Spring 5-15-2022

Candida Species and Bacterial Strains in the Vaginal Microbiome

Brett Tortelli
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS
Division of Biology and Biomedical Sciences
Molecular Microbiology and Microbial Pathogenesis

Dissertation Examination Committee:
Justin C. Fay, Chair
Amanda A. Lewis, Co-Chair
Gautam Dantas
Sarah K. England
Andrew L. Kau

Candida Species and Bacterial Strains in the Vaginal Microbiome

by

Brett A. Tortelli

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2022
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I would like to thank my mentors Justin and Amanda for all of their support and guidance over the course of my thesis work. I would also like to thank the members of my thesis committee for pushing me outside of my comfort zone and working with me to develop my research ideas. A special thank you to all of the members of the Lewis and Kau labs as well as my MSTP classmates for their comradery and advice over the past few years. I would like to acknowledge the CHOICE Project, WIHSC and GAPPS for the clinical samples and data. And lastly a big thank you to all of the women who provided samples used in this thesis work.

This work was supported in part by the March of Dimes Prematurity Research Center at Washington University in St. Louis, the Center for Women's Infectious Disease Research at Washington University in St. Louis and by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the NIH: F30HD094435. The sponsors had no role in the study design, the collection, analysis or interpretation of data, or the decision to submit the article for publication. The content of this thesis is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

Brett Tortelli
Washington University in St. Louis
May 2022

ABSTRACT OF THE DISSERTATION

Candida Species and Bacterial Strains in the Vaginal Microbiome

by

Brett A. Tortelli

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Microbiology and Microbial Pathogenesis

Washington University in St. Louis, 2022

Justin C. Fay, Chair

Amanda L. Lewis, Co-chair

The vaginal environment is a dynamic ecosystem, hosting various microbial species from diverse taxa including bacteria, fungi and viruses. The composition of bacteria within the vaginal microbiome has gained a lot of recent attention and has been associated with reproductive health and disease. The vaginal microbiome of healthy reproductive-aged women is frequently dominated by *Lactobacillus* species and has low species diversity when compared to other anatomic sites. The composition of the bacterial community is often described in terms of five common community types. Four of the five community types are dominated by a single *Lactobacillus* species (*L. crispatus*, *L. iners*, *L. jensenii* or *L. gasseri*). The fifth community type is characterized by a lack of *Lactobacillus* dominance and often exhibits greater community diversity. While recent studies have identified associations between community type and disease, these findings have not been consistent across study populations. In light of these inconsistencies, it is important to consider other aspects of the microbial community that may explain this variance. In this dissertation, I expound on *Candida* colonization and bacterial strain

variation and their relationships with the vaginal microbiome to provide a framework in which future studies may more holistically evaluate the microbiome and its associations with reproductive health.

Candida species frequently colonize the human vagina and colonization may lead to vulvovaginal candidiasis, a significant source of morbidity among women of reproductive age. Furthermore, recent work has suggested that *Candida* colonization may contribute to preterm birth. Despite this, the relationships between *Candida* and bacteria in the vagina are not well understood. To address this gap in knowledge, I designed a nested cohort study using vaginal swab specimens collected from nonpregnant women as part of the Contraceptive CHOICE Project. I then characterized the bacterial composition of the vaginal microbiome using 16S ribosomal profiling and determined *Candida* colonization status using a qPCR assay. I showed that women with *L. iners*-dominant microbiomes were more likely to harbor *Candida* than women with *L. crispatus*-dominant microbiomes. To identify a mechanism for this clinical observation, I evaluated the potential of cell-free supernatants from *L. crispatus* and *L. iners* cultures to inhibit *Candida* growth *in vitro*. *L. crispatus* produced greater concentrations of lactic acid and exhibited significantly more pH-dependent growth inhibition of *C. albicans*. Thus, not all *Lactobacillus*-dominant communities are equally associated with *Candida* colonization and lactic acid production may drive individual species relationships with *Candida*. This work provides additional evidence that *L. iners*-dominant communities are more permissive to vaginal colonization with potential pathogens, including *Candida*. As *Candida* correlates with bacterial community type, I conclude that the incorporation of *Candida* in future studies of the vaginal

microbiome may lead to a better understanding of the relationships between vaginal microbiome and gynecologic health.

Strain differences in vaginal bacteria are believed to be important to reproductive disease, but little is known about the diversity, structure and evolutionary history of vaginal strains. To characterize strain variation in the vagina, I developed a metagenomic approach that utilizes core-genome SNPs to characterize strain variation in the microbiome for six commonly abundant species of bacteria: *G. vaginalis*, *L. crispatus*, *L. iners*, *L. jensenii*, *L. gasseri*, and *A. vaginae*. I showed that with the exception of *L. iners*, strains for all of the species cluster into multiple distinct groups. I also showed that strain diversity is lowest among *Lactobacillus* species. This supports the idea that *Lactobacillus* dominance may have evolved more recently during human evolution and that a diverse community type with abundant *G. vaginalis* and *A. vaginae* may reflect an ancestral state that is more akin to the communities found among other primates. Interestingly, I observed that species diversity is related to strain diversity within a community which indicates that the ecological forces influencing diversity may be similar at both the species and strain levels. Together this work documents a method of characterizing strain variation and provides motivation for the incorporation of strain analysis into future studies of the microbiome. If strain differences reflect functional differences that contribute to either protective or virulent phenotypes, identifying strain group associations with disease may clarify existing discrepancies in the field.

Chapter 1: Introduction

The human body hosts diverse microbial communities that include bacteria, fungi and viruses.[1-3] The term microbiome is often employed to refer specifically to the communities of bacteria that live on and in the human body. Sequencing based surveys of the microbiome have demonstrated that community composition varies by anatomic site and that these site-specific microbial profiles may differ between individual.[2,4] Inter-individual variation in the microbial composition has in many cases been linked with disease (e.g. cancer, inflammatory bowel disease and obesity), motivating further study of the relationship between the microbiome and health.[5]

Like other mucosal sites, the human vagina hosts various microbial species from diverse taxa including bacteria, fungi and viruses.[6-8] Importantly, little is known about many of the organisms that inhabit the vaginal niche. By characterizing the composition and structure of the vaginal microbiome, we have begun to refine our understanding of the ways in which microbial patterns relate to gynecologic and obstetric disease.

1.1 Bacterial composition of the "healthy" and dysbiotic vaginal microbiome

Early studies of the vaginal microbiome were limited by their dependence on culture and microscopic examination of specimens, but recent advances in sequencing technologies have led to a rapid expansion of our knowledge pertaining to the diversity and composition of the microbiota that reside in the human vagina. Both 16S ribosomal profiling and metagenomic shotgun sequencing have shown the vaginal microbiome of healthy reproductive-aged women are frequently dominated by *Lactobacillus* species and have low species diversity when compared to other anatomic sites such as the gut or oral cavity.[2,4] Many different species of *Lactobacillus* are found in human microbial communities,[9,10] but there are four that frequently

dominate the human vagina: *L. crispatus*, *L. iners*, *L. jensenii* and *L. gasseri*.[6,11] In a 2011 study, 16S community profiling of nearly 400 asymptomatic North American women showed that the community composition of the vaginal microbiome could be clustered into five community state types (CSTs).[11] Four of these CSTs were dominated by a single species of *Lactobacillus*, either *L. crispatus*, *L. iners*, *L. jensenii* or *L. gasseri*. The fifth CST (often referred to as Diverse) had fewer *Lactobacillus* and contained a greater abundance of fastidious and anaerobic bacteria and overall higher diversity. Subsequent studies support these five community types.[6]

The "healthy" vaginal microbiome is often viewed through the paradigm of *Lactobacillus* dominance and acidic vaginal pH (<4.5). It is believed that the dominance of *Lactobacillus* species in the vagina provides some protective benefit against pathogens in part through the production of lactic acid and acidification of the vaginal environment.[12,13] A lack of lactobacilli and an abundance of obligate or facultative anaerobes such as *Gardnerella*, *Prevotella* and *Atopobium* is believed to be dysbiotic and is characteristic of bacterial vaginosis (BV).[14,15]

BV is the most common cause of vaginal discharge among reproductive-aged women and while symptoms of BV may include vaginal discharge and odor, it is common for women not to report symptoms,[16,17] BV is often diagnosed by clinical features (Amsel criteria)[18] or by Nugent scoring,[19] a 0-10 scale generated by scoring bacterial morphotypes on Gram stained vaginal smears. A Nugent score of 0-3 is classified as normal (*Lactobacillus* morphotype), a score of 4-7 is classified as intermediate flora and a score of 8-10 is classified as BV. High Nugent score is correlated with greater community diversity and elevated vaginal pH (> 4.5).[11,20] While BV is associated with Diverse communities, it may also occur in *Lactobacillus*-dominant communities,

particularly those dominated by *L. iners*.[11,21] BV is also associated with race/ethnicity which reflects correlations observed between race/ethnicity and the vaginal microbiome.[6,11,14,15,22,23]

Several studies have reported significantly greater bacterial diversity in the microbiomes of Black women when compared to White women.[11,22,24] In terms of community types, both Black and Hispanic women are most likely to have Diverse type communities, while Asian and White women are most likely to harbor communities dominated by *Lactobacillus*.[11] Interestingly, the species of *Lactobacillus* that dominates a community also covaries by race/ethnicity. Among women with *Lactobacillus*-dominant community types, *L. crispatus*-dominant communities are most common among White women, while *L. iners*-dominant communities are most common among Black, Hispanic and Asian women.[11] Studies of European, African and Asian women have largely corroborated these findings outside of North America.[25]

Recently, a few observations have brought into question the idea that *Lactobacillus* dominance and a vaginal pH less than 4.5 are essential to a healthy vaginal microbiome. Many asymptomatic women host Diverse communities with a less acidic vaginal pH (>4.5).[11] Additionally, the vaginal pH associated with different *Lactobacillus*-dominant microbiomes varies and not all women with *Lactobacillus*-dominant communities have a low vaginal pH (<4.5).[13] Lastly, a substantial proportion (approximately 40%) of Black and Hispanic women would be considered to have "unhealthy" microbiota.[11] Thus, our current understanding of what constitutes a "healthy" vaginal microbiome may reflect a Eurocentric perspective that does not hold true in all populations.

Vaginal bacterial community composition not only varies between individuals and populations but may show temporal intraindividual changes as well. Most studies have focused on cross-sectional study designs; however, the few studies that have incorporated longitudinal analysis have shown that community composition can change dramatically over time.[26-28] These longitudinal studies have shown that the microbiome may be influenced by hormonal changes such as those seen during the course of the menstrual cycle and pregnancy.[6]

The stability of the vaginal microbiome changes throughout the menstrual cycle and is least stable during menses.[26] Menstruation can cause rapid changes in community compositions that likely result from changes to the vaginal environment and availability of nutrients (e.g. iron). During menses, an increase in the abundance of *G. vaginalis* and *L. iners* can be observed. During this same period the abundance of other *Lactobacillus* species declines but often rebounds after menstruation is complete.[26,27] In comparison, studies have indicated that the microbiome is more stable and less diverse during pregnancy with *Lactobacillus*-dominant communities being more prevalent.[29] Interestingly, the relative stability of a community appears be related to its composition, as Diverse and *L. iners*-dominant communities are less stable than other community types.[26,28] The stability of the community may have implications for vaginal infection and reproductive health.

1.2 The vaginal microbiome and disease

The vaginal microbiome is important for maintaining gynecologic and obstetric health. One way in which a healthy vaginal microbiome may influence reproductive health is by protecting against sexually transmitted infections (STIs). Vaginal dysbiosis (BV) has been linked to the

transmission of several sexually transmitted infections.[16,30-33] Profiling of the vaginal microbiome using sequencing techniques have supported the link between the vaginal microbiome and STIs, finding associations between bacterial community composition and important sexually transmitted infections like HIV and human papilloma virus (HPV).[25,34,35] These studies have shown that the risk of HIV infection is greater among women with Diverse communities when compared to women with communities dominated by *Lactobacillus* species (particularly *L. crispatus*).[34,36,37] Interestingly, both Diverse and *L. iners*-dominant communities are associated with HPV infection and HPV-related cervical intraepithelial neoplasia (CIN), while other *Lactobacillus*-dominant communities have been linked to quicker viral clearance.[34] In addition to STIs, the vaginal microbiome has been implicated as a contributor to female infertility and other obstetric complications.[38,39]

Preterm birth (PTB) is defined as a live birth prior to 37 weeks of gestational age. It is the leading contributor to infant mortality and morbidity around that world and is estimated to occur in about 10% of all births worldwide.[40] There are clear racial/ethnic disparities when it comes to the prevalence of PTB, with Black women being about twice as likely to deliver preterm when compared to White women.[41] The etiology of PTB varies and is often unknown; however, up to 40% of preterm deliveries have been linked to intrauterine infections.[42-44] It is important to understand that infections can occur at any time during pregnancy and may be asymptomatic.[45,46] Studies of vaginal bacteria during pregnancy have shown that women with BV are at increased risk for delivering prematurely.[46] Although anaerobic bacteria commonly associated with BV are thought to increase a woman's risk of PTB, treating asymptomatic BV does not decrease the

rates of premature delivery.[47] This suggests that the link between BV and PTB may reflect a more complex polymicrobial etiology.

Numerous studies have used sequencing technologies to investigate the associations between the microbiome and PTB, with mixed results. While most studies have shown a correlation between greater community diversity and preterm delivery,[48-50] two reported lower diversity among women delivering preterm.[51,52] The relationship between community type and PTB may be dependent on study population, as Diverse communities and *G. vaginalis* abundance have been correlated with preterm delivery among White women but not Black women.[28,29,53] Interestingly, *G. vaginalis* strains can be grouped into multiple distinct clades and recent analysis suggests that not all clades are correlated with preterm delivery.[53,54] Furthermore, a higher prevalence of non-PTB associated *G. vaginalis* clades has been observed in Black women which may confound the relationship between *G. vaginalis* and preterm delivery at the species level.[53] These findings suggest that strain differences may be important in PTB. Notably, to date studies have been limited by small sample sizes and few have sampled the microbiome longitudinally during pregnancy. Additional work will be needed to clarify the relationships between PTB, race/ethnicity and the microbiome.

1.3 *Candida* and the vaginal microbiome

When it comes to the vaginal microbiome, investigators have primarily focused on resident bacteria. Few studies have also looked at the contributions of fungal taxa to the vaginal microbial community. The lack of investigation into the role fungi play as part of the vaginal microbiome is surprising as asymptomatic yeast colonization is common. The point prevalence of vaginal

yeast colonization has been reported to be around 20%.[7] Yeast colonization can often be transient and an estimated 70% of women are colonized at some point over the course of a single year.[55] A 2013 study of asymptomatic nonpregnant Estonian women of reproductive age characterized the composition of vaginal bacteria and fungi by sequencing the internal transcribed spacer (ITS) of fungal ribosomes and 16S ribosomal sequences from bacteria.[56] This was the first study to characterize the so-called mycobiome colonizing the healthy vagina using barcoded pyrosequencing technology. The study classified 196 functional operational taxonomic units (OTUs), 16 of which belonged to different *Candida spp.* suggesting a more diverse fungal community than previously recognized.[56] However, failures to adequately control for laboratory contamination and differentiate sequencing artifacts are of significant concern when interpreting the fungal profiling presented in this study. As such, robust study of the vaginal mycobiome and its diversity remains to be completed.

Vaginal *Candida* colonization may lead to vulvovaginal candidiasis (VVC), characterized by an aggressive host response to *Candida* overgrowth.[57] A recent survey indicated that up to 40% of women with vaginal complaints presenting to a primary care facility were diagnosed with VVC. It is thought that a woman's lifetime risk of developing VVC is about 75% and an estimated 5% of women with VVC will experience reoccurring infections.[58] Many studies have investigated the host immunologic response and pathogenesis of fungi in the vagina; however, few have considered VVC in relation to vaginal bacteria. The studies that have considered vaginal bacteria have depended mostly on culture-based approaches, which we know to be inadequate in representing the diversity of taxa present. A notable exception is a study that evaluated the bacterial community composition in Chinese women with or without VVC.[59] The study found

that the bacterial communities were more diverse in women with VVC when compared to normal women, but women with VVC maintained higher abundances of *Lactobacillus spp.* compared to women with BV. Women with VVC but no concurrent BV were found to have higher abundances of BV-associated bacteria (*Gardnerella*, *Prevotella* and *Atopobium*). Thus, women with VVC have bacterial communities that are more diverse when compared to healthy women. While this study showed an association between bacterial diversity and VVC, others have found no association between bacterial community type and VVC.[60,61] These conflicting findings highlight the need for additional study to clarify the relationship between VVC and the compositions of the microbiome.

Women are more frequently colonized by *Candida* during pregnancy possibly due to increased circulating estrogens and elevated deposition of glycogen and other substrates in the vaginal environment.[7,62] The extent of health consequences of vaginal *Candida* during pregnancy are yet to be fully elucidated, but increased rates of colonization likely contribute to the association of VVC with pregnancy.[63] In rare cases, *Candida* may even cause chorioamnionitis that can lead to premature rupture of membranes and preterm birth.[64] The relationship between *Candida* and PTB remains unclear. Population based studies in Hungary have reported a 34% to 64% decrease in preterm birth when women with candidiasis were treated with clotrimazole.[65-67] A meta-analysis of two clinical studies revealed the treatment of asymptomatic *Candida* colonization resulted in an overall reduction in spontaneous PTB with a RR=0.36 and a 95% CI=0.17 to 0.75.[68-70] However, two large cohort studies from U.S. showed no association between the presence of moderate to heavy *Candida* growth and PTB among women at 22 to 30 weeks of

gestation.[71,72] These findings indicate that *Candida* may contribute to PTB risk either directly or through some other indirect mechanism, possibly by altering the dynamics of the microbiome.

1.4 Goals of this thesis

While there is mounting evidence for the importance of the vaginal microbiome to gynecologic and obstetric health, there remain significant gaps in our understanding of the ecology and structure of the vaginal microbiome:

1) Women are frequently vaginally colonized by *Candida*, but the relationship between *Candida* colonization and the bacterial community composition of the microbiome remains unclear.

2) Associations between individual bacterial species and reproductive disease may be strain dependent, but our understanding of the structure and diversity of strain among dominant vaginal bacterial species remains limited.

In this thesis, I will address these gaps in our current knowledge. In **Chapter 2**, I show that reproductive aged women are frequently vaginally colonized by *Candida* and examine demographic features associated with *Candida* colonization. I identify associations between *Candida* colonization and the bacterial community type, showing that *Candida* is associated with *L. iner*-dominant but not *L.* crispatus-dominant microbiomes. I then show that *L. crispatus* and *L. iners* differ in their ability to suppress *Candida* growth *in vitro* in a pH dependent manner, reflecting differences in lactic acid production. In **Chapter 3**, I develop and validate a SNP based method to explore strain level variation within vaginal samples using metagenomic data. I

define the population structure for six common vaginal species (*G. vaginalis*, *L. crispatus*, *L. iners*, *L. jensenii*, *L. gasseri* and *A. vaginae*) and identify subpopulation groups for all except *L. iners*. I examine the ecological diversity among vaginal microbiomes and the evolutionary origins of strain diversity within each species. I provide evidence that supports the notion that the diverse vaginal microbiome dominated by *G. vaginalis* and *A. vaginae* predated vaginal lactobacilli in humans. In **Chapter 4**, I discuss the ways in which future work can build on these findings to address outstanding questions in the field. Together these findings provide greater insight into the composition and structure of the vaginal microbiome and provide a framework to incorporate *Candida* and bacterial strain into future studies investigating the role of the vaginal microbiome in gynecologic and obstetric health.

1.5 References

1. Seed PC. The human mycobiome. *Cold Spring Harb Perspect Med.* 2014;5(5):a019810.
2. Human Microbiome Project C. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012;486(7402):207-214.
3. Tan SK, Relman DA, Pinsky BA. The Human Virome: Implications for Clinical Practice in Transplantation Medicine. *J Clin Microbiol.* 2017;55(10):2884-2893.
4. Lloyd-Price J, Mahurkar A, Rahnavard G, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature.* 2017;550(7674):61-66.
5. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. *Nat Med.* 2018;24(4):392-400.
6. Greenbaum S, Greenbaum G, Moran-Gilad J, Weintraub AY. Ecological dynamics of the vaginal microbiome in relation to health and disease. *Am J Obstet Gynecol.* 2019;220(4):324-335.
7. Bradford LL, Ravel J. The vaginal mycobiome: A contemporary perspective on fungi in women's health and diseases. *Virulence.* 2017;8(3):342-351.

8. Wylie KM, Mihindukulasuriya KA, Zhou Y, Sodergren E, Storch GA, Weinstock GM. Metagenomic analysis of double-stranded DNA viruses in healthy adults. *BMC Biol.* 2014;12:71.

9. Mendes-Soares H, Suzuki H, Hickey RJ, Forney LJ. Comparative functional genomics of Lactobacillus spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. *J Bacteriol.* 2014;196(7):1458-1470.

10. Goldstein EJ, Tyrrell KL, Citron DM. Lactobacillus species: taxonomic complexity and controversial susceptibilities. *Clin Infect Dis.* 2015;60 Suppl 2:S98-107.

11. Ravel J, Gajer P, Abdo Z, et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A.* 2011;108 Suppl 1:4680-4687.

12. Tachedjian G, Aldunate M, Bradshaw CS, Cone RA. The role of lactic acid production by probiotic Lactobacillus species in vaginal health. *Res Microbiol.* 2017;168(9-10):782-792.

13. Aldunate M, Srbinovski D, Hearps AC, et al. Antimicrobial and immune modulatory effects of lactic acid and short chain fatty acids produced by vaginal microbiota associated with eubiosis and bacterial vaginosis. *Front Physiol.* 2015;6:164.

14. Onderdonk AB, Delaney ML, Fichorova RN. The Human Microbiome during Bacterial Vaginosis. *Clin Microbiol Rev.* 2016;29(2):223-238.

15. Kenyon C, Colebunders R, Crucitti T. The global epidemiology of bacterial vaginosis: a systematic review. *Am J Obstet Gynecol.* 2013;209(6):505-523.

16. Morris M, Nicoll A, Simms I, Wilson J, Catchpole M. Bacterial vaginosis: a public health review. *BJOG.* 2001;108(5):439-450.

17. Klebanoff MA, Schwebke JR, Zhang J, Nansel TR, Yu KF, Andrews WW. Vulvovaginal symptoms in women with bacterial vaginosis. *Obstet Gynecol.* 2004;104(2):267-272.

18. Amsel R, Totten PA, Spiegel CA, Chen KC, Eschenbach D, Holmes KK. Nonspecific vaginitis. Diagnostic criteria and microbial and epidemiologic associations. *Am J Med.* 1983;74(1):14-22.

19. Nugent RP, Krohn MA, Hillier SL. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *J Clin Microbiol.* 1991;29(2):297-301.

20. Srinivasan S, Hoffman NG, Morgan MT, et al. Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLoS One.* 2012;7(6):e37818.

21. Fredricks DN. Molecular methods to describe the spectrum and dynamics of the vaginal microbiota. *Anaerobe.* 2011;17(4):191-195.

22. Fettweis JM, Brooks JP, Serrano MG, et al. Differences in vaginal microbiome in African American women versus women of European ancestry. *Microbiology.* 2014;160(Pt 10):2272-2282.

23. Zhou X, Brown CJ, Abdo Z, et al. Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. *ISME J.* 2007;1(2):121-133.

24. Serrano MG, Parikh HI, Brooks JP, et al. Racioethnic diversity in the dynamics of the vaginal microbiome during pregnancy. *Nat Med.* 2019;25(6):1001-1011.

