10) and Amsel scores (3 or 4), which together are indicative of symptomatic BV. Considering each clade individually, only clades 1 and 3 were positively associated with high Nugent and Amsel scores. Clade 2 was positively associated only

with intermediate Nugent scores, and clade 4 showed no association. The findings of Balashov *et al.* thus provided the first evidence that some clades described by Ahmed *et al.* and confirmed here were more commonly associated with a BV-positive diagnosis, but the specific metabolic characteristics or the virulence determinants possibly involved remained largely unknown.

Our study adds significant new insight to the potential ecological differences among clades of *G. vaginalis* and supports an emerging view that particular clades—that we here refer to as ecotypes—possess greater virulence potential while others might be relatively benign. In particular, we found that the genomes of strains in clade 1 uniquely encode several glycosidases (e.g., galactosidases, glucosidases and fucosidases) and have expanded capabilities for galactose and pentose sugar metabolism. The most notable feature of strains in clade 2 is the possession of at least two distinct genes encoding sialidase (also a type of glycosidase). This echoes an earlier report of multiple sialidase alleles that were predicted to be functionally similar (Santiago *et al.*, 2011); however, strains in clade 2 were not included in that study. Interestingly, our results indicate that genomes in clades 4 and 5 lack genes for any of these enzymes. This point is especially noteworthy considering the observations of Balashov *et al.* that clade 4 was the most prevalent subtype among both healthy and BV-positive subjects, and clade 1 was positively associated with symptomatic BV. Glycosidases represent a large family of enzymes that are capable of degrading large, glycosylated mucin proteins (Wiggins *et al.*, 2001). Activity of such enzymes may increase susceptibility to infection by thinning the protective layer of vaginal mucus (Briselden *et al.*, 1992; Cauci *et al.*, 1993). Moncla *et al.* (2015) recently demonstrated that enzymatic activities of four glycosidases present in *G. vaginalis*—sialidase, α-galactosidase, β-galactosidase and α-fucosidase—were positively associated with high Nugent scores and may also be influenced by hormonal activity. Clade membership of strains was not assessed in their study, but taken together with our findings, we could reasonably anticipate that strains in clades 1, 2 and 3 might have an enhanced ability to degrade components of vaginal mucus.

An understanding of the serpentine history of *G. vaginalis* classification is needed to appreciate the importance of interpreting diversity within this species. In the first decade following the initial discovery of *Haemophilus vaginalis* in the 1950s, attempts to satisfy Koch's postulates of disease for 'nonspecific bacterial vaginitis' were complicated by differences in reported phenotypes and isolation from both healthy and diseased individuals (reviewed in Catlin, 1992). Eventually, researchers began to cast doubt on its classification as *Haemophilus* due to a number of

nutritional and phenotypic inconsistencies with the genus (Reyn *et al*., 1966; Criswell *et al*., 1971; Vickerstaff and Cole, 1969). In 1963, Zinnemann and Turner recommended reclassification as *Corynebacterium vaginale* solely based on its morphological resemblance to the corynebacteria, although both species names persisted in the literature for many years thereafter. In 1980, Greenwood and Pickett determined using DNA-DNA hybridization assays that strains recognized as *H. vaginalis* or *C. vaginale* should be allocated to a separate genus altogether, leading to the birth of *Gardnerella* as we know it today (Greenwood and Pickett, 1980). Later that year, Piot *et al*. (1980) corroborated this new genus with a battery of DNA-DNA hybridization assays spanning several genera, including *Haemophilus, Corynebacterium* and *Bifidobacterium*. *Gardnerella* remained an 'orphan genus' until the mid-1990s, when researchers first noted the close relationship between strain ATCC 14018 (originally isolated by Gardner and Dukes) and *Bifidobacterium* spp. based on 16S rRNA gene sequences (Embley and Stackebrandt, 1994; Leblond-Bourget *et al*., 1996). Leblond-Bourget *et al*. reasoned that although the similarity between *Gardnerella* and *Bifidobacterium* sequences was great enough to warrant assignment to the same genus, weak DNA-DNA hybridization and non-overlapping ranges in GC content supported keeping the two genera separate. The close phylogenetic relationship of these two genera was further buoyed by the discovery in *G. vaginalis* ATCC 14018 of the gene for fructose-6-phosphate phosphoketolase (F6PPK), a key enzyme of the 'Bifid-shunt' that was previously thought to occur almost exclusively in *Bifidobacterium* spp. (Gavini *et al*., 1996). Thus, 40 years after its discovery, *Gardnerella* found its place within the family Bifidobacteriales, amongst a sea of *Bifidobacterium* spp.

Bifidobacterium spp. are rarely identified as dominant members of the vaginal microbiome (Burton *et al*., 2003), but they are abundant and generally considered 'good players' in the gut microbiome (Ventura *et al*., 2007) and commonly used in cultured yogurt and probiotic supplements (Ventura *et al*., 2009; Lukjancenko *et al*., 2011). It is therefore somewhat surprising that *Gardnerella*, a suspected pathogen, sits in close evolutionary proximity to bacteria that are widely regarded as beneficial to humans. This prompted us to take a closer look at what distinguishes *Gardnerella* and *Bifidobacterium* at a functional genomic level. Our findings indicate that most of *Gardnerella*'s core genome overlaps with the core genome of six *Bifidobacterium* spp., but it has a large accessory genome including many protein families that are unique to *Gardnerella*, and likewise species of *Bifidobacterium* possess hundreds of their own unique features. *G. vaginalis* genomes are substantially smaller and have a significantly lower GC content when compared

to *Bifidobacterium*. Despite having up to 95% similarity in 16S rRNA gene sequence, major differences in the genomic composition of these two genera probably reflect a deep evolutionary split and broadly different ecology.

Phylogenetic analysis of a single marker gene such as 16S rRNA is useful for assigning names to bacterial taxa but masks the genomic and functional diversity among strains. Characterization by phenotype (or biotype) may offer additional insight, but this approach is probably biased toward strains that are easily cultivated in the laboratory. Neither approach offers significant explanatory power of differences in ecological potential among strains because, as Cohan has pointed out (Cohan, 2001), most procedures for classifying bacteria based solely on phenotypic or genotypic characteristics are not grounded in ecological and evolutionary theory. The ecotype concept, in contrast, distills subspecies into evolutionarily cohesive, ecologically meaningful units that contain information about both shared ancestry and functional characteristics. This approach is useful because it identifies groups of strains that have presumably experienced common selective pressures and environmental conditions throughout their evolutionary history.

Demonstration of genetic cohesiveness and ecological distinctness yields a hypothesis of how ecotypes could have evolved and what their key adaptive features may be. This concept predicts that closely related strains can evolve to utilize different resources in the same environment (i.e., niche partitioning) and coexist rather than competitively exclude one another (Cohan, 2001). This may explain the commonly observed phenomenon of multiple clades or biotypes of *G. vaginalis* within a single vaginal community. Although the specific environmental factors that have facilitated their unique adaptations are unknown, delineation of ecotypes can lead to informed predictions of how a particular strain might behave in a particular environment or interact with other species in a community, which for *Gardnerella* may be essential to eliciting host responses and defining pathology (Pybus and Onderdonk, 1997). The ecotype concept can serve as an important basis for making predictions about putative virulence factors influencing the pathogenicity of a species; this could also provide the basis for a nuanced approach to test Koch's postulates. Importantly, variability can still exist within an ecotype for a number of reasons including lateral gene transfer, genetic drift, and geographical separation between subpopulations. One can imagine further subdivisions of ecotypes within ecotypes. As with many aspects of ecology, the appropriate scale on which to characterize diversity is largely dependent on the question being asked.

Recognition of multiple clades as distinct ecological entities can significantly improve our understanding of the role of *G. vaginalis* in health and disease. While others have suggested the designation of new species (Ahmed *et al*., 2012), this may be premature and of limited clinical use, at least until additional studies are performed to demonstrate the clinical significance of differences between clades. Moreover, we think it is critical to recognize and even embrace the within-species diversity of *G. vaginalis* to gain meaningful insight to its ecology. Our study provides preliminary evidence for ecotypes of *G. vaginalis*, but the high degree of diversity even within these groups suggests we have only scratched the surface. Future studies should seek to sample a greater variety of isolates, including those from women with no symptoms of BV and low or intermediate Nugent scores, adolescent girls and postmenopausal women, and women from many geographical regions. To develop further support for 'ecological distinctness' among clades of *G. vaginalis*, studies should measure realized genetic potential with analysis of gene expression, as well as test the effects of interspecies interactions of each ecotype with other bacteria.

4.5 METHODS

4.5.1 *Genome sequences*

We downloaded 35 genome sequences (three complete and 32 draft) of *Gardnerella vaginalis* strains from the PATRIC database (archived at ftp://ftp.patricbrc.org/patric2). Although 36 strains were available at the time of analysis, we excluded strain 6420LIT because it was incompletely sequenced. We downloaded DNA and amino acid sequences of the PATRIC coding DNA sequences (CDS) in FASTA format (file extensions *.PATRIC.ffn and *.PATRIC.faa, respectively, where * represents the strain name), along with relevant tables describing the protein annotations (*.PATRIC.cds.tab, *.PATRIC.features.tab). We also gathered functional annotations directly from PATRIC, including Gene Ontology (GO) function (*.PATRIC.go) and KEGG biochemical pathway assignments (*.PATRIC.path). Counting all 35 genomes, our initial dataset included 44,505 protein CDS, 375 unique GO term annotations covering 14,896 CDS, and 121 unique pathway annotations covering 10,086 CDS. We pieced together available clinical data about the *G. vaginalis* strains from public genome records and relevant publications (Table 4.1). Nugent scores were reported for 26 genomes, but other clinical details were only sparsely available. We designated BV status as 'symptomatic' or 'asymptomatic' if relevant details were available, or

'ambiguous' if samples only had a Nugent score, or the clinical metadata only indicated status as a "patient" with no information about symptoms or antibiotic treatment.

We also selected 20 strains of *Bifidobacterium* spp. (Table C.1), the most closely related genus to *Gardnerella*, to compare genomic differences among the two genera. To narrow our selection, we focused on strains that were isolated from a human, selected complete over draft genome sequences if available, and ignored any strains lacking PATRIC CDS features. The same file types described above were downloaded from PATRIC (ftp://ftp.patricbrc.org/patric2/). Counting all 20 genomes, the initial dataset included 38,276 protein CDS, 537 unique GO term annotations covering 11,999 CDS, and 130 unique pathway annotations covering 8,488 CDS.

4.5.2 *Software and bioinformatic analysis*

We employed several bioinformatic software programs in our computational analyses, including BLASTP (Altschul *et al.*, 1997), BUCKy v1.4.3 (Larget *et al.*, 2010), ClustalW (Larkin *et al.*, 2007), MrBayes v3.2.1 (Ronquist *et al.*, 2012), MUSCLE (Edgar, 2004), OrthoMCL v2.0.9 (Li *et al.*, 2003), and RAxML v8.0.3 (Stamatakis, 2014). Downstream analysis and graphical summarization was performed in R v3.1.0. Data and code to partially reproduce our analysis are found at `http://github.com/roxanahickey/gardnerella`.

4.5.3 *Maximum-likelihood phylogenetic analysis of 16S rRNA genes*

We downloaded DNA sequences of RNA coding genes from the PATRIC database (file extensions *.PATRIC.frn; archived at ftp://ftp.patricbrc.org/patric2/) for all *G. vaginalis* and *Bifidobacterium* genomes. 16S rRNA genes were distinguished by being annotated as 'Small Subunit Ribosomal RNA' (ssuRNA). We considered sequences >1400 bp to be full-length genes and retained them for further analysis; these were available for 18 *G. vaginalis* genomes and all 20 *Bifidobacterium* genomes. Several genomes possessed multiple gene copies, and we removed any within-strain duplicates prior to phylogenetic analysis. We performed multiple sequence alignment in MUSCLE (Edgar, 2004) and computed the maximum-likelihood phylogeny in RAxML (Stamatakis, 2014) under the generalized time-reversible model of substitution rates drawn from a gamma distribution with 1,000 bootstrap replicates.

4.5.4 *Identification of homologous protein families*

In the analysis of 35 *G. vaginalis* genomes, we performed an all-against-all similarity analysis on the amino acid sequences of 44,505 CDS using BLASTP (Altschul *et al.*, 1997) with an E-value cutoff of 1E-10 to construct a database of homologous protein sequences. We then used OrthoMCL to cluster them into protein families specifying a minimum 70% identity threshold and E-value cutoff of 1E-5. We employed the same approach for the analysis of both *G. vaginalis* and *Bifidobacterium* spp. genomes (total of 82,781 CDS among 55 genomes) except that we relaxed the minimum identity threshold to 50% to allow for greater divergence in protein families among genera. Due to differences in the definitions of homologs (i.e., genes evolved from a common ancestral sequence), orthologs (i.e., genes evolved from a common ancestor and separated by speciation) and paralogs (i.e., genes related by duplication within a genome), we agnostically refer to clusters of protein CDS as 'protein families' (or simply genes when we are referring to the DNA sequences). Protein families and singletons were summarized in tables using standard OrthoMCL commands, and data were output as binary values (indicating the presence or absence of each protein family in each genome) and counts (indicating the number of CDS assigned to each protein family in each genome).

4.5.5 *Identification of genome clusters by protein family repertoire*

To assess similarities and differences in genomic composition among strains, we imported the OrthoMCL protein cluster tables into R and computed pairwise differences between genomes using the Bray-Curtis dissimilarity matrix on protein family abundances. To summarize groups of genomes with similar protein repertoires, we performed hierarchical clustering (average linkage method) on the distance matrix, and the clustering result was graphed as a dendrogram. We identified the optimal number of hierarchical clusters according to the maximum silhouette width (Kaufman and Rousseeuw, 2009) and assigned genomes to subgroups by 'cutting' the dendrogram at the specified number of clusters. Thus, genomes within each cluster were deemed more similar to each other in their overall protein repertoire than to genomes outside of that cluster.

4.5.6 *Identification of core, accessory, and unique protein families*

We used the OrthoMCL binary tables to determine which protein families were shared among all genomes (i.e., core proteins), shared among only some genomes (i.e., accessory proteins), or unique to individual genomes. We performed similar analyses on all genomes together as well as for the subgroups of genomes determined as hierarchical genome clusters and phylogenetic clades (described below). We also estimated pangenome size with protein family accumulation curves using the specaccum function in vegan v2.2-1 (Oksanen *et al.*, 2013).

4.5.7 *Identification of clades by phylogenetic concordance analysis of core genes*

We conducted phylogenetic analysis on the single-copy core genes of *G. vaginalis* to assess whether groups of strains were supported as phylogenetic clades. First, we performed multiple sequence alignments on each set of core gene sequences using the transAlign Perl script (Bininda-Emonds, 2005), which facilitates DNA alignment in ClustalW relative to the alignment of translated amino acid sequence. Next, we assessed phylogenetic relationships with Bayesian inference using MrBayes (Ronquist *et al.*, 2012). For each gene, analysis began with random starting trees constructed from the multiple sequence alignment, and posterior probabilities were determined from two independent runs of one million generations of Markov Chain Monte Carlo (MCMC) simulations with the following parameters: nst=6, rates=gamma, ploidy=haploid (i.e., the generalized time-reversible model of substitution rates drawn from a gamma distribution). The program sampled tree topologies every 500 generations, and the first 25% of resulting trees were discarded before summarizing trees and samples of the model substitution parameters.

After generating sets of gene trees for each core protein family, we performed Bayesian concordance analysis (BCA) (Ané *et al.*, 2007) using BUCKy (Larget *et al.*, 2010). BCA incorporates phylogenetic trees from multiple loci (with potentially conflicting topologies) into a single tree with estimates of concordance factors (CFs) representing the proportion of genes for which any given clade (i.e., monophyletic grouping of taxa) is supported. In the first stage of BUCKy, we provided as input the complete tree files generated by MrBayes, ignoring the first 500 trees in each, to obtain the Bayesian posterior probability distributions for individual genes. The second stage implemented MCMC for 100,000 generations to estimate the posterior distribution of gene-

to-tree maps, which were used to estimate CFs and, finally, construct the primary concordance tree using a greedy consensus algorithm.

4.5.8 *Gene set enrichment analysis*

We performed statistical analyses to evaluate protein family and functional category presence, absence, and relative abundance among the subgroups or clades of *G. vaginalis*. To examine over- or underrepresented categories, we calculated odds ratios (OR) and tested their significance using Fisher's exact test. We constructed a two-by-two contingency table for each GO functional category, KEGG pathway, or protein family (as determined by OrthoMCL clustering). Each table included the following parameters: the number of group genomes' protein CDS present in this category (a); the number of group genomes' CDS not in this category (b); the number of other genomes' CDS in this category (c); and the number of other genomes' CDS not in this category (d). We used the odds ratio (defined as ad/bc) to rank the relative overrepresentation (odds ratio > 1) or underrepresentation (odds ratio < 1) of each functional category. Finally, to account for multiple comparisons we adjusted p-values obtained by Fisher's exact test by controlling for the false-discovery rate (Benjamini and Hochberg, 1995) and reported these as q-values.

Bibliography

Ahmed A., Earl J., Retchless A., Hillier S.L., Rabe L.K., Cherpes T.L., Powell E., Janto B., Eutsey R., Hiller N.L., Boissy R., Dahlgren M.E., Hall B.G., Costerton J.W., Post J.C., Hu F.Z., and Ehrlich G.D. 2012. Comparative genomic analyses of 17 clinical isolates of *Gardnerella vaginalis* provide evidence of multiple genetically isolated clades consistent with subspeciation into genovars. Journal of Bacteriology 194:3922–3937.

Alqumber M.A. and Burton J.P. 2008. A species-specific PCR for *Lactobacillus iners* demonstrates a relative specificity of this species for vaginal colonization. Microbial Ecology in Health and Disease 20:135–139.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., and Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389–3402.

Alvarez-Olmos M.I., Barousse M.M., Rajan L., Van Der Pol B.J., Fortenberry D., Orr D., and Fidel Jr. P.L. 2004. Vaginal lactobacilli in adolescents. Sexually Transmitted Diseases 31:393–400.

Amsel R., Totten P.A., Spiegel C.A., Chen K., Eschenbach D., and Holmes K.K. 1983. Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. The American Journal of Medicine 74:14–22.

Ané C., Larget B., Baum D.A., Smith S.D., and Rokas A. 2007. Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution 24:412–426.

Antonio M.A., Hawes S.E., and Hillier S.L. 1999. The identification of vaginal *Lactobacillus* species and the demographic and microbiologic characteristics of women colonized by these species. Journal of Infectious Diseases 180:1950–1956.

Anukam K., Osazuwa E., Ahonkhai I., and Reid G. 2006. *Lactobacillus* vaginal microbiota of women attending a reproductive health care service in Benin city, Nigeria. Sexually Transmitted Diseases 33:59.

Aroutcheva A., Gariti D., Simon M., Shott S., Faro J., Simoes J.A., Gurguis A., and Faro S. 2001a. Defense factors of vaginal lactobacilli. American Journal of Obstetrics and Gynecology 185:375–379.

Aroutcheva A.A., Simoes J.A., Behbakht K., and Faro S. 2001b. *Gardnerella vaginalis* isolated from patients with bacterial vaginosis and from patients with healthy vaginal ecosystems. Clinical Infectious Diseases 33:1022–1027.

Bakken L.R. 1985. Separation and purification of bacteria from soil. Applied and Environmental Microbiology 49:1482–1487.

Balashov S.V., Mordechai E., Adelson M.E., and Gygax S.E. 2014. Identification, quantification and subtyping of *Gardnerella vaginalis* in noncultured clinical vaginal samples by quantitative PCR. Journal of Medical Microbiology 63:162–175.

Ballarini A., Segata N., Huttenhower C., and Jousson O. 2013. Simultaneous quantification of multiple bacteria by the BactoChip microarray designed to target species-specific marker genes. PLoS ONE 8:e55764.

Bates D., Maechler M., Bolker B., and Walker S. 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7.

Beigi R.H., Wiesenfeld H.C., Hillier S.L., Straw T., and Krohn M.A. 2005. Factors associated with absence of H2O2-producing *Lactobacillus* among women with bacterial vaginosis. The Journal of Infectious Diseases 191:924–929.

Benito R., Vazquez J.A., Berron S., Fenoll A., and Saez-Neito J.A. 1986. A modified scheme for biotyping *Gardnerella vaginalis*. Journal of Medical Microbiology 21:357–359.

Benjamini Y. and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57:289–300.

Biavati B., Vescovo M., and Torriani S. 2000. Bifidobacteria: history, ecology, physiology and applications. Annals of Microbiology 50:117–131.

Bik E.M., Eckburg P.B., Gill S.R., Nelson K.E., Purdom E.A., Francois F., Perez-Perez G., Blaser M.J., and Relman D.A. 2006. Molecular analysis of the bacterial microbiota in the human stomach. PNAS 103:732–737.

Bininda-Emonds O.R.P. 2005. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. BMC Bioinformatics 6:156.

Blythe M.J., Fortenberry J.D., and Orr D.P. 2003. Douching behaviors reported by adolescent and young adult women at high risk for sexually transmitted infections. Journal of Pediatric and Adolescent Gynecology 16:95–100.

