Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women

Benjamin J. Callahan,1,2 Daniel B. DiGiulio,3,4 Daniela S. Aliaga Goltzman,5 Christine L. Sun,6 Elizabeth K. Costello,7 Pratheepa Jeganathan,7 Joseph R. Biggio,8 Ronald J. Wong9, Maurice L. Druzin10,9 Gary M. Shaw9, David K. Stevenson9, Susan P. Holmes11, and David A. Relman9,12

Department of Population Health and Pathobiology, College of Veterinary Medicine, North Carolina State University, Raleigh, NC 27607; 2Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305; 3Infectious Diseases Section, Veterans Affairs Palo Alto Health Care System, Palo Alto, CA 94304; 4Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305; 5Department of Obstetrics and Gynecology, Division of Maternal Fetal Medicine, University of Alabama at Birmingham, Birmingham, Birmingham, AL 35249; and 6Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94305

Edited by Lora V. Hooper, The University of Texas Southwestern, Dallas, TX, and approved August 1, 2017 (received for review April 11, 2017)

Preterm birth (PTB) is the leading cause of neonatal morbidity and mortality. Previous studies have suggested that the maternal vaginal microbiota contributes to the pathophysiology of PTB, but conflicting results in recent years have raised doubts. We conducted a study of PTB compared with term birth in two cohorts of pregnant women: one predominantly Caucasian (n = 39) at low risk for PTB, the second predominantly African American and at high-risk (n = 96). We profiled the taxonomic composition of 2,179 vaginal swabs collected prospectively and weekly during gestation using 16S rRNA gene sequencing. Previously proposed associations between PTB and low Lactobacillus and higher Gardnerella abundances replicated in the low-risk cohort, but not in the high-risk cohort. High-resolution bioinformatics enabled taxonomic assignment to the species and subspecies levels, revealing that Lactobacillus crispatus was associated with low risk of PTB in both cohorts, while Lactobacillus iners was not, and that a subspecies clade of Gardnerella vaginalis explained the genus association with PTB. Patterns of cooccurrence between L. crispatus and Gardnerella were highly exclusive, while Gardnerella and L. iners often coexisted at high frequencies. We argue that the vaginal microbiota is better represented by the quantitative frequencies of these key taxa than by classifying communities into five community state types. Our findings extend and corroborate the association between the vaginal microbiota and PTB, demonstrate the benefits of high-resolution statistical bioinformatics in clinical microbiome studies, and suggest that previous conflicting results may reflect the different risk profile of women of black race.

Prenatal care (PTB; delivery at <37 gestational wk) affects ≈12% of US births and is the leading cause of neonatal death and morbidity worldwide. Multiple lines of evidence support a role for the indigenous microbial communities of the mother (the maternal microbiota) in the pathophysiology of PTB. Microbial invasion of the amniotic cavity is one of the most frequent causes of spontaneous PTB (1), and the most common invading taxa are consistent with maternal origin (2–4). Bacterial vaginosis (BV), a condition involving an altered vaginal microbiota, has been consistently identified as a risk factor for PTB (5, 6). Multiple studies have also found chronic periodontitis, another condition associated with an altered microbiota, to be a risk factor for PTB (7, 8).

High-throughput sequencing methods have facilitated new lines of investigation into the microbial etiology of PTB (9, 10). Amplification and high-throughput sequencing of the 16S rRNA gene (metabarcoding) simultaneously measure the presence and relative abundance of thousands of bacterial taxa (composition), and resolves differences to the level of genus and sometimes species or subspecies. To date, metabarcoding studies of the

Significance

Preterm birth (PTB) is a major global public health burden. Previous studies have suggested an association between altered vaginal microbiota composition and PTB, although findings across studies have been inconsistent. To address these inconsistencies, improve upon our previous signature, and better understand the vaginal microbiota’s role in PTB, we conducted a case-control study in two cohorts of pregnant women: one predominantly Caucasian at low risk of PTB, the second predominantly African American at high risk. With the results, we were able to replicate our signature in the first cohort and refine our signature of PTB for both cohorts. Our findings elucidate the ecology of the vaginal microbiota and advance our ability to predict and understand the causes of PTB.
PTB–microbiota association. For example, there is evidence that women of black race, a proxy for host genetics and a correlate to social and demographic factors, have a different range of normal vaginal community compositions (15, 16).

Additional difficulties arise from differences in metabarcoding methodologies. Sample collection, DNA extraction, and PCR primers impact sensitivity to different taxa (17, 18), sometimes so strongly that particular taxa can go undetected (e.g., Gardnerella; ref. 19). Variation in the rate of evolutionary diversification within the 16S rRNA gene makes results from different 16S rRNA gene regions difficult to compare, especially at higher taxonomic resolutions (20). The confounding of methodological differences with population differences across studies is particularly challenging.