25. van de Wijgert JH, Borgdorff H, Verhelst R, et al. The vaginal microbiota: what have we learned after a decade of molecular characterization? *PLoS One.* 2014;9(8):e105998.

26. Gajer P, Brotman RM, Bai G, et al. Temporal dynamics of the human vaginal microbiota. *Sci Transl Med.* 2012;4(132):132ra152.

27. Srinivasan S, Liu C, Mitchell CM, et al. Temporal variability of human vaginal bacteria and relationship with bacterial vaginosis. *PLoS One.* 2010;5(4):e10197.

28. DiGiulio DB, Callahan BJ, McMurdie PJ, et al. Temporal and spatial variation of the human microbiota during pregnancy. *Proc Natl Acad Sci U S A.* 2015;112(35):11060-11065.

29. Romero R, Hassan SS, Gajer P, et al. The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome.* 2014;2(1):4.

30. Paavonen J, Brunham RC. Bacterial Vaginosis and Desquamative Inflammatory Vaginitis. *N Engl J Med.* 2018;379(23):2246-2254.

31. Bautista CT, Wurapa E, Sateren WB, Morris S, Hollingsworth B, Sanchez JL. Bacterial vaginosis: a synthesis of the literature on etiology, prevalence, risk factors, and relationship with chlamydia and gonorrhea infections. *Mil Med Res.* 2016;3:4.

32. Cone RA. Vaginal microbiota and sexually transmitted infections that may influence transmission of cell-associated HIV. *J Infect Dis.* 2014;210 Suppl 3:S616-621.

33. Brotman RM. Vaginal microbiome and sexually transmitted infections: an epidemiologic perspective. *J Clin Invest.* 2011;121(12):4610-4617.

34. Mitra A, MacIntyre DA, Marchesi JR, Lee YS, Bennett PR, Kyrgiou M. The vaginal microbiota, human papillomavirus infection and cervical intraepithelial neoplasia: what do we know and where are we going next? *Microbiome.* 2016;4(1):58.

35. Abdool Karim SS, Baxter C, Passmore JS, McKinnon LR, Williams BL. The genital tract and rectal microbiomes: their role in HIV susceptibility and prevention in women. *J Int AIDS Soc.* 2019;22(5):e25300.

36. Bayigga L, Kateete DP, Anderson DJ, Sekikubo M, Nakanjako D. Diversity of vaginal microbiota in sub-Saharan Africa and its effects on HIV transmission and prevention. *Am J Obstet Gynecol.* 2019;220(2):155-166.

37. McClelland RS, Lingappa JR, Srinivasan S, et al. Evaluation of the association between the concentrations of key vaginal bacteria and the increased risk of HIV acquisition in African women from five cohorts: a nested case-control study. *Lancet Infect Dis.* 2018;18(5):554-564.

38. Vinturache AE, Gyamfi-Bannerman C, Hwang J, Mysorekar IU, Jacobsson B, Preterm Birth International C. Maternal microbiome - A pathway to preterm birth. *Semin Fetal Neonatal Med.* 2016;21(2):94-99.

39. Bernabeu A, Lledo B, Diaz MC, et al. Effect of the vaginal microbiome on the pregnancy rate in women receiving assisted reproductive treatment. *J Assist Reprod Genet.* 2019;36(10):2111-2119.

40. Chawanpaiboon S, Vogel JP, Moller AB, et al. Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob Health.* 2019;7(1):e37-e46.

41. Schaaf JM, Liem SM, Mol BW, Abu-Hanna A, Ravelli AC. Ethnic and racial disparities in the risk of preterm birth: a systematic review and meta-analysis. *Am J Perinatol.* 2013;30(6):433-450.

42. Romero R, Espinoza J, Goncalves LF, Kusanovic JP, Friel L, Hassan S. The role of inflammation and infection in preterm birth. *Semin Reprod Med.* 2007;25(1):21-39.

43. Fiscella K. Racial disparities in preterm births. The role of urogenital infections. *Public Health Rep.* 1996;111(2):104-113.

44. Mysorekar IU, Cao B. Microbiome in parturition and preterm birth. *Semin Reprod Med.* 2014;32(1):50-55.

45. Agrawal V, Hirsch E. Intrauterine infection and preterm labor. *Semin Fetal Neonatal Med.* 2012;17(1):12-19.

46. Hillier SL, Nugent RP, Eschenbach DA, et al. Association between bacterial vaginosis and preterm delivery of a low-birth-weight infant. The Vaginal Infections and Prematurity Study Group. *N Engl J Med.* 1995;333(26):1737-1742.

47. Kahwati LC, Clark R, Berkman N, et al. Screening for Bacterial Vaginosis in Pregnant Adolescents and Women to Prevent Preterm Delivery: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA.* 2020;323(13):1293-1309.

48. Fettweis JM, Serrano MG, Brooks JP, et al. The vaginal microbiome and preterm birth. *Nat Med.* 2019;25(6):1012-1021.

49. Brown RG, Al-Memar M, Marchesi JR, et al. Establishment of vaginal microbiota composition in early pregnancy and its association with subsequent preterm prelabor rupture of the fetal membranes. *Transl Res.* 2019;207:30-43.

50. Freitas AC, Bocking A, Hill JE, Money DM, Group VR. Increased richness and diversity of the vaginal microbiota and spontaneous preterm birth. *Microbiome.* 2018;6(1):117.

51. Stout MJ, Zhou Y, Wylie KM, Tarr PI, Macones GA, Tuuli MG. Early pregnancy vaginal microbiome trends and preterm birth. *Am J Obstet Gynecol.* 2017;217(3):356 e351-356 e318.

52. Nelson DB, Shin H, Wu J, Dominguez-Bello MG. The Gestational Vaginal Microbiome and Spontaneous Preterm Birth among Nulliparous African American Women. *Am J Perinatol.* 2016;33(9):887-893.

53. Callahan BJ, DiGiulio DB, Goltsman DSA, et al. Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women. *Proc Natl Acad Sci U S A.* 2017;114(37):9966-9971.

54. Ahmed A, Earl J, Retchless A, et al. Comparative genomic analyses of 17 clinical isolates of Gardnerella vaginalis provide evidence of multiple genetically isolated clades consistent with subspeciation into genovars. *J Bacteriol.* 2012;194(15):3922-3937.

55. Beigi RH, Meyn LA, Moore DM, Krohn MA, Hillier SL. Vaginal yeast colonization in nonpregnant women: a longitudinal study. *Obstet Gynecol.* 2004;104(5 Pt 1):926-930.

56. Drell T, Lillsaar T, Tummeleht L, et al. Characterization of the vaginal micro- and mycobiome in asymptomatic reproductive-age Estonian women. *PLoS One.* 2013;8(1):e54379.

57. Fidel PL, Jr., Barousse M, Espinosa T, et al. An intravaginal live Candida challenge in humans leads to new hypotheses for the immunopathogenesis of vulvovaginal candidiasis. *Infect Immun.* 2004;72(5):2939-2946.

58. Ilkit M, Guzel AB. The epidemiology, pathogenesis, and diagnosis of vulvovaginal candidosis: a mycological perspective. *Crit Rev Microbiol.* 2011;37(3):250-261.

59. Liu MB, Xu SR, He Y, et al. Diverse vaginal microbiomes in reproductive-age women with vulvovaginal candidiasis. *PLoS One.* 2013;8(11):e79812.

60. Zhou X, Westman R, Hickey R, et al. Vaginal microbiota of women with frequent vulvovaginal candidiasis. *Infect Immun.* 2009;77(9):4130-4135.

61. Macklaim JM, Clemente JC, Knight R, Gloor GB, Reid G. Changes in vaginal microbiota following antimicrobial and probiotic therapy. *Microb Ecol Health Dis.* 2015;26:27799.
62. Sobel JD. Vulvovaginal candidosis. *Lancet.* 2007;369(9577):1961-1971.

63. Aguin TJ, Sobel JD. Vulvovaginal candidiasis in pregnancy. *Curr Infect Dis Rep.* 2015;17(6):462.

64. Maki Y, Fujisaki M, Sato Y, Sameshima H. Candida Chorioamnionitis Leads to Preterm Birth and Adverse Fetal-Neonatal Outcome. *Infect Dis Obstet Gynecol.* 2017;2017:9060138.

65. Banhidy F, Acs N, Puho EH, Czeizel AE. Rate of preterm births in pregnant women with common lower genital tract infection: a population-based study based on the clinical practice. *J Matern Fetal Neonatal Med.* 2009;22(5):410-418.

66. Czeizel AE, Fladung B, Vargha P. Preterm birth reduction after clotrimazole treatment during pregnancy. *Eur J Obstet Gynecol Reprod Biol.* 2004;116(2):157-163.

67. Czeizel AE, Rockenbauer M. A lower rate of preterm birth after clotrimazole therapy during pregnancy. *Paediatr Perinat Epidemiol.* 1999;13(1):58-64.

68. Roberts CL, Algert CS, Rickard KL, Morris JM. Treatment of vaginal candidiasis for the prevention of preterm birth: a systematic review and meta-analysis. *Syst Rev.* 2015;4:31.

69. Kiss H, Petricevic L, Husslein P. Prospective randomised controlled trial of an infection screening programme to reduce the rate of preterm delivery. *BMJ.* 2004;329(7462):371.

70. Roberts CL, Rickard K, Kotsiou G, Morris JM. Treatment of asymptomatic vaginal candidiasis in pregnancy to prevent preterm birth: an open-label pilot randomized controlled trial. *BMC Pregnancy Childbirth.* 2011;11:18.

71. Cotch MF, Hillier SL, Gibbs RS, Eschenbach DA. Epidemiology and outcomes associated with moderate to heavy Candida colonization during pregnancy. Vaginal Infections and Prematurity Study Group. *Am J Obstet Gynecol.* 1998;178(2):374-380.

72. McGregor JA, French JI, Parker R, et al. Prevention of premature birth by screening and treatment for common genital tract infections: results of a prospective controlled evaluation. *Am J Obstet Gynecol.* 1995;173(1):157-167.

Chapter 2: Associations between the vaginal microbiome and *Candida* colonization in women of reproductive age

This chapter was published in the *American Journal of Obstetrics and Gynecology*. BAT wrote this chapter with input from co-authors. BAT performed all experiments and analysis unless otherwise mentioned.

Tortelli, B. A., Lewis, W. G., Allsworth, J. E., Member-Meneh, N., Foster, L. R., Reno, H. E., Peipert, J. F., Fay, J. C., & Lewis, A. L. (2020). Associations between the vaginal microbiome and Candida colonization in women of reproductive age. *American journal of obstetrics and gynecology*, *222*(5), 471.e1–471.e9. https://doi.org/10.1016/j.ajog.2019.10.008

Acknowledgements: We thank the entire Contraceptive CHOICE Project support staff and Jennifer Reed (née Bick, Washington University) for Nugent scoring. We appreciate Dr. Sharon Hillier and her group for verifying Nugent scoring. We also thank Denise Spear WHNP-BC, Valerie Higginbotham WHNP-BC (St. Louis County Department of Public Health) and Courtney Amegashie (Washington University, NIH funding) for collection and swab processing and the St. Louis County Department of Public Health for facilitating sample collection used to validate the *Candida* qPCR assay. We thank Andrew Kau for advice (Washington University, in part NIH funding) on microbiome profiling and Deborah Frank (Washington University) for editorial input on the manuscript. Finally, we appreciate those who provided vaginal samples.

2.1 Introduction

The human vagina is a dynamic ecosystem that hosts microbes from diverse taxa. Profiling 16S ribosomal gene diversity has expanded our understanding of the vaginal microbiome, allowing exploration of links between bacterial composition and reproductive outcomes. Vaginal microbial communities can be clustered into five common community types.[1] Four of these are dominated by a single *Lactobacillus* species: *L. crispatus*, *L. gasseri*, *L. iners*, or *L. jensenii*. The final community type (often described as "Diverse") has few lactobacilli and exhibits greater representation of anaerobic bacteria such as *Gardnerella vaginalis, Atopobium vaginae* and *Predvotella spp.*[1] The prevalence of these community types varies with race and ethnicity; blackand Hispanic women more frequently host *L. iners*-dominant and Diverse communities thanwhite women, who more frequently host *L. crispatus*-dominant communities.[1,2] Diverse communities often harbor bacterial taxa that are abundant during bacterial vaginosis (BV), a condition diagnosed by clinical (Amsel) criteria or by Nugent scoring,[3] a 0-10 scale generated by scoring bacterial morphotypes in Gram-stained vaginal smears (0-3, normal; 4-6, intermediate; 7-10, BV). BV is associated with increased risks of sexually transmitted infections and adverse reproductive outcomes.[4]

Candida (most commonly *C. albicans*) is a common member of the vaginal microbiome (found in ~30% of women[5]). The prevalence of non-*albicans* species among women with vaginal *Candida* varies, ranging from ~10-30%.[5-9] Vaginal *Candida* colonization may lead to vulvovaginal candidiasis (VVC), characterized by an aggressive host response to *Candida* overgrowth.[10] However, *Candida* colonization is frequently asymptomatic and not all women

colonized with *Candida* go on to experience VVC.[5] Vaginal *Candida* colonization has also been linked to other adverse reproductive outcomes.[8,11-16]

Several prior studies have examined relationships between vaginal bacteria and *Candida*. A few of these studies implicate an abundance of lactobacilli with a greater likelihood of harboring *Candida*.[5,6,17] Other studies suggest there may be co-occurrence of *Candida* with some BV-associated bacteria,[18-21] and specifically that *Candida* may be correlated with the simultaneous presence of both lactobacilli and BV-associated bacteria.[19-21] An important limitation is that prior studies, whether using molecular or culture-based techniques, have not distinguished between lactobacilli at the species level. This is a significant limitation, which if resolved, may shed light on why some women are so prone to *Candida* colonization and candidiasis.

Taken together with the prior studies above, several considerations led us to hypothesize that *L. iners* in particular may support the co-occurrence of *Candida, especially compared to L. crispatus*. *L. iners* is unique among the lactobacilli in being prevalent within less stable Nugent intermediate and BV communities[1,22,23] and in producing a cytolytic toxin.[24,25] Furthermore, *L. iners* dominance has been associated with other negative health outcomes such as increased risks of *Chlamydia trachomatis* infection,[26] incident BV,[27] defects in vaginal mucus that compromise antiviral barrier function,[28] and cytokine signatures linked with HIV risk.[29] We performed two types of studies to test our hypothesis that *L. iners* may preferentially support *Candida* colonization 1) a molecular evaluation of clinical specimens, and 2) *in vitro* growth inhibition studies.

2.2 Methods

Study design

This nested cross-sectional study uses samples and questionnaire data collected by the Contraceptive CHOICE Project (CHOICE)[30] according to Washington University IRB-approved protocol 201108155. In total, 9256 women from the St. Louis-area gave informed consent from August 2007 through September 2011. For this nested study, 299 women enrolled from 08/2008-06/2009 were selected based on power calculations made from preliminary data. Women enrolled in the CHOICE study were between the ages of 14 and 45, reported sexual activity in the past six months or anticipated sexual activity with a male partner and were seeking contraception. Women with a history of tubal ligation or hysterectomy were excluded. All women underwent a pregnancy test. Vaginal swab specimens were self-collected in the vast majority of cases, then stored at -80℃ until analysis. Of the swabs used in the final analysis, one was collected by a clinician and the collection method was missing for five samples.

Women who completed a baseline survey (including Sociodemographic data) and had a vaginal swab available were eligible for inclusion. Samples from all participants underwent Nugent scoring to determine BV status.[3,31,32] Unfortunately, vaginal pH and data regarding menstrual cycle and recent sexual activity was only available for a subset of women and were inadequate for analysis. Overall, the distribution of self-reported race/ethnicity of women in the CHOICE study were representative of the St. Louis region; few women reported a race other than "black or African-American" (hereafter referred to as "black") or "white." Due to small numbers of other groups, only women who reported "black" or "white" race were eligible for inclusion in this sub-study.

Composition of the vaginal microbiota has been previously associated with race.[1] To test whether *Candida* was associated with vaginal niches occupied by particular bacterial communities, we sought a strategy to avoid inadequate representation of less common community types in the different demographic groups so that we would be powered to ask whether *Candida* is associated with particular microbial patterns. We used frequency matching to similarly represent black and white women in each of the three Nugent categories. We used a normal:intermediate:BV ratio of 2:2:1 to ensure that we had samples represented across the Nugent spectrum, while balancing the practical reality that relatively few BV specimens were available from white women. Of the 299 subjects selected, 35 were pregnant at the time of swab collection and excluded from final analysis. Additionally, 9 specimens were excluded due to low bacterial biomass. See Supplemental Methods.

Microbiome analysis and *Candida* colonization status

DNA was extracted from eluted vaginal swabs and 16S ribosomal profiling of the V4 hypervariable region was performed as described in the Supplemental Methods. The microbiome was classified based on the dominant *Lactobacillus* species present, defined as 50% relative abundance or greater and referred to as, "*L. crispatus*-, *L. iners*-, *L. gasseri*-, or *L. jensenii*-dominant" microbiomes. Communities without a single *Lactobacillus* species reaching 50% were referred to as Diverse communities. A pan-*Candida* qRT-PCR[33] that amplifies the internally transcribed spacer 1 (ITS1) was used to determine *Candida* colonization status using isolated DNA as template. Prior to analysis we validated this assay among vaginal specimens

collected from a second cohort of women enrolled at a different site. See Supplemental Methods for details.

Candida growth inhibition

Candida strains were grown in yeast extract-peptone-dextrose (YPD) media. *C. albicans* strain SC5314 was obtained from the American Type Culture Collection. Vaginal strains of *Candida* (*C. albicans*: BAT8133, BAT8135, BAT8143, BAT8152, BAT8154, BAT3353A; *C. glabrata*: BAT8139, BAT3353B) were isolated from women as described in the Supplemental Methods. *L. crispatus* (MV-1A-US, JV-V01, MV-3A-US, 125-2-CHN) and *L. iners* (UP II 143-D, Lactin V09V1-C, LEAF 2032-Ad, LEAF 3008-A) strains were obtained from BEI resources and cultured in De Man, Rogosa and Sharpe (MRS) media for 48 hours to make cell free supernatants (CFS). All *Candida* growth inhibition experiments were conducted in 96-well plates. Each well contained a 1:1 ratio of CFS and YPD inoculated with ~10^6 *C. albicans* colony-forming units (CFU)/mL. YPD was buffered with 300 mM sodium bicarbonate and 300 mM HEPES sodium salt for neutralization assays. For lactic acid growth inhibition assays, fresh MRS was supplemented with racemic lactic acid. A micro pH electrode was used to measure pH of each mixture and lactate was measured with a colorimetric assay. Protonated lactic acid concentrations were calculated using lactate molarity and pH using the Henderson-Hasselbalch equation (pKa = 3.9). See Supplemental Methods for more details about *Candida* growth inhibition experiments.

Statistical analysis

Statistical analyses and data representation were completed in R (v3.5.1) and Prism (v7). Fisher's Exact Tests (Fisher) were used to assess for associations between cohort characteristics and race, with odds ratios (OR) determined by a conditional maximum likelihood estimate. Unless otherwise noted, we used an extension of the generalized linear model (GLM) method that included race as a potentially confounding covariate to test for associations between cohort characteristics and *Candida* colonization status, using the exponent of the coefficient from the logistic regression to calculate ORs. Note that because *Candida* colonization incidence is >10% the odds ratios may not be an accurate approximation of the relative risk; see [34] for conversion between the two.

We used type-II analysis of variance (ANOVA-II) with Wald test and Tukey's Honestly Significant Different Test (Tukey) to evaluate significance in these models. In instances where multiple statistical tests were performed, we relied on GLM accounting for race. Mann-Whitney tests were used to test for associations with *Candida* abundance and effect size (r) was calculated from the Z value. Statistical tests for *in vitro* experiments included one-way ANOVA with Tukey's correction for multiple comparisons and Mann-Whitney tests as appropriate. Regardless of the statistical method used, P-values < 0.05 were considered significant.

2.3 Results

Description of the clinical cohort

Two-hundred fifty-five non-pregnant women of reproductive age were included in our analysis. In this cohort, 53% of women identified as "white" and 47% identified as "black". Forty-four (17%) women had BV, while 109 (43%) and 102 (40%) had intermediate and normal vaginal

flora respectively. About half of the women (54%) reported using public assistance or having trouble meeting daily needs and were classified as having low socioeconomic status. Body mass index (BMI) was calculated and categorized using standard methods and definitions. Most women (64.3%) reported at least one prior pregnancy. Seventy-two women (28.2%) reported vaginal douching in the last 180 days. Race was found to be associated with socioeconomic status ($p < 0.0001$), BMI ($p = 0.003$), gravidity ($p < 0.0001$) and vaginal douching ($p < 0.0001$). A summary of demographic data and cohort characteristics by race is presented in Table 2.S1.

Forty-two (16%) women were vaginally colonized with *Candida*. Of these, most (90%) were colonized by *C. albicans*. *C. glabrata* was less common (~10%). Sequencing of the vaginal microbiome revealed that fifty-two women (20%) had *L. crispatus*-dominant microbiomes, 99 (39%) had *L. iners*-dominant microbiomes and 98 (38%) had microbiomes that were not dominated by a single *Lactobacillus species* (Diverse). We were not powered to test associations between *Candida* and microbiomes dominated by *Lactobacillus jensenii* or *gasseri since* few women (n=6) exhibited these microbiomes. Black women were more likely than white women to have *L. iners*-dominant communities (46.7% vs 31.9% Fisher's Exact; OR = 1.87, 95% CI: 1.10 to 3.14, $p = 0.020$) and less likely to have *L. crispatus*-dominant communities (11.9% vs. 22.1% Fisher's Exact; OR = 0.380, 95% CI: 0.185 to 0.747, $p = 0.003$).

Associations between *Candida* and cohort characteristics

Forty-two (16%) women were vaginally colonized with *Candida*. Of these, most (90%) were colonized by *C. albicans*. *C. glabrata* colonization was less common (~10%). Table 2.1 contains a summary of *Candida* status by sociodemographic and other cohort characteristics. Only race

was significantly correlated with vaginal *Candida*; black women were more likely to be colonized compared to white women (OR =2.05, 95% CI: 1.03 to 4.25, Fisher's Exact, p = 0.042). Based on these findings, race was considered to be a potential confounder and incorporated into subsequent analyses using generalized linear models (GLM) to evaluate factors associated with *Candida* colonization.

Associations between *Candida* and cohort characteristics

Candida colonization rates did not differ based on Nugent-defined BV status (GLM; ANOVA-II, p = 0.897). We did not find any association between a woman's socioeconomic status and vaginal *Candida* colonization. *Candida* colonization did not differ significantly among underweight (20% *Candida*), normal weight (18%) and overweight (23%) women. However, obese women were less likely to be colonized compared to non-obese women (GLM; OR = 0.322, 95% CI: 0.123 to 0.744; Tukey's HSD, p = 0.013, see Supplement for comment). Women reporting current use of hormonal contraceptives containing estrogen and progestin were *Candida*-colonized at higher rates than women reporting non-hormonal methods, although this did not reach statistical significance (GLM; OR = 1.77, 95% CI: 0.858 to 3.58; Tukey's HSD, p = 0.237, see Supplement for details). Women who reported vaginal douching in the last 180 days were less likely to be *Candida* positive compared to women who reported no vaginal douching (GLM; OR = 0.364, 95% CI: 0.143 to 0.838; Tukey's HSD, p = 0.047).