Borcard D., Gillet F., and Legendre P. 2011. Numerical ecology with R. Springer, New York.

Boskey E.R., Cone R.A., Whaley K.J., and Moench T.R. 2001. Origins of vaginal acidity: high D/L lactate ratio is consistent with bacteria being the primary source. Human Reproduction 16:1809–1813.

Boskey E.R., Telsch K.M., Whaley K.J., Moench T.R., and Cone R.A. 1999. Acid production by vaginal flora *in vitro* is consistent with the rate and extent of vaginal acidification. Infection and Immunity 67:5170–5175.

Bottacini F., Medini D., Pavesi A., Turroni F., Foroni E., Riley D., Giubellini V., Tettelin H., Van Sinderen D., and Ventura M. 2010. Comparative genomics of the genus *Bifidobacterium*. Microbiology 156:3243–3254.

Brabin L., Roberts S.A., Fairbrother E., Mandal D., Higgins S.P., Chandiok S., Wood P., Barnard G., and Kitchener H.C. 2005. Factors affecting vaginal pH levels among female adolescents attending genitourinary medicine clinics. Sexually Transmitted Infections 81:483–487.

Bradshaw C.S., Tabrizi S.N., Fairley C.K., Morton A.N., Rudland E., and Garland S.M. 2006. The association of *Atopobium vaginae* and *Gardnerella vaginalis* with bacterial vaginosis and recurrence after oral metronidazole therapy. The Journal of Infectious Diseases 194:828–836.

Bray J.R. and Curtis J.T. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs 27:325–349.

Briselden A.M. and Hillier S.L. 1990. Longitudinal study of the biotypes of *Gardnerella vaginalis*. Journal of Clinical Microbiology 28:2761–2764.

Briselden A.M., Moncla B.J., Stevens C.E., and Hillier S.L. 1992. Sialidases (neuraminidases) in bacterial vaginosis and bacterial vaginosis-associated microflora. Journal of Clinical Microbiology 30:663–666.

Brook I. 2002. Microbiology and management of polymicrobial female genital tract infections in adolescents. Journal of Pediatric and Adolescent Gynecology 15:217–226.

Brotman R.M., Erbelding E.J., Jamshidi R.M., Klebanoff M.A., Zenilman J.M., and Ghanem K.G. 2007. Findings associated with recurrence of bacterial vaginosis among adolescents attending sexually transmitted diseases clinics. Journal of Pediatric and Adolescent Gynecology 20:225–231.

Brotman R.M., Klebanoff M.A., Nansel T.R., Andrews W.W., Schwebke J.R., Zhang J., Yu K.F., Zenilman J.M., and Scharfstein D.O. 2008. A longitudinal study of vaginal douching and bacterial vaginosis—a marginal structural modeling analysis. American Journal of Epidemiology 168:188–196.

Brown C.J., Wong M., Davis C.C., Kanti A., Zhou X., and Forney L.J. 2007. Preliminary characterization of the normal microbiota of the human vulva using cultivation-independent methods. Journal of Medical Microbiology 56:271–276.

Bump R.C., Sachs L.A., and Buesching W.J. 1986. Sexually transmissible infectious agents in sexually active and virginal asymptomatic adolescent girls. Pediatrics 77:488–494.

Burton J.P., Dixon J.L., and Reid G. 2003. Detection of *Bifidobacterium* species and *Gardnerella vaginalis* in the vagina using PCR and denaturing gradient gel electrophoresis (DGGE). International Journal of Gynecology & Obstetrics 81:61–63.

Cailliez F., Pages J.P., Morlat G., and Amiard J.C. 1976. Introduction à l'analyse des données. Société de mathématiques appliquées et de sciences humaines, Paris, France.

Carvalho B.S. and Irizarry R.A. 2010. A framework for oligonucleotide microarray preprocessing. Bioinformatics 26:2363–2367.

Catlin B.W. 1992. *Gardnerella vaginalis*: characteristics, clinical considerations, and controversies. Clinical Microbiology Reviews 5:213–237.

Cauci S., Driussi S., De Santo D., Penacchioni P., Iannicelli T., Lanzafame P., De Seta F., Quadrifoglio F., de Aloysio D., and Guaschino S. 2002. Prevalence of bacterial vaginosis and vaginal flora changes in peri- and postmenopausal women. Journal of Clinical Microbiology 40:2147–2152.

Cauci S., Monte R., Ropele M., Missero C., Not T., Quadrifoglio F., and Menestrina G. 1993. Pore-forming and haemolytic properties of the *Gardnerella vaginalis* cytoiysin. Molecular Microbiology 9:1143–1155.

Chaijareenont K., Sirimai K., Boriboonhirunsarn D., and Kiriwat O. 2004. Accuracy of Nugent's score and each Amsel's criteria in the diagnosis of bacterial vaginosis. Journal of the Medical Association of Thailand 87:1270–1274.

Chase D.J., Schenkel B.P., Fahr A.M., Eigner U., and Group T.S. 2007. A prospective, randomized, double-blind study of vaginal microflora and epithelium in women using a tampon with an apertured film cover compared with those in women using a commercial tampon with a cover of nonwoven fleece. Journal of Clinical Microbiology 45:1219–1224.

Cohan F.M. 2001. Bacterial species and speciation. Systematic Biology 50:513–524.

Cohan F.M. 2002. What are Bacterial Species? Annual Review of Microbiology 56:457–487.

Cohan F.M. 2006. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. Philosophical Transactions of the Royal Society B: Biological Sciences 361:1985–1996.

Collins M.D. and Wallbanks S. 1992. Comparative sequence analyses of the 16S rRNA genes of *Lactobacillus minutus*, *Lactobacillus rimae* and *Streptococcus parvulus*: proposal for the creation of a new genus *Atopobium*. FEMS Microbiology Letters 74:235–240.

Connon S.A. and Giovannoni S.J. 2002. High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. Applied and Environmental Microbiology 68:3878–3885.

Coolen M.J.L., Post E., Davis C.C., and Forney L.J. 2005. Characterization of microbial communities found in the human vagina by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes. Applied and Environmental Microbiology 71:8729–8737.

Crielaard W., Zaura E., Schuller A.A., Huse S.M., Montijn R.C., and Keijser B.J. 2011. Exploring the oral microbiota of children at various developmental stages of their dentition in the relation to their oral health. BMC Medical Genomics 4:22.

Criswell B.S., Marston J.H., Stenback W.A., Black S.H., and Gardner H.L. 1971. *Haemophilus vaginalis* 594, a Gram-negative organism? Canadian Journal of Microbiology 17:865–869.

Cruciani F., Biagi E., Severgnini M., Consolandi C., Calanni F., Donders G., Brigidi P., and Vitali B. 2015. Development of a microarray-based tool to characterize vaginal bacterial fluctuations: application to a novel antibiotic treatment for bacterial vaginosis. Antimicrobial Agents and Chemotherapy pages AAC.00225–15–34.

Culhane J. 2002. Exposure to chronic stress and ethnic differences in rates of bacterial vaginosis among pregnant women. American Journal of Obstetrics and Gynecology 187:1272–1276.

Dabney A., Storey J.D., and Warnes G.R. 2004. qvalue: Q-value estimation for false discovery rate control. R package version 1.34.0.

De Backer E., Verhelst R., Verstraelen H., Burton J.P., Temmerman M., and Vaneechoutte M. 2007. Quantitative determination by real-time PCR of four vaginal *Lactobacillus* species, *Gardnerella vaginalis* and *Atopobium vaginae* indicates an inverse relationship between *L. gasseri* and *L. iners*. BMC Microbiology 7:115.

De Cáceres M. and Legendre P. 2009. Associations between species and groups of sites: indices and statistical inference. Ecology 90:3566–3574.

De Cáceres M., Legendre P., and Moretti M. 2010. Improving indicator species analysis by combining groups of sites. Oikos 119:1674–1684.

DeAngelis K.M., Wu C.H., Beller H.R., Brodie E.L., Chakraborty R., DeSantis T.Z., Fortney J.L., Hazen T.C., Osman S.R., Singer M.E., Tom L.M., and Andersen G.L. 2011. PCR amplification-independent methods for detection of microbial communities by the high-density microarray PhyloChip. Applied and Environmental Microbiology 77:6313–6322.

Dekio I. 2005. Detection of potentially novel bacterial components of the human skin microbiota using culture-independent molecular profiling. Journal of Medical Microbiology 54:1231–1238.

Didham R.K., Tylianakis J.M., Hutchison M.A., Ewers R.M., and Gemmell N.J. 2005. Are invasive species the drivers of ecological change? Trends in Ecology & Evolution 20:470–474.

Dols J.A.M., Smit P.W., Kort R., Reid G., Schuren F.H.J., Tempelman H., Romke Bontekoe T., Korporaal H., and Boon M.E. 2011. Microarray-based identification of clinically relevant vaginal bacteria in relation to bacterial vaginosis. American Journal of Obstetrics and Gynecology 204:1.e1–1.e7.

Dominguez-Bello M.G., Costello E.K., Contreras M., Magris M., Hidalgo G., Fierer N., and Knight R. 2010. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. PNAS 107:11971–11975.

Donders G.G.G. 2007. Definition and classification of abnormal vaginal flora. Best Practice & Research Clinical Obstetrics & Gynaecology 21:355–373.

Doolittle W.F. and Papke R.T. 2006. Genomics and the bacterial species problem. Genome Biology 7:116.

Du Plessis E.M. and Dicks L.M. 1995. Evaluation of random amplified polymorphic DNA (RAPD)-PCR as a method to differentiate *Lactobacillus acidophilus*, *Lactobacillus crispatus*, *Lactobacillus amylovorus*, *Lactobacillus gallinarum*, *Lactobacillus gasseri*, and *Lactobacillus johnsonii*. Current Microbiology 31:114–118.

Dufrêne M. and Legendre P. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs 67:345–366.

Dykhuizen D.E. and Green L. 1991. Recombination in *Escherichia coli* and the definition of biological species. Journal of Bacteriology 173:7257–7268.

Eckburg P.B. 2005. Diversity of the human intestinal microbial flora. Science 308:1635–1638.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792–1797.

Eitinger T., Rodionov D.A., Grote M., and Schneider E. 2011. Canonical and ECF-type ATP-binding cassette importers in prokaryotes: diversity in modular organization and cellular functions. FEMS Microbiology Reviews 35:3–67.

Elsner P. and Maibach H.I. 1990. Microbiology of specialized skin: the vulva. Seminars in Dermatology 9:300–304.

Embley T.M. and Stackebrandt E. 1994. The molecular phylogeny and systematics of the actinomycetes. Annual Review of Microbiology 48:257–289.

Eren A.M., Zozaya M., Taylor C.M., Dowd S.E., Martin D.H., and Ferris M.J. 2011. Exploring the diversity of *Gardnerella vaginalis* in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation. PLoS ONE 6:e26732.

Eschenbach D.A., Davick P.R., Williams B.L., Klebanoff S.J., Young-Smith K., Critchlow C.M., and Holmes K.K. 1989. Prevalence of hydrogen peroxide-producing *Lactobacillus* species in normal women and women with bacterial vaginosis. Journal of Clinical Microbiology 27:251–256.

Eschenbach D.A., Thwin S.S., Patton D.L., Hooton T.M., Stapleton A.E., Agnew K., Winter C., Meier A., and Stamm W.E. 2000. Influence of the normal menstrual cycle on vaginal tissue, discharge, and microflora. Clinical Infectious Diseases 30:901–907.

Falcon S., Carvalho B., Carey V., Settles M., and de Beuf K. 2009. pdInfoBuilder: platform design information package builder. R package version 1.24.0.

Falsen E., Pascual C., Sjödén B., Ohlén M., and Collins M.D. 1999. Phenotypic and phylogenetic characterization of a novel *Lactobacillus* species from human sources: description of *Lactobacillus iners* sp. nov. International Journal of Systematic Bacteriology 49:217–221.

Farage M. and Maibach H. 2006. Lifetime changes in the vulva and vagina. Archives of Gynecology and Obstetrics 273:195–202.

Ferris M.J., Norori J., Zozaya-Hinchliffe M., and Martin D.H. 2007. Cultivation-independent analysis of changes in bacterial vaginosis flora following metronidazole treatment. Journal of Clinical Microbiology 45:1016–1018.

Fethers K., Twin J., Fairley C.K., Fowkes F.J.I., Garland S.M., Fehler G., Morton A.M., Hocking J.S., Tabrizi S.N., and Bradshaw C.S. 2012. Bacterial vaginosis (BV) candidate bacteria: associations with bv and behavioural practices in sexually-experienced and inexperienced women. PLoS ONE 7:e30633.

Forney L.J., Gajer P., Williams C.J., Schneider G.M., Koenig S.S.K., McCulle S.L., Karlebach S., Brotman R.M., Davis C.C., Ault K., and Ravel J. 2010. Comparison of self-collected and physician-collected vaginal swabs for microbiome analysis. Journal of Clinical Microbiology 48:1741–1748.

Frank J.A., Reich C.I., Sharma S., Weisbaum J.S., Wilson B.A., and Olsen G.J. 2008. Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. Applied and Environmental Microbiology 74:2461–2470.

Fredricks D.N., Fiedler T.L., and Marrazzo J.M. 2005. Molecular identification of bacteria associated with bacterial vaginosis. The New England Journal of Medicine 353:1899–1911.

Gajer P., Brotman R.M., Bai G., Sakamoto J., Schutte U.M.E., Zhong X., Koenig S.S.K., Fu L., Ma Z., Zhou X., Abdo Z., Forney L.J., and Ravel J. 2012. Temporal dynamics of the human vaginal microbiota. Science Translational Medicine 4:132ra52.

Gardner H.L. and Dukes C.D. 1955. *Haemophilus vaginalis* vaginitis: a newly defined specific infection previously classified non-specific vaginitis. American Journal of Obstetrics and Gynecology 69:962–976.

Gavini F., Van Esbroeck M., Touzel J.P., and Fourment A. 1996. Detection of fructose-6-phosphate phosphoketolase (F6PPK), a key enzyme of the bifid-shunt, in *Gardnerella vaginalis*. Anaerobe 2:191–193.

Gelber S.E., Aguilar J.L., Lewis K.L.T., and Ratner A.J. 2008. Functional and phylogenetic characterization of vaginolysin, the human-specific cytolysin from *Gardnerella vaginalis*. Journal of Bacteriology 190:3896–3903.

Gentry T.J., Wickham G.S., Schadt C.W., He Z., and Zhou J. 2006. Microarray applications in microbial ecology research. Microbial Ecology 52:159–175.

Gerstner G.J., Grünberger W., Boschitsch E., and Rotter M. 1982. Vaginal organisms in prepubertal children with and without vulvovaginitis. Archives of Gynecology 231:247–252.

Gevers D., Cohan F.M., Lawrence J.G., Spratt B.G., Coenye T., Feil E.J., Stackebrandt E., Van de Peer Y., Vandamme P., Thompson F.L., and Swings J. 2005. Opinion: Re-evaluating prokaryotic species. Nature Reviews Microbiology 3:733–739.

Gilbert J.A. and Dupont C.L. 2011. Microbial metagenomics: beyond the genome. Annual Review of Marine Science 3:347–371.

Gilbert N.M., Lewis W.G., and Lewis A.L. 2013. Clinical features of bacterial vaginosis in a murine model of vaginal infection with *Gardnerella vaginalis*. PLoS ONE 8:e59539.

Ginkel P.D., Soper D.E., Bump R.C., and Dalton H.P. 1993. Vaginal flora in postmenopausal women: the effect of estrogen replacement. Infectious Diseases in Obstetrics and Gynecology 1:94–97.

Goldenberg R.L., Klebanoff M.A., Nugent R., Krohn M.A., Hillier S., and Andrews W.W. 1996. Bacterial colonization of the vagina during pregnancy in four ethnic groups. Vaginal Infections and Prematurity Study Group. American Journal of Obstetrics and Gynecology 174:1618–1621.

Gower J.C. 1983. Comparing Classifications. *In* Numerical Taxonomy, pages 137–155, Springer-Verlag, Berlin/Heidelberg, Germany.

Greenwood J.R. and Pickett M.J. 1980. Transfer of *Haemophilus vaginalis* Gardner and Dukes to a new genus, *Gardnerella*: *G. vaginalis* (Gardner and Dukes) comb. nov. International Journal of Systematic and Evolutionary Microbiology 30:170–178.

Gunderson L. 2000. Ecological resilience—in theory and application. Annual review of ecology and systematics 31:425–439.

Gupta K., Hillier S.L., Hooton T.M., Roberts P.L., and Stamm W.E. 2000. Effects of contraceptive method on the vaginal microbial flora: a prospective evaluation. The Journal of Infectious Diseases 181:595–601.

Haggerty C.L., Hillier S.L., Bass D.C., Ness R.B., and PID Evaluation and Clinical Health study investigators. 2004. Bacterial vaginosis and anaerobic bacteria are associated with endometritis. Clinical Infectious Diseases 39:990–995.

Hammerschlag M.R., Alpert S., Onderdonk A.B., Thurston P., Drude E., McCormack W.M., and Bartlett J.G. 1978a. Anaerobic microflora of the vagina in children. American Journal of Obstetrics and Gynecology 131:853–856.

Hammerschlag M.R., Alpert S., Rosner I., Thurston P., Semine D., McComb D., and McCormack W.M. 1978b. Microbiology of the vagina in children: normal and potentially pathogenic organisms. Pediatrics 62:57–62.

Harwich M.D., Alves J.M., Buck G.A., Strauss J.F., Patterson J.L., Oki A.T., Girerd P.H., and Jefferson K.K. 2010. Drawing the line between commensal and pathogenic *Gardnerella vaginalis* through genome analysis and virulence studies. BMC Genomics 11:375.

Hawes S.E., Hillier S.L., Benedetti J., Stevens C.E., Koutsky L.A., Wolner-Hanssen P., and Holmes K.K. 1996. Hydrogen peroxide-producing lactobacilli and acquisition of vaginal infections. The Journal of Infectious Diseases 174:1058–1063.

He Z., Gentry T.J., Schadt C.W., Wu L., Liebich J., Chong S.C., Huang Z., Wu W., Gu B., Jardine P., Criddle C., and Zhou J. 2007. GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. ISME Journal 1:67–77.

Hickey R.J. and Forney L.J. 2014. *Gardnerella vaginalis* does not always cause bacterial vaginosis. The Journal of Infectious Diseases 210:1682–1683.

Hickey R.J., Zhou X., Pierson J.D., Ravel J., and Forney L.J. 2012. Understanding vaginal microbiome complexity from an ecological perspective. Translational Research 160:267–282.

Hickey R.J., Zhou X., Settles M.L., Erb J., Malone K., Hansmann M.A., Shew M.L., Van Der Pol B., Fortenberry J.D., and Forney L.J. 2015. Vaginal microbiota of adolescent girls prior to the onset of menarche resemble those of reproductive-age women. mBio 6:e00097–15–14.

Hill G.B., St Claire K.K., and Gutman L.T. 1995. Anaerobes predominate among the vaginal microflora of prepubertal girls. Clinical Infectious Diseases 20 Suppl 2:S269–70.

Hill J.E., Goh S.H., Money D.M., Doyle M., Li A., Crosby W.L., Links M., Leung A., Chan D., and Hemmingsen S.M. 2005. Characterization of vaginal microflora of healthy, nonpregnant women by chaperonin-60 sequence-based methods. American Journal of Obstetrics and Gynecology 193:682–692.

Hillier S.L., Nugent R.P., Eschenbach D.A., Krohn M.A., Gibbs R.S., Martin D.H., Cotch M.F., Edelman R., Pastorek J.G., and Rao A.V. 1995. Association between bacterial vaginosis and preterm delivery of a low-birth-weight infant. New England Journal of Medicine 333:1737–1742.

Hobbs R. and Huenneke L. 1992. Disturbance, diversity, and invasion: implications for conservation. Conservation Biology 6:324–337.

Hou D., Zhou X., Zhong X., Settles M.L., Herring J., Wang L., Abdo Z., Forney L.J., and Xu C. 2013. Microbiota of the seminal fluid from healthy and infertile men. Fertility and Sterility 100:1261–1269.

Hugenholtz P., Goebel B.M., and Pace N.R. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. Journal of Bacteriology 180:4765–4774.

Hyman R.W., Fukushima M., Diamond L., Kumm J., Giudice L.C., and Davis R.W. 2005. Microbes on the human vaginal epithelium. PNAS 102:7952–7957.

Ingianni A., Petruzzelli S., Morandotti G., and Pompei R. 1997. Genotypic differentiation of *Gardnerella vaginalis* by amplified ribosomal DNA restriction analysis (ARDRA). FEMS Immunology and Medical Microbiology 18:61–66.

Irizarry R.A., Hobbs B., Collin F., Beazer Barclay Y.D., Antonellis K.J., Scherf U., and Speed T.P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264.

Jain A.K. and Dubes R.C. 1988. Algorithms for Clustering Data. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Jaksić F. 1981. Abuse and misuse of the term “guild" in ecological studies. Oikos 37:397.

Janda J.M. and Abbott S.L. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of Clinical Microbiology 45:2761–2764.