Here, we present a study designed with these challenges in mind. Using a uniform methodology, we attempt to replicate and refine a small number of previously generated hypotheses in two different populations differing dramatically in racial composition.

In our 2015 study (11), we measured the composition of the vaginal microbiota longitudinally over the course of pregnancy in 49 women and tested for associations with PTB. At the taxonomic level of genus, we reported that PTB was associated with (i) lower abundance of Lactobacillus, (ii) higher abundance of Gardnerella, and (iii) higher abundance of Ureaplasma. Here, we test those associations in two new study cohorts: a “replication” (Stanford) cohort of 39 low-risk women (30 term, 9 preterm deliveries) enrolled from the predominantly African-American population that provided the hypothesis-generating cohort in our 2015 study, and a “high-risk” (UAB, University of Alabama at Birmingham) cohort of 96 women (55 term, 41 preterm) enrolled from the predominantly African-American population of pregnant women with a prior history of PTB who received prenatal care and progesterone treatment at the UAB. Our deep sequencing and weekly sampling approach provides the most accurate quantification yet of the vaginal community compositions (15, 16).

Our bioinformatics approach grants subgenus taxonomic resolution for many taxa. We reconsidered the Lactobacillus association for each of the four principal Lactobacillus species of the vaginal microbiome and the Gardnerella association for each of the nine unique G. vaginalis 16S rRNA sequence variants detected in our study.

We found striking differences among the associations of different Lactobacillus species with PTB, especially between the highly abundant L. crispatus and L. iners species. A lower abundance of L. crispatus was significantly associated with PTB in both cohorts, while no significant association was detected for L. iners (Fig. 2A). In the UAB cohort, decreased abundance of the less common species L. jensenii and L. gasseri was also associated with PTB.

The association between the Gardnerella genus and PTB was driven by a single variant, Gardnerella sequence variant 2 (G2) had a strong significant association in the Stanford cohort (P = 0.0033, Wilcoxon rank-sum test), while no other variants, nor the Gardnerella genus excluding G2, significantly associated with PTB in either cohort (Fig. 2A). The most abundant Gardnerella sequence variants (G1, G2, and G3) differ at just one or two nucleotides. However, when mapped onto whole-genome phylogenies (Fig. 2B and SI Materials and Methods), these variants differentiate between significantly diverged “genovars” within the G. vaginalis species (21). The functional classifications of clade-specific genes are largely redundant but transporters, membrane proteins, and toxin-antitoxin systems appear overrepresented in G2 variant genomes (Figs. S1 and S2).

We also considered the refined Lactobacillus and Gardnerella associations in the Stanford cohort for samples restricted to fall within certain time windows of pregnancy (Fig. S3). These results suggest that the replicated associations hold even early in the second trimester (13–18 gestational wk).

**Ecology.** Four “landmark” samples with the highest observed proportion of G1, G2, L. crispatus, and L. iners, are highlighted in a multidimensional scaling (MDS) ordination of all samples (Fig. 2C). The landmarks sit at the four locations with the highest concentration of samples, indicating that these taxa drive or track community-wide variation in the vaginal microbiota.

**Fig. 1.** Average gestational frequencies of Lactobacillus, Gardnerella, and Ureaplasma for women who delivered preterm and at term. Each point shows the average frequency of the genus-of-interest across gestational samples from one woman. Upper shows the Stanford cohort (n = 39), and Lower shows the UAB cohort (n = 96). The significance of the differences between average gestational frequencies in term and preterm births, in the previously reported directions, were assessed by one-sided Wilcoxon rank-sum test; **P < 0.01.

---

**Results**

**Replication.** The average gestational frequencies of the Lactobacillus, Gardnerella, and Ureaplasma genera, stratified by birth outcome, are shown in Fig. 1. Our previously reported associations between less Lactobacillus and PTB, and between more Gardnerella and PTB, replicated in the Stanford cohort (Wilcoxon rank-sum test; Lactobacillus, P = 0.0093; Gardnerella, P = 0.0070). However, neither association was significant in the UAB cohort, and our previously hypothesized association between more Ureaplasma and PTB was not supported in either cohort. For characteristics of these cohorts, see Tables S1 and S2.

We calculated odds ratios (ORs) for the Lactobacillus and Gardnerella associations in the Stanford cohort. Using a somewhat arbitrary frequency threshold of 70% to delineate low Lactobacillus (high risk) and high Lactobacillus (low risk), we obtain an OR of 5.81 (95% CI range: 1.12–33.7, P = 0.0093; Gardnerella, P = 0.0070). However, neither association was significant in the UAB cohort, and our previously hypothesized association between more Ureaplasma and PTB was not supported in either cohort. For characteristics of these cohorts, see Tables S1 and S2.