Relationships between *Candida* colonization and the vaginal microbiome

Next, we investigated relationships between *Candida* colonization and dominant members of the vaginal microbiome based on 16S ribosomal gene profiling. *Candida* prevalence did not differ

between *Lactobacillus* dominated (50% or greater *Lactobacillus*) and non-*Lactobacillus* dominated microbiomes (GLM; ANOVA-II, p = 0.327). Although the absolute abundance of *Candida* as measured by qPCR did not differ within *L. iners*-dominant communities compared to other community types (Mann-Whitney, r = 0.046, p = 0.617), *L. iners*-dominant communities were more likely to harbor *Candida* than non-*L. iners*-dominant communities (GLM; OR = 2.00, 95% CI: 1.02 to 3.98; Tukey's HSD, p = 0.045; see Table 2.S2). Further analysis specifically showed that *L. iners*-dominant communities were more likely to be colonized than *L. crispatus*-dominant communities (OR = 2.85, 95% CI: 1.03 to 7.21; Fisher's Exact, p = 0.048). Among *Candida* positive women, higher levels of *Candida* (by qRT-PCR) were observed among black women compared to white women, although not statistically significant (Mann-Whitney test, r = 0.173, p = 0.131).

In vitro studies: inhibition of *Candida* growth by lactobacilli

Both *L. crispatus* and lactic acid have been shown to thwart the growth of *C. albicans*.[35-37] Next, we compared the inhibitory potential of *L. crispatus* and *L. iners* on *Candida* growth *in vitro*. *C. albicans* was cultured together with cell free supernatants (CFS) from *L. crispatus* and *L. iners* (8 strains total), followed by *Candida* CFU enumeration. Compared to *L. iners* CFS*, L. crispatus* CFS resulted in lower pH (pH = 4.0 vs. pH = 4.6, p < 0.0001) and correspondingly higher levels of protonated lactic acid in CFS-YPD (55 mM vs. 11 mM, p < 0.0001) (Figure 2.2). Buffering CFS-YPD to a neutral pH reduced levels of protonated lactic acid to below appreciable levels, ablated *Candida* growth inhibition, and eliminated the difference in *C. albicans* growth observed between *L. crispatus* and *L. iners* (Figure 2.2). Further, lactic acid was sufficient to inhibit *Candida* growth. In particular, significantly more growth inhibition was observed at 49 mM

protonated lactic acid compared to 11 mM, levels comparable to the *L. crispatus* and *L. iners* CFS-YPD respectively. Similar findings were seen using *vaginal* isolates of *C. albicans*. In contrast, *C. glabrata* exhibited only modest growth inhibition (Figure 2.2). Together, these data suggest that lactic acid is both necessary and sufficient for growth inhibition of *C. albicans in vitro*.

2.4 Comment

Principal Findings

We demonstrate that *Candida* colonization is associated with characteristics of the vaginal microbiome (dominance of *L. iners* compared to *L. crispatus*). Results in clinical specimens are consistent with *in vitro* data, which show that *L. crispatus* produces a pH-dependent factor that inhibits *C. albicans* growth more effectively compared to secreted factors of *L. iners* grown under the same conditions.

Results

As a relatively common vaginal microbial community member, *Candida* may influence reproductive health. Previous studies suggested vaginal *Lactobacillus* colonization as a risk factor for *Candida* colonization or VVC,[5,6,17] but seem inconsistent with other reports of *Candida*-bacteria associations.[18-21] Here we provide more taxonomic resolution, showing that that not all *Lactobacillus*-dominant communities are equally associated with *Candida* colonization.

Clinical Implications

Clinicians often group all lactobacilli together. This study adds to the growing body of evidence suggesting that *L. iners*-dominant communities are more permissive to vaginal colonization with potential pathogens, including *Candida*.

Research Implications

Of interest, black race was associated with obesity and vaginal douching as in prior studies. But surprisingly, the correlation between *Candida* and black race cannot be accounted for by obesity or douching because obese women and those who douche were actually ***less*** likely to be colonized with *Candida* (OR = 0.322 and 0.364 respectively). The literature contains inconsistent reports regarding the role of *Lactobacillus* colonization as a risk factor for *Candida* colonization or VVC.[5,6,17,18-21] We show that that not all *Lactobacillus*-dominant communities are equally associated with *Candida*. *In vitro* data provide one possible explanation, showing that *L. iners* strains do not produce the same magnitude of lactic acid compared to *L. crispatus* strains. An alternative, albeit not mutually exclusive explanation, is that vaginal *Candida* colonization may shift the microbiome to favor *L. iners*.

Interestingly, we observed similar rates of *Candida* colonization in *L. crispatus*-dominant and Diverse communities. With fewer lactic acid producing bacteria present, the vaginal pH of women with Diverse microbiome is less acidic.[1] These findings indicate that Diverse communities resist *Candida* by lactic acid-independent mechanisms.

Additional studies are needed to evaluate potential mechanisms governing these relationships and apply these findings in clinical settings.

Strengths and Limitations

Key strengths of our study design were the validation of a *Candida*-specific qPCR assay[33] for laboratory testing for *Candida* colonization, offering flexibility in settings where archived frozen vaginal swabs are more practical. We acknowledge that the specimens selected for this study are not a naturalistic representation of vaginal microbiomes. Rather, the frequency matching of black and white women across the Nugent spectrum is a strength that enabled power to test associations between yeast and bacteria in different racial groups. Limitations include: 1) the sample size and number of *Candida*-positive women were relatively small, limiting power to model multiple potential confounders, 2) this cohort may not be representative of the U.S. population, 3) clinical data were not available to examine the relationship between *Candida* colonization and VVC, and 4) our *in vitro* findings may not be representative of *in vivo* relationships.

Conclusion

These data suggest that *L. iners*-dominant vaginal communities may support the co-occurrence of *Candida.*

2.5 Supplemental Methods

Coding Survey Data

Data pertaining to age, socioeconomic status (SES), body mass index (BMI), current birth control method, vaginal douching and gravidity were extracted from survey response data and categorized. Age in years was converted to the following categorical variable: less than 20, 20 to 29, 30 to 39, and 40 or more. Low socioeconomic status (Low SES) was defined as reporting any

current receipt of public assistance (food stamps; Special Supplemental Nutrition Program for Women, Infants and Children; welfare; or unemployment) or trouble paying for necessities (transportation, housing, health or medical care, or food). BMI (kg/m^2) was converted to categorical variables as follows: underweight (< 18.5), normal weight (18.5 to 24.9), overweight (25-30) and obese (>30). Current birth control method was categorized into one of three categories: hormonal contraceptives containing a combination of estrogen and progestin (Estrogen + Progestin), hormonal contraceptives containing progestin alone (Progestin), or non-hormonal contraceptive methods (Non-hormonal). Responses indicating the use of oral contraceptive pills or birth control ring were grouped as "Estrogen + Progestin". Responses indicating the use of a levonorgestrel containing intrauterine device or depot medroxyprogesterone acetate were grouped as "Progestin". Responses indicating the use of condoms, rhythm/natural family planning method, abstinence, withdrawal or nothing were classified as "Non-hormonal". The number of times a patient reported vaginal douching in the past 180 days was converted to a categorical variable: "yes" if the number was 1 or more, "no" if it was 0 and "don't know" if the patient reported not knowing. Gravidity was converted to a categorical variable with "3+" designating a response of 3 or more.

Vaginal Swab Processing and Controls

Frozen vaginal swabs (CHOICE study) were arrayed in deep well 96 well plates (Eppendorf, Hauppauge, NY) in 1 mL of 0.1 M sodium acetate buffer (pH 5.5). To minimize cross-contamination during swab elution, each 96 well plate of swabs was arrayed on two plates in a checkerboard fashion, such that empty wells were present between samples. Swabs were incubated for one hour on ice and agitated every 20 minutes manually. Swabs were removed, and

the two plates merged by transferring suspensions into a single deep well 96 well plate. The plate was then centrifuged at 32,000 x g at 4℃ for 20 minutes and the supernatants removed from the samples. The pelleted material was resuspended in 250 microliters of a buffer containing 200 mM Tris-HCl (pH 8.0), 200 mM NaCl and 20 mM EDTA and then transferred to a 2 mL screwcap tube (Axygen, Oneonta, NY) containing 250 µL of (0.1 mm) zirconia/silica beads and 105 µL of 20% SDS. 250 µL of a solution of phenol:chloroform:isoamyl alcohol (25:24:1) saturated with 10 mM Tris (pH 8.0) and 1mM EDTA (Sigma-Aldrich, St. Louis, MO) was added. Samples were lysed by mechanical disruption with a bead beater (Biospec Products, Bartelsville, OK) for 3 minutes at room temperature before being centrifuged at 32,000 x g at 4℃ for 5 minutes. DNA was cleaned and concentrated from the aqueous layer using a QIAquick 96-well PCR Purification Kit (Qiagen, Germantown, MD) with some modifications to the manufacturer instructions. The extraction process was automated with an EPMotion that performed all pipetting steps. The binding buffer was modified by supplementing 500 mL of Buffer PM with 33.3 mL of 3M sodium acetate (pH 5.5). DNA was eluted from the columns with 50 µL of water into 96-well PCR plates (Phenix Research Products, Candler, NC). Each 96-well plate of samples contained the following reagent controls: eight wells of sodium acetate-eluted sterile swabs and eight wells of sodium acetate buffer used for swab elution. DNA was normalized to 5 ng/µL and all samples diluted 1:5 after normalization to dilute PCR inhibitors. All PCR plates were sealed with Biomek aluminum foil seals (Becker Coulter, Brea, CA). To avoid cross-contamination, plates were centrifuged at 32,000 x g prior to removal of the seal and resealed after each use. Also, caution was exercised when using a multichannel pipettor to mix samples, microscale splashes and aerosol that could cause cross-contamination were avoided by gentle pipetting and expelling material only to the soft stop.

16S Sequencing

The V4 hypervariable region was PCR amplified by adding 6.4 μL normalized genomic DNA (dilution process described above) to a PCR master mix containing primers with integrated barcodes for multiplexing as previously described.[38] PCR product was then quantified with a Quant-iT dsDNA Assay (Invitrogen, Carlsbad, CA) and pooled into quartiles based on abundance prior to size selection by AMPure XP magnetic beads (Beckman Coulter, Pasadena, CA). Each purified quartile was then quantified and pooled into a library for 2 x 250 paired-end sequencing on an Illumina MiSeq platform through the Center for Genome Science at Washington University in St. Louis.

Microbiome Analysis

Reads were trimmed to a length of 200 base pairs and mate-pairs merged with a minimum overlap of 18 bases. All analysis with Qiime software was completed with version 1.9.0. Reads were demultiplexed and OTUs clustered as previously described.[38] Taxonomy was assigned to OTUs using RDP 2.4 trained on a custom database as previously described.[39] Taxa were assigned with a confidence of 0.7 or greater. Because the V4 region among some common vaginal *Lactobacillus species* (i.e. *Lactobacillus crispatus*) share high sequence similarity with other *Lactobacillus species* that rarely colonize the vagina, a modified approach to classifying *Lactobacillus* OTUs to the species level was completed. OTUs assigned to the genus *Lactobacillus* were aligned to the NCBI 16S database using BLASTn. The top ranked species returned with a sequence homology of 97% or greater was identified as the OTU species. If the

top BLASTn hit was less than 97% identical, the OTU was not assigned to the species. Read data was then rarefied so that each sample contained 1000 reads.

Inclusion Criteria for Analysis

Low bacterial biomass samples are at increased risk of having endogenous signal masked by contamination. To avoid the inclusion of low bacterial biomass samples, we used the abundance of the V4 amplicon after 16s PCR as a proxy of bacterial biomass. V4 amplicon abundance was quantified after 16s PCR and reagent control samples were used to determine the threshold for inclusion. The maximum V4 amplicon concentration from all 64 reagent controls quantified was chosen as the cutoff for inclusion in analysis (6.1 ng/μL). We removed 9 samples from analysis due to low V4 amplicon abundance.

***Candida* qPCR Validation**

A separate cohort was needed to validate the qPCR assay (see below) we later used for determining *Candida* colonization status. Women were recruited from the North Central Community Health Center according to Washington University IRB-approved protocol number 201704121. Women underwent a speculum exam by a clinician, during which mid-vaginal swabs were collected. Two double-headed anaerobic swabs were collected and transported using the Starswab Anaerobic Transport System (Starplex Scientific Inc, Cleveland, TN). Two standard aerobic Starplex double headed rayon swabs (S09D, Starplex Scientific Inc, Cleveland, TN) were also collected. Anaerobic swabs were transported to the laboratory for same day processing and aerobic swabs frozen at -80°C. Anaerobic swabs were eluted in 2X NYCIII media, and "fresh frozen" (i.e. "0 passage," without growth or amplification of any kind) in the presence of sterile glycerol (20% final).

Aliquots of fresh frozen material were then stored at -80℃. Fresh frozen aliquots were thawed on ice and centrifuged at 3,000 x g for 5 minutes. The media was removed, the pellet resuspended in 200 µL of YPD, and plated on CHROMagar *Candida* semi-selective plates (DRG, Springfield, NJ). Plates were incubated for 48 hours aerobically at 37℃. Plates were then examined and specimens were considered to be culture positive if *Candida* colonies distinguished by a distinctive green color were observed. Specimens that were culture positive for *Candida* were considered to be true positives and this information was used to determine the sensitivity and specificity of the *Candida* (ITS1) qPCR assay described below (conducted on DNA that was isolated from the eluted aerobic swab).

Determination of *Candida* Colonization Status

A pan-*Candida* qPCR[33] designed to detect medically relevant *Candida* species in the presence of human genomic DNA, was validated for use on DNA extracted from vaginal swabs as described above. The primers 18S-1F (GCAAGTCATCAGCTTGCGTT) and 5.8S-1R (TGCGTTCTTCATCGATGCGA) amplify the internally transcribed spacer 1 (ITS1). Power SYBR Green PCR Master Mix (Applied Biosystems, Waltham, MA) was used for the qPCR reaction and each reaction contained 2 ng of genomic DNA as template. All reactions were run in triplicate. CT values were converted to ng of *Candida* DNA based off a standard curve of genomic DNA extracted from *C. albicans* strain SC5314. A sample was denoted as "*Candida* Positive" if the mean of the replicates was one standard deviation greater than the reported detection limit of 10 fg of *Candida* DNA. Candida DNA quantities were adjusted for initial genomic DNA normalization to 5 ng/uL and used as a proxy for *Candida* abundance. *Candida*

species identification was confirmed by Sanger sequencing the ITS1 amplicon and BLASTing the sequence against the NCBI database.

Candida Strains

The *C. albicans* strain SC5314 was obtained from the American Type Culture Collection in Manassas, VA. Vaginal *Candida* strains that were used in the *in vitro* assays were isolated from vaginal swab specimens originally collected from pregnant women as part of a different study, in accordance with Washington University IRB-approved protocol number 201610121. Vaginal swabs were rolled on CHROMagar *Candida* plates to isolate *Candida* colonies and species identification was confirmed by sequencing.

Preparation of *Lactobacillus* Cell Free Supernatant (CFS)

Four strains of *L. crispatus* (MV-1A-US, JV-V01, MV-3A-US, 125-2-CHN) and four strains of *L. iners* (UP II 143-D, Lactin V09V1-C, LEAF 2032-Ad, LEAF 3008-A) were cultured to make CFS. *Lactobacilli* were grown in 10 mL of De Man, Rogosa and Sharpe (MRS) media (pH 6.5) for 48 hours in 10 mL cell culture flasks (GBO, Monroe, NC) at 37℃ in an anaerobic chamber (Coy, Grass Lake, MI). Cultures were centrifuged at 3200 x g for 20 minutes at 4℃ and the supernatants filtered through a 0.22 μm filter to remove residual bacteria. CFS was aliquoted in microcentrifuge tubes and stored at -20℃. Our findings were reproduced with two different batches of CFS from *L. crispatus* and *L. iners.*

CFS Growth Inhibition Assays

All CFS growth inhibition experiments were conducted in 96-well microplates (GBO, Monroe, NC). A mixture of 50 µL *Lactobacillus* CFS, 40 µL YPD, and 10 µL *C. albicans* suspension, ~10^6 colony-forming units (CFU)/mL, were added to each well. Unconditioned MRS media was added instead of *Lactobacillus* CFS as a control. The plates were sealed with breathable seals (Diversified Biotech, Dedham, MA) and incubated aerobically for 16 hours at 37°C with constant shaking at 300 rpm. Suspensions were then plated for CFU on YPD agarose plates. For CFS neutralization assay, YPD was buffered to by adding 300 mM sodium bicarbonate and 300 mM HEPES resulting in a final pH of 8.6. Lactate concentrations of CFS supplemented YPD medium were measured with a colorimetric assay adapted for microplate use (Megazyme, Chicago, IL). A micro pH meter (S220-MIC, Mettler-Toledo, Columbus, OH) was used to determine the pH of each mixture and the protonated lactic acid concentration calculated using lactate molarity and pH using the Henderson-Hasselbalch equation (pKa = 3.9). Each growth inhibition experiment was conducted in triplicate and repeated at least twice.

Lactic Acid Inhibition Assays

MRS was supplemented with racemic lactic acid (Sigma-Aldrich, St. Louis, MO) at the following final concentrations: 100 mM, 200 mM, 300 mM and 400 mM. A mixture of 50 µL lactic acid supplemented MRS, 40 µL YPD, and 10 µL *Candida* suspension (~10^7 CFU/mL) were added to each well. Fresh non-conditioned MRS media was added instead of lactic acid supplemented MRS as a control. Suspensions were then plated for CFU on YPD agarose plates. Lactate concentrations and pH of lactic acid supplemented YPD medium were measured as described above and used to determine the protonated lactic acid concentration. Each growth experiment was conducted in triplicate and repeated at least twice.

2.6 Supplemental Comment

The effect of obesity on the composition of the vaginal microbiome is not well understood. Recent studies showed that Nugent score was positively associated with BMI[40,41] and obesity has also been associated with greater overall diversity and colonization by particular BV-associated taxa.[42] Previously it was found that obese women were less likely to be heavily colonized with *Candida*.[6] Our data support the same conclusion, although using a much more sensitive detection method. We found that regardless of race, obese women were more likely to have *L.i.*-dominant than *L.c.*-dominant communities. Taken together with our finding that *L.i.*-dominant communities were more likely than *L.c.*-dominant communities to harbor Candida, the data suggest a more complex and multifactorial interaction that cannot be explained by the dominant species of *Lactobacillus* present in the vagina. Further study is required to understand the interplay between obesity, the microbiome and *Candida* colonization. Factors that could contribute to this interplay may include disturbances in host metabolic, hormonal, and/or immune function associated with obesity. A higher prevalence of menstrual irregularity in obese women could also contribute to changes in the microbiome. Behaviors could also play a role, for example, obese women may be more likely to engage in vaginal douching.[40,41] Previous links between the gut microbiome and obesity could also be involved, especially given findings that the gut microbiome can be a reservoir of vaginal community members,[43] including *Candida*.

Table 2.1: Characteristics of subjects with vaginal *Candida* compared with those without vaginal *Candida.*

Characteristics	Total Cohort	*Candida* Positive	*Candida* Negative	P-value
Total Number of Subjects	255	42 (16.5)	213 (83.5)	
Age				0.811
< 20	28 (11.0)	6 (14.3)	22 (10.3)	
20 to 29	178 (69.8)	29 (69.0)	149 (70.0)	
30 to 39	44 (17.3)	7 (16.7)	37 (17.4)	
40 +	5 (2.0)	0 (0.0)	5 (2.3)	
Race				0.042
Black	120 (47.1)	26 (61.9)	94 (44.1)	
White	135 (52.9)	16 (38.1)	119 (55.9)	
Nugent-defined Vaginal Flora				0.833
Normal	102 (40.0)	15 (35.7)	87 (40.8)	
Intermediate	109 (42.7)	19 (45.2)	90 (42.3)	
BV	44 (17.3)	8 (19.0)	36 (16.9)	
Socioeconomic Status (SES)				1
Low SES	138 (54.1)	23 (54.8)	115 (54.0)	
Not Low SES	117 (45.9)	19 (45.2)	98 (46.0)	
Body Mass Index (kg/m2)				0.127
Underweight (< 18.5)	15 (5.9)	3 (7.1)	12 (5.6)	
Normal Weight (18.5 - 24.9)	103 (40.4)	19 (45.2)	84 (39.4)	
Overweight (25 - 30)	48 (18.8)	11 (26.2)	37 (17.4)	
Obese (> 30)	78 (30.6)	7 (16.7)	71 (33.3)	

Not Documented	11 (4.3)	2 (4.8)	9 (4.2)	
Current Birth Control Method				0.320
Estrogen + Progestin[a]	72 (28.2)	16 (38.1)	56 (26.3)	
Progestin[b]	12 (4.7)	1(2.4)	11 (5.2)	
Non-Hormonal[c]	171 (67.1)	25 (59.5)	146 (68.5)	
Vaginal Douching in Last 180 Days				0.323
Yes	72 (28.2)	8 (19.0)	64 (30.0)	
No	182 (71.4)	34 (81.0)	148 (69.5)	
Don't Know	1 (0.4)	0 (0.0)	1 (0.5)	
Gravidity				0.160
None	91 (35.7)	15 (35.7)	76 (35.7)	
1	58 (22.7)	6 (14.3)	52 (24.4)	
2	47 (18.4)	6 (14.3)	41 (19.2)	
3+	59 (23.1)	15 (35.7)	44 (20.7)	
Community Type				0.113
L. crispatus-dominant	52 (20.4)	5 (11.9)	47 (22.1)	
L. iners-dominant	99 (38.8)	23 (54.8)	76 (35.7)	
L. jensenii-dominant	3 (1.2)	1 (2.4)	2 (0.9)	
L. gasseri-dominant	3 (1.2)	0 (0.0)	3 (1.4)	
Diverse	98 (38.4)	13 (31.0)	85 (39.9)	

Values are n (%). Fisher's Exact Tests were used to determine p-values for each set of variables without adjusting for race. Note that p-values given in the text use GLM (accounting for race as a potential confounder).

[a]Women who reported the oral contraceptive pill or the birth control ring;

[b]Women who reported the levonorgestrel-containing intrauterine device or depot medroxyprogesterone acetate;

[c]Women who reported condoms, rhythm/natural family planning, abstinence, withdrawal or nothing.