Jayaprakash T.P., Schellenberg J.J., and Hill J.E. 2012. Resolution and characterization of distinct cpn60-based subgroups of *Gardnerella vaginalis* in the vaginal microbiota. PLoS ONE 7:e43009.

Johnson J., Phelps C., Cummins C., London J., and Gasser F. 1980. Taxonomy of the *Lactobacillus acidophilus* group. International Journal of Systematic and Evolutionary Microbiology 30:53–68.

Joishy M., Ashtekar C.S., Jain A., and Gonsalves R. 2005. Do we need to treat vulvovaginitis in prepubertal girls? BMJ 330:186–188.

Jones F.R., Miller G., Gadea N., Meza R., Leon S., Perez J., Lescano A.G., Pajuelo J., Caceres C.F., Klausner J.D., Coates T.J., and NIMH Collaborative HIV/STI Prevention Trial Group. 2007. Prevalence of bacterial vaginosis among young women in low-income populations of coastal Peru. International Journal of STD & AIDS 18:188–192.

Kang S., Denman S.E., Morrison M., Yu Z., Dore J., Leclerc M., and McSweeney C.S. 2010. Dysbiosis of fecal microbiota in Crohn’s disease patients as revealed by a custom phylogenetic microarray. Inflammatory Bowel Diseases 16:2034–2042.

Karlsson C.L.J., Molin G., Cilio C.M., and Ahrné S. 2011. The pioneer gut microbiota in human neonates vaginally born at term-a pilot study. Pediatric Research 70:282–286.

Kaufman L. and Rousseeuw P.J. 2009. Finding groups in data: an introduction to cluster analysis, vol. 344. John Wiley & Sons, Hoboken, NJ, USA.

Klebanoff M.A., Andrews W.W., Zhang J., Brotman R.M., Nansel T.R., Yu K.F., and Schwebke J.R. 2010. Race of male sex partners and occurrence of bacterial vaginosis. Sexually Transmitted Diseases 37:184–190.

Koenig J.E., Spor A., Scalfone N., Fricker A.D., Stombaugh J., Knight R., Angenent L.T., and Ley R.E. 2011. Succession of microbial consortia in the developing infant gut microbiome. PNAS 108 Suppl 1:4578–4585.

Konstantinidis K.T., Ramette A., and Tiedje J.M. 2006. The bacterial species definition in the genomic era. Philosophical Transactions of the Royal Society B: Biological Sciences 361:1929–1940.

Koumans E.H., Sternberg M., Bruce C., McQuillan G., Kendrick J., Sutton M., and Markowitz L.E. 2007. The prevalence of bacterial vaginosis in the united states, 2001–2004; associations with symptoms, sexual behaviors, and reproductive health. Sexually Transmitted Diseases 34:864–869.

Kuznetsova A., Brockhoff P.B., and Christensen R. 2014. lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package version 2.0-11.

Lambert J.A., John S., Sobel J.D., and Akins R.A. 2013. Longitudinal analysis of vaginal microbiome dynamics in women with recurrent bacterial vaginosis: recognition of the conversion process. PLoS ONE 8:e82599.

Lan R. and Reeves P.R. 2000. Intraspecies variation in bacterial genomes: the need for a species genome concept. Trends in Microbiology 8:396–401.

Langfelder P. and Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.

Langmead B., Trapnell C., Pop M., and Salzberg S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25.

Larget B.R., Kotha S.K., Dewey C.N., and Ané C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26:2910–2911.

Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J., and Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.

Larsen B., Goplerud C.P., Petzold C.R., Ohm-Smith M.J., and Galask R.P. 1982. Effect of estrogen treatment on the genital tract flora of postmenopausal women. Obstetrics and Gynecology 60:20–24.

Larsen B. and Monif G.R. 2001. Understanding the bacterial flora of the female genital tract. Clinical Infectious Diseases 32:e69–77.

Lauer E., Helming C., and Kandler O. 1980. Heterogeneity of the species *Lactobacillus acidophilus* (Moro) Hansen and Moquot as revealed by biochemical characteristics and DNA-DNA hybridisation. Zentralblatt für Bakteriologie: I. Abt. Originale C: Allgemeine, angewandte und ökologische Mikrobiologie 1:150–168.

Leblond-Bourget N., Philippe H., Mangin I., and Decaris B. 1996. 16S rRNA and 16S to 23S internal transcribed spacer sequence analyses reveal inter- and intraspecific *Bifidobacterium* phylogeny. International Journal of Systematic Bacteriology 46:102–111.

Lee Y.J., Van Nostrand J.D., Tu Q., Lu Z., Cheng L., Yuan T., Deng Y., Carter M.Q., He Z., Wu L., Yang F., Xu J., and Zhou J. 2013. The PathoChip, a functional gene array for assessing pathogenic properties of diverse microbial communities. ISME Journal 7:1974–1984.

Legendre P. and Gallagher E. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129:271–280.

Legendre P. and Legendre L. 2012. Cluster analysis. *In* Numerical Ecology, pages 337–424, Elsevier, Amsterdam, Netherlands.

Leitich H., Bodneradler B., BrunBauer M., Kaider A., Egarter C., and Husslein P. 2003. Bacterial vaginosis as a risk factor for preterm delivery: A meta-analysis. American Journal of Obstetrics and Gynecology 189:139–147.

Levison M.E., Corman L.C., Carrington E.R., and Kaye D. 1977. Quantitative microflora of the vagina. American Journal of Obstetrics and Gynecology 127:80–85.

Li L., Stoeckert C.J., and Roos D.S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 13:2178–2189.

Li W. and Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

Linhares I.M., Giraldo P.C., and Baracat E.C. 2010a. New findings about vaginal bacterial flora. Rev Assoc Med Bras 56:370–374.

Linhares I.M., Summers P.R., Larsen B., Giraldo P.C., and Witkin S.S. 2010b. Contemporary perspectives on vaginal pH and lactobacilli. American Journal of Obstetrics and Gynecology 203:1.e1–1.e5.

Liu C.M., Aziz M., Kachur S., Hsueh P.R., Huang Y.T., Keim P., and Price L.B. 2012. BactQuant: An enhanced broad-coverage bacterial quantitative real-time PCR assay. BMC Microbiology 12:1–1.

Lopes dos Santos Santiago G., Cools P., Verstraelen H., Trog M., Missine G., Aila N.E., Verhelst R., Tency I., Claeys G., Temmerman M., and Vaneechoutte M. 2011. Longitudinal study of the dynamics of vaginal microflora during two consecutive menstrual cycles. PLoS ONE 6:e28180.

Lopes dos Santos Santiago G., Tency I., Verstraelen H., Verhelst R., Trog M., Temmerman M., Vancoillie L., Decat E., Cools P., and Vaneechoutte M. 2012. Longitudinal qPCR study of the dynamics of *L. crispatus, L. iners, A. vaginae*, (sialidase positive) *G. vaginalis*, and *P. bivia* in the vagina. PLoS ONE 7:e45281.

Lukjancenko O., Ussery D.W., and Wassenaar T.M. 2011. Comparative genomics of *Bifidobacterium*, *Lactobacillus* and related probiotic genera. Microbial Ecology 63:651–673.

Marrazzo J.M., Koutsky L.A., Eschenbach D.A., Agnew K., Stine K., and Hillier S.L. 2002. Characterization of vaginal flora and bacterial vaginosis in women who have sex with women. The Journal of Infectious Diseases 185:1307–1313.

Marrazzo J.M., Thomas K.K., Fiedler T.L., Ringwood K., and Fredricks D.N. 2008. Relationship of specific vaginal bacteria and bacterial vaginosis treatment failure in women who have sex with women. Annals of Internal Medicine 149:20–28.

Marshall W.A. and Tanner J.M. 1969. Variations in pattern of pubertal changes in girls. Archives of Disease in Childhood 44:291–303.

Martin H.L., Richardson B.A., Nyange P.M., Lavreys L., Hillier S.L., Chohan B., Mandaliya K., Ndinya-Achola J.O., Bwayo J., and Kreiss J. 1999. Vaginal lactobacilli, microbial flora, and risk of human immunodeficiency virus type 1 and sexually transmitted disease acquisition. Journal of Infectious Diseases 180:1863–1868.

McCann K.S. 2000. The diversity-stability debate. Nature 405:228–233.

McCormack W.M., Hayes C.H., Rosner B., Evrard J.R., Crockett V.A., Alpert S., and Zinner S.H. 1977. Vaginal colonization with *Corynebacterium vaginale* (*Haemophilus vaginalis*). Journal of Infectious Diseases 136:740–745.

Mendes-Soares H., Suzuki H., Hickey R.J., and Forney L.J. 2014. Comparative functional genomics of *Lactobacillus* spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. Journal of Bacteriology 196:1458–1470.

Merchant J.S., Oh K., and Klerman L.V. 1999. Douching: a problem for adolescent girls and young women. Archives of Pediatrics & Adolescent Medicine 153:834–837.

Mirmonsef P., Gilbert D., Zariffard M.R., Hamaker B.R., Kaur A., Landay A.L., and Spear G.T. 2011. The effects of commensal bacteria on innate immune responses in the female genital tract. American Journal of Reproductive Immunology 65:190–195.

Moncla B.J., Chappell C.A., Mahal L.K., Debo B.M., Meyn L.A., and Hillier S.L. 2015. Impact of bacterial vaginosis, as assessed by Nugent criteria and hormonal status on glycosidases and lectin binding in cervicovaginal lavage samples. PLoS ONE 10:e0127091–11.

Moncla B.J. and Pryke K.M. 2009. Oleate lipase activity in *Gardnerella vaginalis* and reconsideration of existing biotype schemes. BMC Microbiology 9:78.

Morris M., Nicoll A., Simms I., Wilson J., and Catchpole M. 2001. Bacterial vaginosis: a public health review. British Journal of Obstetrics and Gynaecology 108:439–450.

Mutschler H., Gebhardt M., Shoeman R.L., and Meinhart A. 2011. A novel mechanism of programmed cell death in bacteria by toxin–antitoxin systems corrupts peptidoglycan synthesis. PLoS Biology 9:e1001033–12.

Muzny C.A. and Schwebke J.R. 2013. *Gardnerella vaginalis*: still a prime suspect in the pathogenesis of bacterial vaginosis. Current Infectious Disease Reports 15:130–135.

Myhre A.K., Bevanger L.S., Berntzen K., and Bratlid D. 2002. Anogenital bacteriology in non-abused preschool children: a descriptive study of the aerobic genital flora and the isolation of anogenital *Gardnerella vaginalis*. Acta Paediatrica 91:885–891.

Myhre A.K., Myklestad K., and Adams J.A. 2010. Changes in genital anatomy and microbiology in girls between age 6 and age 12 years: a longitudinal study. Journal of Pediatric and Adolescent Gynecology 23:77–85.

Nam H., Whang K., and Lee Y. 2007. Analysis of vaginal lactic acid producing bacteria in healthy women. Journal of Microbiology 45:515–520.

Nansel T.R., Riggs M.A., Yu K.F., Andrews W.W., Schwebke J.R., and Klebanoff M.A. 2006. The association of psychosocial stress and bacterial vaginosis in a longitudinal cohort. American Journal of Obstetrics and Gynecology 194:381–386.

Newton E.R., Piper J.M., Shain R.N., Perdue S.T., and Peairs W. 2001. Predictors of the vaginal microflora. American Journal of Obstetrics and Gynecology 184:845–853.

Nugent R.P., Krohn M.A., and Hillier S.L. 1991. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of Gram stain interpretation. Journal of Clinical Microbiology 29:297–301.

Numanović F., Hukić M., Nurkić M., Gegić M., Delibegović Z., Imamović A., and Pasić S. 2008. Importance of isolation and biotypization of *Gardnerella vaginalis* in diagnosis of bacterial vaginosis. Bosnian Journal of Basic Medical Sciences 8:270–276.

O'Hanlon D.E., Lanier B.R., Moench T.R., and Cone R.A. 2010. Cervicovaginal fluid and semen block the microbicidal activity of hydrogen peroxide produced by vaginal lactobacilli. BMC Infectious Diseases 10:120.

O'Hanlon D.E., Moench T.R., and Cone R.A. 2011. In vaginal fluid, bacteria associated with bacterial vaginosis can be suppressed with lactic acid but not hydrogen peroxide. BMC Infectious Diseases 11:200.

Oksanen J., Blanchet F.G., Kindt R., and Legendre P. 2013. Vegan: community ecology package version 2.0-2. R package version 2.0-10.

Onderdonk A.B., Zamarchi G.R., Rodriguez M.L., Hirsch M.L., Muñoz A., and Kass E.H. 1987. Quantitative assessment of vaginal microflora during use of tampons of various compositions. Applied and Environmental Microbiology 53:2774–2778.

Ott M.A., Ofner S., and Fortenberry J.D. 2009. Beyond douching: use of feminine hygiene products and STI risk among young women. The Journal of Sexual Medicine 6:1335–1340.

Pages H., Aboyoun R., Gentleman R., and DebRoy S. 2013. Biostrings: string objects representing biological sequences and matching algorithms. R package version 2.28.0.

Paliy O. and Agans R. 2012. Application of phylogenetic microarrays to interrogation of human microbiota. FEMS Microbiology Ecology 79:2–11.

Palmer C., Bik E.M., DiGiulio D.B., Relman D.A., and Brown P.O. 2007. Development of the human infant intestinal microbiota. PLoS Biology 5:e177.

Park J., Kerner A., Burns M.A., and Lin X.N. 2011. Microdroplet-enabled highly parallel co-cultivation of microbial communities. PLoS ONE 6:e17019.

Patterson J.L., Stull-Lane A., Girerd P.H., and Jefferson K.K. 2010. Analysis of adherence, biofilm formation and cytotoxicity suggests a greater virulence potential of *Gardnerella vaginalis* relative to other bacterial-vaginosis-associated anaerobes. Microbiology 156:392–399.

Pimm S. 1984. The complexity and stability of ecosystems. Nature 307:321–326.

Piot P., Van Dyck E., Goodfellow M., and Falkow S. 1980. A taxonomic study of *Gardnerella vaginalis* (*Haemophilus vaginalis*) Gardner and Dukes 1955. Journal of General Microbiology 119:373–396.

Piot P., Van Dyck E., Peeters M., Hale J., Totten P.A., and Holmes K.K. 1984. Biotypes of *Gardnerella vaginalis*. Journal of Clinical Microbiology 20:677–679.

Pleckaityte M., Janulaitiene M., Lasickiene R., and Zvirbliene A. 2012. Genetic and biochemical diversity of *Gardnerella vaginalis* strains isolated from women with bacterial vaginosis. FEMS Immunology and Medical Microbiology 65:69–77.

Priestley C.J., Jones B.M., Dhar J., and Goodwin L. 1997. What is normal vaginal flora? Genitourinary Medicine 73:23–28.

Pybus V. and Onderdonk A.B. 1997. Evidence for a commensal, symbiotic relationship between *Gardnerella vaginalis* and *Prevotella bivia* involving ammonia: potential significance for bacterial vaginosis. The Journal of Infectious Diseases 175:406–413.

Qin J., Li R., Raes J., Arumugam M., Burgdorf K.S., Manichanh C., Nielsen T., Pons N., Levenez F., Yamada T., Mende D.R., Li J., Xu J., Li S., Li D., Cao J., Wang B., Liang H., Zheng H., Xie Y., Tap J., Lepage P., Bertalan M., Batto J.M., Hansen T., Le Paslier D., Linneberg A., Nielsen H.B., Pelletier E., Renault P., Sicheritz-Ponten T., Turner K., Zhu H., Yu C., Li S., Jian M., Zhou Y., Li Y., Zhang X., Li S., Qin N., Yang H., Wang J., Brunak S., Dore J., Guarner F., Kristiansen K., Pedersen O., Parkhill J., Weissenbach J., MetaHIT Consortium, Bork P., Ehrlich S.D., and Wang J. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65.

Rajilić-Stojanović M., Heilig H.G.H.J., Molenaar D., Kajander K., Surakka A., Smidt H., and De Vos W.M. 2009. Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environmental Microbiology 11:1736–1751.

Ralph S.G., Rutherford A.J., and Wilson J.D. 1999. Influence of bacterial vaginosis on conception and miscarriage in the first trimester: cohort study. BMJ 319:220–223.

Randjelović G., Kocić B., Stojanović M., and Mišić M. 2005. Bacteriological findings of the vulvar swab specimens from girls with vulvovaginitis. Facta Universitatis: Medicine and Biology 12:159–163.

Rao C.R. 1995. A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. Qüestiió 19:23–63.

Ravel J., Brotman R.M., Gajer P., Ma B., Nandy M., Fadrosh D.W., Sakamoto J., Koenig S.S., Fu L., Zhou X., Hickey R.J., Schwebke J.R., and Forney L.J. 2013. Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. Microbiome 1:29.

Ravel J., Gajer P., Abdo Z., Schneider G.M., Koenig S.S.K., McCulle S.L., Karlebach S., Gorle R., Russell J., Tacket C.O., Brotman R.M., Davis C.C., Ault K., Peralta L., and Forney L.J. 2011. Vaginal microbiome of reproductive-age women. PNAS 108 Suppl 1:4680–4687.

Redondo-Lopez V., Cook R.L., and Sobel J.D. 1990. Emerging role of lactobacilli in the control and maintenance of the vaginal bacterial microflora. Reviews of Infectious Diseases 12:856–872.

Reid G., McGroarty J.A., Tomeczek L., and Bruce A.W. 1996. Identification and plasmid profiles of *Lactobacillus* species from the vagina of 100 healthy women. FEMS Immunology and Medical Microbiology 15:23–26.

Reyn A., Birch-Andersen A., and Lapage S.P. 1966. An electron microscope study of thin sections of *Haemophilus vaginalis* (Gardner and Dukes) and some possibly related species. Canadian Journal of Microbiology 12:1125–1136.

Riley M.A. and Lizotte-Waniewski M. 2009. Population genomics and the bacterial species concept. *In* Horizontal gene transfer: genomes in flux, pages 367–377, Humana Press, Totowa, NJ.

Robbins C.L., Fortenberry J.D., Roth A.M., and Ott M.A. 2012. Premenarchal girls' genital examination experiences. The Journal of Adolescent Health 51:179–183.

Roche NimbleGen. 2011. NimbleGen Arrays User's Guide: CGH and CNV Arrays (version 8.1).

Rodionov D.A., Hebbeln P., Eudes A., ter Beek J., Rodionova I.A., Erkens G.B., Slotboom D.J., Gelfand M.S., Osterman A.L., Hanson A.D., and Eitinger T. 2009. A novel class of modular transporters for vitamins in prokaryotes. Journal of Bacteriology 191:42–51.

Rogosa M. and Sharpe M.E. 1960. Species differentiation of human vaginal lactobacilli. Journal of General Microbiology 23:197–201.

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., and Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539–542.

Rosenstein I.J., Fontaine E.A., Morgan D.J., Sheehan M., Lamont R.F., and Taylor-Robinson D. 1997. Relationship between hydrogen peroxide-producing strains of lactobacilli and vaginosis-associated bacterial species in pregnant women. European Journal of Clinical Microbiology & Infectious Diseases 16:517–522.

Santiago G.L.D.S., Deschaght P., El Aila N., Kiama T.N., Verstraelen H., Jefferson K.K., Temmerman M., and Vaneechoutte M. 2011. *Gardnerella vaginalis* comprises three distinct genotypes of which only two produce sialidase. American Journal of Obstetrics and Gynecology 204:450.e1–7.

Sautter R.L. and Brown W.J. 1980. Sequential vaginal cultures from normal young women. Journal of Clinical Microbiology 11:479–484.

Schellenberg J., Ball T.B., Lane M., Cheang M., and Plummer F. 2008. Flow cytometric quantification of bacteria in vaginal swab samples self-collected by adolescents attending a gynecology clinic. Journal of Microbiological Methods 73:216–226.

Schellenberg J., Links M.G., Hill J.E., Dumonceaux T.J., Peters G.A., Tyler S., Ball T.B., Severini A., and Plummer F.A. 2009. Pyrosequencing of the chaperonin-60 universal target as a tool for determining microbial community composition. Applied and Environmental Microbiology 75:2889–2898.

Schleifer K.H. and Ludwig W. 1995. Phylogeny of the genus *Lactobacillus* and related genera. Systematic and Applied Microbiology 18:461–467.

Schloss P.D., Westcott S.L., Ryabin T., Hall J.R., Hartmann M., Hollister E.B., Lesniewski R.A., Oakley B.B., Parks D.H., Robinson C.J., Sahl J.W., Stres B., Thallinger G.G., Van Horn D.J., and Weber C.F. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75:7537–7541.

Schmid G., Markowitz L., Joesoef R., and Koumans E. 2000. Bacterial vaginosis and HIV infection. Sexually Transmitted Infections 76:3–4.

Schwebke J. 2000. Asymptomatic bacterial vaginosis: Response to therapy. American Journal of Obstetrics and Gynecology 183:1434–1439.

Schwebke J.R., Desmond R.A., and Oh M.K. 2004. Predictors of bacterial vaginosis in adolescent women who douche. Sexually Transmitted Diseases 31:433–436.

Schwebke J.R., Flynn M.S., and Rivers C.A. 2014a. Prevalence of *Gardnerella vaginalis* among women with *Lactobacillus*-predominant vaginal flora. Sexually Transmitted Infections 90:61–63.