We calculated odds ratios (ORs) for the Lactobacillus and Gardnerella associations in the Stanford cohort. Using a somewhat arbitrary frequency threshold of 70% to delineate low Lactobacillus (high risk) and high Lactobacillus (low risk), we obtain an OR of 5.81 (95% CI range: 1.12–33.7) for the low Lactobacillus category. Using a frequency threshold of 0.1% for Gardnerella produced an OR of 5.12 (95% CI range: 1.05–31.1) for the high Gardnerella (high risk) category. These ORs are specific to the Stanford population.

**Refinement.** Our bioinformatics approach grants subgenus taxonomic resolution for many taxa. We reconsidered the Lactobacillus association for each of the four principal Lactobacillus species of the vaginal microbiome and the Gardnerella association for each of the nine unique G. vaginalis 16S rRNA sequence variants detected in our study.

We found striking differences among the associations of different Lactobacillus species with PTB, especially between the highly abundant L. crispatus and L. iners species. A lower abundance of L. crispatus was significantly associated with PTB in both cohorts, while no significant association was detected for L. iners (Fig. 2A). In the UAB cohort, decreased abundance of the less common species L. jensenii and L. gasseri was also associated with PTB.

The association between the Gardnerella genus and PTB was driven by a single variant, Gardnerella sequence variant 2 (G2) had a strong significant association in the Stanford cohort (P = 0.0033, Wilcoxon rank-sum test), while no other variants, nor the Gardnerella genus excluding G2, significantly associated with PTB in either cohort (Fig. 2A). The most abundant Gardnerella sequence variants (G1, G2, and G3) differ at just one or two nucleotides. However, when mapped onto whole-genome phylogenies (Fig. 2B and SI Materials and Methods), these variants differentiate between significantly diverged “genovars” within the G. vaginalis species (21). The functional classifications of clade-specific genes are largely redundant but transporters, membrane proteins, and toxin-antitoxin systems appear overrepresented in G2 variant genomes (Figs. S1 and S2).

We also considered the refined Lactobacillus and Gardnerella associations in the Stanford cohort for samples restricted to fall within certain time windows of pregnancy (Fig. S3). These results suggest that the replicated associations hold even early in the second trimester (13–18 gestational wk).
both the Stanford and UAB cohorts (Fig. 3A). Notable differences between the cohorts are the higher frequency of *Gardnerella* in the UAB cohort (8.3%) relative to the Stanford cohort (4.7%) and the lower frequency of *L. crispatus* in UAB (15%) relative to Stanford (45%).

The pattern of cooccurrence between *L. crispatus* and *Gardnerella* is different from between *L. iners* and *Gardnerella* (Fig. 3B). The ratio of *L. crispatus* to *Gardnerella* is extremely skewed toward one predominating over the other when the two make up a large fraction of the community, a pattern consistent with an exclusionary interaction. In marked contrast, *L. iners* and *Gardnerella* often coexist at comparable frequencies, at least in the UAB cohort, even at high combined frequencies.

We quantified community instability as the Bray–Curtis dissimilarity between communities sampled in consecutive weeks. The gestational average of community instability was inversely correlated with *Lactobacillus* dominance as expected (Spearman $\rho = -0.76$), but intriguingly community instability was associated with PTB in the Stanford cohort (Wilcoxon rank-sum test; $P = 0.0018$) even more strongly than was *Lactobacillus* frequency (22).

**Exploration and Contamination.** We tested for associations between PTB and increased gestational frequencies of several genera frequently linked with BV (*Gardnerella*; *Prevotella*; *Atopobium*; *Mobiluncus*; *Megasphaera*; *Diaister*; *Peptoniphilus*; and *Mycoplasma*) and three sequence variants that exactly matched 16S rRNA genes from the recently identified bacterial vaginosis-associated bacteria BVAB1, BVAB2, and BVAB3 (23, 24). *Prevotella* was associated with PTB in both cohorts ($P < 0.05$), while most of the remaining genera were significantly associated only in the Stanford cohort (Table 1). BVAB sequence variants were not significantly associated with PTB in either cohort.