Table 2.S1: Characteristics of subjects of black race compared with those of white race

Characteristics	Total Cohort	White Race	Black Race	P-value
Total Number of Subjects	255	135 (52.9)	120 (47.1)	
Age				0.3073
< 20	28 (11.0)	12 (8.9)	16 (13.3)	
20 to 29	178 (69.8)	101 (74.8)	77 (64.2)	
30 to 39	44 (17.3)	20 (14.8)	24 (20.0)	
40 +	5 (2.0)	2 (1.5)	3 (2.5)	
Nugent-defined Vaginal Flora				0.420
BV	44 (17.3)	21 (15.6)	23 (19.2)	
Intermediate	109 (42.7)	55 (40.7)	54 (45.0)	
Normal	102 (40.0)	59 (43.7)	43 (35.8)	
Socioeconomic Status (SES)				< 0.0001
Not Low SES	117 (45.9)	82 (60.7)	35 (29.2)	
Low SES	138 (54.1)	53 (39.3)	85 (70.8)	
Body Mass Index (kg/m2)				0.003
Underweight (< 18.5)	15 (5.9)	9 (6.7)	6 (5.0)	
Normal Weight (18.5 - 24.9)	103 (40.4)	68 (50.4)	35 (29.2)	
Overweight (25 - 30)	48 (18.8)	22 (16.3)	26 (21.7)	
Obese (> 30)	78 (30.6)	31 (23.0)	47 (39.2)	
Not Documented	11 (4.3)	5 (3.7)	6 (5.0)	
Current Birth Control Method				0.108
Estrogen + Progestin[a]	72 (28.2)	42 (31.1)	30 (25.0)	
Progestin[b]	12 (4.7)	3 (2.2)	9 (7.5)	
Non-Hormonal[c]	171 (67.1)	90 (66.7)	81 (67.5)	
Vaginal Douching in Last 180 Days				< 0.0001
Yes	72 (28.2)	17 (12.6)	55 (45.8)	
No	182 (71.4)	117 (86.7)	65 (54.2)	
Don't Know	1 (0.4)	1 (0.7)	0 (0.0)	
Gravidity				< 0.0001
None	91 (35.7)	64 (47.4)	27 (22.5)	
1	58 (22.7)	28 (20.7)	30 (25.0)	
2	47 (18.4)	25 (18.5)	22 (18.3)	
3 +	59 (23.1)	18 (13.3)	41 (34.2)	
Community Type				0.004
L. crispatus-dominant	52 (20.4)	37 (27.4)	15 (12.5)	

L. gasseri-dominant	3 (1.2)	1 (0.7)	2 (1.7)	
L. iners-dominant	99 (38.8)	43 (31.9)	56 (46.7)	
Diverse	98 (38.4)	51 (37.8)	47 (39.2)	
L. jensenii-dominant	3 (1.2)	3 (2.2)	0 (0.0)	
Candida				0.042
Positive	42 (16.5)	16 (11.9)	26 (21.7)	
Negative	213 (83.5)	119 (88.1)	94 (78.3)	

Values are n (%). Fisher's Exact Tests were used to determine p-values.
[a]Women who reported the oral contraceptive pill or the birth control ring;
[b]Women who reported the levonorgestrel-containing intrauterine device or depot medroxyprogesterone acetate;
[c]Women who reported condoms, rhythm/natural family planning, abstinence, withdrawal or nothing.

Table 2.S2: Vaginal *Candida* colonization by race and community type

Characteristics	*Candida* Positive	*Candida* Negative
Total Number of Subjects	42 (16.5)	213 (83.5)
Black Race	26 (21.7)	94 (78.3)
L. crispatus-dominant	2 (13.3)	13 (86.7)
L. gasseri-dominant	0 (0.0)	2 (100.0)
L. iners-dominant	17 (30.4)	39 (69.6)
Diverse	7 (14.9)	40 (85.1)
L. jensenii-dominant	0 (0.0)	0 (0.0)
White Race	16 (11.9)	119 (88.1)
L. crispatus-dominant	3 (8.1)	34 (91.9)
L. gasseri.-dominant	0 (0.0)	1 (100.0)
L. iners-dominant	6 (14.0)	37 (86.0)
Diverse	6 (11.8)	45 (88.2)
L. jensenii-dominant	1 (33.3)	2 (66.6)

Values are n (% of characteristic).

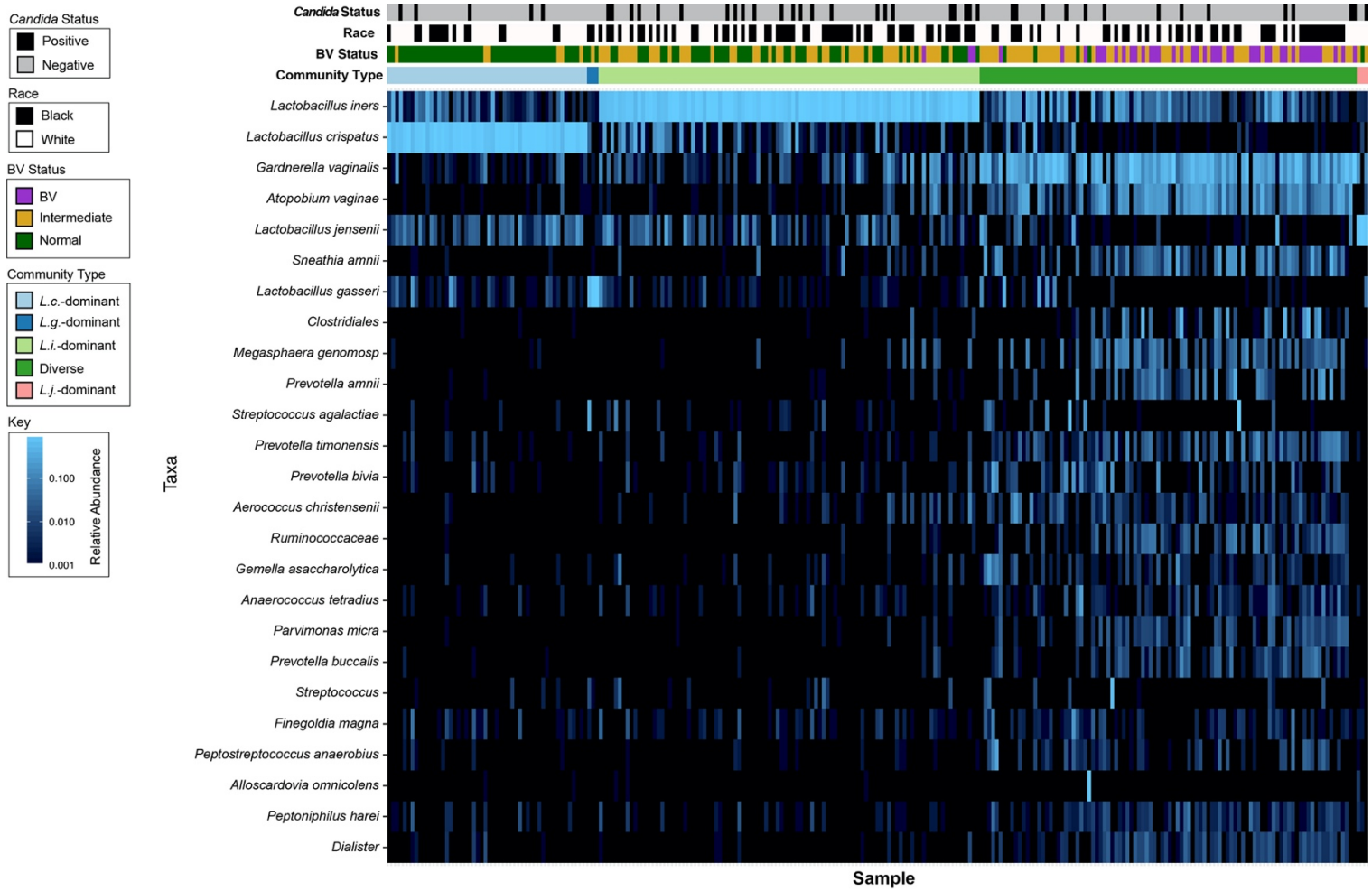

Figure 2.1: Heatmap of all samples in the cohort clustered by community type.
Heat map of samples clustered by community type showing the top 25 taxa observed across the cohort. The bars above the heatmap indicate community type, BV status by Nugent score, race and *Candida* status. In the heat map, *light blue* indicates the highest abundance, *darker blues* indicate lower abundance and *black* indicates very low abundance or not present. Black race ($p = 0.037$) and *L. iners*-dominant communities ($p = 0.045$) were associated with *Candida* colonization.

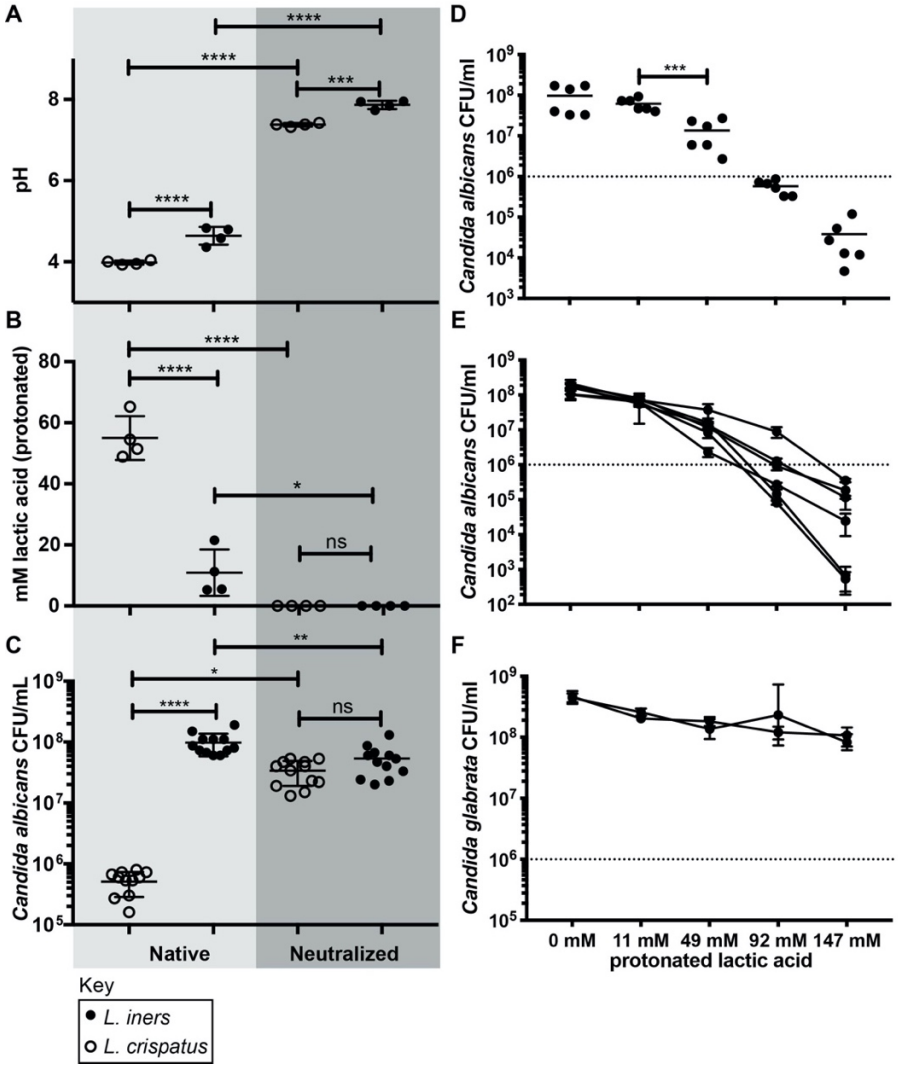

Figure 2.2: *In vitro* inhibition of *Candida* by *Lactobacillus* CFS and lactic acid.
A-B, Characterization of *Candida* growth medium supplemented with *Lactobacillus* CFS (YPD-CFS) in native and buffered states from four *L. crispatus* and four *L. iners* strains, prior to *Candida* inoculation. **A**, pH of YPD-CFS; **B**, Concentration of protonated lactic acid in YPD-CFS; **C,** Growth inhibition of *Candida* laboratory strain SC5314, showing three technical replicates for each *Lactobacillus* YPD-CFS. Analysis by one-way ANOVA with Tukey's correction for multiple comparisons. **D-F**, Characterization of the inhibitory effect of lactic acid supplemented medium on *Candida* growth. Three technical replicates from two biological experiments are shown. **D**, Growth inhibition of SC5314 by lactic acid showing Mann-Whitney test comparison of 11 mM to 49 mM protonated lactic acid; **E**, Lactic acid growth inhibition of 6 vaginal *C. albicans* isolates; **F**, Lactic acid growth inhibition of 2 vaginal *C. glabrata* isolates. Data points in panel D reflect 6 replicates from two experiments for each condition. Error bars in E-F show the standard deviation from the mean of three replicates for each isolate. Approximate starting inoculum for growth assays is indicated by a dashed line. Statistical significance: ns (not significant), *$p < 0.05$, **$p < 0.01$, ***$p < 0.005$, ****$p < 0.0001$.

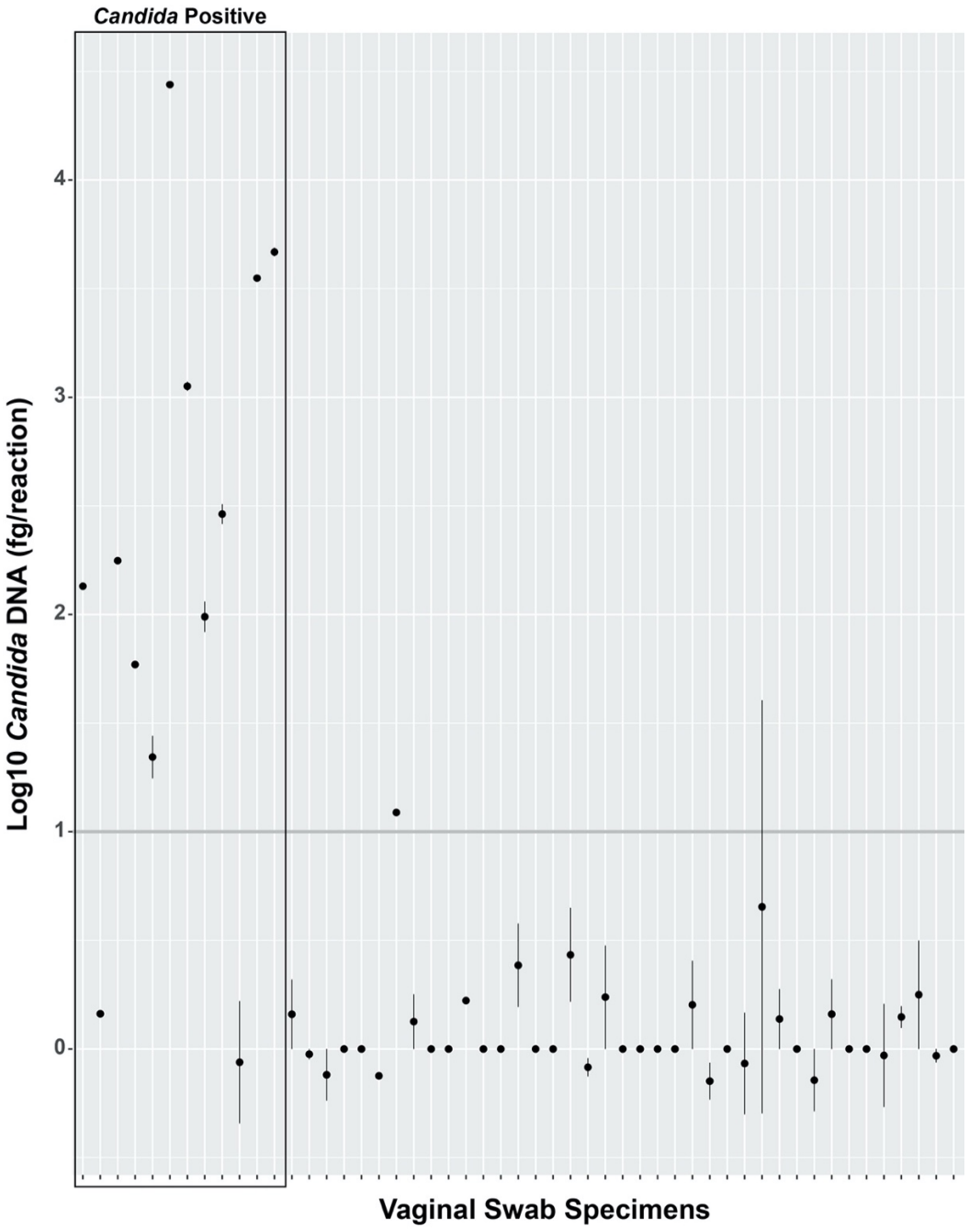

Figure 2.S1: Validation of *Candida* qPCR on vaginal samples
Fifty-one women were assessed for vaginal *Candida* colonization by culture and qPCR. The mean abundance and standard deviation of *Candida* DNA for each specimen are plotted. Culture positive specimens are indicated by the black box. Twelve specimens were culture positive for *Candida*, ten of which were also qPCR positive for *Candida.* The sensitivity of the qPCR diagnostic was 83.3% and the specificity was 100% for this set of samples.

2.7 References

1. Ravel J, Gajer P, Abdo Z, et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A.* 2011;108 Suppl 1:4680-4687.

2. Fettweis JM, Brooks JP, Serrano MG, et al. Differences in vaginal microbiome in African American women versus women of European ancestry. *Microbiology.* 2014;160(Pt 10):2272-2282.

3. Nugent RP, Krohn MA, Hillier SL. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of Gram stain interpretation. *J Clin Microbiol.* 1991;29(2):297-301.

4. Anahtar MN, Gootenberg DB, Mitchell CM, Kwon DS. Cervicovaginal Microbiota and Reproductive Health: The Virtue of Simplicity. *Cell Host Microbe.* 2018;23(2):159-168.

5. Beigi RH, Meyn LA, Moore DM, Krohn MA, Hillier SL. Vaginal yeast colonization in nonpregnant women: a longitudinal study. *Obstet Gynecol.* 2004;104(5 Pt 1):926-930.

6. Cotch MF, Hillier SL, Gibbs RS, Eschenbach DA. Epidemiology and outcomes associated with moderate to heavy *Candida* colonization during pregnancy. Vaginal Infections and Prematurity Study Group. *Am J Obstet Gynecol.* 1998;178(2):374-380.

7. Brandolt TM, Klafke GB, Goncalves CV, et al. Prevalence of *Candida* spp. in cervical-vaginal samples and the *in vitro* susceptibility of isolates. *Braz J Microbiol.* 2017;48(1):145-150.

8. Payne MS, Ireland DJ, Watts R, et al. *Ureaplasma parvum* genotype, combined vaginal colonisation with *Candida albicans*, and spontaneous preterm birth in an Australian cohort of pregnant women. *BMC Pregnancy Childbirth.* 2016;16(1):312.

9. Payne MS, Cullinane M, Garland SM, et al. Detection of *Candida* spp. in the vagina of a cohort of nulliparous pregnant women by culture and molecular methods: Is there an association between maternal vaginal and infant oral colonisation? *Aust N Z J Obstet Gynaecol.* 2016;56(2):179-184.

10. Fidel PL, Jr., Barousse M, Espinosa T, et al. An intravaginal live *Candida* challenge in humans leads to new hypotheses for the immunopathogenesis of vulvovaginal candidiasis. *Infect Immun.* 2004;72(5):2939-2946.

11. Maki Y, Fujisaki M, Sato Y, Sameshima H. *Candida* Chorioamnionitis Leads to Preterm Birth and Adverse Fetal-Neonatal Outcome. *Infect Dis Obstet Gynecol.* 2017;2017:9060138.

12. To V, Gurberg J, Krishnamurthy S. Tubo-Ovarian Abscess Caused by *Candida albicans* in an Obese Patient. *J Obstet Gynaecol Can.* 2015;37(5):426-429.

13. Menon S, Broeck DV, Rossi R, Ogbe E, Harmon S, Mabeya H. Associations Between Vaginal Infections and Potential High-risk and High-risk Human Papillomavirus Genotypes in Female Sex Workers in Western Kenya. *Clin Ther.* 2016;38(12):2567-2577.

14. Holzer I, Farr A, Kiss H, Hagmann M, Petricevic L. The colonization with *Candida* species is more harmful in the second trimester of pregnancy. *Arch Gynecol Obstet.* 2017;295(4):891-895.

15. Farr A, Kiss H, Holzer I, Husslein P, Hagmann M, Petricevic L. Effect of asymptomatic vaginal colonization with *Candida albicans* on pregnancy outcome. *Acta Obstet Gynecol Scand.* 2015;94(9):989-996.

16. Babu G, Singaravelu BG, Srikumar R, Reddy SV, Kokan A. Comparative Study on the Vaginal Flora and Incidence of Asymptomatic Vaginosis among Healthy Women and in Women with Infertility Problems of Reproductive Age. *J Clin Diagn Res.* 2017;11(8):DC18-DC22.

17. McClelland RS, Richardson BA, Hassan WM, et al. Prospective study of vaginal bacterial flora and other risk factors for vulvovaginal candidiasis. *J Infect Dis.* 2009;199(12):1883-1890.

18. Gonzalez-Fernandez D, Koski KG, Sinisterra OT, Del Carmen Pons E, Murillo E, Scott ME. Interactions among urogenital, intestinal, skin, and oral infections in pregnant and lactating Panamanian Ngabe women: a neglected public health challenge. *Am J Trop Med Hyg.* 2015;92(6):1100-1110.

19. Liu MB, Xu SR, He Y, et al. Diverse vaginal microbiomes in reproductive-age women with vulvovaginal candidiasis. *PLoS One.* 2013;8(11):e79812.

20. Vahidnia A, Tuin H, Bliekendaal H, Spaargaren J. Association of sexually transmitted infections, *Candida* species, Gram-positive flora and perianal flora with bacterial vaginosis. *New Microbiol.* 2015;38(4):559-563.

21. Rathod SD, Klausner JD, Krupp K, Reingold AL, Madhivanan P. Epidemiologic features of Vulvovaginal Candidiasis among reproductive-age women in India. *Infect Dis Obstet Gynecol.* 2012;2012:859071.

22. Fredricks DN, Fiedler TL, Marrazzo JM. Molecular identification of bacteria associated with bacterial vaginosis. *N Engl J Med.* 2005;353(18):1899-1911.

23. Fredricks DN. Molecular methods to describe the spectrum and dynamics of the vaginal microbiota. *Anaerobe.* 2011;17(4):191-195.

24. Rampersaud R, Lewis EL, LaRocca TJ, Ratner AJ. Environmental pH modulates inerolysin activity via post-binding blockade. *Sci Rep.* 2018;8(1):1542.

25. Rampersaud R, Planet PJ, Randis TM, et al. Inerolysin, a cholesterol-dependent cytolysin produced by *Lactobacillus iners*. *J Bacteriol.* 2011;193(5):1034-1041.

26. van Houdt R, Ma B, Bruisten SM, Speksnijder A, Ravel J, de Vries HJC. *Lactobacillus iners*-dominated vaginal microbiota is associated with increased susceptibility to *Chlamydia trachomatis* infection in Dutch women: a case-control study. *Sex Transm Infect.* 2018;94(2):117-123.

27. Muzny CA, Blanchard E, Taylor CM, et al. Identification of Key Bacteria Involved in the Induction of Incident Bacterial Vaginosis: A Prospective Study. *J Infect Dis.* 2018;218(6):966-978.

28. Nunn KL, Wang YY, Harit D, et al. Enhanced Trapping of HIV-1 by Human Cervicovaginal Mucus Is Associated with *Lactobacillus crispatus*-Dominant Microbiota. *MBio.* 2015;6(5):e01084-01015.

29. Joag V, Obila O, Gajer P, et al. Impact of Standard Bacterial Vaginosis Treatment on the Genital Microbiota, Immune Milieu, and Ex Vivo Human Immunodeficiency Virus Susceptibility. *Clin Infect Dis.* 2018.

30. Secura GM, Allsworth JE, Madden T, Mullersman JL, Peipert JF. The Contraceptive CHOICE Project: reducing barriers to long-acting reversible contraception. *Am J Obstet Gynecol.* 2010;203(2):115 e111-117.

31. Lewis WG, Robinson LS, Perry J, et al. Hydrolysis of secreted sialoglycoprotein immunoglobulin A (IgA) in ex vivo and biochemical models of bacterial vaginosis. *J Biol Chem.* 2012;287(3):2079-2089.

32. Amegashie CP, Gilbert NM, Peipert JF, Allsworth JE, Lewis WG, Lewis AL. Relationship between nugent score and vaginal epithelial exfoliation. *PLoS One.* 2017;12(5):e0177797.

33. Zhang J, Hung GC, Nagamine K, Li B, Tsai S, Lo SC. Development of *Candida*-Specific Real-Time PCR Assays for the Detection and Identification of Eight Medically Important *Candida* Species. *Microbiol Insights.* 2016;9:21-28.

34. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA.* 1998;280(19):1690-1691.