Schwebke J.R., Muzny C.A., and Josey W.E. 2014b. Role of *Gardnerella vaginalis* in the pathogenesis of bacterial vaginosis: a conceptual model. The Journal of Infectious Diseases 210:338–343.

Schwebke J.R., Richey C.M., and Weiss H.L. 1999. Correlation of behaviors with microbiological changes in vaginal flora. The Journal of Infectious Diseases 180:1632–1636.

Sha B.E., Chen H.Y., Wang Q.J., Zariffard M.R., Cohen M.H., and Spear G.T. 2005. Utility of Amsel Criteria, Nugent Score, and quantitative PCR for *Gardnerella vaginalis, Mycoplasma hominis*, and *Lactobacillus* spp. for diagnosis of bacterial vaginosis in human immunodeficiency virus-infected women. Journal of Clinical Microbiology 43:4607–4612.

Shafer M.A., Sweet R.L., Ohm-Smith M.J., Shalwitz J., Beck A., and Schachter J. 1985. Microbiology of the lower genital tract in postmenarchal adolescent girls: differences by sexual activity, contraception, and presence of nonspecific vaginitis. The Journal of Pediatrics 107:974–981.

Shi Y., Chen L., Tong J., and Xu C. 2009. Preliminary characterization of vaginal microbiota in healthy Chinese women using cultivation-independent methods. The Journal of Obstetrics and Gynaecology Research 35:525–532.

Shiraishi T., Fukuda K., Morotomi N., Imamura Y., Mishima J., Imai S., Miyazawa K., and Taniguchi H. 2011. Influence of menstruation on the microbiota of healthy women's labia minora as analyzed using a 16S rRNA gene-based clone library method. Japanese Journal of Infectious Diseases 64:76–80.

Simpson T., Merchant J., Grimley D.M., and Oh M.K. 2004. Vaginal douching among adolescent and young women: more challenges than progress. Journal of Pediatric and Adolescent Gynecology 17:249–255.

Smith C.B., Noble V., Bensch R., Ahlin P.A., Jacobson J.A., and Latham R.H. 1982. Bacterial flora of the vagina during the menstrual cycle: findings in users of tampons, napkins, and sea sponges. Annals of Internal Medicine 96:948–951.

Smyth G.K. 2005. Limma: linear models for microarray data. Statistics for Biology and Health, Springer New York, New York, NY.

Sobel J.D. 2000. Bacterial vaginosis. Annual Review of Medicine 51:349–356.

Srinivasan S. and Fredricks D.N. 2008. The human vaginal bacterial biota and bacterial vaginosis. Interdisciplinary Perspectives on Infectious Diseases 2008:750479.

Srinivasan S., Hoffman N.G., Morgan M.T., Matsen F.A., Fiedler T.L., Hall R.W., Ross F.J., McCoy C.O., Bumgarner R., Marrazzo J.M., and Fredricks D.N. 2012. Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. PLoS ONE 7:e37818.

Srinivasan S., Liu C., Mitchell C.M., Fiedler T.L., Thomas K.K., Agnew K.J., Marrazzo J.M., and Fredricks D.N. 2010. Temporal variability of human vaginal bacteria and relationship with bacterial vaginosis. PLoS ONE 5:e10197.

Srinivasan S., Morgan M.T., Liu C., Matsen F.A., Hoffman N.G., Fiedler T.L., Agnew K.J., Marrazzo J.M., and Fredricks D.N. 2013. More than meets the eye: associations of vaginal bacteria with Gram stain morphotypes using molecular phylogenetic analysis. PLoS ONE 8:e78633.

Stackebrandt E., Frederiksen W., Garrity G.M., Grimont P.A.D., Kämpfer P., Maiden M.C.J., Nesme X., Rosselló-Móra R., Swings J., Trüper H.G., Vauterin L., Ward A.C., and Whitman W.B. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. International Journal of Systematic and Evolutionary Microbiology 52:1043–1047.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stingl U., Tripp H.J., and Giovannoni S.J. 2007. Improvements of high-throughput culturing yielded novel SAR11 strains and other abundant marine bacteria from the Oregon coast and the Bermuda Atlantic Time Series study site. ISME Journal 1:361–371.

Storey J.D. 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64:479–498.

Sweet R.L. 1995. Role of bacterial vaginosis in pelvic inflammatory disease. Clinical Infectious Diseases 20 Suppl 2:S271–5.

Swidsinski A., Mendling W., Loening-Baucke V., Ladhoff A., Swidsinski S., Hale L.P., and Lochs H. 2005. Adherent biofilms in bacterial vaginosis. Obstetrics and Gynecology 106:1013–1023.

Tabrizi S.N., Fairley C.K., Bradshaw C.S., and Garland S.M. 2006. Prevalence of *Gardnerella vaginalis* and *Atopobium vaginae* in virginal women. Sexually Transmitted Diseases 33:663–665.

Tamrakar R., Yamada T., Furuta I., Cho K., Morikawa M., Yamada H., Sakuragi N., and Minakami H. 2007. Association between *Lactobacillus* species and bacterial vaginosis-related bacteria, and bacterial vaginosis scores in pregnant Japanese women. BMC Infectious Diseases 7:128.

Tanner J.M. 1962. Growth at Adolescence. 2 ed., Thomas, Springfield, IL.

Tärnberg M., Jakobsson T., Jonasson J., and Forsum U. 2002. Identification of randomly selected colonies of lactobacilli from normal vaginal fluid by pyrosequencing of the 16S rDNA variable V1 and V3 regions. APMIS 110:802–810.

Thies F.L., König W., and König B. 2007. Rapid characterization of the normal and disturbed vaginal microbiota by application of 16S rRNA gene terminal RFLP fingerprinting. Journal of Medical Microbiology 56:755–761.

Thoma M.E., Gray R.H., Kiwanuka N., Aluma S., Wang M.C., Sewankambo N., and Wawer M.J. 2011. Longitudinal changes in vaginal microbiota composition assessed by Gram stain among never sexually active pre- and postmenarcheal adolescents in Rakai, Uganda. Journal of Pediatric and Adolescent Gynecology 24:42–47.

Thomas S. 1928. Döderlein's bacillus: *Lactobacillus acidophilus*. Journal of Infectious Diseases 43:218–227.

Totten P.A., Amsel R., Hale J., Piot P., and Holmes K.K. 1982. Selective differential human blood bilayer media for isolation of *Gardnerella* (*Haemophilus*) *vaginalis*. Journal of Clinical Microbiology 15:141–147.

Tottey W., Denonfoux J., Jaziri F., Parisot N., Missaoui M., Hill D., Borrel G., Peyretaillade E., Alric M., Harris H.M.B., Jeffery I.B., Claesson M.J., O'Toole P.W., Peyret P., and Brugère J.F. 2013. The Human Gut Chip "HuGChip", an explorative phylogenetic microarray for determining gut microbiome diversity at family level. PLoS ONE 8:e62544.

Turnbaugh P.J., Ley R.E., Hamady M., Fraser-Liggett C.M., Knight R., and Gordon J.I. 2007. The Human Microbiome Project. Nature 449:804–810.

Vaca M., Guadalupe I., Erazo S., Tinizaray K., Chico M., Cooper P., and Hay P. 2009. High prevalence of bacterial vaginosis in adolescent girls in a tropical area of Ecuador. British Journal of Obstetrics and Gynaecology 117:225–228.

Vallor A.C., Antonio M.A., Hawes S.E., and Hillier S.L. 2001. Factors associated with acquisition of, or persistent colonization by, vaginal lactobacilli: role of hydrogen peroxide production. The Journal of Infectious Diseases 184:1431–1436.

Valore E.V., Park C.H., Igreti S.L., and Ganz T. 2002. Antimicrobial components of vaginal fluid. American Journal of Obstetrics and Gynecology 187:561–568.

van der Lelie D., Lesaulnier C., McCorkle S., Geets J., Taghavi S., and Dunn J. 2006. Use of single-point genome signature tags as a universal tagging method for microbial genome surveys. Applied and Environmental Microbiology 72:2092–2101.

Vásquez A., Jakobsson T., Ahrné S., Forsum U., and Molin G. 2002. Vaginal *Lactobacillus* flora of healthy Swedish women. Journal of Clinical Microbiology 40:2746–2749.

Ventura M., Canchaya C., Tauch A., Chandra G., Fitzgerald G.F., Chater K.F., and van Sinderen D. 2007. Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. Microbiology and Molecular Biology Reviews 71:495–548.

Ventura M., O'Flaherty S., Claesson M.J., Turroni F., Klaenhammer T.R., van Sinderen D., and O'Toole P.W. 2009. Genome-scale analyses of health-promoting bacteria: probiogenomics. Nature Reviews Microbiology 7:61–71.

Verhelst R., Verstraelen H., Claeys G., Verschraegen G., Delanghe J., Van Simaey L., De Ganck C., Temmerman M., and Vaneechoutte M. 2004. Cloning of 16S rRNA genes amplified from normal and disturbed vaginal microflora suggests a strong association between *Atopobium vaginae*, *Gardnerella vaginalis* and bacterial vaginosis. BMC Microbiology 4:16.

Verhelst R., Verstraelen H., Claeys G., Verschraegen G., Van Simaey L., De Ganck C., De Backer E., Temmerman M., and Vaneechoutte M. 2005. Comparison between Gram stain and culture for the characterization of vaginal microflora: definition of a distinct grade that resembles grade I microflora and revised categorization of grade I microflora. BMC Microbiology 5:61.

Verstraelen H. and Swidsinski A. 2013. The biofilm in bacterial vaginosis: implications for epidemiology, diagnosis and treatment. Current Opinion in Infectious Diseases 26:86–89.

Verstraelen H., Verhelst R., Claeys G., De Backer E., Temmerman M., and Vaneechoutte M. 2009. Longitudinal analysis of the vaginal microflora in pregnancy suggests that *L. crispatus* promotes the stability of the normal vaginal microflora and that *L. gasseri* and/or *L. iners* are more conducive to the occurrence of abnormal vaginal microflora. BMC Microbiology 9:116.

Verstraelen H., Verhelst R., Claeys G., Temmerman M., and Vaneechoutte M. 2004. Culture-independent analysis of vaginal microflora: the unrecognized association of *Atopobium vaginae* with bacterial vaginosis. American Journal of Obstetrics and Gynecology 191:1130–1132.

Vickerstaff J.M. and Cole B.C. 1969. Characterization of *Haemophilus vaginalis*, *Corynebacterium cervicis*, and related bacteria. Canadian Journal of Microbiology 15:587–594.

Wagner M., Smidt H., Loy A., and Zhou J. 2007. Unravelling microbial communities with DNA-microarrays: challenges and future directions. Microbial Ecology 53:498–506.

Walker B. 1995. Conserving biological diversity through ecosystem resilience. Conservation Biology 9:747–752.

Wang Q., Garrity G.M., Tiedje J.M., and Cole J.R. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73:5261–5267.

Warnes G.R., Bolker B., Bonebakker L., Gentleman R., Huber W., Liaw A., Lumley T., Maechler M., Magnusson A., and Moeller S. 2014. gplots: Various R programming tools for plotting data. R package version 2.

Warton D.I. and Hui F.K.C. 2011. The arcsine is asinine: the analysis of proportions in ecology. Ecology 92:3–10.

White P. and Jentsch A. 2001. The search for generality in studies of disturbance and ecosystem dynamics. Progress in Botany 62:399–450.

Wiesenfeld H.C., Hillier S.L., Krohn M.A., Amortegui A.J., Heine R.P., Landers D.V., and Sweet R.L. 2002. Lower genital tract infection and endometritis: insight into subclinical pelvic inflammatory disease. Obstetrics and Gynecology 100:456–463.

Wiesenfeld H.C., Hillier S.L., Krohn M.A., Landers D.V., and Sweet R.L. 2003. Bacterial vaginosis is a strong predictor of *Neisseria gonorrhoeae* and *Chlamydia trachomatis* infection. Clinical Infectious Diseases 36:663–668.

Wiggins R., Hicks S.J., Soothill P.W., Millar M.R., and Corfield A.P. 2001. Mucinases and sialidases: their role in the pathogenesis of sexually transmitted infections in the female genital tract. Sexually Transmitted Infections 77:402–408.

Wilks M. and Tabaqchali S. 1987. Quantitative bacteriology of the vaginal flora during the menstrual cycle. Journal of Medical Microbiology 24:241–245.

Wilks M., Wiggins R., Whiley A., Hennessy E., Warwick S., Porter H., Corfield A., and Millar M. 2004. Identification and H_2O_2 production of vaginal lactobacilli from pregnant women at high risk of preterm birth and relation with outcome. Journal of Clinical Microbiology 42:713–717.

Wira C.R., Fahey J.V., Sentman C.L., Pioli P.A., and Shen L. 2005. Innate and adaptive immunity in female genital tract: cellular responses and interactions. Immunological Reviews 206:306–335.

Witkin S.S., Mendes-Soares H., Linhares I.M., Jayaram A., Ledger W.J., and Forney L.J. 2013. Influence of vaginal bacteria and D- and L-lactic acid isomers on vaginal extracellular matrix metalloproteinase inducer: implications for protection against upper genital tract infections. mBio 4:e00460–13.

Workowski K.A., Berman S., and Centers for Disease Control and Prevention (CDC). 2010. Sexually transmitted diseases treatment guidelines, 2010. MMWR 59 No. RR-12:1–109.

Wylie J.G. and Henderson A. 1969. Identity and glycogen-fermenting ability of lactobacilli isolated from the vagina of pregnant women. Journal of Medical Microbiology 2:363–366.

Yamamoto T., Zhou X., Williams C.J., Hochwalt A., and Forney L.J. 2009. Bacterial populations in the vaginas of healthy adolescent women. Journal of Pediatric and Adolescent Gynecology 22:11–18.

Yen S., Shafer M.A., Moncada J., Campbell C.J., Flinn S.D., and Boyer C.B. 2003. Bacterial vaginosis in sexually experienced and non-sexually experienced young women entering the military. Obstetrics and Gynecology 102:927–933.

Yeoman C.J., Thomas S.M., Miller M.E.B., Ulanov A.V., Torralba M., Lucas S., Gillis M., Cregger M., Gomez A., Ho M., Leigh S.R., Stumpf R., Creedon D.J., Smith M.A., Weisbaum J.S., Nelson K.E., Wilson B.A., and White B.A. 2013. A multi-omic systems-based approach reveals metabolic markers of bacterial vaginosis and insight into the disease. PLoS ONE 8:e56111.

Yeoman C.J., Yildirim S., Thomas S.M., Durkin A.S., Torralba M., Sutton G., Buhay C.J., Ding Y., Dugan-Rocha S.P., Muzny D.M., Qin X., Gibbs R.A., Leigh S.R., Stumpf R., White B.A., Highlander S.K., Nelson K.E., and Wilson B.A. 2010. Comparative genomics of *Gardnerella vaginalis* strains reveals substantial differences in metabolic and virulence potential. PLoS ONE 5:e12411.

Yilmaz A.E., Celik N., Soylu G., Donmez A., and Yuksel C. 2012. Comparison of clinical and microbiological features of vulvovaginitis in prepubertal and pubertal girls. Journal of the Formosan Medical Association 111:392–396.

Yuan S., Cohen D.B., Ravel J., Abdo Z., and Forney L.J. 2012. Evaluation of methods for the extraction and purification of DNA from the human microbiome. PLoS ONE 7:e33865.

Zariffard M.R., Saifuddin M., Sha B.E., and Spear G.T. 2002. Detection of bacterial vaginosis-related organisms by real-time PCR for Lactobacilli, *Gardnerella vaginalis* and *Mycoplasma hominis*. FEMS Immunology and Medical Microbiology 34:277–281.

Zhou A., He Z., Qin Y., Lu Z., Deng Y., Tu Q., Hemme C.L., Van Nostrand J.D., Wu L., Hazen T.C., Arkin A.P., and Zhou J. 2013. StressChip as a high-throughput tool for assessing microbial community responses to environmental stresses. Environmental Science & Technology 47:9841–9849.

Zhou J. 2003. Microarrays for bacterial detection and microbial community analysis. Current Opinion in Microbiology 6:288–294.

Zhou X., Bent S.J., Schneider M.G., Davis C.C., Islam M.R., and Forney L.J. 2004. Characterization of vaginal microbial communities in adult healthy women using cultivation-independent methods. Microbiology 150:2565–2573.

Zhou X., Brown C.J., Abdo Z., Davis C.C., Hansmann M.A., Joyce P., Foster J.A., and Forney L.J. 2007. Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. ISME Journal 1:121–133.

Zhou X., Hansmann M.A., Davis C.C., Suzuki H., Brown C.J., Schütte U., Pierson J.D., and Forney L.J. 2010. The vaginal bacterial communities of Japanese women resemble those of women in other racial groups. FEMS Immunology and Medical Microbiology 58:169–181.

Zinnemann K. and Turner G.C. 1963. The taxonomic position of "*Haemophilus vaginalis*" [*Corynebacterium vaginale*]. The Journal of Pathology and Bacteriology 85:213–219.

Zozaya-Hinchliffe M., Lillis R., Martin D., and Ferris M. 2010. Quantitative PCR assessments of bacterial species in women with and without bacterial vaginosis. Journal of Clinical Microbiology 48:1812–1819.

APPENDIX A

Supplementary Information to Chapter 2

A.1 VCHIP PROBE SET DESIGN

We downloaded complete and draft genomes of selected bacterial species from the Genomes Online Database (`http://www.genomesonline.org`) in November 2011. Initially our design of VChip probe sets included 336 bacterial strains (200 species). Among these were 812,653 coding DNA sequences (CDS) of which 119,091 were redundant. 240 CDS were shorter than 10 bp or longer than 9,999 bp and were excluded from further analysis. The remaining 693,322 CDS were clustered based on 80% identity and coverage using the software program Cd-hit (Li and Godzik, 2006), yielding 473,709 clusters (364,808 singletons). Multiple sequence alignments were performed in MUSCLE (Edgar, 2004) to generate a representative consensus sequence for each cluster, replacing ambiguous or heterogeneous sites with ambiguity code 'N'. Consensus sequences were submitted to Roche NimbleGen (Madison, WI, USA) to design unique 60-mer probes for each cluster. Initially five probes per cluster were designed, but because the total number exceeded the capacity of the array (~2.4 million probes versus 1.4 million available per sub-array), we manually reduced the initial set of 336 bacterial strain genomes to down to 313 (184 species). After excluding the removed strains, 307,860 of the original 473,709 gene clusters were represented in the final array design in addition to 716 human immunity genes (3,580 probes). 74.4% of the bacterial gene clusters (n=229,131) had five probes, and all had at least two probes. Of the 313 strains represented on the array, 246 (78.6%) had strain-specific probe sets while the rest were represented by probe sets shared with other strains or species. 165 of the 184 species on the array (89.7%) had species-specific probe sets.

The NimbleGen Design File (NDF) containing the probe sequences and detailed information about the design of the array is available at `http://github.com/roxanahickey/vchip`. An explanation of the variables within this file can be found in the NimbleScan Software User's Guide (version 2.6), available at `http://www.nimblegen.com/downloads/support/NimbleScan_v2p6_UsersGuide.pdf`. Additionally, spreadsheets including bacterial and human gene cluster information and annotations are available at `http://github.com/roxanahickey/vchip`.

A.2 RESOLUTION OF UNEXPECTED *LACTOBACILLUS CRISPATUS* HYBRIDIZATION SIGNAL FROM MC-1

We utilized a set of comparisons (Figure A.1) to resolve the unexpected hybridization of *Lactobacillus crispatus*-specific probes in the mock community MC-1 (Figure 2.2), which should have consisted of only *Anaerococcus hyrogenalis, Anaerococcus tetradius, Atopobium vaginae, Finegoldia magna* and *Gardnerella vaginalis* in equal proportions (Table 2.1). To verify whether the observed anomaly was due to contamination of the sample with *L. crispatus* or cross-hybridization of species-specific probe sets on the VChip, we compared the array hybridization results of several mock community and vaginal samples against those of mock community MC-3, which we knew should contain *L. crispatus* (95% of the total genomic DNA) along with the other five species listed above (1% each of total genomic DNA).

First consider Figure A.1, panel S1, in which array MC-8 (hybridized with 100% *L. crispatus*) on the y-axis is compared to array MC-3 on the x-axis (hybridized with 95% *L. crispatus* + 5% mixed sample, where 'mixed' refers to the mixture of five non-*Lactobacillus* bacterial species in equal proportions). The set of purple points represents *L. crispatus* and shows high expression values in both arrays. Additionally, the large set of red points (representing all probe sets associated with the five species in the mixed sample) shows high expression values on the expected array MC-3, but not in the 100% *L. crispatus* array MC-8. This clearly illustrates that very little cross hybridization between *L. crispatus* probes and the five species of the mixed sample occurs.