We detected sequence variants corresponding to the causative organisms for three sexually transmitted infections (STIs) in the UAB cohort: *Chlamydia trachomatis* (5 gestations), *Neisseria gonorrhoeae* (4), and *Trichomonas vaginalis* (28). We also detected sequence variants corresponding to *Trichomonas*-associated Candidatus Mycoplasma gierdhii (25–27), and *Trichomonas* was indeed present in all samples with frequency $>10^{-3.5}$. No STI organism was significantly associated with PTB (Fig. S4), but this may reflect insufficient power due to the small numbers of subjects in which they were observed.

We also performed an exploratory analysis in which the gestational frequency of each detected genus was tested for an association with PTB (Fig. 4 and Table S3). Many genera met a common standard for statistical significance in exploratory analyses (FDR $< 0.1$) (Table S3). However, closer inspection of the four genera which stand out the most from the bulk of the $P$ value distribution (*Yersinia*, *Abiotrophia*, *Stomatobaculum*, *Tumebacillus*) suggests a potentially artifactual signal. These taxa were present in a small fraction of samples (0.5–7.0%) and at low overall frequency ($<2 \times 10^{-5}$), and were detected in at least one negative control (attempted PCR of extraction blank or DNA-free water). Furthermore, the prevalence of *Yersinia* and *Tumebacillus* in vaginal and control samples was strongly correlated across runs (Fig. 4B), suggesting these taxa were contaminants in the vaginal samples (SI Discussion).

**Fig. 2.** Associations between *Gardnerella*/*Lactobacillus* variants and PTB in two cohorts of women. (A) Associations between average gestational frequency and PTB were tested for the nine detected *Gardnerella* 16S rRNA sequence variants, and for the four major *Lactobacillus* species in the vaginal microbiota. Testing was performed separately for the Stanford and UAB cohorts. Dashed lines indicate $P = 0.05$ (one-sided Wilcoxon rank-sum test). (B) The three most abundant *Gardnerella* sequence variants were mapped onto a phylogenetic tree constructed from sequenced *G. vaginalis* genomes (SI Materials and Methods). (C) MDS was performed on the Bray–Curtis distances between all samples. The two MDS dimensions that explain the greatest amounts of variation in the data are displayed. Stanford ($n = 897$) and UAB ($n = 1,282$) samples are plotted separately. Landmark samples (magenta) contained the highest proportion of the labeled taxa: G1 (66% frequency in landmark sample); G2 (96%); *L. crispatus* (>99%); and *L. iners* (>99%).
Alternative exploratory analyses restricted to sequence variants and genera with frequency greater than 10^{-3} (to avoid contaminants) yielded results largely consistent with BV-associated PTB risk (Figs. S5 and S6).

**Discussion**

**Replication and Refinement.** The profusion of exploratory microbiome studies in recent years has yielded a surfeit of reported associations between health outcomes and microbial taxa, and a lack of confidence in each individually. Here, we focused on retesting the associations we discovered and published in our previous exploratory study (11). The replication of the associations between PTB and the *Lactobacillus* and *Gardnerella* genera substantially increases our confidence that these are true associations within the Stanford population that could serve as biomarkers of PTB risk.

More precise characterization of PTB-associated alterations in the vaginal microbiota is important given the nonspecific diagnostic criteria of bacterial vaginosis (BV), and the previous failures of antibiotic treatment of BV to reduce PTB (28, 29).

Our second focus here was on refining the replicated associations with new methods that resolve exact sequence variants from amplicon data, rather than lumping sequences within an arbitrary dissimilarity threshold into operational taxonomic units (OTUs) (30). This high-resolution approach localized all of the risk associated with the *Gardnerella* genus to a subspecies clade, and differentiated *N. gonorrhoeae* from other *Neisseria* species within 3% Hamming distance over the sequenced gene region. In clinical applications, we see little reason to choose OTUs over exact sequence variants, as health impacts can vary significantly between closely related taxa.

**An Ecological Model of the Vaginal Microbiota in Health.** The vaginal microbiota is commonly classified into five discrete CSTs (community state types): CST1, communities dominated by *L. crispatus*; CST2, by *L. gasseri*; CST3, by *L. iners*; CST5, by *L. jensenii*; and CST4, diverse communities not dominated by *Lactobacillus* (15).

We argue that the lack of a clear distinction between CST3 and CST4 is a critical shortcoming of the CST approach. In this study, *L. iners* and *Gardnerella* often coexisted at near equal frequencies. Discrete CSTs do not well-represent that variation, and will obscure the relationship between the vaginal microbiota and health.

We propose instead a simplified ecological model of the vaginal microbiota based on the quantitative frequencies of three key taxa: *G. vaginalis*, *L. crispatus*, and *L. iners*. The frequency of *G. vaginalis* correlates with adverse health outcomes, i.e., PTB and symptomatic BV. *Gardnerella* and *L. crispatus* strongly exclude each other; exclusion between *L. iners* and *Gardnerella* is weak or absent. Additional taxa can be added as their ecology and health importance are established (e.g., *Prevotella*).