35. Krasner RI, Young G, Yudkofsky PL. Interactions of oral strains of *Candida albicans* and lactobacilli. *J Bacteriol.* 1956;72(4):525-529.

36. Wang S, Wang Q, Yang E, Yan L, Li T, Zhuang H. Antimicrobial Compounds Produced by Vaginal *Lactobacillus crispatus* are able to Strongly Inhibit *Candida albicans* Growth, Hyphal Formation and Regulate Virulence-related Gene Expressions. *Front Microbiol.* 2017;8:564.

37. Parolin C, Marangoni A, Laghi L, et al. Isolation of Vaginal lactobacilli and Characterization of Anti-*Candida* Activity. *PLoS One.* 2015;10(6):e0131220.

38. Kau AL, Planer JD, Liu J, et al. Functional characterization of IgA-targeted bacterial taxa from undernourished Malawian children that produce diet-dependent enteropathy. *Sci Transl Med.* 2015;7(276):276ra224.

39. Ridaura VK, Faith JJ, Rey FE, et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science.* 2013;341(6150):1241214.

40. Koumans EH, Sternberg M, Bruce C, et al. The prevalence of bacterial vaginosis in the United States, 2001-2004; associations with symptoms, sexual behaviors, and reproductive health. *Sex Transm Dis.* 2007;34(11):864-869.

41. Brookheart RT, Lewis WG, Peipert JF, Lewis AL, Allsworth JE. Association between obesity and bacterial vaginosis as assessed by Nugent score. *Am J Obstet Gynecol.* 2019.

42. Si J, You HJ, Yu J, Sung J, Ko G. *Prevotella* as a Hub for Vaginal Microbiota under the Influence of Host Genetics and Their Association with Obesity. *Cell Host Microbe.* 2017;21(1):97-105.

43. Marrazzo JM, Fiedler TL, Srinivasan S, et al. Extravaginal reservoirs of vaginal bacteria as risk factors for incident bacterial vaginosis. *J Infect Dis.* 2012;205(10):1580-1588.

Chapter 3: The structure and diversity of strain level variation in vaginal bacteria

This chapter was published in the *Journal of Microbial Genomics*. BAT wrote this chapter with input from co-authors. BAT performed all experiments and analysis.

Tortelli BA, Lewis AL, Fay JC. The structure and diversity of strain-level variation in vaginal bacteria. *Microb Genom.* 2021;7(3)

3.1 Introduction

A diverse range of microbial communities have been found to be associated with human anatomical sites, including the skin, gastrointestinal tract, oral cavity and vagina.[1] Surveys of these microbial communities have demonstrated significant differences between anatomical sites but also variation among individuals.[1,2] Inter-individual variation in microbial communities has in many instances been associated with a variety of host factors including human health and disease, e.g. obesity and inflammatory bowel disease,[3] leading to continued investigation of the implications of microbial variation.

Inter-individual variation has largely been explored by means of characterizing differences in species presence or relative abundance. However, prevalent species not only show differences in relative abundance but also exhibit appreciable strain level variation.[1,2] Individual strains may be unique to a person's microbiome and bacterial strains of a species isolated from the same individual have been noted to be more similar to each other than strains isolated from different individuals.[2,4] When examined, strain level variation is characterized by functional differences, prominent among these are differences in metabolic potential and antibiotic resistance.[2,5,6] This suggests that strain level variation may contribute to phenotypic differences in personal microbiomes observed between individuals. However, knowing the extent to which strain-level differences translate to functionally distinct strains remains an open and important question. Currently, most comparisons of microbial communities utilize operational taxonomic units as means of grouping similar strains together and differentiating them from other groups.

Strain differences and their relationships define the population structure of a species. Population structure is relevant for both grouping strains but also making inferences about their history. In the absence of recombination, strains continually diversify, but those lineages that are most successful will expand and others will be lost. Eventually such lineages can diverge in function and even establish new species. Under the ecological species concept, two species can't stably coexist unless they differ in their niche.[7] However, population structure can also be established by limited migration, in which case subpopulations may have the same functions in the community but diverge (neutrally) in their genome.[8] Although distinguishing functional populations from neutral populations is difficult, population structure remains an important component of describing groups of strains with shared functional differences or shared population history. As each human may carry or enable the formation of unique microbial strains, the characterization of population structures and their determinants is important to addressing the role of strain level variation in the human microbiome.[9]

Among human microbial communities, the vaginal microbiome differs in its community composition. Both 16S ribosomal profiling and metagenomic community profiling have indicated that the vaginal microbiome often exhibits lower community diversity when compared to other anatomical sites, frequently being dominated by a single species.[1,2] The composition of the bacterial community is often described in terms of five common community types.[10] Four of the five community types are dominated by a single *Lactobacillus* species (*L. crispatus*, *L. iners*, *L. jensenii* or *L. gasseri*). The fifth community type is characterized by a lack of *Lactobacillus* dominance and often exhibits higher community diversity. This diverse

community has been correlated with a high vaginal pH (> 4.5) and bacterial vaginosis (BV),[10] a dysbiosis associated with the overgrowth of anaerobic bacteria including *G. vaginalis* and *A. vaginae*. The prevalence of these vaginal communities vary by self-reported race/ethnicity[10,11] and have been associated with reproductive health.[12] While community type classification offers a convenient method for categorizing the overall composition of the vaginal microbiome, the significance of strain level variation is of increasing interest.

For certain vaginal bacterial species, functional, phenotypic and genomic differences have been described among isolated strains. An example of this is the classification of *G. vaginalis* into distinct phylogenetic clades (groups) through genomic approaches such as gene ontology and genome-wide single-nucleotide polymorphism (SNP) analysis.[13-15] Characterization of individual strains have shown functional differences (including sialidase activity) between groups with phenotypic consequences.[16-19] Such functional differences may explain why some groups but not others have been associated with BV.[20-22] While less is known about other vaginal bacterial species, genomic analysis of *L. crispatus* and *L. iners* strains has provided some insight into the population structure of these species.[4,23,24] An examination of 41 strains found that *L. crispatus* may be comprised of two closely related groups,[4] but identification of phenotypic differences between these groups is lacking. Additionally, the population structure of *L. iners* appears to lack strain groupings, but rather each strain appears to be distinct.[4,24] These assessments of strain level variation have focused on isolated strains, and assessments of strain level variation within the vaginal microbiome have been limited.[4] The use of variable regions of the 16S gene to define genovariants has been used by some as a proxy for strain diversity.[4,25,26] However, the

use of 16S genovariants to explore strain level variation and associations with health is limited by the resolution of genovariants and their correspondence to phylogeny.

A critical factor in evaluating strain level variation is how it is measured. Early studies employed multilocus sequence typing (MLST),[27] but recombination and horizontal gene transfer (HGT) can cause results to differ depending on the loci employed. Strain level variation has also been examined using gene ontology or copy number variation (CNV) analysis,[4,6] which has the advantage that many CNVs are functionally important. However, CNV can be hard to detect in low coverage samples and HGT can stimulate CNV.[28] Genome-wide SNP analysis has also been used,[2,4,5] but limited reference genomes for some species and variation in genome content present challenges.[29] Furthermore, widely divergent species may have limited core genomes and alignment methods for such divergent species present significant difficulties.[30] Additionally, it can be challenging to distinguish between strains with mixed ancestry from multi-strain samples.

The goal of the present study was to define and compare the population structure of common vaginal bacteria and identify patterns of strain level variation among vaginal microbiomes. We developed and validated a genome-wide SNP analysis based on available reference genomes. We applied this approach to metagenomic data from vaginal samples and found that diversity present among the vaginal samples was well represented by the available reference strains. We found species-specific differences in strain variation and structure, identifying clear groupings within most of the species. Although our power was limited, no strong associations between

strain and host factors were identified. Together, our results provide insight into how vaginal microbiome community types developed over the course of human history and lay the groundwork for assessing the importance of strain level variation in the vaginal microbiome and human health.

3.2 Methods

Metagenomic sequencing of vaginal samples

We obtained 197 cervicovaginal swabs from 195 pregnant women: 25 cervical swabs and 142 vaginal swabs (collectively referred to as vaginal samples) through the Global Alliance to Prevent Prematurity and Stillbirth (GAPPS) biobank and 30 vaginal swabs from the Women and Infants Health Consortium (WIHSC) at Washington University in St. Louis (IRB #201610121). When selecting samples from the GAPPS biobank, efforts were made to: 1) select all available specimens from women who delivered preterm (< 37 weeks of gestational age), 2) increase the representation of specimens from women of non-White race/ethnicity among the cohort, and 3) balance samples across all 3 trimesters. We augmented the samples selected from GAPPS with samples obtained from women currently enrolled in other studies with WIHSC. Patient data including gestational age at time of swab collection, gestational age at the time of birth, birthweight, maternal age and race/ethnicity were obtained from GAPPS and WIHSC. Women who delivered prior to 37 weeks of gestational age were considered preterm and represented both spontaneous and indicated preterm delivery. To extract genomic DNA, frozen vaginal swabs were eluted in 250 µL of an enzyme solution containing 0.5 mg ml^{-1} lysosozyme, 150 units ml^{-1} mutanolysin, 12 units ml^{-1} lysostaphin, 0.025 units ml^{-1} zymolase in 0.05 M potassium

phosphate buffer (pH 7.5) and incubated for 1 hour at 37°C. A ZR Fungal/Bacterial DNA MicroPrep Kit (Zymo Research) was used to extract and purify genomic DNA from swab elutions. Metagenomic sequencing libraries were prepared with a Nextera DNA Sample Prep Kit (Illumina) using 5 ng of genomic DNA and a small volume protocol.[31] PCR was performed using KAPA Hi-Fi HotStart ReadyMix (KAPA Biosystems) and libraries were purified with AMPure XP magnetic beads (Beckman Coulter). Libraries were pooled and sequenced on an Illumina NextSeq platform (75 cycles).

Sequence processing and classification of the vaginal microbiome

Sequence reads from our vaginal samples were trimmed and quality filtered using fqtrim (version 0.9.7) to remove reads less than 50 basepairs in length and trim read ends where quality scores drop below 10. Reads were then aligned to the human genome with Bowtie2[32] (version 2.3.4) and human reads were discarded. Metagenomic data from 128 vaginal specimens collected as part of the Human Microbiome Project[1] were obtained from NCBI's Sequence Read Archive (SRP002163). Data were filtered to remove reads less than 50 basepairs in length and remove human reads comprised of Ns using fastq-mcf (version 1.04.803). Taxonomic profiling was performed on non-human reads using MetaPhlAn2[33] (version 2.6.0). Each microbiome was classified into community types based on the dominant *Lactobacillus* species present, defined as 50% relative abundance or greater and referred to as, "*L. crispatus*-dominant", *"L. iners*-dominant"*, "L. gasseri*-dominant", or "*L. jensenii*-dominant". Communities without a single *Lactobacillus* species reaching 50% were referred to as "diverse", as most

communities without an abundant *Lactobacillus* species have high species diversity.[10] Read data for all vaginal samples were aligned with BWA and Stampy as described below.

Description of our clinical cohort

For the 195 women in our study, we obtain clinical and demographic data. Data on self-reported race/ethnicity showed most (96%) reported White, Black, Hispanic or Asian. The remaining women (4%) reported either American Indian/Alaskan Native, multiple races or their race/ethnicity was unknown. Maternal age (years), gestational age at sample collection (days), birthweight (grams) and gestational age at delivery (days) was also collected. Preterm delivery was defined as delivery prior to 37 weeks. Sixty-nine (35%) women had *L. crispatus*-dominant microbiomes, 53 (27%) had *L. iners*-dominant microbiomes, 9 (5%) had *L. jensenii*-dominant microbiomes, 9 (5%) had *L. gasseri*-dominant microbiomes and 55 (28%) had diverse microbiomes. A summary of clinical and demographic data can be found in Table 3.S1.We noted a higher prevalence of *L. crispatus*-dominant microbiomes among White (42%) than Hispanic (32%) or Black (20%) women; a higher prevalence of *L. iners*-dominant microbiomes among Hispanic (39%) and Black (35%) than White women (19%); and a higher prevalence of diverse microbiomes among Black (45%) than White (25%) or Hispanic (25%) women (Table 3.S2).

Reference strain analysis and validation

As reference for the vaginal samples and to identify the core genome we obtained genome assemblies for reference strains from NCBI for the six bacterial species of interest. A total of 101 *G. vaginalis*, 60 *L. crispatus*, 21 *L. iners*, 18 *L. jensenii*, 31 *L. gasseri*, and 5 *A. vaginae*

assemblies were obtained. Two *G. vaginalis* strains were not included in our analysis: UMB0388 which mapped extensively to other genomes, suggesting the assembly was not a pure isolate; and 6420LIT which had a particularly small genome when compared to all other *G. vaginalis* genomes. ART-MountRainer[34] (version 2.5.8) was used to generate simulated Illumina data (75 basepair reads, NextSeq500 platform v2) at 20x coverage for each assembly.

Simulated read data were aligned to a concatenated reference database containing a representative assembly for each species (Table 3.S3). A concatenated database was used in order to eliminate reads with low mapping quality due to equivalent mappings to multiple species. Alignments were performed with Stampy[30] (version 1.0.32) using the BWA-facilitated option and with an expected divergence of 0.05. We used this divergence parameter since higher rates of divergence (0.10) in some cases decreased the size of our filtered dataset, presumably due to higher rates of reads mapping to multiple reference sequences. Alignment data was filtered to remove reads with a mapping quality score of less than 10 using Samtools[35] (version 1.9).

We identified single nucleotide polymorphisms (SNPs) among all reference strains for a species and removed all variants not present within the core genome. The core genome was defined as all sites that had coverage for all reference strains in each species, and represented 12% (*G. vaginalis*), 47% (*L. crispatus*), 82% (*L. iners*), 72% (*L. jensenii*), 72% (*L. gasseri*) and 24% (*A. vaginae*) of the genome. Single nucleotide polymorphisms (SNPs) were called using GATK UnifiedGenotyper[36,37] (version 3.6.0). Variant calls were filtered using GATK to remove variants

that met the following criteria: QD < 50, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0. A small number of variants were removed due to cross-species read mapping. These sites were called based on reads from one species mapping to an incorrect reference genome, e.g. variant calls in the *L. crispatus* genome based on simulated reads from an *L. iners* reference assembly. Variant selection and removal was completed using VCFtools[38] (version 0.1.14). Among the reference strains, variant sites represented 15% (*G. vaginalis*), 3% (*L. crispatus*), 3% (*L. iners*), 2% (*L. jensenii*), 4% (*L. gasseri*) and 17% (*A. vaginae*) of the core genome.

To determine whether the choice of reference genome for read mapping affected the relationship between reference strains we aligned all simulated reference strain data to a second set of alternative reference genomes (Table 3.S3) and genotyped variants as described above. Variant sites with more than 2 alleles were then removed from both the original and alternate call set using VCFtools[38] (version 0.1.14) and genotypes were extracted using a custom script. A Euclidean distance matrix of strains for each species and each mapping reference was generated and compared by a Mantel test in R (version 3.5.1). The distance matrices for each mapping reference were found to be significantly correlated (Pearson's correlation coefficient > 0.9, $p < 0.05$) for each species, indicating that the choice of mapping reference genome did not alter the relationships among strains. Based on this finding we utilized a single mapping reference set (Primary Reference Set) for all analyses.

Alignment and SNP calling for vaginal samples

Sequence reads from vaginal samples obtained for this study (N = 197) and HMP samples (N = 128) were aligned to the reference genomes as described above. SNPs were independently called from alignment files for all 325 vaginal samples and 234 reference strains and filtered using GATK as described above. SNPs outside of the core genome and genotype calls with less than 4x coverage were removed with VCFtools. We removed vaginal samples with > 50% missing sites, and subsequently variant sites with > 10% missing genotypes across all samples. The resulting SNP dataset had a sample size of 668, with 234 from reference strains and 434 from vaginal samples (some vaginal samples had sufficient genotype calls in multiple species). Most samples (574/668) had an average read coverage over 10x at genotyped sites, and the average coverage was 53x in the vaginal samples.

Population structure

Population structure was inferred using both principal component analysis (PCA) and ADMIXTURE[39] and compared to phylogenetic trees. ADMIXTURE assumes free recombination among sites via sexual reproduction but PCA does not infer subpopulations and admixture proportions.[39] Our rationale was that both methods can handle genetic exchange between populations and should thus complement one another if they produce similar results. Additionally, because recombination can disrupt phylogenetic relationships, we employed the results of ADMIXTURE/PCA to eliminate recombinant samples from our phylogenetic analysis (next section).

The core genome variants were filtered to select for biallelic sites with VCFtools. Heatmaps of the hierarchically clustered variants and samples were generated in R using the heatmap function. Principal component analysis (PCA) was performed on the variant data in R with the package 'FactoMineR'. All principal components (PCs) explaining > 5% variance were assessed for associations with host factors (see Statistical Methods). For *L. iners* where no PC explained > 5% of the variance, we assessed the PC that explained the most variance.

VCF files containing core genome SNPs for reference strains and vaginal samples were converted to PLINK format with PLINK[40] (version 1.9). ADMIXTURE[39] (version 1.3.0) was then used to identify populations and infer ancestry. The number of populations (groups) was estimated based on the cross validation (CV) error for the number of groups {K1…10}. The estimated number of groups (K) was identified as the point at which the CV error plateaued to a minimum. Reference strains and samples with < 90% of ancestry estimated to be derived from a single group were classified as mixed ancestry.

For multi-strain *G. vaginalis* and *L. crispatus* samples we estimated subpopulation (group) abundance within vaginal samples using the relative allele depth at group-specific SNPs. Group-specific SNPs were defined as those with an allele frequency of 80% or more in one group but 20% or less in all other groups. Based on this designation we identified 22 (Group 1), 210 (Group 2), 19 (Group 3) and 178 (Group 4) SNPs out of 4,884 SNPs in the 88 *G. vaginalis* reference strains with less than 10% mixed ancestry. For *L. crispatus* we identified 604 (Group 1), 55 (Group 2) and 75 (Group 3) SNPs out of 4,469 SNPs. Using these SNPs we extracted the

allele depth supporting each nominally heterozygous genotype. The abundance of each group in each mixed sample was estimated by the average proportion of allele depth for each group-specific SNP.

Phylogenetic Analysis

Reference strains without mixed ancestry were used for phylogenetic analysis. Four-fold degenerate synonymous SNPs were selected using SNPeff[41] (version 4.3T) and SNPsift[42] (version 4.0). The SNPs for each sample were concatenated into FASTA format with VCF-kit.[43] The number of 4-fold synonymous sites surveyed was determined by identifying all 4-fold synonymous sites within the core genome using the same filters as described above except both variant and non-variant sites were retained. A distance matrix of pairwise differences in 4-fold synonymous SNPs/4-fold synonymous sites surveyed (4π) was used to generate a neighbor joining (NJ) tree in R with the packages 'ape' and 'phangorn'. We calculated Watterson's estimator of diversity[44] with the formula: $\hat{\theta}_w = \frac{S}{a_n}$, where S= the number of SNPs, $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$ and n is the sample size.

In parallel we identified 0-fold degenerate non-synonymous SNPs and 0-fold non-synonymous sites surveyed. The average pairwise difference in 0-fold non-synonymous variant sites/0-fold non-synonymous sites surveyed (0π) were determined. Tajima's D was calculated as previously described.[45]

For each species we estimated the average time to the most recent common ancestor in generations as $t=d/(2\mu)$, where μ is the mutation rate and d is the average or maximum pairwise distance between strains at synonymous sites. We used a bacterial mutation rate of 2×10^{-10} mutations per base pair per replication from *E. coli*.[46] We used an *in vitro* doubling rate of *G. vaginalis* (7.1 hours)[47] to estimate a replication rate of approximately 3.38 generations per day for all of the species and convert time in generations to time in years.

Statistical analysis

All principal components (PCs) explaining > 5% variance were assessed for associations with host factors. For *L. iners* where no PC explained > 5% of the variance, we assessed the PC that explained the most variance. Kruskal-Wallis and Spearman rank correlation tests were used to test for associations between PCs and host factors as appropriate. Due to multiple comparisons, a p-value below 0.001 was considered significant for associations with principal components and a p-value below 0.05 was considered significant for associations with microbiome community type. Statistical analysis was conducted in R.

3.3 Results

To characterize the diversity and structure of variation within common vaginal bacterial species, we generated metagenomic data from 197 vaginal swabs. A median of 94% of metagenomic reads per sample mapped to the human genome (range 50%-99%). After removing human reads, a median of 8.43×10^5 reads remained per sample, providing adequate coverage of the

microbial genome. An analysis of the composition of the microbiome using MetaPhlAn2 [33] indicated compositions similar to those described in prior studies with *Lactobacillus*-dominant and diverse community types: 71 (36.0%) were *L. crispatus*-dominant, 9 (4.6%) were *L. jensenii*-dominant, 53 (26.9%) were *L. iners*-dominant, 55 (27.9%) were diverse and 9 (4.6%) were *L. gasseri*-dominant (Table 3.S1). The prevalence of these community types differed by self-reported race/ethnicity (Table 3.S2).

Using MetaPhlAn2 community composition data, we identified species for strain analysis that were well represented among our vaginal samples. Six bacterial species (*L. crispatus*, *L. iners*, *L. jensenii*, *L. gasseri*, *G. vaginalis* and *A. vaginae*) showed a minimum relative abundance of 10% in at least 10 samples. When metagenomic data was mapped to a set of reference genomes (Table 3.S3), most of the samples (60%) showed more than 4x coverage to one or more of the bacterial species. This indicated that many of the samples had sufficient shotgun metagenomic data to identify single nucleotide polymorphisms (SNPs) and examine strain level variation. To increase the number of metagenomic samples for strain analysis, we included data from an additional 128 vaginal samples collected as part of the Human Microbiome Project (HMP).[1] A summary of the number of metagenomic samples included in the strain analysis is presented in Table 3.1.

A significant challenge to SNP identification from mixed metagenomic samples is being able to reliably call variants for the correct species. This is complicated by HGT, close relationships among the *Lactobacillus* species and variation in genome content. To address these issues, we

generated simulated metagenomic read data from publicly available reference genomes for each of the bacterial species (99 *G. vaginalis*, 60 *L. crispatus*, 21 *L. iners*, 18 *L. jensenii*, 31 *L. gasseri* and 5 *A. vaginae*). We aligned the simulated metagenomic reads to the reference set and found low mean misalignment rates for each species: 0.011 (*G. vaginalis*), 0.020 (*L. crispatus*), 0.003 (*L. iners*), 0.018 (*L. jensenii*), 0.017 (*L. gasseri*) and 0.092 (*A. vaginae*). While infrequent, misalignment did result in a small number of SNPs being called to the wrong species (e.g. variant calls in the *L. crispatus* genome based on simulated reads from an *L. iners* reference assembly). These invalid SNPs were excluded from our analysis. Next we identified the core genome SNPs for each species, based on sites represented among all reference strains. Within the core genome, we identified thousands of SNPs for each of the species (Table 3.2).

To evaluate strain variation among vaginal samples, we independently called SNPs among all vaginal samples and reference strains within core genomic regions. When the data were hierarchically clustered based on SNP profiles, clusters of strains were present for all of the species except *L. iners*. Notably, many of the vaginal samples contained numerous genotype calls with both alleles present (nominally heterozygotes), indicating the presence of multiple strains in a single sample (Figure 3.1). We conservatively defined samples as having multiple strains present if more than 10% of the SNPs were called heterozygous. Using this definition, we identified a high proportion of multi-strain samples for *G. vaginalis* (63%) and *A. vaginae* (77%) and a lower proportion (<22%) for the Lactobacillus species (Table 3.1).