As a second check, and to further confirm that minimal cross-hybridization was occurring, we compared array MC-9 (hybridized with 100% human DNA) to array MC-3 (Figure A.1, panel S2). It is clear from this plot that only a handful of bacterial probes have high expression values when hybridized only with human DNA. This small number may represent random noise or limited cross-hybridization in a few probe sets. Figure A.1 panel S3 presents a comparison of array MC-1 (hybridized with 100% mixed sample) compared against array MC-3. The set of high-expression *L. crispatus* probes (purple) in this plot are common to both arrays, despite the expectation that no *L. crispatus* was present in MC-1. Panel S3 is very similar to panel S4 in which array MC-2 (hybridized with 50% *L. crispatus*, 50% mixed) is compared to array MC-3. The high similarity of these two plots suggests that the two samples were also very similar in composition, and that the sample hybridized to array MC-1 most likely contained *L. crispatus* DNA.

Finally, we considered two comparisons against samples collected from vaginal communities in which community composition was determined using 16S rRNA amplicon sequencing (Figure A.1, panels S5 and S6). In panel S5, expression signals from sample VM-1, dominated almost entirely by *L. iners*, show almost no overlapping signal with array MC-3. In panel S6, 16S rRNA results indicating the presence of *L. crispatus* are corroborated by high expression values for *L. crispatus* probes. Taken as a whole, this set of comparisons provides strong evidence that the VChip is functioning properly with minimal cross-hybridization and good fidelity to 16S rRNA-based methods.

A.3 SUPPLEMENTARY TABLES AND FIGURES FOR CHAPTER 2

In addition to the two supplementary tables and one supplementary figure below, Supplementary File 1 (differential gene expression results for VM-1 and VM-2, XLSX file) is available at `http://github.com/roxanahickey/dissertation`.

Table A.1: Number of total and species-specific gene cluster probe sets of 184 bacterial species on VChip

Species	Total	Unique	Species	Total	Unique	Species	Total	Unique
Actinobacillus minor	1	0	*Fusobacterium nucleatum*	1294	791	*Porphyromonas uenonis*	784	433
Actinomyces urogenitalis	9	0	*Fusobacterium periodonticum*	681	650	*Prevotella amnii*	1498	1444
Aerococcus urinae	1572	1570	*Fusobacterium ulcerans*	2688	2686	*Prevotella bergensis*	2308	2256
Aerococcus viridans	1765	1760	*Gardnerella vaginalis*	1723	329	*Prevotella bivia*	1426	1360
Alistipes putredinis	2231	2197	*Gemella haemolysans*	1355	1346	*Prevotella buccalis*	1856	1678
Anaerococcus hydrogenalis	1284	1261	*Gordonia bronchialis*	4210	4208	*Prevotella copri*	2766	2760
Anaerococcus lactolyticus	1506	1464	*Lactobacillus acidophilus*	2	0	*Prevotella denticola*	2069	1998
Anaerococcus prevotii	1061	1046	*Lactobacillus amylolyticus*	5	0	*Prevotella disiens*	2024	1952
Anaerococcus tetradius	1160	1099	*Lactobacillus antri*	146	0	*Prevotella melaninogenica*	1690	1656
Arcanobacterium haemolyticum	1552	1550	*Lactobacillus brevis*	10	0	*Prevotella oralis*	4586	0
Arcobacter butzleri	1145	1143	*Lactobacillus buchneri*	1	0	*Prevotella oris*	2510	2489
Arcobacter nitrofigilis	2208	2205	*Lactobacillus casei*	3034	0	*Prevotella ruminicola*	2562	2562
Atopobium parvulum	1148	1127	*Lactobacillus crispatus*	2837	381	*Prevotella tannerae*	2039	2021
Atopobium rimae	1302	1279	*Lactobacillus delbrueckii*	4088	1092	*Prevotella timonensis*	1669	1523
Atopobium vaginae	3207	3196	*Lactobacillus fermentum*	2920	496	*Prevotella veroralis*	1906	1873
Bacteroides eggerthii	2930	2900	*Lactobacillus gasseri*	1252	554	*Propionibacterium acnes*	8283	447
Bacteroides finegoldii	2531	2446	*Lactobacillus helveticus*	85	0	*Propionibacterium freudenreichii*	2065	2065
Bacteroides fragilis	3670	3567	*Lactobacillus iners*	7032	354	*Propionibacterium sp.*	3594	94
Bacteroides sp.	3440	870	*Lactobacillus jensenii*	2929	208	*Proteus penneri*	1	0
Bacteroides thetaiotaomicron	3107	3021	*Lactobacillus johnsonii*	946	576	*Roseburia intestinalis*	3598	3431
Bifidobacterium adolescentis	1958	966	*Lactobacillus oris*	786	650	*Roseburia inulinivorans*	3569	3390
Bifidobacterium angulatum	1186	1141	*Lactobacillus paracasei*	1995	392	*Roseomonas cervicalis*	4393	4393
Bifidobacterium bifidum	1319	1281	*Lactobacillus plantarum*	7582	564	*Rothia dentocariosa*	1863	1860
Bifidobacterium breve	2071	702	*Lactobacillus reuteri*	5738	461	*Rothia mucilaginosa*	1447	1444
Bifidobacterium catenulatum	1078	991	*Lactobacillus rhamnosus*	8667	509	*Ruminococcus albus*	3441	3390
Bifidobacterium dentium	6688	268	*Lactobacillus ruminis*	1818	1815	*Ruminococcus flavefaciens*	3717	3667
Bifidobacterium gallicum	1600	1600	*Lactobacillus sakei*	1607	1607	*Ruminococcus gnavus*	3182	3131
Bifidobacterium longum	3946	1440	*Lactobacillus salivarius*	2	0	*Ruminococcus lactaris*	2154	2119
Brevibacterium mcbrellneri	2209	2206	*Lactobacillus ultunensis*	1120	1095	*Ruminococcus obeum*	3489	3422
Bulleidia extructa	1141	1141	*Lactobacillus vaginalis*	1408	1401	*Ruminococcus torques*	2166	2107
Capnocytophaga gingivalis	2060	1997	*Lactococcus lactis*	1	0	*Segniliparus rotundus*	2758	2758
Capnocytophaga ochracea	1042	905	*Leptotrichia buccalis*	883	687	*Selenomonas flueggei*	1	0
Capnocytophaga sputigena	1365	1185	*Leptotrichia goodfellowii*	1582	1581	*Selenomonas noxia*	1	0
Catenibacterium mitsuokai	2740	2738	*Leptotrichia hofstadii*	1198	1005	*Selenomonas sputigena*	1	0
Cellvibrio japonicus	3390	3390	*Megasphaera genomosp.*	1291	125	*Serratia odorifera*	3838	3698
Chlamydia trachomatis	2242	2242	*Megasphaera sp.*	1225	70	*Serratia proteamaculans*	1274	1169
Chryseobacterium gleum	4534	4512	*Micrococcus luteus*	2	0	*Sphingobacterium spiritivorum*	12938	0
Clostridiales genomosp.	1264	1249	*Mobiluncus curtisii*	8918	89	*Sphingomonas wittichii*	4653	4653
Clostridium beijerinckii	4369	4369	*Mobiluncus mulieris*	8452	343	*Staphylococcus aureus*	3518	519
Clostridium difficile	3093	3072	*Mycobacterium parascrofulaceum*	5809	5807	*Staphylococcus hominis*	1450	1448
Collinsella aerofaciens	1896	1880	*Neisseria cinerea*	386	337	*Staphylococcus lugdunensis*	1786	1776
Collinsella intestinalis	1183	1128	*Neisseria elongata*	2074	2039	*Staphylococcus saprophyticus*	1915	1913
Collinsella stercoris	1852	1795	*Neisseria flavescens*	1327	667	*Stenotrophomonas maltophilia*	4078	4077
Coprococcus comes	3302	3237	*Neisseria gonorrhoeae*	1099	189	*Streptococcus agalactiae*	8061	700
Coprococcus eutactus	2660	2625	*Neisseria lactamica*	521	400	*Streptococcus bovis*	2236	0
Corynebacterium aurimucosum	1897	1867	*Neisseria meningitidis*	1171	654	*Streptococcus gordonii*	1193	1014
Corynebacterium genitalium	1904	1895	*Neisseria mucosa*	1001	750	*Streptococcus infantarius*	1015	940
Corynebacterium glucuronolyticum	4617	610	*Neisseria polysaccharea*	449	329	*Streptococcus mitis*	745	659
Corynebacterium jeikeium	3231	440	*Neisseria sicca*	1405	1157	*Streptococcus mutans*	2	0
Corynebacterium lipophiloflavum	1993	1982	*Neisseria subflava*	802	480	*Streptococcus oralis*	91	0
Corynebacterium pseudogenitalium	2048	2005	*Paenibacillus larvae*	3797	3797	*Streptococcus parasanguinis*	1467	1444
Corynebacterium striatum	2068	2034	*Parabacteroides distasonis*	3101	595	*Streptococcus pseudoporcinus*	1637	1596
Dialister invisus	1605	1603	*Parabacteroides johnsonii*	2412	1783	*Streptococcus salivarius*	19	0
Dialister microaerophilus	1119	1117	*Parabacteroides merdae*	1975	1289	*Streptococcus sanguinis*	1463	1293
Eggerthella lenta	2887	2880	*Pasteurella multocida*	1	0	*Treponema pallidum*	2512	61
Enterococcus faecalis	17130	337	*Peptoniphilus duerdenii*	3288	0	*Treponema phagedenis*	4734	0
Eremococcus coleocola	1560	1553	*Peptoniphilus lacrimalis*	1076	1054	*Ureaplasma parvum*	237	16
Escherichia coli	9500	0	*Peptostreptococcus anaerobius*	1508	1491	*Ureaplasma urealyticum*	627	40
Finegoldia magna	2028	539	*Porphyromonas asaccharolytica*	837	482	*Veillonella atypica*	1082	1022
Flavobacterium johnsoniae	3969	3968	*Porphyromonas endodontalis*	1574	1567	*Veillonella dispar*	451	380
Flavobacterium psychrophilum	1637	1636	*Porphyromonas gingivalis*	3243	493	*Veillonella parvula*	1124	505
Fusobacterium gonidiaformans	1168	1162						

Table A.2: Reference genomes used to calculate genome copy equivalents of species in mock communities

Species	Estimated genome size[a] (Mb)	NCBI reference genome(s) used to estimate genome size (Mb / GenBank ID)
Anaerococcus hydrogenalis (vaginal isolate)	1.94	*A. hydrogenalis* ACS-025-V-Sch4 (1.98 / GCA_000191745.2)
		A. hydrogenalis DSM7454 (1.89 / GCA_000173355.1)
Anaerococcus tetradius (vaginal isolate)	2.15	*A. tetradius* ATCC 35098 (2.15 / GCA_000159095.1)
Atopobium vaginae ATCC BAA-55	1.44	*A. vaginae* DSM 15829 (1.43 / GCA_000159235.2)
		A. vaginae PB189-T1-4 (1.45 / GCA_000179715.1)
Finegoldia magna (vaginal isolate)	1.93	*F. magna* ATCC 53516 (1.92 / GCA_000159695.1)
		F. magna ACS-171-V-Col3 (1.83 / GCA_000179495.1)
		F. magna SY403409CC001050417 (2.03 / GCA_000221585.2)
Gardnerella vaginalis ATCC 14018	1.6	*G. vaginalis* ATCC 14018 = JCM 11026 (1.60 / GCA_000178355.1)
Lactobacillus crispatus ATCC 33820	1.98	*L. crispatus* ST1 (2.04 / GCA_000091765.1)
		L. crispatus 214-1 (2.07 / GCA_000177575.1)
		L. crispatus EM-LC1 (1.83 / GCA_000497065.1)
Homo sapiens (female genomic DNA)	2.90E+03	*H. sapiens* (2.84E+03 / GCA_000002125.2)
		H. sapiens (2.86E+03 / GCA_000002115.2)
		H. sapiens (2.99E+03 / GCA_000365445.1)

[a] Estimated genome size was calculated by taking the average of genome sizes listed in the third column for each species.

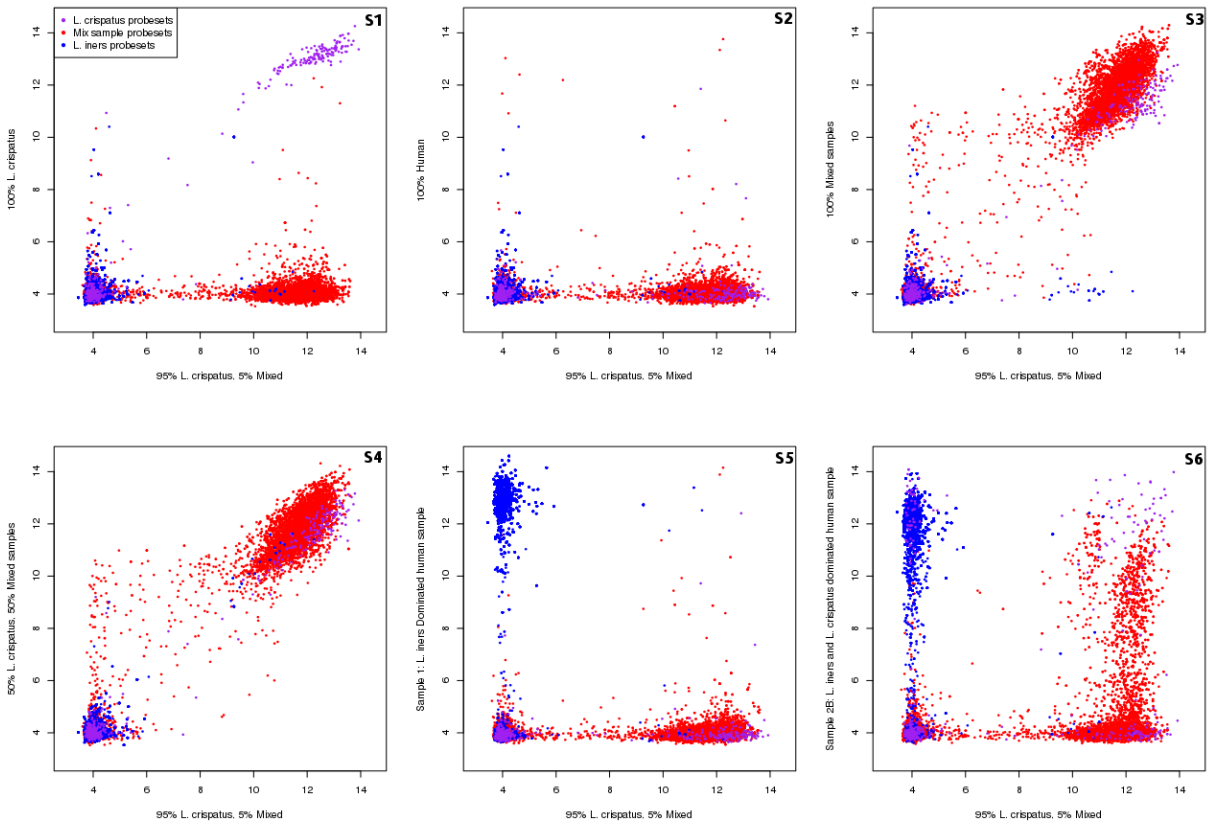

Figure A.1: Comparisons of probe set hybridization with selected mock community samples and vaginal swabs. Array hybridization intensities were compared across samples to determine whether an anomalous result in mock community MC-1 was due to contamination or failure of the microarray. Hybridization signal less than < 4–5 is considered background level.

APPENDIX B

Supplementary Information to Chapter 3

B.1 Estimation of 16S rRNA gene copy number in low-pH and high-pH vaginal microbiota

B.1.1 *Pan-bacterial 16S rRNA qPCR assay*

We observed many vaginal microbiota samples from girls that had a high relative abundance of lactobacilli along with a vaginal pH higher than what is typically considered healthy in adults (i.e., ≤ 4.5). We hypothesized this might be due to lower total numbers of bacteria, and thus lower levels of lactic acid production, which contributes to the low vaginal pH seen in most healthy women. We performed a coarse test of this hypothesis using the bacterial BactQuant qPCR assay, developed by Liu *et al.* (2012), to estimate the number of 16S rRNA gene copies present in vaginal samples from girls with high proportions of lactobacilli (> 0.75) and either low (< 5.0, n=62) or high (≥ 5.0, n=40) vaginal pH. Twelve genomic DNA samples from each of the low and high groups were selected by a random sampling function in R. DNA samples had been stored in AE buffer (Qiagen, Venlo, Netherlands) at -20°C since the time of extraction and purification, ranging from December 2010 to August 2012 (the qPCR was done in June 2014). Each qPCR reaction consisted of 5.0 μL of 2X qPCR SuperMix (Invitrogen, Carlsbad, CA, USA), 0.45 μL of 40 μM forward primer, 0.45 μL of 40 μM reverse primer, 0.113 μL of 20 μM 6-FAM-labeled TaqMan probe (Applied Biosystems, Foster City, CA, USA), 2.99 μL molecular-grade water, and 1 μL genomic DNA template, for a total reaction volume of 10 μL. The primer and probe sequences were as described by Liu *et al.*:

Primer/probe	Sequence	*E. coli* region (16S rRNA)
Forward primer	5'-CCTACGGG**D**GGC**W**GCA-3'	341-356
Reverse primer	5'-GGACTAC**HV**GGGT**M**TCTAATC-3'	786-806
Probe	(6FAM) 5'-CAGCAGCCGCGGTA-3' (MGBNFQ)	519-532

where bold font (in the primer sequences) denotes degenerate bases. Samples, along with an in-run standard curve (103-109 in 10-fold serial dilutions) and no-template controls, were processed

in triplicate. Amplification and real-time fluorescence detections were performed on the AB StepOnePlus Real-Time PCR System (Applied Biosystems) with the following PCR conditions: 3 minutes at 50°C for UNG treatment, 10 minutes at 95°C for *Taq* activation, 15 seconds at 95°C for denaturation, and 1 minute at 60°C for annealing and extension, repeated over 40 cycles. Cycle threshold values (i.e., Ct value) were determined, and 16S rRNA gene copy number estimates generated from the standard curve, using StepOne Software v2.1 (Applied Biosystems).

B.1.2 *Comparison of 16S rRNA gene copies in low-pH and high-pH vaginal microbiota*

Results of the pan-bacterial 16S rRNA qPCR assay are shown in Figure B.5. Out of 24 samples, 23 were successfully amplified. Estimated 16S rRNA copy numbers ranged from 6.7E+07 to 3.0E+09 copies/μL (mean 9.1E+08 copies/μL) in the low-pH samples, and 8.5E+06 to 2.3E+09 copies/μL (mean 4.7E+08 copies/μL) in the high-pH samples. These were compared using a one-tailed t-test (H_0: low-pH 16S rRNA copies/μL > high-pH 16S rRNA copies/μL) and found to be not significantly different (p=0.14). However, the distribution of the low-pH group appears to be skewed higher than in the high-pH group, providing at least some indication that bacterial counts may indeed be higher in low-pH microbiota, even if the difference is not statistically significant in the subset of samples we tested. Linear modeling also did not detect any statistically significant associations of 16S rRNA copy number or pH with individual *Lactobacillus* spp. present in the communities (data not shown).

Although we failed to detect a significant difference in 16S rRNA copies (as a rough proxy for bacterial cells) in relation to vaginal pH, we should not rule out the possibility based on this analysis. We note that factors such as sample quantity and quality may have posed significant limitations to the qPCR assay, particularly since it was performed well after genomic DNA had been extracted and archived. Furthermore, the amount of material collected on the vaginal swabs was not strictly controlled for, other than rotating each swab in the vaginal introitus three times during collection. Differences in the amount of vaginal secretions or sloughed epithelial cells collected on the swabs could greatly impact the amount of DNA recovered. This is why we typically analyze proportions rather than absolute numbers of bacterial taxa. A more thorough evaluation of bacterial counts during puberty would ideally utilize fresh samples and some means of controlling for sample quantity (e.g., weighing swab material, quantifying vaginal secretions, or determining the proportion of bacterial cells or DNA relative to human).

B.2 GENOMIC DNA EXTRACTION AND 16S RRNA PYROSEQUENCING

B.2.1 *Genomic DNA extraction and purification*

Genomic DNA was extracted from vaginal and vulvar swabs as previously described (Yuan *et al.*, 2012) by cell lysis using enzymatic and bead-beating treatments followed by purification using QIAamp DNA Mini Kit (Qiagen). Briefly, vaginal swabs suspended in Amies medium (Copan Diagnostics, Murrietta, CA, USA) were thawed on ice and vigorously vortexed for 5 minutes to dislodge and resuspend cells. A 500 μL aliquot was transferred to a clean sterile beading-beating tube (MP Biomedicals, Santa Ana, CA, USA) and kept on ice. A lytic enzyme cocktail was prepared at the time of extraction and added to each sample as follows: 50 μL of 10 mg/mL lysozyme, 6 μL of 25,000 U/mL mutanolysin (Sigma-Aldrich, St. Louis, MO, USA), 3 μL of 4,000 U/mL lysostaphin in sodium acetate (Sigma-Aldrich), and 41 μL TE50 buffer for a final volume of 100 μL per sample. Samples were digested by incubation at 37°C for 60 minutes in a dry heat block. 750 mg sterile 0.1 mm diameter zirconia/silica beads (Biospec Products, Bartlesville, OK, USA) were added to each digested sample. Bead-beating was performed for 1 minute at 36 oscillations per second (2,100 rpm) with the use of a Mini-Beadbeater-96 (Biospec Products). Following cell disruption, the tubes were centrifuged at 1,200 rpm for 1 minute. Aliquots of crude lysate from each sample were transferred to new sterile microcentrifuge tubes, and 50 μL proteinase K (20 mg/mL [>600 mAU/mL]) and 500 μL Qiagen buffer AL were added. Samples were mixed by pulse-vortexing for 15 seconds and then incubated at 56°C for 30 minutes. After this step, 50 μL 3 mol/L sodium acetate (pH 5.5) was added, followed by 500 μLL 100% ethanol at each sample. Vortexing was repeated for an additional 15 seconds before briefly centrifuging. From this point onward, purification of genomic DNA was done with QIAamp DNA Mini Kits following the manufacturer's instructions.