To compare strain diversity among vaginal samples and isolated reference strains we performed principal component analysis (PCA) of the core genome SNPs. Plotting vaginal samples and reference strains by principal components (PCs) revealed that most reference strains formed clusters and the reference strains represent much of the variation observed among the vaginal samples (Figure 3.2).

PCA can distinguish different strain groups but it does not identify strains of mixed ancestry, which can occur through conjugation, transduction and transformation. To examine the structure of strain diversity within each species we used ADMIXTURE[39] to classify samples into groups (subpopulations) and identify samples with mixed ancestry to multiple groups. This analysis identified multiple groups for most of the vaginal species we studied: four *G. vaginalis*, three *L. crispatus*, three *L. gasseri* and two each for *L. jensenii* and *A. vaginae*. We did not identify any population structure for *L. iners*. These groups closely correspond to the clusters observed by PCA (Figure 3.2).

Consistent with the high proportion of multi-strain samples, many of the vaginal samples were inferred to be mixtures of groups. While ADMIXTURE is unable to distinguish vaginal samples with distinct strains from multiple groups, from samples with a single strain of mixed ancestry, it can identify reference strains with mixed ancestry. Among our reference panel, mixed ancestry was present but uncommon, representing 11 of 99 (11.1%) *G. vaginalis*, 8 of 60 (13.3%) *L. crispatus*, 0 of 18 (0.0%) *L. jensenii*, 1 of 31 (3.2%) *L. gasseri*, and 0 of 5 (0.0%) *A. vaginae* strains.

Among the vaginal samples inferred to have mixed ancestry, many were multi-strain samples. In these samples, the relative abundance of reads supporting each allele should be indicative of strain frequency in a sample. We thus used allele-specific read counts of population-specific SNPs to quantify relative abundance of each group in mixed vaginal samples. This was done for *G. vaginalis* and *L. crispatus*, which both have well defined groups based on multiple reference samples and an appreciable number of multi-strain vaginal samples. For each of the multi-strain samples the relative abundance of each group ranged from 10-90%, indicating inter-individual variation in strain frequency as well as strain type (Figure 3.S1). Additionally, the frequencies of different groups are inter-related in *G. vaginalis*. The frequency of Group 4 is negatively correlated with all other groups, whereas among the other groups only Group 1 and Group 3 are negatively correlated with each other (Table 3.S4).

Patterns and levels of strain level variation and population structure are shaped by historical effective population sizes, relationships among groups and strength of selection. To identify the relationships among strains and inferred groups we generated phylogenetic trees for reference strains without mixed ancestry. We constructed phylogenetic trees using synonymous four-fold degenerate sites for 88 *G. vaginalis*, 52 *L. crispatus*, 21 *L. iners*, 18 *L. jensenii*, 30 *L. gasseri* and five *A. vaginae* reference strains (Figure 3.3). Groups identified by ADMIXTURE could clearly be identified in the phylogenetic trees (Figure 3.3). We then estimated the genetic diversity within each species using the Watterson's estimator ($\hat{\theta}_w$)[44]. $\hat{\theta}_w$was high for both *G. vaginalis* (0.068) and *A. vaginae* (0.206) reflecting greater genetic diversity when compared to the lower values

observed for *L. crispatus* (0.012), *L. iners* (0.026), *L. jensenii* (0.011) and *L. gasseri* (0.028) (Table 3.2). Tajima's D measures the relative abundance of common versus rare alleles.[45] Population bottlenecks and population structure are expected to generate positive Tajima's D values and historical expansion of population size is expected to generate negative Tajima's D values. Tajima's D was negative for *G. vaginalis* (-1.062), *A. vaginae* (-0.610) and *L. crispatus* (-0.273) and positive for *L. iners* (0.806), *L. jensenii* (2.650) and *L. gasseri* (0.749) (Table 3.2).

The ratio of nonsynonymous to synonymous diversity is indicative of past selective pressure on a species. Species with higher constraints have lower ratios whereas species with low or altered constraints have higher ratios. We measured diversity at 0-fold degenerate sites (nonsynonymous) and compared it to 4-fold degenerate sites (synonymous). Two of the species, *L. crispatus* and *A. vaginae* had a much higher ratio of 0-fold to 4-fold diversity, 0.172 and 0.167 respectively, compared to the other species (Table 3.2).

The population structure of vaginal species raises the possibility that subpopulations may exhibit associations with their human host, similar to those associations present at the species level. To test for such associations, principal component values were extracted as a proxy for strain relationships (including groups) for 195 vaginal samples in our study and tested for associations with host factors (race/ethnicity, age and microbiome community type) and birth outcomes (preterm delivery and birth weight). We did not observe any statistically significant correlations (Table 3.S5); however, power analysis suggests that our study was only powered to detect large effect sizes.

3.4 Discussion

Strain level variation is thought to be functional, motivating fine scale measurement of strain variation and testing for its association with reproductive health. In this study we developed and validated a reference genome-based analysis of metagenomic vaginal samples to study the structure of strain level variation within and between individuals. We find reference genomes encompass the majority of strain level vaginal samples, thereby providing a means interpreting strain level variation and structure in vaginal samples. Despite occupying the same environment, we find differences in strain level variation, multi-strain samples, population structure and strength of selection among the vaginal species. Below, we discuss these results in relation to prior studies of strain level variation, and the ecology and evolutionary history of vaginal strains relevant to identifying functional differences among groups and their role in human health.

Strain level variation

Our analysis of core genomic SNPs provides fine-scale measures of strain level variation and captures known and new aspects of population structure present in vaginal species. Previous genomic studies encompassed only reference genomes, were limited to smaller sample sizes, or did not accommodate multi-strain or admixed samples (Table 3.S6). We made use of combined metagenomic and reference genomes to 1) survey metagenomic SNPs in core genomic regions and establish that most metagenomic variation is captured in reference genomes, and 2) identify groups or subpopulations within each species while accounting for a number of mixed

ancestry genomes and numerous multi-strain samples. Mixed ancestry due to genetic exchange can confound phylogenetic analysis of strain level variation,[48] and multi-strain samples are difficult to resolve due to the challenges of accurate assemblies from metagenomic data.[49,50] While our approach does not resolve multi-strain samples into individual lineages, we make use of multiple allele genotype calls to estimate relative abundance of groups within multi-strain samples.

Inference of population structure and admixture can depend on the methods and data used. Our dataset includes some samples (94/668) with low (4-10x) coverage. Although these samples may have higher rates of genotyping error, they were intermingled with high coverage samples in each subpopulation and represented a small fraction of admixed strains (17/117). Thus, while the branch tips leading to low coverage samples may be slightly longer due to genotyping error, the exclusion of these samples would not affect the population structure of strain level variation, Our analyses used PCA, ADMIXTURE and phylogenetics to identify population structure. While our use of ADMIXTURE violates the program's assumption of free recombination via sexual reproduction, the results were consistent with PCA. Additionally, population groups were further supported by phylogenetic groups after removing mixed ancestry strains. Thus, our results support a consistent picture of population structure with genetic exchange.

With the exception of strains showing mixed ancestry, the structure of bacterial strain diversity is largely consistent with prior studies (Table 3.S6). We find no population structure of *L. iners*,

consistent with previous genomic analyses that showed a highly conserved genome among *L. iners* strains with little difference in gene content.[4,24] A lack of population structure does not convey a lack of strain diversity. Indeed, most *L.* iners strains identified appeared to be unique and average nucleotide diversity among *L. iners* strains was greater than that seen among *L. crispatus* strains (Table 3.1). The three *L. gasseri* groups we identified correspond to two previously defined groups with distinct gene content.[51] Notably, recent studies[52,53] suggest our *L. gasseri* Group 3 represents the closely related *L. paragasseri*. The three *L. crispatus* groups correlate with two previously reported groups described in a genomic analysis of 41 strains.[4] Notably, most vaginal samples harbored Group 2 or Group 3 strains, while reference isolates from avian hosts were common in the more diverse Group 1. This suggests that Group 1 strain colonization of the human vagina may be rare.

The four groups of *G. vaginalis* that we found encompass and are largely consistent with prior groups (Table 3.S7).[4,13-15] However, a number of these previously defined groups correspond to strains we find to have mixed ancestry. *Ahmed et al.* proposed the division of the species into four groups after a phylogenetic analysis of the core genome of 17 isolated strains.[13] Subsequent studies of the *cpn60* gene[17,22,54] as well as our strain group assignments are consistent with those described by *Ahmed et al.*[13] One exception is that strains 1400E and 55152, which were assigned to Group 1 but our analysis suggested were of mixed ancestry (mostly Groups 1 and 2). However, assemblies may give the appearance of mixed ancestry if unknowingly generated using a mixture of two or more strains. More recent studies[4,14,15] have expanded the number of strains as well as the number of groups (Table 3.S6). However, many

of these new groups are comprised of strains our analysis indicates are of mixed ancestry. The placement of strains with mixed ancestry into a separate group is not incorrect; such groups may be functionally distinct. While our approach to inferring population structure does not place mixed ancestry strains into separate groups, our results provide insight into the historical origin of these mixed ancestry groups.

Mixed ancestry strains likely derive from genetic exchange between population groups. While we did not investigate the mechanism of exchange, *G. vaginalis* encodes predicted competence-promoting proteins, prophages, and transposable elements. Moreover, prior studies have provided evidence of HGT.[13,55-57] A recent analysis of *G. vaginalis* core and accessory genomes within and between populations suggests that population structure is maintained through barriers to genetic exchange.[57] Patterns of HGT appear to differ between population groups, with some groups displaying more evidence of genetic exchange than others. Interestingly, this analysis suggests that intergroup HGT may have been more common in the remote past and indicates that at least some genes (e.g. vaginolysin) are freely exchanged within and between groups.[57]

Ecological diversity of the vaginal microbiome

Vaginal microbial diversity has been correlated with reproductive health and determinants of this diversity are of significant clinical interest. Species diversity within the vaginal niche is determined through ecological interactions within that environment. A key correlate of species diversity is vaginal pH. A vaginal pH less than 4.5 is thought of as healthy and is associated with

low diversity, *Lactobacillus*-dominated communities.[10,58] It is believed that through the production of lactic acid, *Lactobacillus* species are able to outcompete other vaginal bacteria and dominate that niche.[58] Among *Lactobacillus*-dominated microbiomes, multiple *Lactobacillus* species may be present but a single *Lactobacillus* species usually dominates.[10] This suggests that these species may occupy very similar niches within the vagina. According to ecological theory, multiple species cannot occupy the same niche indefinitely and one species will eventually outcompete the others.[59] Conversely, a more neutral pH correlates with greater diversity and an abundance of BV-associated anaerobes including *G. vaginalis* and *A. vaginae*.[10] These polymicrobial communities support multiple species which may be explained by the theory of resource partitioning in which competing species utilize different subsets of resources to occupy niche divisions within an environment.[60] However, such a co-occurrence could also result from spatial structure within the vagina.[61]

We find that patterns of strain level variation mimic those of species level diversity. We observed greater strain level diversity within *G. vaginalis* and *A. vaginae*, which are found in more species-diverse communities. Furthermore, most samples with *G. vaginalis* (62.7%) and *A. vaginae* (76.7%) harbor multiple strains from different groups, while multi-strain samples among lactobacilli are much less common (2-21%). The high frequency of multi-strain samples is consistent with prior studies of *G. vaginalis*[4,20-22] and *A. vaginae*.[62] The co-occurrence of different groups can be explained by ecotype theory which suggests that different strains of the same species may occupy the same niche if they function as different ecological species (ecotypes), exploiting different resources.[7,18]

The co-occurrence of differentiated groups within vaginal communities is important for understanding group associations with health. The presence of multiple groups of *G. vaginalis* has been correlated with BV.[20,22] We find that the frequency of Group 4 is negatively correlated with the frequency of all other groups in mixed vaginal samples, potentially indicating that it competitively excludes these groups. In contrast, Group 1, 2 and 3 co-occur but only 1 and 3 are negatively correlated. This is particularly interesting as the co-occurrence of multiple groups of *G. vaginalis* has been correlated with BV.[20,22] Additionally, prior studies have failed to show an associated between Group 4 and BV,[20,22] which may indicate that Group 4 strains are less pathogenic. These findings suggest that mixed group communities may confound *G. vaginalis* group associations with vaginal health and should be accounted for in future models.

While a low pH may enable *Lactobacillus* species to exclude high pH species from the vaginal niche, this does not explain why multiple *Lactobacillus* strains are not observed more frequently in the same sample. If *Lactobacillus* groups represented distinct ecotypes, one would expect groups to co-occur as observed with *G. vaginalis* and *A. vaginae*. One potential explanation is that there has not yet been enough time for *Lactobacillus* strains to diversify and evolve resource partitioning strategies.

Evolutionary origins of strain level diversity

Strain level diversity is indicative of the species' demographic history, including past changes in population size, population structure and migration between host microbiomes or other environments. Some insight into the evolutionary origins of the vaginal microbiome may be gleaned by comparing it to microbiome composition of other primates. While vaginal microbial signatures of non-human primates are unique to each species, community compositions more closely resemble the diverse structure associated with *G. vaginalis* and *A. vaginae*.[63-65] Only humans are dominated by *Lactobacillus* species.[63] While *Lactobacillus* species are closely associated with food and agriculture,[66] the species that dominate the human vagina have reduced genome sizes when compared to other *Lactobacillus* species suggesting adaptation to the host environment.[67]

Among the species studied here, *A. vaginae* and *G. vaginalis* exhibited the greatest strain divergence which is consistent with large, long-term populations of *G. vaginalis* and *A. vaginae* in humans as part of an ancestral state. The *Lactobacillus* species showed much less diversity, which may reflect a smaller historic population size or more recent colonization during human history. Among the *Lactobacillus* species, *L. iners* and *L. gasseri* are the most divergent. The diversity of *L. iners* may reflect a larger historic population size and earlier association with human vaginal microbiomes compared to other *Lactobacillus* species. This idea is consistent with it being the most prevalent (most frequently detected) of the *Lactobacillus* species.(10) *L. gasseri* strain divergence is also greater than the other *Lactobacillus* species, but this is partly caused by strong divisions between groups, which could predate colonization of the human vaginal microbiome. *L. jensenii* has low diversity and a large positive Tajima's D, consistent with

a recent bottleneck, potentially related to colonization of the vaginal niche. Together, these differences in strain divergence support the hypothesis that the vaginal microbiome of ancestral humans was more similar to modern diverse communities, and that the emergence of *Lactobacillus* species is a more recent event in the evolution of the human vaginal microbiome.

The relative differences in divergence prompted us to examine the timescale over which strain divergence may have occurred and align this timescale with the migration of human populations out of Africa. Using an experimental estimate of mutation rate from *E. coli*[46] and an *in vitro* estimate of *G. vaginalis* replication rate[47] we estimated average time to most recent common ancestor using pairwise divergence between the two most distantly related strains of a species. This analysis gave a divergence time of 800 (*A. vaginae*), 180 (*G. vaginalis*), 144 (*L. gasseri*), 72 (*L. jensenii*), 71 (*L. iners*) and 31 (*L. crispatus*) thousand years. However, these divergence times could be off by a factor of two or more since they depend on a general estimate for replication time and mutation rate that likely differ from the true values at the species and possibly even the strain level. As such, these numbers should be interpreted with caution. Our estimates indicated *A. vaginae* and *G. vaginalis* groups likely diverged prior to the migration of modern humans out of Africa.[68] Of the two most commonly found *Lactobacillus* species, our estimates suggest that *L. iners* diverged around the time when modern humans were beginning to disperse out of Africa,[68] while *L. crispatus* diverged after the time that it is believed modern humans settled Europe.[68] These observations seem to parallel earlier findings from many research groups showing that vaginal microbiomes with an abundance of *G. vaginalis* and/or *L. iners* are more common in women of African descent, whereas an

abundance of *L. crispatus* within the vaginal microbiome is more commonly found in women of European descent.[10,11,69]

Conclusion

Our results show that most species are characterized by multiple distinct groups of strains, and that strain diversity and the frequency of multi-strain samples is related to species-level diversity of the microbiome in which they occur. Future work will need to uncover the ecological variables that impact variation within and between communities at the strain level, and the historical genomic and functional differentiation that led to extant population structure. Doing so will not only help resolve the role of strain variation in the vaginal microbiome as it relates to reproductive disease, but could also provide insight into the establishment and subsequent changes in community composition as a function of important gynecologic and obstetric events including: sexual development, the menstrual cycle, pregnancy and menopause.[69]

Table 3.1: Metagenomic vaginal samples included in strain analysis for each species.

	G. vaginalis	*L. crispatus*	*L. iners*	*L. jensenii*	*L. gasseri*	*A. vaginae*
Total samples	67 (100)	137 (100)	115 (100)	53 (100)	32 (100)	30 (100)
From this study	42 (62.7)	71 (51.8)	70 (60.9)	20 (37.7)	16 (50.0)	16 (53.3)
From HMP	25 (37.3)	66 (48.2)	45 (39.1)	33 (62.3)	16 (50.0)	14 (46.7)
Multi-strain samples	42 (62.7)	29 (21.2)	22 (19.1)	1 (1.9)	3 (9.4)	23 (76.7)

Values are n (%)

Table 3.2: SNP counts and nucleotide diversity measures.

	G. vaginalis	*L. crispatus*	*L. iners*	*L. jensenii*	*L. gasseri*	*A. vaginae*
Core genome sites	2.01E+05	1.00E+06	1.06E+06	1.19E+06	1.41E+06	3.52E+05
Core genome SNPs	3.12E+04	2.61E+04	3.24E+04	1.84E+04	6.01E+04	6.14E+04
4-fold degenerate SNPs	1.04E+04	6.60E+03	1.10E+04	5.49E+03	1.85E+04	2.33E+04
0-fold degenerate SNPs	7.33E+03	8.36E+03	7.08E+03	3.04E+03	1.23E+04	1.70E+04
Average π4	0.047	0.011	0.031	0.019	0.033	0.190
Average π0	0.003	0.002	0.003	0.001	0.003	0.032
Average π0/π4	0.063	0.172	0.100	0.068	0.077	0.167
θ_w	0.068	0.012	0.026	0.011	0.028	0.206
Tajima's D	-1.06	-0.273	0.806	2.65	0.749	-0.610

Table 3.S1: Summary of clinical and demographic data for women in this study

Characteristic	N	% of Cohort
Total	195	-
Race/ethnicity		
White	97	50%
Hispanic	56	29%
Black	20	10%
Asian	14	7%
Other	8	4%
Community Type		
L. crispatus-dominant	69	35%
L. jensenii-dominant	9	5%
L. iners-dominant	53	27%
Diverse	55	28%
L. gasseri-dominant	9	5%
Preterm Delivery (<37 weeks)		
Yes	41	21%
No	154	79%
Age (years)		
Median	29	-
IQR	25 - 33	-
Gestational Age at Collection (days)		
Mean	164	-
IQR	88 - 247	-
Gestational Age at Birth (days)		
Mean	274	-
IQR	262 - 280	-
Birth Weight (grams)		
Mean	3270	-
IQR	2904 - 3692	-

Table 3.S2: Summary of vaginal microbiome community type by self-reported race

Race/ethnicity	*L. crispatus*-dominant	*L. gasseri*-dominant	*L. iners*-dominant	Diverse	*L. jensenii*-dominant
White	41 (42)	7 (7)	18 (19)	24 (25)	7 (7)
Hispanic	18 (32)	0 (0)	22 (39)	14 (25)	2 (4)
Black	4 (20)	0 (0)	7 (35)	9 (45)	0 (0)
Asian	5 (36)	1 (7)	2 (14)	6 (43)	0 (0)
Other	1 (13)	1 (13)	4 (50)	2 (25)	0 (0)

Values are n (%)

Table 3.S3: Reference strains used in reference databases

Primary Reference Set		
Species	Strain	Assembly
L. crispatus	ST1	GCA_000091765.1
L. iners	DSM 13335	GCA_000160875.1
L. jensenii	SNUV360	GCA_001936235.1
L. gasseri	ATCC 33323	GCA_000014425.1
G. vaginalis	ATCC 14019	GCA_000159155.2
A. vaginae	PB189-T1-4	GCA_000179715.1
Secondary Reference Set		
Species	Strain	Assembly
L. crispatus	JV-V01	GCA_000160515.1
L. iners	LactinV 03V1-b	GCA_000149105.2
L. jensenii	115-3-CHN	GCA_000162435.1
L. gasseri	AL3	GCA_002007185.1
G. vaginalis	6420B	GCA_000263575.1
A. vaginae	DSM 15829	GCA_000159235.2

Table 3.S4: Correlation between *G. vaginalis* groups within multi-strain vaginal samples

	Group 1	Group 2	Group 3	Group 4
Group 1		0.19531	-0.36107	-0.65611
Group 2	0.24000		0.05515	-0.56748
Group 3	0.03309	0.73200		-0.49208
Group 4	0.00002	0.00013	0.00170	

Pearson's correlation coefficient above, p-value below diagonal

Table 3.S5a: Summary of statistical analysis of strain association with host factors for *G. vaginalis*

PCs assessed

PC	% Variance
PC1	23.93
PC2	14.33
PC3	7.19

Races/ethnicities assessed

Race	N
Black	9
Hispanic	14
White	16

Community types assessed

Community Type	N
L. iners-dominant	6
Diverse	33

P-values

Factor	PC1 (P)	PC2 (P)	PC3 (P)	N
Race/ethnicity	0.72	0.85	0.84	39
Age	0.33	0.92	0.15	42
Community Type	0.15	0.64	0.17	39
Preterm delivery	0.45	0.97	0.66	42
Birthweight	0.24	0.96	0.94	41

Table 3.S5.b: Summary of statistical analysis of strain association with host factors for L. crispatus

PCs assessed

PC	% Variance
PC1	50.47
PC2	9.46

Races/ethnicities assessed

Race	N
Asian	6
Hispanic	20
White	39
Black	5

Community types assessed

Community Type	N
L. crispatus-dominant	67

P-values

Factor	PC1 (P)	PC2 (P)	N
Race/ethnicity	0.04	0.05	70
Age	0.40	0.57	72
Community Type	-	-	-
Preterm delivery	0.29	0.31	72
Birthweight	0.14	0.57	71

Table 3.S5c: Summary of statistical analysis of strain association with host factors for *L. iners*

PCs assessed

PC	% Variance
PC1	4.05

Races/ethnicities assessed

Race	N
Asian	7
Black	8
Hispanic	25
White	26

Community types assessed

Community Type	N
L. iners-dominant	49
Diverse	15
L. crispatus-dominant	6

P-values

Factor	PC1 (P)	N
Race/ethnicity	0.08	66
Age	0.45	70
Community Type	0.74	70
Preterm delivery	0.19	70
Birthweight	0.07	69

Table 3.S5d: Summary of statistical analysis of strain association with host factors for *L. jensenii*

PCs assessed

PC	% Variance
PC1	92.31

Races/ethnicities assessed

Race	N
Hispanic	9
White	9

Community types assessed

Community Type	N
L. iners-dominant	6
L. jensenii-dominant	8

P-values

Factor	PC1 (P)	N
Race/ethnicity	0.31	18
Age	0.37	20
Community type	0.61	14
Preterm delivery	0.06	20
Birthweight	0.48	20

Table 3.S5e: Summary of statistical analysis of strain association with host factors for *L. gasseri*

PCs assessed

PC	% Variance
PC1	69.55
PC2	13.30

Races/ethnicities assessed

Race	N
White	9

Community types assessed

Community Type	N
L. gasseri-dominant	6

P-values

Factor	PC1 (P)	PC2 (P)	N
Race/ethnicity	-	-	-
Age	0.07	0.81	15
Community type	-	-	-
Preterm delivery	-	-	-
Birthweight	0.36	0.37	15

Table 3.S5f: Summary of statistical analysis of strain association with host factors for *A. vaginae*

PCs assessed

PC	% Variance
PC1	54.18
PC2	8.15

Races/ethnicities assessed

Race	N
Black	6
White	5

Community types assessed

Community Type	N
Diverse	11

P-values

Factor	PC1 (P)	PC2 (P)	N
Race/ethnicity	0.10	0.27	11
Age	0.15	0.11	11
Community type	-	-	-
Preterm delivery	-	-	-
Birthweight	0.98	0.89	11

Table 3.S6: Summary of strain group literature by species

G. vaginalis

PMID	Author	Published	Number of strains in analysis	Groups
22609915	*Ahmed et al.*	2012	17 reference strain genomes	4
30232199	*Goltsman et al.*	2018	34 reference strain genomes, 6 strain genomes reconstructed from metagenomic samples	5
30648938	*Vaneechoutte et al.,*	2019	81 reference strain genomes	13
31462445	*Potter et al.*	2019	103 reference strain genomes	9

L. iners

PMID	Author	Published	Number of strains in analysis	Groups
27694231	*France et al.*	2016	15 reference strain genomes	NA
30232199	*Goltsman et al.*	2018	16 reference strain genomes, 8 strain genomes reconstructed from metagenomic samples	No groups identified

L. crispatus

PMID	Author	Published	Number of strains in analysis	Groups
25480015	*Ojala et al.*	2014	10 reference strain genomes	NA
27694231	*France et al.*	2016	15 reference strain genomes	NA
30232199	*Goltsman et al.*	2018	37 reference strain genomes, 6 strain genomes reconstructed from metagenomic samples	2

L. gasseri

PMID	Author	Published	Number of strains in analysis	Groups
29038772	*Tada et al.*	2017	75 reference strain genomes	2

Table 3.S7: Comparing *G. vaginalis* group assignment in our study to others

Strain Name	Our analysis	*Ahmed et al.*[13]	*Goltsman et al.*[4]	*Vaneechoutte et al.*[14]	*Potter et al.*[15]
55152	Mixed Ancestry	Group 1	Clade 1	2	GS01
75712	Group 1	Group 1	Clade 1	1	GS01
3549624	Group 1	NA	NA	1	GS01
0288E	Group 1	Group 1	Clade 1	1	GS01
1400E	Mixed Ancestry	Group 1	Clade 1	2	GS01
14019_MetR	Group 1	NA	NA	1	GS01
18-4	Group 1	NA	NA	1	GS01
23-12	Group 1	NA	NA	1	GS01
284V	Group 1	Group 1	Clade 1	1	GS01
315-A	Group 1	NA	Clade 1	1	GS01
41V	Mixed Ancestry	NA	Clade 1	2	GS01
ATCC 14018	Group 1	Group 1	Clade 1	1	GS01
ATCC 14019	Group 1	Group 1	Clade 1	1	GS01
ATCC 49145	Group 1	NA	NA	1	GS01
DNF01149	Group 1	NA	NA	1	GS01
DSM 4944	Group 1	NA	NA	1	GS01
FDAARGOS_296	Group 1	NA	NA	1	GS01
GH015	Group 1	NA	NA	NA	GS01
GH021	Group 1	NA	NA	NA	GS01
HMP9231	Group 1	NA	Clade 1	1	GS01
JCM 11026	Group 1	NA	NA	1	GS01
JCP7275	Group 1	NA	Clade 1	1	GS01
JCP7276	Group 1	NA	Clade 1	1	GS01
JCP7672	Group 1	NA	Clade 1	1	GS01
JCP8108	Mixed Ancestry	NA	Clade 1	2	GS01
N165	Group 1	NA	NA	NA	GS01
NR001	Group 1	NA	NA	NA	GS01
NR037	Group 1	NA	NA	NA	GS01
NR038	Group 1	NA	NA	NA	GS01
NR039	Group 1	NA	NA	NA	GS01
S2_012_000_R3_92	Group 1	NA	NA	NA	GS01
S2_012_000_R3_93	Group 1	NA	NA	1	GS01
UGent 09.01	Group 1	NA	NA	1	GS01
UGent 09.07	Group 1	NA	NA	1	GS01
UGent 25.49	Group 1	NA	NA	1	GS01
UMB0032A	Group 1	NA	NA	1	GS01

UMB0032B	Group 1	NA	NA	1	GS01
UMB0061	Group 1	NA	NA	1	GS01
UMB0233	Group 1	NA	NA	1	GS01
UMB0298	Group 1	NA	NA	1	GS01
UMB0386	Group 1	NA	NA	1	GS01
UMB0768	Group 1	NA	NA	1	GS01
UMB0770	Group 1	NA	NA	1	GS01
UMB0775	Group 1	NA	NA	1	GS01
WP023	Group 1	NA	NA	NA	GS01
00703Bmash	Group 2	Group 2	Clade 2	3	GS02
00703C2mash	Group 2	Group 2	Clade 2	3	GS02
GED7275B	Group 2	NA	NA	3	GS02
GH007	Group 2	NA	NA	NA	GS02
GH020	Group 2	NA	NA	NA	GS02
JCP7659	Group 2	NA	Clade 2	3	GS02
JCP7719	Group 2	NA	Clade 2	3	GS02
JCP8017A	Group 2	NA	Clade 2	3	GS02
JCP8017B	Group 2	NA	Clade 2	3	GS02
JCP8066	Group 2	NA	Clade 2	4 (*G. piotti*)	GS02
JCP8070	Group 2	NA	Clade 2	4 (*G. piotti*)	GS02
JCP8151A	Group 2	NA	Clade 2	4 (*G. piotti*)	GS02
JCP8151B	Group 2	NA	Clade 2	4 (*G. piotti*)	GS02
JCP8522	Group 2	NA	Clade 2	4 (*G. piotti*)	GS02
N101	Group 2	NA	NA	NA	GS02
N144	Group 2	NA	NA	NA	GS02
N153	Group 2	NA	NA	NA	GS02
N95	Group 2	NA	NA	NA	GS02
UGent 18.01	Group 2	NA	NA	4 (*G. piotti*)	GS02
UGent 21.28	Group 2	NA	NA	4 (*G. piotti*)	GS02
UMB0830	Group 2	NA	NA	3	GS02
UMB0833	Group 2	NA	NA	3	GS02
W11	Group 2	NA	NA	NA	GS02
5-1	Group 4	Group 4	Clade 4	6 (*G. swidsinskii)*	GS03
409-05	Group 4	Group 4	Clade 4	6 (*G. swidsinskii)*	GS03
6420B	Group 4	Group 4	Clade 4	5 (*G. leopoldi*)	GS03
AMD	Group 4	Group 4	Clade 4	5 (*G. leopoldi*)	GS03
GS 10234	Group 4	NA	NA	6 (*G. swidsinskii)*	GS03
GS 9838-1	Group 4	NA	NA	6 (*G. swidsinskii)*	GS03

GV37	Group 4	NA	NA	6 (*G. swidsinskii)*	GS03
N72	Group 4	NA	NA	NA	GS03
UGent 06.41	Group 4	NA	NA	5 (*G. leopoldi*)	GS03
UGent 09.48	Group 4	NA	NA	5 (*G. leopoldi*)	GS03
UMB0170	Group 4	NA	NA	6 (*G. swidsinskii)*	GS03
UMB0264	Group 4	NA	NA	6 (*G. swidsinskii)*	GS03
UMB0682	Group 4	NA	NA	5 (*G. leopoldi*)	GS03
UMB0912	Group 4	NA	NA	5 (*G. leopoldi*)	GS03
UMB0913	Group 4	NA	NA	5 (*G. leopoldi*)	GS03
UMB1642	Group 4	NA	NA	6 (*G. swidsinskii)*	GS03
JCP8481A	Mixed Ancestry	NA	Clade 5	7	GS04
JCP8481B	Mixed Ancestry	NA	Clade 5	7	GS04
PSS_7772B	Mixed Ancestry	NA	NA	7	GS04
101	Group 3	NA	Clade 3	8	GS05
00703Dmash	Group 3	Group 3	Clade 3	8	GS05
1500E	Group 3	Group 3	NA	10	GS05
6119V5	Group 3	Group 3	Clade 3	9	GS05
N160	Group 3	NA	NA	NA	GS05
UMB1686	Group 3	NA	NA	8	GS05
GED7760B	Mixed Ancestry	NA	NA	11	GS06
CMW7778B	Mixed Ancestry	NA	NA	12	GS07
KA00225	Mixed Ancestry	NA	NA	13	GS08
NR010	Mixed Ancestry	NA	NA	NA	GS09
FDAARGOS_568	Group 1	NA	NA	NA	NA
GH019	Group 2	NA	NA	NA	NA

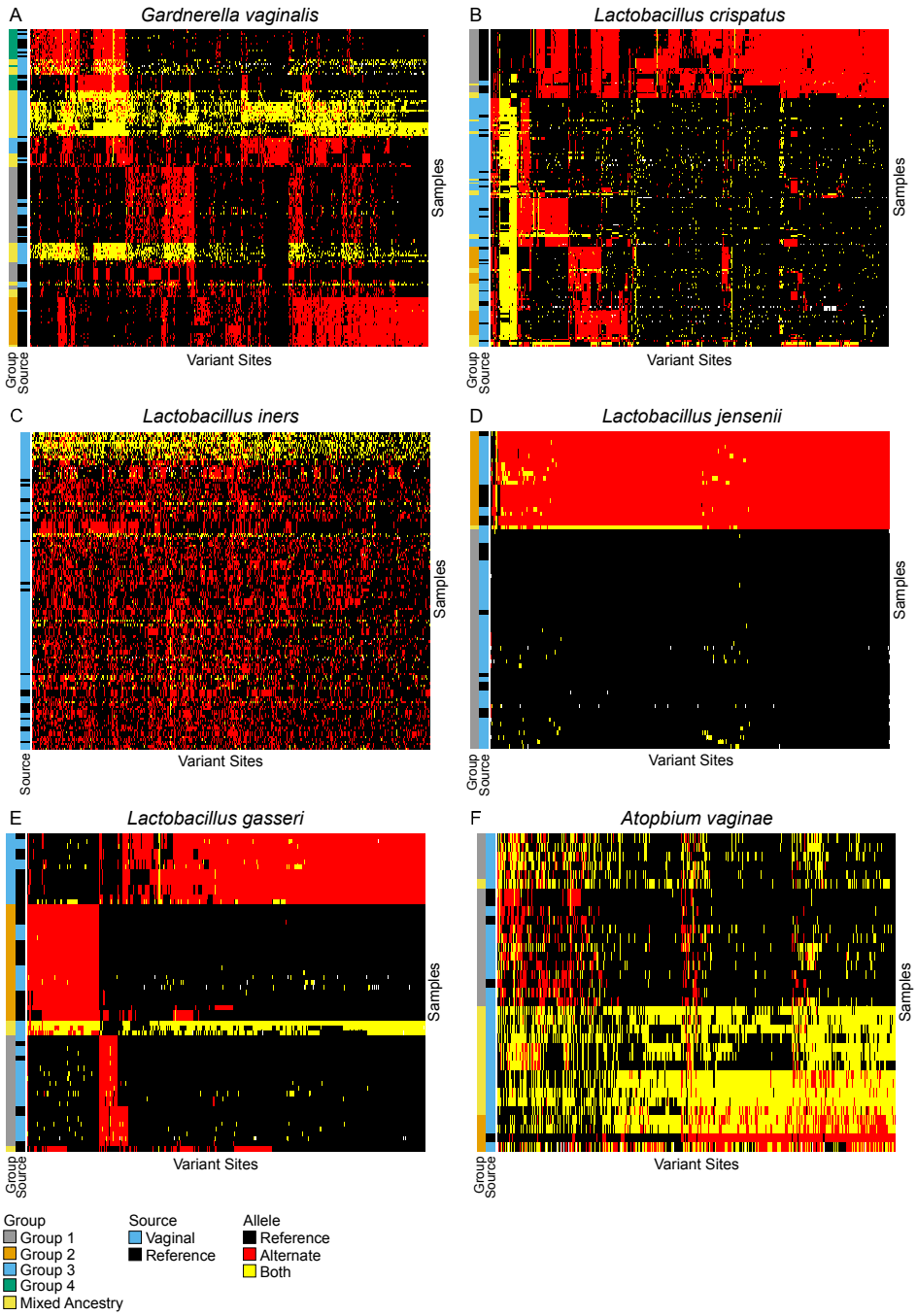

Figure 3.1: Heatmaps of core genome SNPs

Hierarchically clustered core genome SNPs for reference and vaginal samples for A) *G. vaginalis*, B) *L. crispatus*, C) *L. iners*, D) *L. jensenii*, E) *L. gasseri* and F) *A. vaginae.* The reference allele is indicated by Black, the alternate allele is indicated by red and the presence of both alleles is indicated by yellow, missing data is indicated by white. To the left of each heatmap is a bar indicating the source of the samples either vaginal or reference strain, and the group to which the strain for each sample was assigned using ADMIXTURE. Samples with mixed ancestry or multi-strain samples from multiple groups are identified as “Mixed Ancestry”.

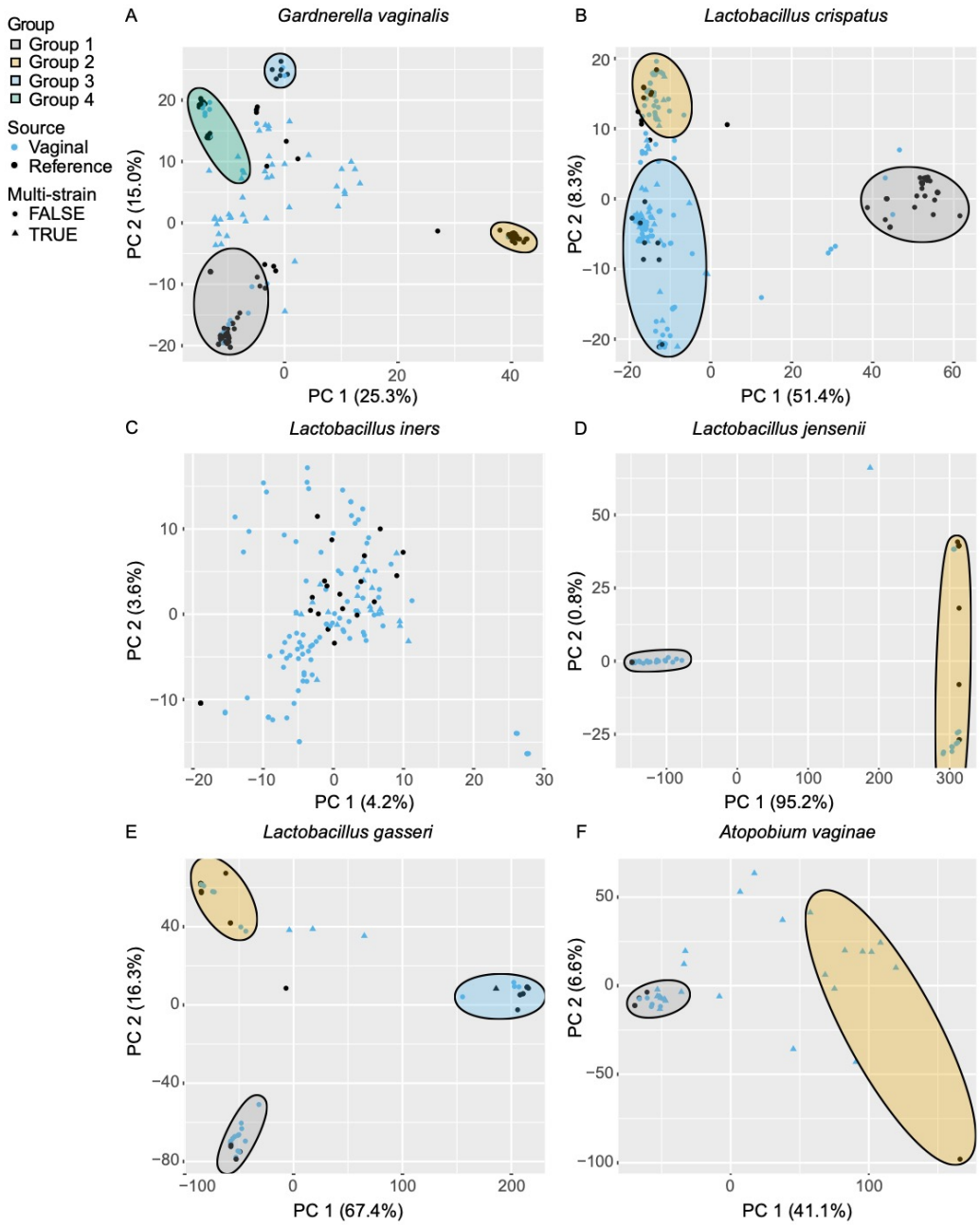

Figure 3.2: Principal component analysis of core genome SNPs
Principal component analysis (PCA) of SNP data for A) *G. vaginalis*, B) *L. crispatus*, C) *L. iners*, D) *L. jensenii*, E) *L. gasseri* and F) *A. vaginae.* Each vaginal sample and reference strain is shown by their PC1 and PC2 coordinates. Vaginal samples are indicated by blue points and reference strains are indicated by black points. Samples that were identified as multi-strain are represented by triangles and single strain samples are represented by circles. Subpopulation groups were determined by ADMIXTURE analysis and ellipses drawn to show vaginal samples and reference strains belonging to a single group.

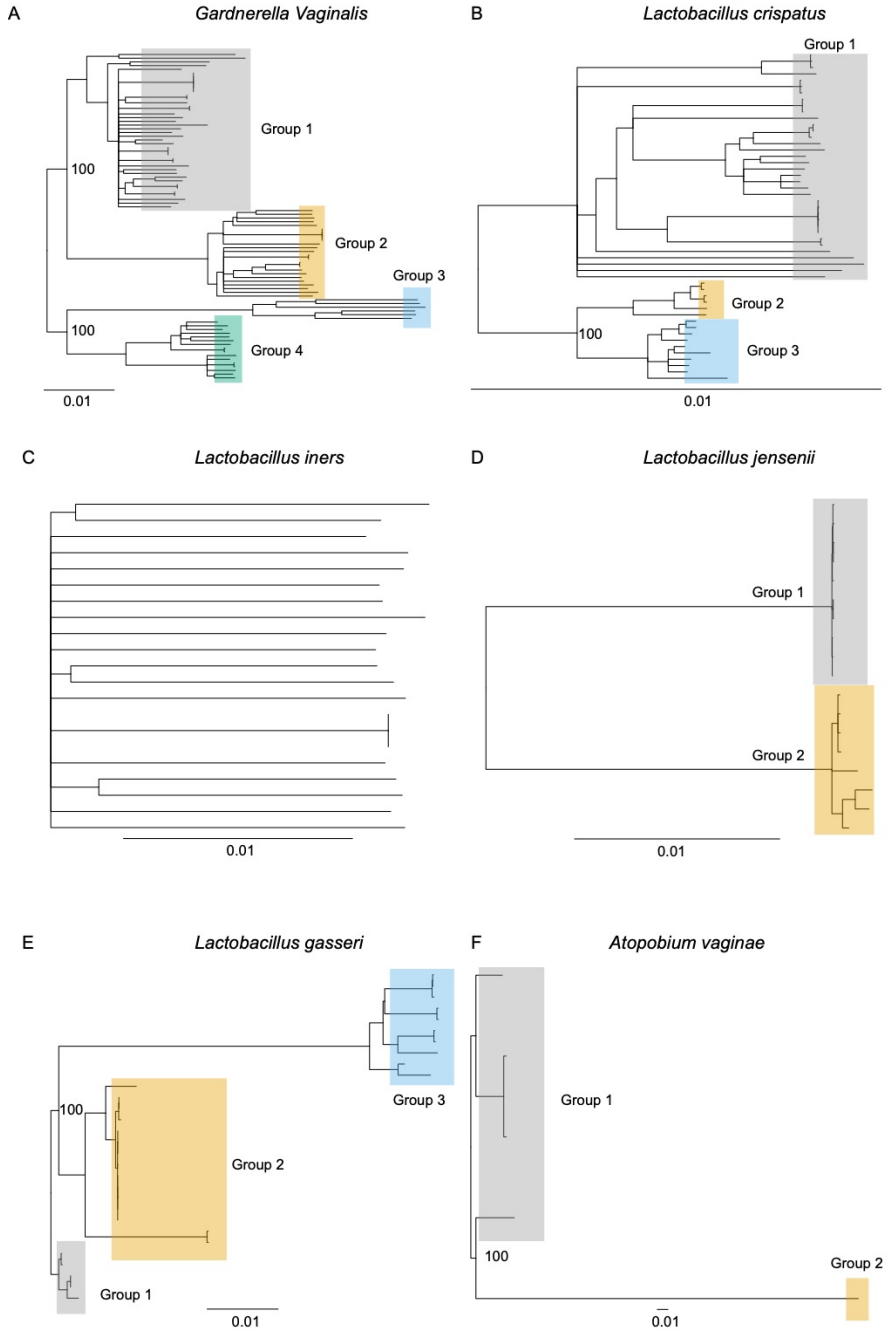

Figure 3.3: Phylogenetic trees of reference strains
Neighbor joining trees of reference strains created from synonymous sites for A) *G. vaginalis*, B) *L. crispatus*, C) *L. iners*, D) *L. jensenii*, E) *L. gasseri*, F) *A. vaginae*. Branch lengths represent pairwise differences per site surveyed. Groups were determined by ADMIXTURE analysis. Select bootstrap values for nodes separating groups are shown. Bootstrap values indicate the number of supporting iterations out of 100 as calculated by resampling with replacement.