B.2.2 *Pyrosequencing V1–V3 regions of 16S rRNA genes*

To characterize the composition and structure of bacterial communities in vaginal and vulvar samples, we sequenced the V1–V3 regions of 16S rRNA genes amplified from each sample. The amplicons were obtained by PCR using primers that flanked hypervariable regions 1 and 3 of bacterial 16S rRNA genes (*Escherichia coli* positions 27–534). The sequences of the primers used

were as follows:

Primer	Sequence
454_27F-YM	5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGTC**AGAGTTTGATYMTGGCTCAG**-3'
454_27F-Bif	5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGTC**AGGGTTCGATTCTGGCTCAG**-3'
454_27F-Bor	5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGTC**AGAGTTTGATCCTGGCTTAG**-3'
454_27F-Chl	5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGTC**AGAATTTGATCTTGGTTCAG**-3'
454_534R	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG*NNNNNNNN*TC**ATTACCGCGGCTGCTGGCA**-3'

where the underlined sequences are Roche 454 fusion adapters B and A in 27F and 534R, respectively, and the bold font denotes the universal 16S rRNA primers 27F and 534R. The four 27F primers were combined in equal amounts and designated 27F*. The 534R primer included a unique sequence tag to barcode each of the samples denoted by the 8 italicized Ns. This allowed us to sequence the amplicons from all samples simultaneously, and afterwards assign each sequence to the sample they were obtained from. Each PCR contained 34.4 μL of molecular-grade water, 5.0 μL of 10X buffer (Applied Biosystems), 6.0 μL of 25 mM $MgCl_2$ (Applied Biosystems), 0.4 μL of 25 mM dNTP (Amersham Bioscience, Amersham, UK), 0.5 μL of 20 μM forward primer 454_27F*, and 0.5 μL of 20 μM reverse primer 454_534R, 0.2 μL of 5 U/μL Taq DNA polymerase (Applied Biosystems), and 1.0 μL of DNA template, in a total volume of 50 μL. Amplification of fragments was done using an initial denaturation step at 94°C for 4 min, followed by 30 cycles of denaturation at 94°C for 1 min, annealing at 55°C for 1 min, and an extension at 72°C for 2 min. A final extension step of 10 minutes at 72°C was done.

Concentrations of amplicons were estimated with the use of a fluorometric Picogreen assay on a SpectraMax GeminiXPS 96-well plate reader (Molecular Devices, Sunnyvale, CA, USA), and roughly equal amounts (~100 ng) were mixed in a single microfuge tube. Amplification primers and reaction buffer were removed by processing the amplicon mixture with the Agencourt AMPure Kit (Beckman Coulter, Brea, CA, USA). To determine the final quality we amplified the resulting amplicon pool with 454 adapter-specific primers in order to mimic the emulsion PCR (emPCR) process and processed the PCR product on a DNA 1000 chip for the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The final amplicon pool was deemed acceptable only if no short fragments were identified after PCR; otherwise the procedure was repeated again. The cleaned amplicon pool was then quantified using the KAPA 454 library quantification kit (KAPA Biosciences, Wilmington, MA, USA) and the Applied Biosystems StepOne

plus real-time PCR system (Applied Biosystems). Emulsion PCRs were performed on each quantified pool and final sequences were obtained using a Roche 454 GS-FLX+ pyrosequencer (Roche 454 Life Sciences, Branford, CT, USA).

B.2.3 *DNA sequence data analysis and taxonomic classification*

Raw, unclipped DNA sequence reads were cleaned, assigned, and filtered in the following manner. Raw SFF files were processed using the R package rSFFreader (`http://www.bioconductor.org/packages/release/bioc/html/rSFFreader.html`). Full-length (unclipped) sequence reads were used for the identification of Roche 454 adapters, barcodes and amplicon primers sequence using Cross Match (v1.080806; parameters: minimum matches 8, minimum score 16) from the Phred/Phrap/Consed application suite. Cross Match alignment information was then read into R and processed to identify alignment quality, directionality, barcode assignment, and sequence quality clip points. Base-quality clipping was done with the use of the Lucy application (v1.20p; parameters: maximum average error 0.002, maximum error at ends 0.002), and the clipped reads were aligned with the SILVA bacterial sequence database with the use of mothur v1.27 (Schloss *et al.*, 2009). Alignment end points were identified and used in subsequent filtering. Sequence reads were filtered to only those that met the following criteria: (a) sequences at least 100 bp in length; (b) maximum Hamming distance of barcode = 1; (c) maximum number of matching error to forward primer sequences = 2; (d) < 2 ambiguous bases (Ns); (e) < 10 bp homopolymer run in sequence; (f) alignment to the SILVA bacterial database (`http://www.arb-silva.de`) was within 75 bp of the expected alignment start position (507 bp); and (g) read alignment started within the first 5 bp and extended through read to within the final 5 bp. Each partial 16S rRNA gene sequence was classified with the use of the Ribosomal Database Project (RDP) Naïve Bayesian Classifier (Wang *et al.*, 2007). Reads were assigned to the first RDP level with a bootstrap score ≥ 50.

In this study, reads assigned to *Lactobacillus*, *Streptococcus* or *Gardnerella* were further assigned to the species level. Reference 16S rRNA sequences from the target genera were first extracted from the Patric database of bacterial genomes (`http://patricbrc.org`). These were aligned to each other using mothur and the SILVA database to generate a species reference alignment. Reads matching the target genus were extracted from the larger dataset and clustered using Cd-hit (Li and Godzik, 2006) at 99.5% identity to reduce redundancy in the reads as well

as the overall size of the dataset. Cluster representatives were then aligned to the SILVA database using mothur, and the species reference alignment was merged into the read alignment. Pairwise distances for the alignments were calculated using mothur and loaded into R for clustering and species identification. The pairwise distance matrix was clustered using hierarchical clustering with average linkage. Cluster modules were identified using the function cutreeDynamic from the WGCNA package (Langfelder and Horvath, 2008), varying the deepSplit parameter from 1 to 4. The final deepSplit parameter was selected based upon visual inspection of the hierarchical tree and module designations. Each read was then assigned to a cluster module based on its cluster representative's module assignment. If the module contained species representative sequences, the module was assigned this species name; otherwise the module was assigned with an ambiguous species identifier. Downstream analysis of taxonomic composition was performed in R as described in the main text.

B.3 COMMUNITY RICHNESS AND DIVERSITY ANALYSES

B.3.1 *Rarefaction analysis*

We performed analyses of community richness and diversity to qualitatively assess how they varied in relation to pubertal development and vaginal pH. These analyses are detailed on GitHub at `https://github.com/roxanahickey/adolescent`. Because there was considerable variability in the sequence read count among samples, it was necessary to perform a rarefaction analysis and randomly subsample counts at the same depth across all samples prior to calculating richness and diversity. Rarefaction curves were generated using the 'rarecurve' function in the vegan R package. From this plot we determined that subsampling at 2,000 observations per sample would sufficiently detect the observed genera without having to disregard a significant portion of our samples. Using this threshold we excluded 39 samples with less than 2,000 sequence reads from further analysis. We also generated genus accumulation curves using the 'specaccum' function in vegan. These curves indicated that most genera in our dataset could be observed from 50 or more samples, well below the number of samples we had available. A new genus abundance matrix was generated by subsampling at a depth of 2,000 observations using the 'rrarefy' function in vegan.

B.3.2 *Genus richness and diversity*

A considerable caveat to our analysis is that any taxa that could not be classified at the genus level were combined into an 'Other' category and treated as a single genus. Our estimates of richness and diversity are therefore likely to be quite conservative based on these numbers. For this reason we elected not to perform quantitative analyses of diversity; an OTU analysis would perhaps be better suited for that purpose. Richness was calculated using the 'rarefy' function on the subsampled data. Diversity indices (Shannon, Simpson, inverse Simpson) were calculated using the 'diversity' function. All three diversity indices showed similar trends, but because Shannon's index is more sensitive to large differences in richness, we chose to favor Simpson's index. Genus richness and diversity varied in both the vagina and vulva across Tanner breast stages and menarcheal stages as shown in Figure B.7. As perhaps expected, the vulva generally exhibited higher richness and diversity than the vagina but still experienced downward trends with progressive Tanner stages. Differences between premenarche and postmenarche were less pronounced. Richness and diversity of the vaginal microbiota varied with respect to vaginal pH as shown Figure B.8. The positive relationship between increasing richness or diversity and vaginal pH is interesting, but we caution against extrapolating biologically meaningful conclusions from these data for reasons described above.

B.4 SUPPLEMENTARY TABLES AND FIGURES FOR CHAPTER 3

In addition to the three supplementary tables and six supplementary figures below, Supplementary File 2 (graphical summaries of vaginal and vulvar community dynamics for each adolescent subject, PDF) is available at `http://github.com/roxanahickey/dissertation`.

Table B.1: Characteristics of all enrolled adolescent study participants

Subject ID	Race	Ethnicity	Age at enrollment (yr)	Tanner stage at enrollment (breast/pubic)	Participation duration (yr)	Achieved menarche during study	Participating mother ID
101	Black	Non-Hispanic	10.7	1/1	3.0	Yes	201
102	Caucasian	Non-Hispanic	10.8	3/3	2.9	Yes	202
103	Black	Non-Hispanic	10.7	2/2	2.7	Yes	203
104	Caucasian	Non-Hispanic	10.5	3/4	2.9	Yes	-
105	Caucasian	Hispanic	10.2	3/2	1.4	No	205
106	Black	Non-Hispanic	12.3	3/3	2.7	Yes	-
107	Black	Non-Hispanic	12.4	3/4	2.5	Yes	207
108	Black	Non-Hispanic	11.5	2/3	2.6	Yes	208
109	Black	Non-Hispanic	10.6	2/2	2.4	Yes	209
110	Black	Non-Hispanic	10.1	5/4	0.1	Yes	210
111	Black	Non-Hispanic	10.5	2/2	2.2	Yes	211
112	Black	Non-Hispanic	10.1	2/NA	1.9	Yes	212
113	Black	Non-Hispanic	11.6	3/3	1.2	Yes	213
114	Black	Non-Hispanic	11.0	3/3	2.0	Yes	214
115	Caucasian	Non-Hispanic	11.0	2/1	1.3	Yes	215
116	Black	Non-Hispanic	11.2	2/3	0.5	No	216
118	Caucasian	Non-Hispanic	11.3	3/2	Single visit	No	-
120	Black	Non-Hispanic	12.0	3/3	1.1	Yes	220
121	Black	Non-Hispanic	10.1	2/2	1.6	No	-
123	Black	Non-Hispanic	10.1	2/2	1.8	Yes	223
124	Black	Non-Hispanic	10.1	2/2	1.4	Yes	224
125	Caucasian	Non-Hispanic	10.9	2/3	1.4	Yes	-
126	Caucasian	Non-Hispanic	10.6	3/2	1.5	Yes	226
127	Native American	Non-Hispanic	12.1	3/2	1.3	Yes	227
128	Native American	Non-Hispanic	11.0	3/3	1.3	Yes	228
129	Caucasian	Non-Hispanic	10.6	1/2	Single visit	No	229
132	Black	Non-Hispanic	10.3	2/2	1.4	No	232
133	Black	Non-Hispanic	11.6	3/3	1.2	No	233
134	Black	Non-Hispanic	10.8	2/1	1.0	No	-
135	Black	Non-Hispanic	11.1	2/2	1.0	No	235
136	Black	Non-Hispanic	10.6	2/2	0.9	No	-

Table B.2: Linear mixed effects modeling of lactic acid bacteria and vaginal pH using Tanner pubic stage

Model and parameters[a]	Result for model				
LAB model: logit(LAB) ~ TP + age + TP:age + 1|subject + ε					
Random effects	Variance	SD	No. of observations	No. of groups	
Subject (intercept)	6.4	2.5	183	28	
Residual	6.8	2.6			
Fixed effects/contrasts[b]	Coefficient[c]	SE	df	*p*-value[d]	
Intercept	-19.8	6.1	163.6	1.31E-03	**
TP 2 vs. 1	-1.7	3.1	171.5	5.90E-01	
TP 3 vs. 2	32.3	11.4	164.5	5.23E-03	**
TP 4 vs. 3	16.6	7.4	150.3	2.58E-02	*
TP 5 vs. 4	-6.5	13.5	157.8	6.31E-01	
Age	2.0	0.5	163.2	1.71E-04	***
TP 3 vs. 2 : Age	-2.8	1.0	165.8	7.00E-03	**
TP 4 vs. 3 : Age	-1.4	0.6	150.4	2.94E-02	*
TP 5 vs. 4 : Age	0.5	1.0	156.6	6.34E-01	
Vaginal pH model: pH ~ TP + menarche status + age + 1|subject + ε					
Random effects	Variance	SD	No. of observations	No. of groups	
Subject (intercept)	0.6	0.8	117	20	
Residual	0.3	0.6			
Fixed effects/contrasts	Coefficient	SE	df	*p*-value	
Intercept	9.6	1.5	107.8	3.71E-09	***
TP 3 vs. 2	-0.8	0.2	104.3	6.00E-04	***
TP 4 vs. 3	0.0	0.2	99.3	8.13E-01	
TP 5 vs. 4	0.1	0.2	101.8	5.10E-01	
Postmenarche vs. premenarche	-0.1	0.2	110.0	4.86E-01	
Age	-0.4	0.1	108.5	6.49E-03	**

[a] LAB, lactic acid bacterium proportion; TP, Tanner pubic stage; ε, random error; SD, standard deviation; SE, standard error; df, degrees of freedom.

[b] Contrasts between successive Tanner stages were made, excluding stage 1 (not represented).

[c] Marginal slope of the fixed effect on the response.

[d] Significance is indicated as follows: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Table B.3: Indicator taxa for groups of vaginal and vulvar microbiota of girls

Sample groups[a]	Taxon	Classification level	Indicator value[b]	*p*-value[c]	Strength of association across groups[d]			
					Vagina pre	Vagina post	Vulva pre	Vulva post
Vag pre + vul pre	*Mobiluncus*	Genus	0.605	5.00E-03	0.275	0.032	0.575	0.111
Vul pre + vul post	*Segniliparus*	Genus	0.609	5.00E-03	0.163	0.173	0.390	0.475
	Murdochiella	Genus	0.590	5.00E-03	0.141	0.051	0.409	0.417
	Fusobacterium	Genus	0.577	5.00E-03	0.078	0.032	0.296	0.459
Vag pre + vul pre + vul post	*Prevotella*	Genus	0.904	5.00E-03	0.410	0.131	0.707	0.416
	Peptoniphilus	Genus	0.878	5.00E-03	0.386	0.106	0.706	0.383
	Anaerococcus	Genus	0.873	5.00E-03	0.463	0.121	0.684	0.296
	Dialister	Genus	0.818	5.00E-03	0.340	0.146	0.554	0.517
	Clostridiales	Order	0.805	5.00E-03	0.326	0.115	0.667	0.370
	Corynebacterium	Genus	0.782	5.00E-03	0.268	0.161	0.528	0.555
	Anaerosphaera	Genus	0.772	5.00E-03	0.333	0.090	0.624	0.343
	Clostridiales_Incertae_Sedis_XI	Family	0.748	5.00E-03	0.363	0.044	0.628	0.242
	Porphyromonas	Genus	0.682	5.00E-03	0.304	0.058	0.579	0.271
	Campylobacter	Genus	0.663	5.00E-03	0.301	0.051	0.577	0.227
	Gallicola	Genus	0.600	5.00E-03	0.255	0.024	0.548	0.200
	Facklamia	Genus	0.572	5.00E-03	0.264	0.005	0.522	0.169
	Streptococcus_Other	Genus	0.567	5.00E-03	0.264	0.129	0.444	0.242
	Varibaculum	Genus	0.562	5.00E-03	0.251	0.047	0.505	0.187
	Peptococcus	Genus	0.540	5.00E-03	0.271	0.038	0.420	0.192
	Porphyromonadaceae	Family	0.529	5.00E-03	0.258	0.044	0.447	0.190
	Anaerovorax	Genus	0.503	5.00E-03	0.199	0.044	0.463	0.188
Vag post + vul pre + vul post	Actinomycetales	Order	0.684	5.00E-03	0.157	0.215	0.472	0.481
	Staphylococcus	Genus	0.604	1.00E-02	0.149	0.202	0.356	0.473

[a] Groups of samples collected from girls are separated by premenarcheal (Pre) and postmenarcheal (Post) vagina (Vag) and vulva (Vul).

[b] Indicator value of each taxon for the associated group(s), calculated using "IndVal.g" with the "multipatt" function in the R package indicspecies.

[c] *p*-value for associated indicator value statistic (only taxa with $p \leq 0.01$ are listed).

[d] Association values between each taxon and group (equal to indicator value per group), calculated using the "strassoc" function of indicspecies with 100 bootstrap replicates.

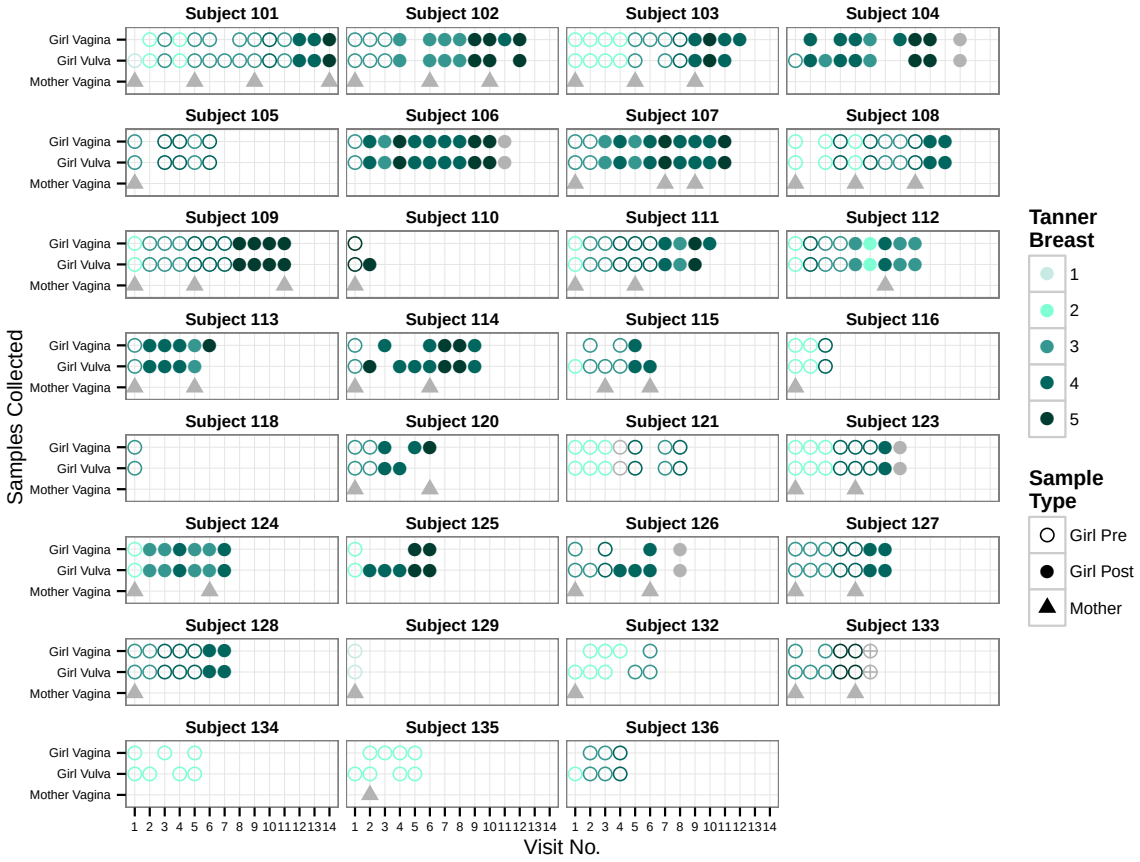

Figure B.1: Summary of all vaginal and vulvar samples collected from girls and mothers. Each panel shows all of the vaginal and vulvar samples collected from an individual participant (circles), as well as vaginal samples collected from her mother (triangles) when applicable. The x-axis indicates the clinical visit at which each sample was collected; visits occurred approximately every three months. Open circles signify premenarcheal status, and filled circles signify postmenarcheal status in girls. The menarcheal status was unknown for subject 133 at visit 6, indicated by an open circle with crosshatch. Points are color-coded to signify the Tanner breast stage of the girls as shown in the legend at top right (mother samples and those with missing values are colored gray).