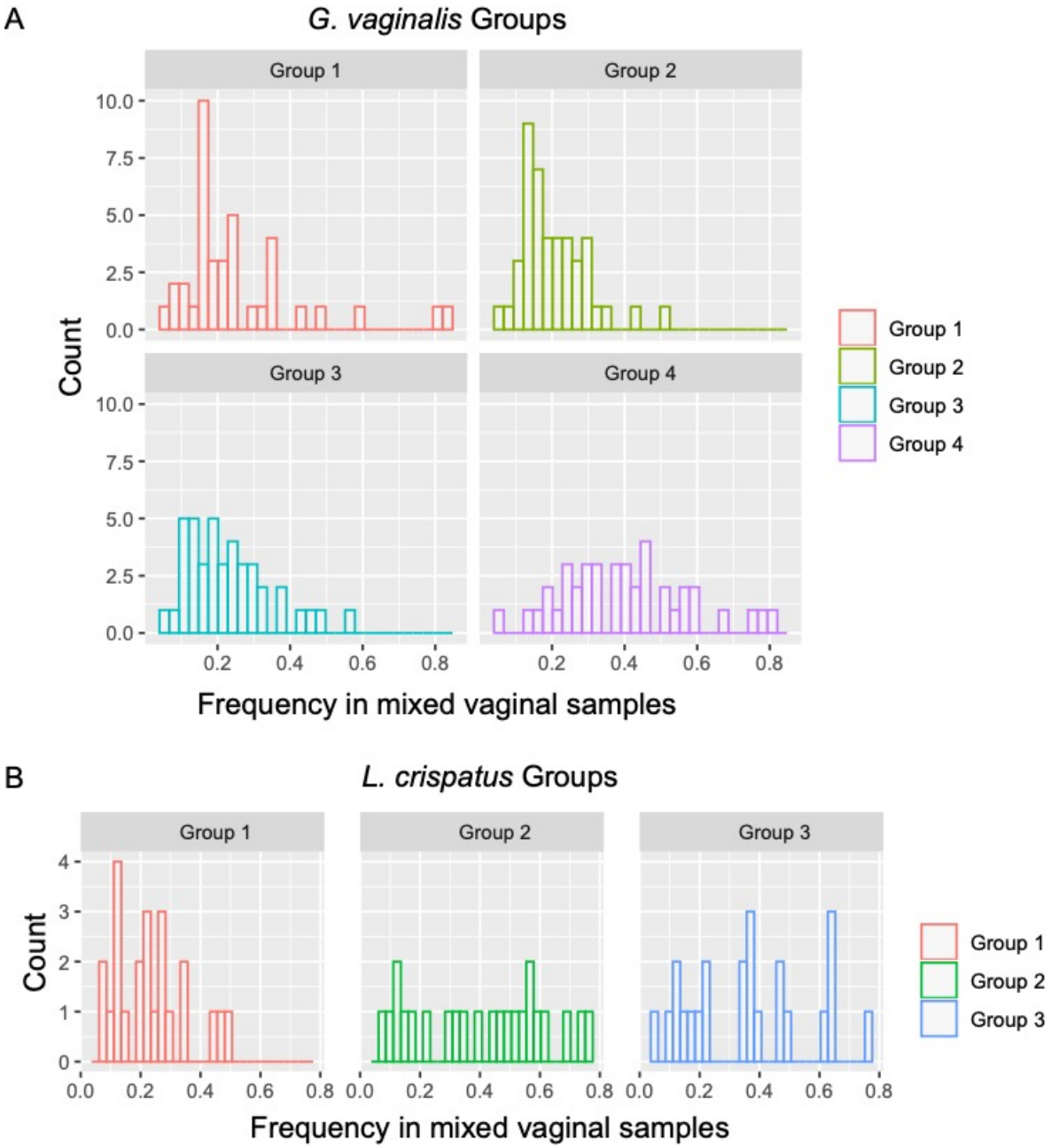

Figure 3.S1: Group relative abundance in mixed vaginal samples
The relative abundance of A) *G. vaginalis* and B) *L. crispatus* groups in mixed samples based on group-specific SNPs. Each panel shows a histogram of the inferred frequency of the four *G. vaginalis* groups and three *L. crispatus* groups in vaginal samples designated as mixed populations by ADMIXTURE.

3.5 References

1. Human Microbiome Project C. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012;486(7402):207-214.

2. Lloyd-Price J, Mahurkar A, Rahnavard G, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature.* 2017;550(7674):61-66.

3. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. *Nat Med.* 2018;24(4):392-400.

4. Goltsman DSA, Sun CL, Proctor DM, et al. Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. *Genome Res.* 2018;28(10):1467-1480.

5. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* 2017;27(4):626-638.

6. Greenblum S, Carr R, Borenstein E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell.* 2015;160(4):583-594.

7. Cohan FM. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos Trans R Soc Lond B Biol Sci.* 2006;361(1475):1985-1996.

8. Sheppard SK, Guttman DS, Fitzgerald JR. Population genomics of bacterial host adaptation. *Nat Rev Genet.* 2018;19(9):549-565.

9. Schloissnig S, Arumugam M, Sunagawa S, et al. Genomic variation landscape of the human gut microbiome. *Nature.* 2013;493(7430):45-50.

10. Ravel J, Gajer P, Abdo Z, et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A.* 2011;108 Suppl 1:4680-4687.

11. Fettweis JM, Brooks JP, Serrano MG, et al. Differences in vaginal microbiome in African American women versus women of European ancestry. *Microbiology.* 2014;160(Pt 10):2272-2282.

12. Ma B, Forney LJ, Ravel J. Vaginal microbiome: rethinking health and disease. *Annu Rev Microbiol.* 2012;66:371-389.

13. Ahmed A, Earl J, Retchless A, et al. Comparative genomic analyses of 17 clinical isolates of Gardnerella vaginalis provide evidence of multiple genetically isolated clades consistent with subspeciation into genovars. *J Bacteriol.* 2012;194(15):3922-3937.

14. Vaneechoutte M, Guschin A, Van Simaey L, Gansemans Y, Van Nieuwerburgh F, Cools P. Emended description of Gardnerella vaginalis and description of Gardnerella leopoldii sp. nov., Gardnerella piotii sp. nov. and Gardnerella swidsinskii sp. nov., with delineation of 13 genomic species within the genus Gardnerella. *Int J Syst Evol Microbiol.* 2019;69(3):679-687.

15. Potter RF, Burnham CD, Dantas G. In Silico Analysis of Gardnerella Genomospecies Detected in the Setting of Bacterial Vaginosis. *Clin Chem.* 2019;65(11):1375-1387.

16. Li W, Raoult D, Fournier PE. Bacterial strain typing in the genomic era. *FEMS Microbiol Rev.* 2009;33(5):892-916.

17. Schellenberg JJ, Paramel Jayaprakash T, Withana Gamage N, Patterson MH, Vaneechoutte M, Hill JE. Gardnerella vaginalis Subgroups Defined by cpn60 Sequencing and Sialidase Activity in Isolates from Canada, Belgium and Kenya. *PLoS One.* 2016;11(1):e0146510.

18. Cornejo OE, Hickey RJ, Suzuki H, Forney LJ. Focusing the diversity of Gardnerella vaginalis through the lens of ecotypes. *Evol Appl.* 2018;11(3):312-324.

19. Janulaitiene M, Gegzna V, Baranauskiene L, Bulavaite A, Simanavicius M, Pleckaityte M. Phenotypic characterization of Gardnerella vaginalis subgroups suggests differences in their virulence potential. *PLoS One.* 2018;13(7):e0200625.

20. Janulaitiene M, Paliulyte V, Grinceviciene S, et al. Prevalence and distribution of Gardnerella vaginalis subgroups in women with and without bacterial vaginosis. *BMC Infect Dis.* 2017;17(1):394.

21. Hill JE, Albert AYK, Group VR. Resolution and Cooccurrence Patterns of Gardnerella leopoldii, G. swidsinskii, G. piotii, and G. vaginalis within the Vaginal Microbiome. *Infect Immun.* 2019;87(12).

22. Balashov SV, Mordechai E, Adelson ME, Gygax SE. Identification, quantification and subtyping of Gardnerella vaginalis in noncultured clinical vaginal samples by quantitative PCR. *J Med Microbiol.* 2014;63(Pt 2):162-175.

23. Ojala T, Kankainen M, Castro J, et al. Comparative genomics of Lactobacillus crispatus suggests novel mechanisms for the competitive exclusion of Gardnerella vaginalis. *BMC Genomics.* 2014;15:1070.

24. France MT, Mendes-Soares H, Forney LJ. Genomic Comparisons of Lactobacillus crispatus and Lactobacillus iners Reveal Potential Ecological Drivers of Community Composition in the Vagina. *Appl Environ Microbiol.* 2016;82(24):7063-7073.

25. Callahan BJ, DiGiulio DB, Goltsman DSA, et al. Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women. *Proc Natl Acad Sci U S A.* 2017;114(37):9966-9971.

26. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods.* 2016;13(7):581-583.

27. Maiden MC. Multilocus sequence typing of bacteria. *Annu Rev Microbiol.* 2006;60:561-588.

28. Lind PA, Tobin C, Berg OG, Kurland CG, Andersson DI. Compensatory gene amplification restores fitness after inter-species gene replacements. *Mol Microbiol.* 2010;75(5):1078-1089.

29. Kraal L, Abubucker S, Kota K, Fischbach MA, Mitreva M. The prevalence of species and strains in the human microbiome: a resource for experimental efforts. *PLoS One.* 2014;9(5):e97279.

30. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* 2011;21(6):936-939.

31. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One.* 2015;10(5):e0128036.

32. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357-359.

33. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods.* 2015;12(10):902-903.

34. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics.* 2012;28(4):593-594.

35. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078-2079.

36. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297-1303.

37. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43(5):491-498.

38. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27(15):2156-2158.

39. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 2009;19(9):1655-1664.

40. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559-575.

41. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* 2012;6(2):80-92.

42. Cingolani P, Patel VM, Coon M, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front Genet.* 2012;3:35.

43. Cook DE, Andersen EC. VCF-kit: assorted utilities for the variant call format. *Bioinformatics.* 2017;33(10):1581-1582.

44. Watterson GA. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 1975;7(2):256-276.

45. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* 1989;123(3):585-595.

46. Sung W, Ackerman MS, Dillon MM, et al. Evolution of the Insertion-Deletion Mutation Rate Across the Tree of Life. *G3 (Bethesda).* 2016;6(8):2583-2591.

47. Khan S, Voordouw MJ, Hill JE. Competition Among Gardnerella Subgroups From the Human Vaginal Microbiome. *Front Cell Infect Microbiol.* 2019;9:374.

48. Didelot X, Maiden MC. Impact of recombination on bacterial evolution. *Trends Microbiol.* 2010;18(7):315-322.

49. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 2017;27(5):824-834.

50. Quince C, Delmont TO, Raguideau S, et al. DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* 2017;18(1):181.

51. Tada I, Tanizawa Y, Endo A, Tohno M, Arita M. Revealing the genomic differences between two subgroups in Lactobacillus gasseri. *Biosci Microbiota Food Health.* 2017;36(4):155-159.

52. Putonti C, Akhnoukh V, Anagnostopoulos Z, et al. Draft Genome Sequences of Six Lactobacillus gasseri and Three Lactobacillus paragasseri Strains Isolated from the Female Bladder. *Microbiol Resour Announc.* 2019;8(37).

53. Tanizawa Y, Tada I, Kobayashi H, et al. Lactobacillus paragasseri sp. nov., a sister taxon of Lactobacillus gasseri, based on whole-genome sequence analyses. *Int J Syst Evol Microbiol.* 2018;68(11):3512-3517.

54. Paramel Jayaprakash T, Schellenberg JJ, Hill JE. Resolution and characterization of distinct cpn60-based subgroups of Gardnerella vaginalis in the vaginal microbiota. *PLoS One.* 2012;7(8):e43009.

55. Yeoman CJ, Yildirim S, Thomas SM, et al. Comparative genomics of Gardnerella vaginalis strains reveals substantial differences in metabolic and virulence potential. *PLoS One.* 2010;5(8):e12411.

56. Malki K, Shapiro JW, Price TK, et al. Genomes of Gardnerella Strains Reveal an Abundance of Prophages within the Bladder Microbiome. *PLoS One.* 2016;11(11):e0166757.

57. Bohr LL, Mortimer TD, Pepperell CS. Lateral Gene Transfer Shapes Diversity of Gardnerella spp. *Front Cell Infect Microbiol.* 2020;10:293.

58. Tachedjian G, Aldunate M, Bradshaw CS, Cone RA. The role of lactic acid production by probiotic Lactobacillus species in vaginal health. *Res Microbiol.* 2017;168(9-10):782-792.

59. Gause GF. Experimental Analysis of Vito Volterra's Mathematical Theory of the Struggle for Existence. *Science.* 1934;79(2036):16-17.

60. Schoener TW. Resource partitioning in ecological communities. *Science.* 1974;185(4145):27-39.

61. Kim TK, Thomas SM, Ho M, et al. Heterogeneity of vaginal microbial communities within individuals. *J Clin Microbiol.* 2009;47(4):1181-1189.

62. Mendes-Soares H, Krishnan V, Settles ML, Ravel J, Brown CJ, Forney LJ. Fine-scale analysis of 16S rRNA sequences reveals a high level of taxonomic diversity among vaginal Atopobium spp. *Pathog Dis.* 2015;73(4).

63. Yildirim S, Yeoman CJ, Janga SC, et al. Primate vaginal microbiomes exhibit species specificity without universal Lactobacillus dominance. *ISME J.* 2014;8(12):2431-2444.

64. Stumpf RM, Wilson BA, Rivera A, et al. The primate vaginal microbiome: comparative context and implications for human health and disease. *Am J Phys Anthropol.* 2013;152 Suppl 57:119-134.

65. Rivera AJ, Stumpf RM, Wilson B, Leigh S, Salyers AA. Baboon vaginal microbiota: an overlooked aspect of primate physiology. *Am J Primatol.* 2010;72(6):467-474.

66. Duar RM, Lin XB, Zheng J, et al. Lifestyles in transition: evolution and natural history of the genus Lactobacillus. *FEMS Microbiol Rev.* 2017;41(Supp_1):S27-S48.

67. Mendes-Soares H, Suzuki H, Hickey RJ, Forney LJ. Comparative functional genomics of Lactobacillus spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. *J Bacteriol.* 2014;196(7):1458-1470.

68. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. *Nature.* 2017;541(7637):302-310.

69. Greenbaum S, Greenbaum G, Moran-Gilad J, Weintraub AY. Ecological dynamics of the vaginal microbiome in relation to health and disease. *Am J Obstet Gynecol.* 2019;220(4):324-335.

Chapter 4: Conclusions and Future Directions

4. 1 Conclusions

In this dissertation I sought to address gaps in our current understanding of the ecology and structure of the vaginal microbiome. In Chapter 2, I evaluated the relationship between vaginal *Candida* colonization and the bacterial community. Subsequently, I explored vaginal bacterial strain variation, diversity and population structure in Chapter 3. Through this work, I made several key observations:

1) Race is associated with *Candida* colonization. I found that *Candida* colonization was more common among Black women when compared to White women. Among women who were *Candida*-positive, I observed higher levels of *Candida* among Black women than White women. Importantly, race also correlated with the composition of the microbiome and may in part explain the association between race and *Candida* colonization. In my analysis, Black women were more likely to have *L. iners*-dominant communities and less likely to have *L. crispatus*-dominant communities when compared to White women. The association between *Candida* colonization and race may also reflect differences in sociodemographic characteristics or behaviors that were not accounted for in my models.

2) Vaginal *Candida* colonization is associated with bacterial community type. I showed that women with *L. iners*-dominant microbiomes were most likely to be colonized with *Candida* when compared to women with *L. crispatus*-dominant communities. I then proposed that a greater production of lactic acid and inhibition of *Candida* growth by *L. crispatus* may explain this association. I also observed a similarly low rate of *Candida* colonization in women with Diverse communities as *L. crispatus*-dominant communities. As vaginal pH in women with

Diverse communities is often high (>4.5) reflecting low levels of lactic acid,[1] this finding suggests that Diverse communities have mechanisms of inhibiting *Candida* growth that are independent of lactic acid.

3) Most commonly abundant species of vaginal bacteria have population structure. I showed that *G. vaginalis*, *L. crispatus*, *L. jensenii*, *L. gasseri* and *A. vaginae* have population structure and strains of each species can be grouped into subpopulations. The sources of population structure in vaginal bacteria are unknown but may include host-associated population structure.[2] Subpopulations might diverge neutrally through genetic drift and population bottlenecks or by adaptive diversification as a result of selective pressures.[2] In contrast, no subpopulation groups were observed among *L. iners* strains.

4) Strain diversity is related to species diversity. I showed that strain diversity was greatest in *G. vaginalis* and *A. vaginae*, which are commonly found in species diverse communities. Strain diversity in the *Lactobacillus* species was low in comparison, corresponding with lower species diversity commonly observed in *Lactobacillus* dominated communities. Additionally, most samples of *G. vaginalis* and *A. vaginae* harbor multiple strains, but the presence of multiple strains is less common for the *Lactobacillus* species. This relates to strain diversity and may reflect strains that function as different ecological species or ecotypes.

Together this work provides additional insight into the ecology of the vaginal microbiome and provides practical methods by which to incorporate *Candida* and bacterial strain into microbiome analysis. Next I will discuss ways in which *Candida* and bacterial strain analysis might be

incorporated into future studies of the vaginal microbiome to address existing outstanding questions in the field.

4.2 Future Directions

1) Temporal dynamics of *Candida* and the microbiome – The composition of the vaginal microbiome is associated with *Candida* colonization, but the temporal dynamics between the microbiome and *Candida* remain unknown. Previous work has shown that women are often transiently colonized by *Candida*,[3] and vaginal bacteria may play an important role in *Candida* clearance or persistence. Through longitudinal sample collection, future studies could explore changes in bacterial community composition and how they influence *Candida* colonization over time. This might clarify whether *L. iners*-dominant communities indeed permit *Candida* colonization and growth or alternatively whether *Candida* colonization induces shifts in the bacterial community that favor *L. iners*-dominant microbiomes. Additionally, we might learn how disruptions of the bacterial community influence *Candida* colonization and/or abundance and whether this in turn affects the development of vulvovaginal candidiasis (VVC).

2) Role of the microbiome in VVC – VVC is characterized by a robust host immune response to *Candida* overgrowth.[4] Most women will experience at least one episode of VVC and for some women, recurrent episodes can be a significant source of morbidity.[5,6] Symptoms in VVC are believed to be mediated by the host immune response and may include itching, burning, redness and discharge.[7] While significant effort has been invested into characterizing the immune mediators of VVC, little is known about how the microbiome influences symptoms and immune activation in the setting of VVC. Future work might relate the composition of the microbiome to

VVC symptoms and immune mediators (i.e. chemokines and cytokines) prior to and during the development of *Candidiasis*. Such analysis might provide the basis for future manipulations of the microbiome to reduce the occurrence of VVC in susceptible women.

3) Role of *Candida* in PTB – The contribution of vaginal *Candida* to preterm birth (PTB) remains unclear. While some studies have found no association between vaginal *Candida* and PTB,[8,9]others have indicated that treatment of vaginal *Candida* with Clotrimazole reduces rates of PTB.[10-15] Notably, none of these studies incorporated the microbiome into their analysis. A possible explanation for this discrepancy could be that *Candida* contributes to PTB indirectly through interaction with the microbiome. Longitudinal study designs that incorporate analysis of *Candida* and the microbiome over the course of pregnancy may identify interactions that contribute to PTB. Additionally, the treatment of vaginal *Candida* with clotrimazole may alter the microbiome in yet unknown ways. Vaginal treatment with an aminoglycoside antibiotic was shown to alter the inflammatory milieu in a microbiome independent fashion and it is possible that clotrimazole may modify PTB risk through a similar mechanism.[16] Future work evaluating the effects of clotrimazole on *Candida,* the microbiome and immune mediators during pregnancy might provide insight into their relationships with PTB.

4) Global bacterial strain variation – My analysis of bacterial strains in the vaginal microbiome was limited to women in North America. As such, it remains unknown whether the strain groups identified in my work represent the global diversity of strains. It is possible that unique strain groups exist among under sampled populations of humans. Additionally, characterizing global strain variation would clarify whether strain prevalence is influenced by

geography or ancestry. This type of analysis may provide unique insights into the evolutionary origins of strain variation. It could also help us to better understand why microbiome community types differ among some racial and ethnic groups.

5) Bacterial strain and reproductive outcomes – The role of strain variation in reproductive health is understudied. Prior research has suggested *G. vaginalis* strain may be important in the development of BV and PTB,[17-19]but still these relationships remain unclear. Furthermore, the role of strain in poor reproductive outcomes has not been studied for other species. Future investigation should evaluate strain variation over the course of pregnancy and its relationship with outcomes including PTB. Insights gained from such work may shed light on the different associations between the microbiome and PTB reported in the literature.[20]

4.3 Concluding remarks

The vaginal microbiome has garnered a lot of interest in recent years for its relationship with reproductive health. To develop novel therapeutic approaches aimed at manipulating the vaginal microbiome, we must first understand its ecology. While advancements in sequencing technologies have led to significant progress in our understanding of the composition of the microbial community, the study of the vaginal microbiome is still in its infancy. This thesis provides insight into the ecology of the microbial communities inhabiting the human vagina and establishes a framework for the inclusion of *Candida* and bacterial strain in future study of the vaginal microbiome and its role in gynecologic and obstetric health.

4.3 References

1. Ravel J, Gajer P, Abdo Z, et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A.* 2011;108 Suppl 1:4680-4687.

2. Sheppard SK, Guttman DS, Fitzgerald JR. Population genomics of bacterial host adaptation. *Nat Rev Genet.* 2018;19(9):549-565.

3. Beigi RH, Meyn LA, Moore DM, Krohn MA, Hillier SL. Vaginal yeast colonization in nonpregnant women: a longitudinal study. *Obstet Gynecol.* 2004;104(5 Pt 1):926-930.

4. Fidel PL, Jr., Barousse M, Espinosa T, et al. An intravaginal live Candida challenge in humans leads to new hypotheses for the immunopathogenesis of vulvovaginal candidiasis. *Infect Immun.* 2004;72(5):2939-2946.

5. Bradford LL, Ravel J. The vaginal mycobiome: A contemporary perspective on fungi in women's health and diseases. *Virulence.* 2017;8(3):342-351.

6. Donders GG, Sobel JD. Candida vulvovaginitis: A store with a buttery and a show window. *Mycoses.* 2017;60(2):70-72.

7. Fidel PL, Jr. Immunity in vaginal candidiasis. *Curr Opin Infect Dis.* 2005;18(2):107-111.

8. Cotch MF, Hillier SL, Gibbs RS, Eschenbach DA. Epidemiology and outcomes associated with moderate to heavy Candida colonization during pregnancy. Vaginal Infections and Prematurity Study Group. *Am J Obstet Gynecol.* 1998;178(2):374-380.

9. McGregor JA, French JI, Parker R, et al. Prevention of premature birth by screening and treatment for common genital tract infections: results of a prospective controlled evaluation. *Am J Obstet Gynecol.* 1995;173(1):157-167.

10. Roberts CL, Algert CS, Rickard KL, Morris JM. Treatment of vaginal candidiasis for the prevention of preterm birth: a systematic review and meta-analysis. *Syst Rev.* 2015;4:31.

11. Kiss H, Petricevic L, Husslein P. Prospective randomised controlled trial of an infection screening programme to reduce the rate of preterm delivery. *BMJ.* 2004;329(7462):371.

12. Roberts CL, Rickard K, Kotsiou G, Morris JM. Treatment of asymptomatic vaginal candidiasis in pregnancy to prevent preterm birth: an open-label pilot randomized controlled trial. *BMC Pregnancy Childbirth.* 2011;11:18.

13. Banhidy F, Acs N, Puho EH, Czeizel AE. Rate of preterm births in pregnant women with common lower genital tract infection: a population-based study based on the clinical practice. *J Matern Fetal Neonatal Med.* 2009;22(5):410-418.

14. Czeizel AE, Fladung B, Vargha P. Preterm birth reduction after clotrimazole treatment during pregnancy. *Eur J Obstet Gynecol Reprod Biol.* 2004;116(2):157-163.

15. Czeizel AE, Rockenbauer M. A lower rate of preterm birth after clotrimazole therapy during pregnancy. *Paediatr Perinat Epidemiol.* 1999;13(1):58-64.

16. Gopinath S, Kim MV, Rakib T, et al. Topical application of aminoglycoside antibiotics enhances host resistance to viral infections in a microbiota-independent manner. *Nat Microbiol.* 2018;3(5):611-621.

17. Balashov SV, Mordechai E, Adelson ME, Gygax SE. Identification, quantification and subtyping of Gardnerella vaginalis in noncultured clinical vaginal samples by quantitative PCR. *J Med Microbiol.* 2014;63(Pt 2):162-175.

18. Janulaitiene M, Paliulyte V, Grinceviciene S, et al. Prevalence and distribution of Gardnerella vaginalis subgroups in women with and without bacterial vaginosis. *BMC Infect Dis.* 2017;17(1):394.

19. Callahan BJ, DiGiulio DB, Goltsman DSA, et al. Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women. *Proc Natl Acad Sci U S A.* 2017;114(37):9966-9971.

20. Greenbaum S, Greenbaum G, Moran-Gilad J, Weintraub AY. Ecological dynamics of the vaginal microbiome in relation to health and disease. *Am J Obstet Gynecol.* 2019;220(4):324-335.