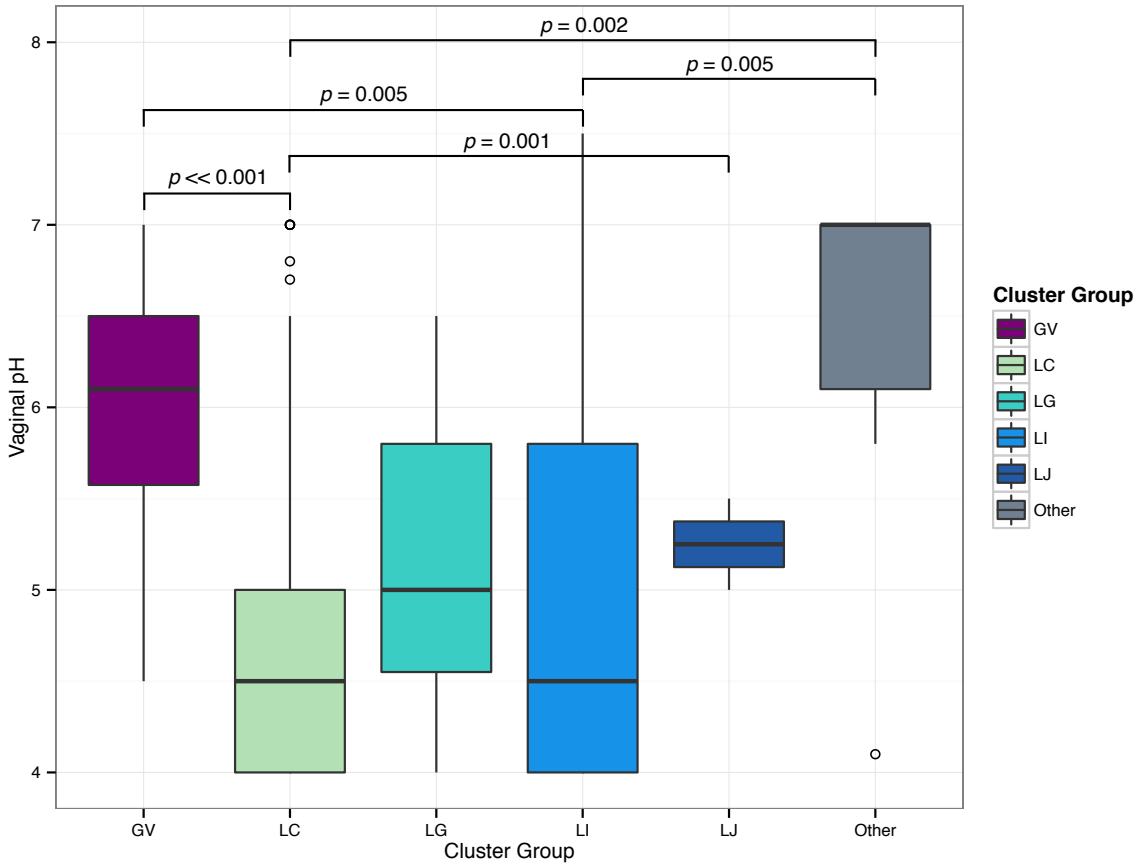

Figure B.2: Vaginal pH across hierarchical cluster groups. Vaginal microbiota were grouped into several hierarchical clusters listed in the legend at right (see also Figure 3.1). A multiple comparisons test using Tukey's method was used to identify significant differences in vaginal pH between clusters. *P*-values less than 0.05 are shown on the plot with a connector indicating the two groups being compared. Each box represents the interquartile range, the whiskers represent the upper and lower quartiles, the horizontal line represents the median, and open circles represent outliers.

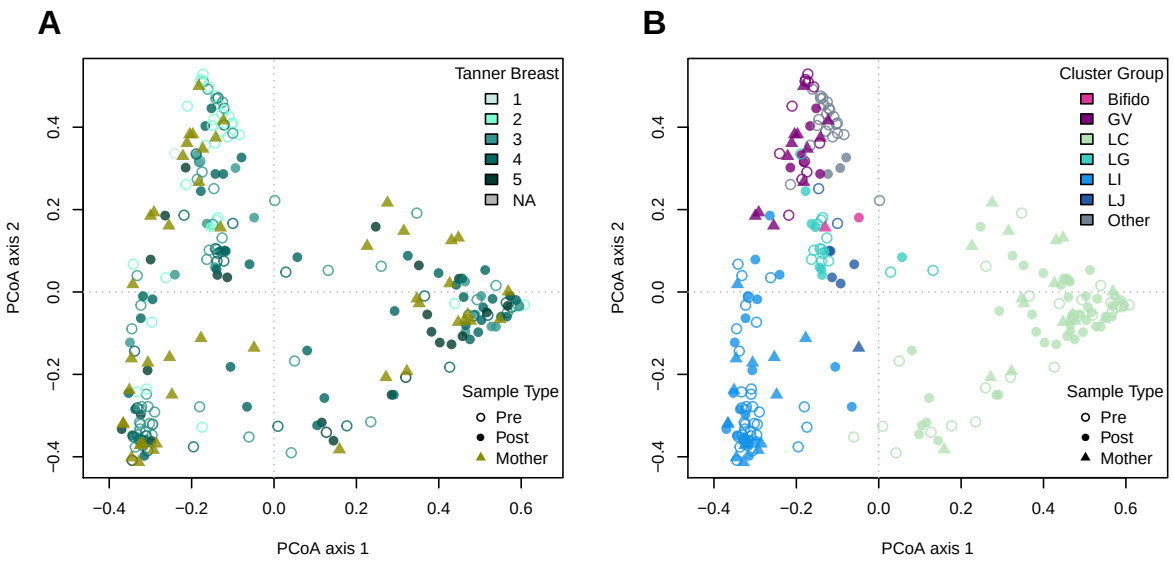

Figure B.3: PCoA of vaginal microbiota from girls and mothers. Principal Coordinates analysis (PCoA) was performed on the Bray-Curtis dissimilarity matrix computed from Hellinger-standardized taxon abundance data. Each point represents the vaginal microbiota sampled from a single individual at a single point in time (198 samples from girls and 47 from mothers). (A) Points color-coded according to Tanner breast stage (mother samples are colored green). (B) Points color-coded according to groups determined by UPGMA hierarchical clustering. After applying a Cailliez correction to adjust for negative eigenvalues, the corrected R^2-like ratios (essentially percent variance accounted for) for the first and second PCoA axes are 0.168 and 0.110 (16.8% and 11.0%), respectively.

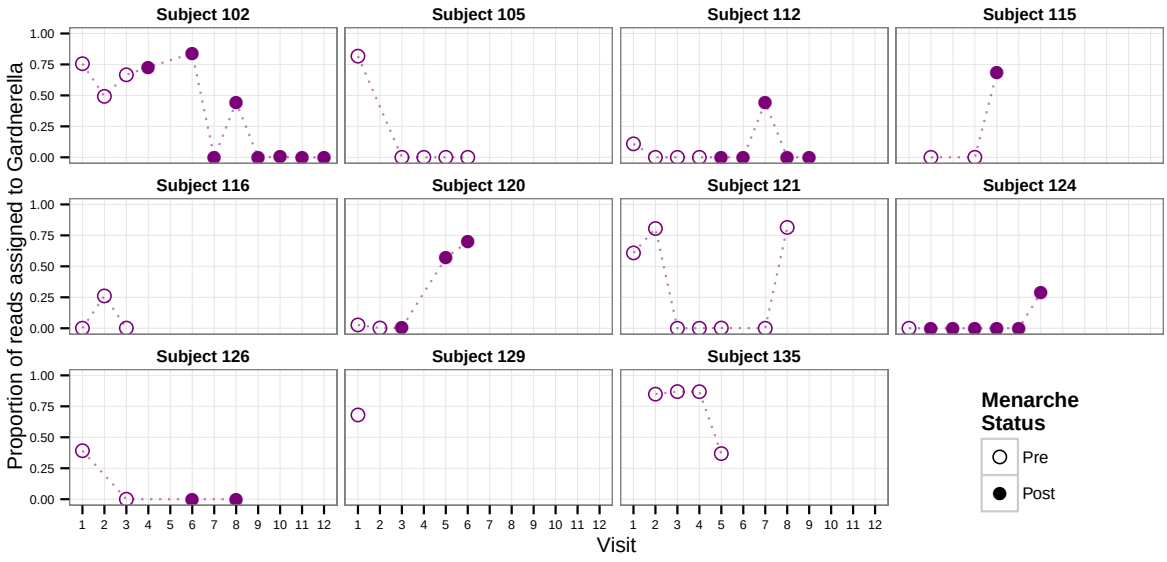

Figure B.4: Proportion of *Gardnerella* over time in the vaginal microbiota of 11 girls. *Gardnerella* was present in the vaginal microbiota at a relative abundance of 10% or greater at least once in 11/31 adolescent participants. Each panel shows the proportion of *Gardnerella* (encompassing sequence reads assigned to either the species level as *G. vaginalis* or genus level as *Gardnerella*) in the vaginal microbiota of a single participant at each clinical visit. Open circles represent premenarcheal samples, and filled circles represent postmenarcheal samples. The x-axis indicates the clinical visit at which each sample was collected; visits occurred approximately every three months.

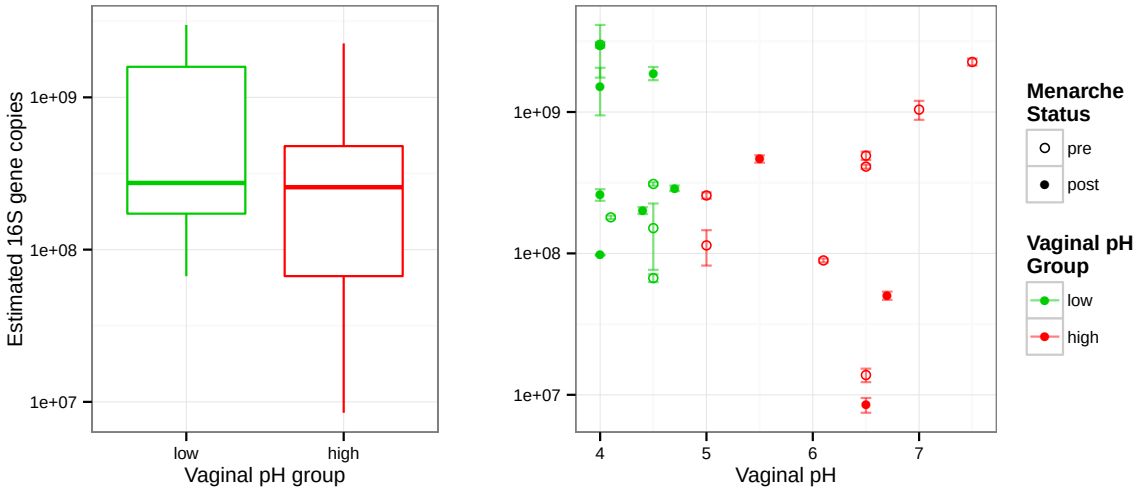

Figure B.5: Estimated number of 16S rRNA gene copies in low-pH vs. high-pH vaginal microbiota samples from girls. Pan-bacterial 16S rRNA qPCR was performed to estimate the number of gene copies per μL of genomic DNA in vaginal microbiota samples from girls. 24 samples were chosen at random from the subset of girl vaginal microbiota samples containing at least 75% lactic acid bacteria (LAB, including *Lactobacillus*, *Streptococcus*, *Aerococcus* and *Facklamia*); 12 samples had a vaginal pH < 5.0 ('low'), and 12 had a vaginal pH ≥ 5.0 ('high'). Each sample was prepared in triplicate, and a standard curve was generated from quantified and normalized standards. All samples from the 'low' group and 11/12 samples from the 'high' group successfully amplified. The left plot shows box plots for the estimated number of gene copies for the 'low' (green) and 'high' (red) groups. The right plot shows the vaginal pH (x-axis) and estimated number of gene copies (y-axis) for each sample with standard error bars. In this plot, open circles represent premenarcheal samples and filled circles represent postmenarcheal samples. A one-way *t*-test failed to detect a statistically significant difference between the mean estimates for 'low' and 'high' groups (p=0.14).

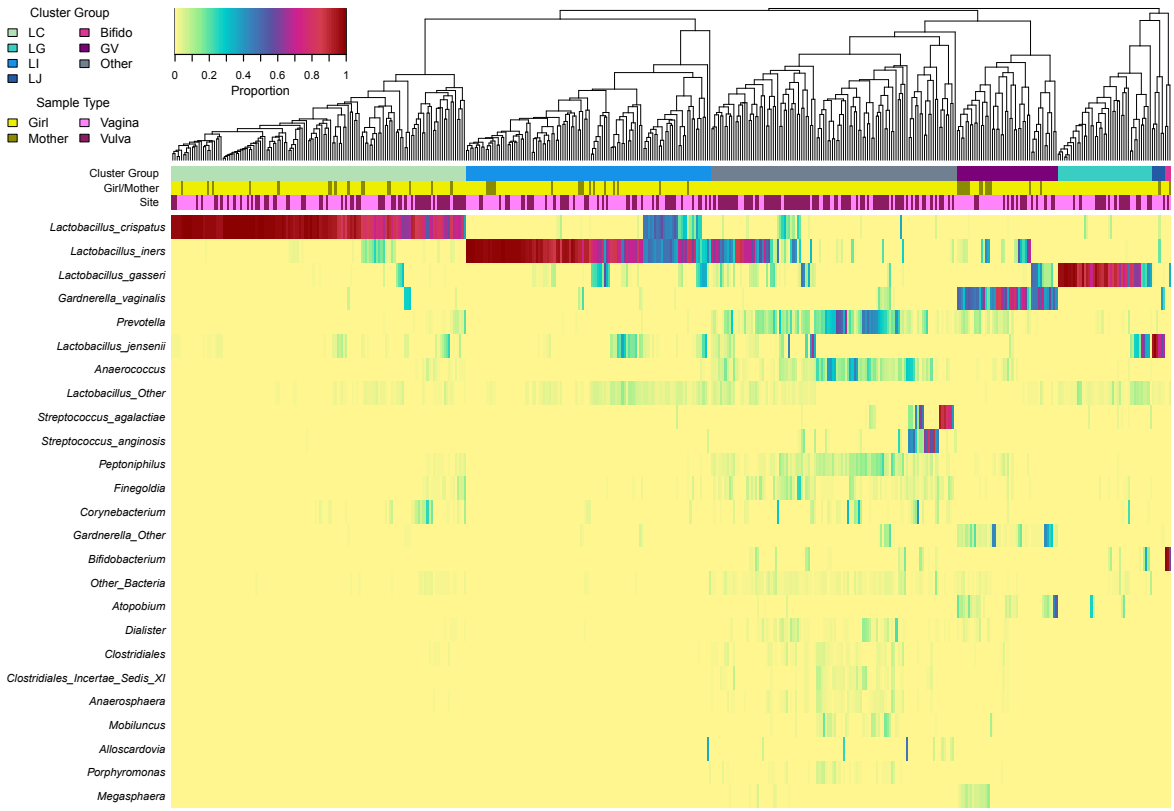

Figure B.6: Bacterial community composition of the vulvar and vaginal microbiota of girls and vaginal microbiota of mothers sampled longitudinally. Each column in the dendrogram and heatmap represents the vulvar or vaginal microbiota sampled from a single individual at a single point in time. In total 456 samples are represented: 198 vaginal samples and 211 vulvar samples from 31 girls, and 47 vaginal samples from 24 mothers. The dendrogram represents the average linkage (UPGMA) hierarchical clustering of samples based on the Bray-Curtis dissimilarity matrix computed from Hellinger standardized taxon abundance data. The colored bars below the dendrogram represent sample type (girl/mother, vagina/vulva) and hierarchical cluster assignments. Clusters are named to signify the most abundant taxon, when applicable: LC (*Lactobacillus crispatus* dominant, n=134), 'Other' (n=117), LI (*L. iners*, n=107), LG (*L. gasseri*, n=49), GV (*Gardnerella vaginalis*, n=47), and 'Bifido' (Bifidobacterium, n=3). The heatmap represents proportions (prior to Hellinger standardization) of the 25 overall most abundant taxa within each community as indicated by the legend at top right.

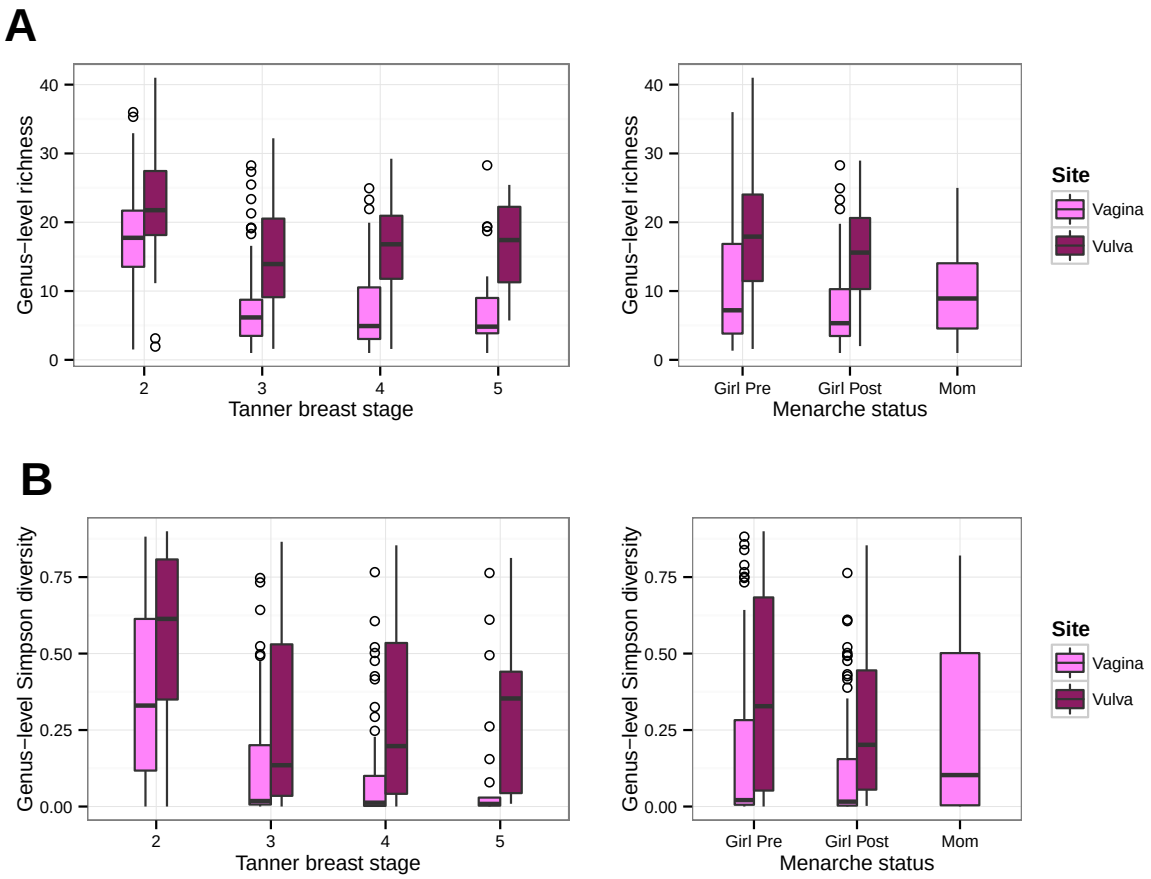

Figure B.7: Trends in genus-level richness and Simpson's diversity index of vaginal and vulvar microbiota with pubertal development and menarche status. Genus counts were subsampled at a depth of 2000 observations per sample and used to calculate richness (total number of genera observed) and Simpson's diversity index. In both cases all taxa that could not be classified to the genus level are combined in a single 'Other' category; richness and diversity are therefore underestimated. 415 samples are represented: 181 vaginal samples and 190 vulvar samples from 30 girls, and 44 vaginal samples from 23 mothers. (A) Genus richness in relation to Tanner breast stage (girls only) on the left; richness in relation to menarche status (girls and mothers) on the right. Vagina and vulva samples are represented by light pink and dark magenta coloring, respectively. (B) Simpson's diversity in relation to Tanner breast stage on left, and menarche status on right. Each box represents the interquartile range, the whiskers represent the upper and lower quartiles, the horizontal line represents the median, and open circles represent outliers.

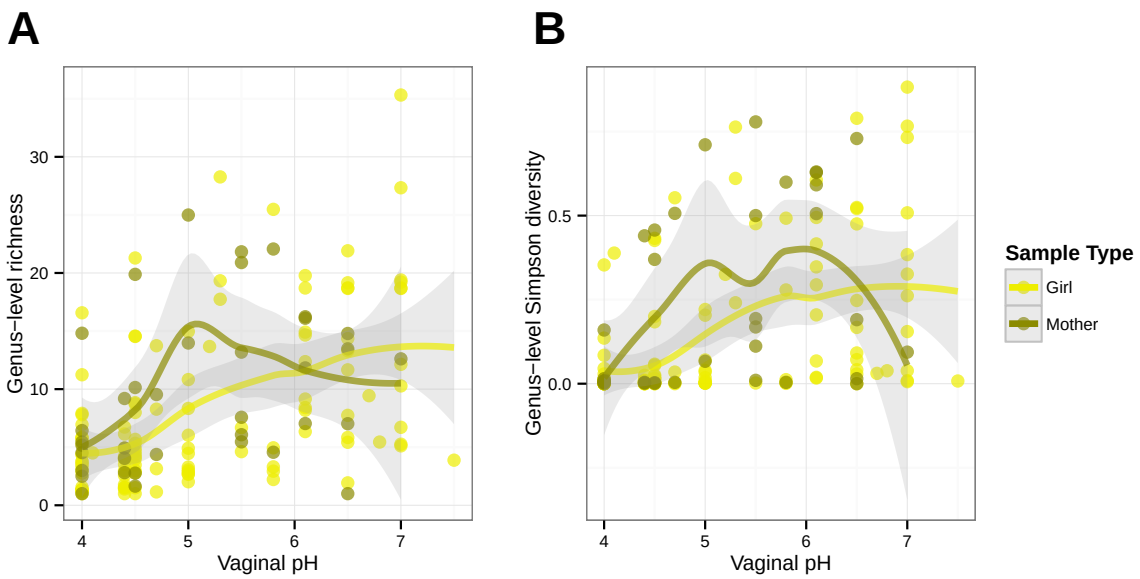

Figure B.8: Genus-level richness, Simpson's diversity index and vaginal pH. Each point represents a vaginal sample from either a girl or a mother. Vaginal pH was recorded for 120 samples from 24 girls and 37 samples from 21 mothers. (A) Genus-level richness plotted against vaginal pH. A locally weighted scatterplot smoothing (LOESS) function was applied separately to girl and mother data points, with 95% confidence intervals indicated by the light grey regions. (B) Genus-level Simpson's diversity index plotted against vaginal pH with LOESS curves for girls and mothers.

APPENDIX C

Supplementary Information to Chapter 4

C.1 supplementary tables and figures for Chapter 4

In addition to the three supplementary tables and two supplementary figures below, Supplementary File 3 (gene set enrichment results for *Gardnerella vaginalis*, XLSX) and Supplementary File 4 (gene set enrichment results for *Gardnerella vaginalis* and *Bifidobacterium* spp., XLSX) are available at `http://github.com/roxanahickey/dissertation`.

Table C.1: Genomic and clinical characteristics of 20 *Bifidobacterium* spp. strains

Strain	GenBank accession(s)	Genomic characteristics[a]					Clinical and phenotypic characteristics[b]	
		Size (Mb)	Contigs	Plasmids	GC%	CDS	Source	Specific Comments
Bifidobacterium adolescentis L2-32	AAXD00000000	2.389	35	0	59.2	2,031	GI tract	Infant feces
Bifidobacterium animalis subsp. lactis B420	CP003497.1	1.939	1	0	60.5	1,604	Unknown	NR
Bifidobacterium animalis subsp. lactis Bi-07	CP003498.1	1.939	1	0	60.5	1,601	Unknown	NR
Bifidobacterium animalis subsp. lactis V9	CP001892	1.944	1	0	60.5	1,613	GI tract	Infant feces
Bifidobacterium bifidum BGN4	CP001361.1	2.224	1	0	62.6	1,828	GI tract	Feces
Bifidobacterium breve 12L	CP006711.1	2.245	1	0	58.9	1,907	Mammary gland	Human milk
Bifidobacterium breve 689b	CP006715.1	2.332	1	0	58.7	1,991	GI tract	Infant feces
Bifidobacterium breve ACS-071-V-Sch8b	CP002743	2.327	1	0	58.7	1,972	Urogenital tract	NR
Bifidobacterium breve JCM 7017	CP006712.1	2.289	1	0	58.7	1,944	GI tract	Infant feces
Bifidobacterium breve JCM 7019	CP006713.1	2.359	1	0	58.6	2,073	GI tract	Adult feces
Bifidobacterium breve JCP7499	AWSX00000000	2.367	91	0	58.6	2,073	Urogenital tract	Vagina
Bifidobacterium breve NCFB 2258	CP006714.1	2.316	1	0	58.7	1,978	GI tract	Infant feces
Bifidobacterium breve S27	CP006716.1	2.294	1	0	58.7	1,948	GI tract	Infant feces
Bifidobacterium breve UCC2003	CP000303	2.423	1	0	58.7	2,071	GI tract	Nursing infant feces
Bifidobacterium longum NCC2705	AE014295, AF540971	2.26	1	1	60.1	1,868	GI tract	Infant feces
Bifidobacterium longum subsp. infantis 157F	AP010890, AP010891, AP010892	2.409	1	2	60.1	2,058	GI tract	NR
Bifidobacterium longum subsp. longum BBMN68	CP002286	2.266	1	0	59.9	1,882	GI tract	Long-lived adult male
Bifidobacterium longum subsp. longum F8	FP929034	2.385	1	0	59.6	1,997	GI tract	NR
Bifidobacterium longum subsp. longum JCM 1217	AP010888	2.385	1	0	60.3	1,998	GI tract	NR
Bifidobacterium thermophilum RBL67	CP004346.1	2.292	1	0	60.1	1,839	GI tract	Infant feces

[a] Genomes were downloaded from the PATRIC database in March 2015 (ftp://ftp.patricbrc.org/patric2). CDS = coding DNA sequence.

[b] Gathered from information available in PATRIC (`http://patricbrc.org`).

Table C.2: Gene Ontology categories differentially represented in *Gardnerella* compared to *Bifidobacterium* spp.

Category	Odds ratio[a]	*q*-value [b]	
GO:0004617 phosphoglycerate dehydrogenase activity	0.025	7.81E-09	***
GO:0004565 β-galactosidase activity	0.106	1.20E-20	***
GO:0004834 tryptophan synthase activity	0.146	1.98E-06	***
GO:0008982 protein-N(PI)-phosphohistidine-sugar phosphotransferase activity	0.157	1.47E-10	***
GO:0003961 O-acetylhomoserine aminocarboxypropyltransferase activity	0.167	5.82E-06	***
GO:0004557 α-galactosidase activity	0.202	1.75E-04	***
GO:0004129 cytochrome-c oxidase activity	0.251	2.38E-02	*
GO:0003848 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase activity	0.264	3.63E-02	*
GO:0004069 aspartate transaminase activity	0.515	1.97E-04	***
GO:0004558 α-glucosidase activity	0.520	2.61E-02	*
GO:0003887 DNA-directed DNA polymerase activity	1.509	1.41E-03	**
GO:0008444 CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase activity	1.967	2.37E-03	**
GO:0008955 peptidoglycan glycosyltransferase activity	2.052	3.46E-02	*
GO:0008477 purine nucleosidase activity	2.109	3.14E-04	***
GO:0004616 phosphogluconate dehydrogenase (decarboxylating) activity	2.194	2.12E-02	*
GO:0004022 alcohol dehydrogenase activity	2.208	1.37E-02	*
GO:0009002 serine-type D-Ala-D-Ala carboxypeptidase activity	2.391	2.76E-03	**
GO:0008930 methylthioadenosine nucleosidase activity	2.403	6.12E-03	**
GO:0008782 adenosylhomocysteine nucleosidase activity	2.403	6.12E-03	**
GO:0008834 di-trans,poly-cis-decaprenylcistransferase activity	2.672	4.29E-02	*
GO:0008829 dCTP deaminase activity	2.672	4.29E-02	*
GO:0004308 exo-α-sialidase activity	3.593	7.88E-04	***
GO:0004013 adenosylhomocysteinase activity	4.198	2.40E-02	*
GO:0008897 phosphopantetheinyltransferase activity	4.762	2.01E-03	**
GO:0004512 inositol-3-phosphate synthase activity	5.546	7.52E-04	***
GO:0008677 2-dehydropantoate 2-reductase activity	5.884	4.87E-04	***
GO:0050118 N-acetyldiaminopimelate deacetylase activity	22.686	3.88E-05	***
GO:0009001 serine O-acetyltransferase activity	29.436	1.29E-06	***
GO:0004850 uridine phosphorylase activity	Inf	2.50E-08	***
GO:0004496 mevalonate kinase activity	Inf	1.13E-06	***
GO:0004452 isopentenyl-diphosphate delta-isomerase activity	Inf	1.13E-06	***
GO:0004421 hydroxymethylglutaryl-CoA synthase activity	Inf	3.53E-07	***
GO:0004420 hydroxymethylglutaryl-CoA reductase (NADPH) activity	Inf	1.22E-05	***
GO:0004163 diphosphomevalonate decarboxylase activity	Inf	5.82E-06	***
GO:0004124 cysteine synthase activity	Inf	7.29E-08	***
GO:0004105 choline-phosphate cytidylyltransferase activity	Inf	4.02E-02	*
GO:0004103 choline kinase activity	Inf	4.02E-02	*
GO:0003988 acetyl-CoA C-acyltransferase activity	Inf	3.53E-07	***
GO:0003985 acetyl-CoA C-acetyltransferase activity	Inf	3.53E-07	***

[a] Odds ratio (OR) >1, overrepresentation; OR < 1, underrepresentation; OR = Inf (Infinite), category unique to group; 78 GO categories had OR=0 (i.e., category missing from *Gardnerella*) and are not listed here.

[b] False discovery rate adjusted *p*-value (obtained by Fisher's exact test). Significance < 0.05 *, < 0.01 **, < 0.001 ***

Table C.3: KEGG biochemical pathways differentially represented in *Gardnerella* compared to *Bifidobacterium* spp.

Category	Odds ratio[a]	*q*-value[b]	
930 Caprolactam degradation	0	1.00E-03	**
460 Cyanoamino acid metabolism	0	3.71E-27	***
410 β-Alanine metabolism	0	1.24E-05	***
364 Fluorobenzoate degradation	0	3.02E-02	*
1040 Biosynthesis of unsaturated fatty acids	0	6.07E-06	***
290 Valine, leucine and isoleucine biosynthesis	0.132	3.24E-38	***
660 C5-Branched dibasic acid metabolism	0.226	4.17E-15	***
4150 mTOR signaling pathway	0.229	4.77E-04	***
531 Glycosaminoglycan degradation	0.342	4.29E-15	***
750 Vitamin B6 metabolism	0.466	4.87E-03	**
500 Starch and sucrose metabolism	0.526	2.15E-20	***
600 Sphingolipid metabolism	0.532	6.44E-08	***
340 Histidine metabolism	0.604	4.68E-07	***
910 Nitrogen metabolism	0.619	2.03E-03	**
561 Glycerolipid metabolism	0.634	3.27E-03	**
940 Phenylpropanoid biosynthesis	0.647	2.20E-03	**
790 Folate biosynthesis	0.652	1.74E-03	**
53 Ascorbate and aldarate metabolism	0.664	9.02E-08	***
260 Glycine, serine and threonine metabolism	0.676	1.55E-06	***
401 Novobiocin biosynthesis	0.687	2.45E-02	*
330 Arginine and proline metabolism	0.705	1.40E-05	***
604 Glycosphingolipid biosynthesis - ganglio series	0.709	3.79E-03	**
400 Phenylalanine, tyrosine and tryptophan biosynthesis	0.744	4.77E-04	***
760 Nicotinate and nicotinamide metabolism	0.76	1.82E-02	*
52 Galactose metabolism	0.778	1.98E-06	***
20 Citrate cycle (TCA cycle)	0.867	1.63E-02	*
30 Pentose phosphate pathway	1.182	4.66E-06	***
190 Oxidative phosphorylation	1.271	4.51E-02	*
230 Purine metabolism	1.274	1.68E-07	***
564 Glycerophospholipid metabolism	1.278	4.39E-02	*
860 Porphyrin and chlorophyll metabolism	1.37	3.70E-03	**
240 Pyrimidine metabolism	1.37	8.17E-09	***
650 Butanoate metabolism	1.387	4.42E-03	**
550 Peptidoglycan biosynthesis	1.41	4.82E-05	***
970 Aminoacyl-tRNA biosynthesis	1.502	1.83E-09	***
983 Drug metabolism - other enzymes	1.548	1.09E-03	**
231 Puromycin biosynthesis	1.641	1.64E-02	*
626 Naphthalene and anthracene degradation	1.672	3.70E-03	**
592 α-Linolenic acid metabolism	1.76	4.43E-03	**
642 Ethylbenzene degradation	1.908	1.44E-03	**
982 Drug metabolism - cytochrome P450	2.217	1.37E-02	*
980 Metabolism of xenobiotics by cytochrome P450	2.217	1.37E-02	*
281 Geraniol degradation	5.4	1.16E-03	**

[a] Odds ratio (OR) >1, overrepresentation; OR < 1, underrepresentation; OR=0, pathway absent from group; OR = Inf (Infinite), pathway unique to group.

[b] False discovery rate adjusted *p*-value (obtained by Fisher's exact test). Significance < 0.05 *, < 0.01 **, < 0.001 ***

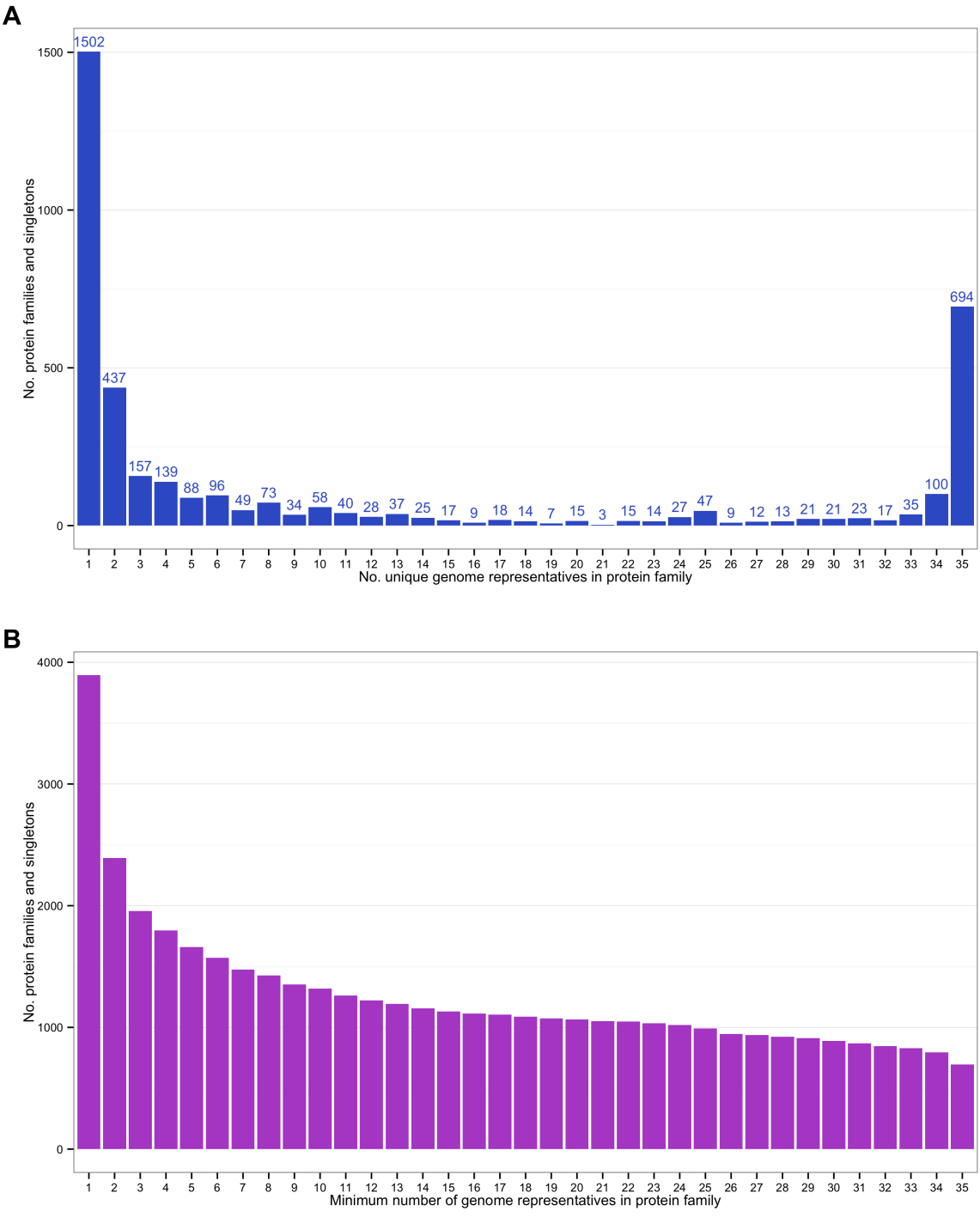

Figure C.1: Protein family counts among *G. vaginalis* genomes. 44,505 protein-coding gene sequences from 35 *G. vaginalis* genomes were clustered into 2,399 protein families and 1,495 singletons based on >70% similarity in amino acid sequence. (A) Number of singletons or protein families present in exactly *n* genomes shown on the x-axis. (B) Number of singletons and protein families present in at least *n* genomes shown on the x-axis.

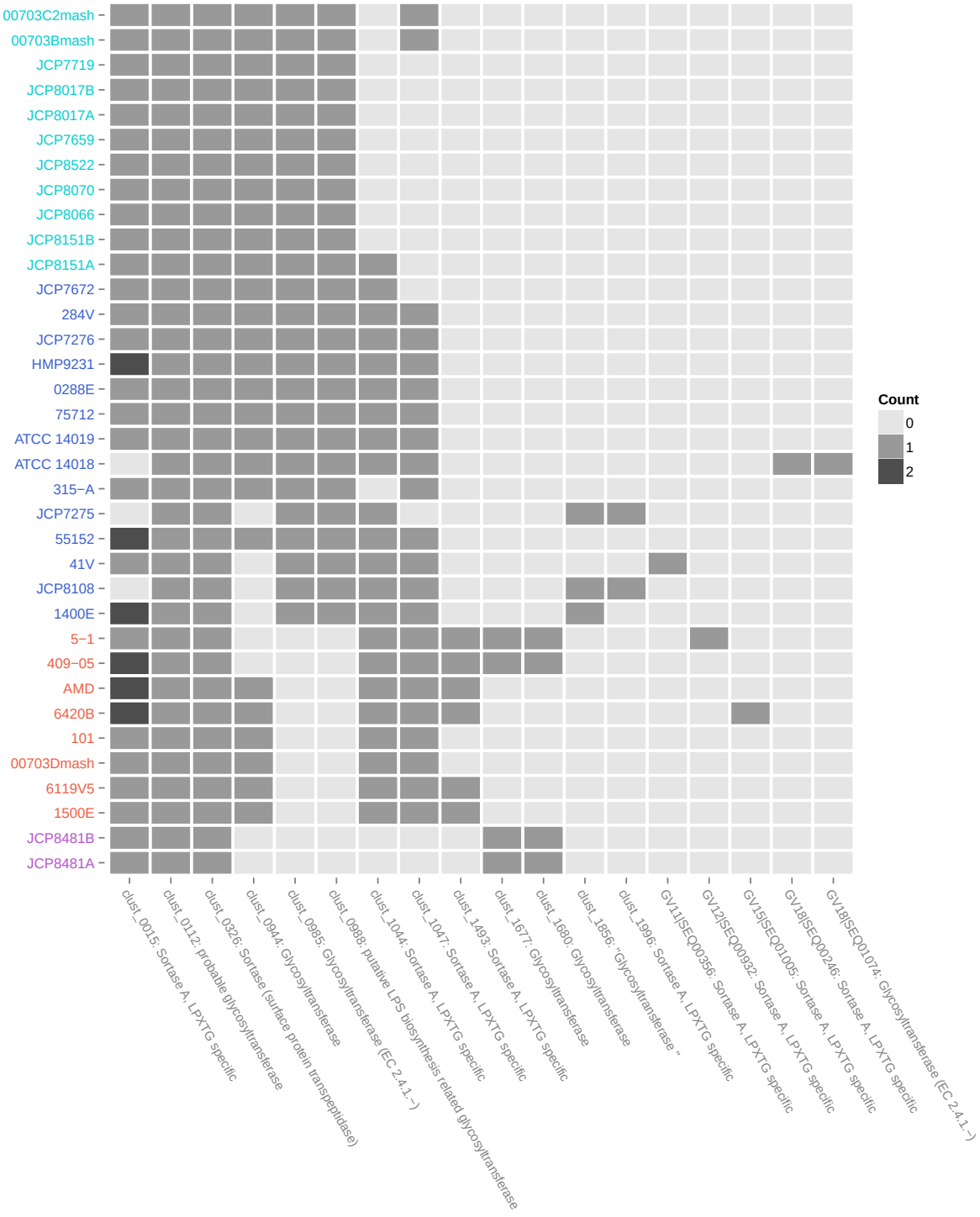

Figure c.2: Prevalence of glycosyltransferase and sortase protein families in *G. vaginalis*. The heatmap reflects the count of coding sequences in each genome assigned to a particular OrthoMCL protein family. The strain names of 35 *G. vaginalis* genomes are listed along the y-axis and colored by genome cluster as shown in Figure 4.3. The most prevalent annotations for each protein family are listed along the x-axis with the OrthoMCL cluster identifier.