The association between *Gardnerella* and PTB in the Stanford cohort could be explained by a gain of pathologic function, such as the stimulation of a detrimental host immune response, a loss of protective community function such as colonization resistance against pathogens by *L. crispatus*, or a cooccurrence between *Gardnerella* and other risk factors, or combinations thereof. That one *Gardnerella* variant (G2) drives the association for the entire genus is suggestive of a causative role, but is still consistent with indirect association if evolutionary changes in the G2 clade modified host specificity and/or ecological behavior.

The *Gardnerella* genus deserves further investigation. Different associations between PTB and *Gardnerella* sequence variants,

![Fig. 4](image-url) **Fig. 4.** Statistical association between the average gestational frequencies of detected genera and preterm birth in two cohorts of women. (A) We tested the association between PTB and increased gestational frequency for all detected genera (for *Lactobacillus*, decreased frequency) in each cohort by the Wilcoxon rank-sum test. Genera in red have a significant composite $P$ value after controlling FDR < 0.1 (Methods). Text size scales with the square root of study-wide frequency. (B) The fraction of samples in which the four highlighted genera were present among vaginal samples and negative controls, stratified by sequencing run.

---

**Table 1. Association between PTB and BV-relevant taxa**

<table>
<thead>
<tr>
<th>Taxon</th>
<th>${P}$ value</th>
<th>$P$ value (UAB)</th>
<th>Frequency</th>
<th>Frequency (UAB)</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stanford</td>
<td>UAB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BVA1</td>
<td>0.907</td>
<td>0.339</td>
<td>0.0017</td>
<td>0.034</td>
<td>SV</td>
</tr>
<tr>
<td>BVA2</td>
<td>0.823</td>
<td>0.475</td>
<td>0.00049</td>
<td>0.00081</td>
<td>SV</td>
</tr>
<tr>
<td>BVA3</td>
<td>0.729</td>
<td>0.280</td>
<td>1.4e-06</td>
<td>3.9e-05</td>
<td>SV</td>
</tr>
<tr>
<td>Gardnerella</td>
<td>0.007</td>
<td>0.193</td>
<td>0.051</td>
<td>0.083</td>
<td>Genus</td>
</tr>
<tr>
<td>Prevotella</td>
<td>0.008</td>
<td>0.024</td>
<td>0.023</td>
<td>0.062</td>
<td>Genus</td>
</tr>
<tr>
<td>Atopobium</td>
<td>0.043</td>
<td>0.476</td>
<td>0.0072</td>
<td>0.012</td>
<td>Genus</td>
</tr>
<tr>
<td>Mobiluncus</td>
<td>0.034</td>
<td>0.619</td>
<td>0.00017</td>
<td>0.00059</td>
<td>Genus</td>
</tr>
<tr>
<td>Megabacillus</td>
<td>0.108</td>
<td>0.320</td>
<td>0.010</td>
<td>0.054</td>
<td>Genus</td>
</tr>
<tr>
<td>Dialister</td>
<td>0.012</td>
<td>0.140</td>
<td>0.0033</td>
<td>0.0069</td>
<td>Genus</td>
</tr>
<tr>
<td>Peptoniphilus</td>
<td>0.016</td>
<td>0.043</td>
<td>0.0038</td>
<td>0.0089</td>
<td>Genus</td>
</tr>
<tr>
<td>Mycoplasma</td>
<td>0.042</td>
<td>0.283</td>
<td>0.00058</td>
<td>0.0058</td>
<td>Genus</td>
</tr>
</tbody>
</table>

Associations were tested between PTB and taxa previously observed to have increased frequency during BV. BVAB sequence variants (SV) were identified by exact matching to the 16S rRNA gene sequence in a BVAB genome. $P < 0.05$ are in bold.
which tracked genovars identified from whole genome sequences, suggest health-relevant differentiation. Characterizing more Gardnerella strains will help elucidate their role in health, especially coupled with the health status of the women from whom they are recovered. Diagnostic information more specific than Nugent score is needed: The signs and symptoms of the women must also be reported. Diversity within L. iners deserves similar scrutiny (31).

The striking difference we found between the exclusion of Gardnerella by L. crispatus, and the coexistence of Gardnerella with L. iners, was also seen in previous population surveys (32, 33). Laboratory experiments support direct exclusion between Gardnerella and L. crispatus, but not L. iners, via reciprocal interference in epithelial adhesion and biofilm formation (34, 35). Some but not all Lactobacillus species can produce lacticolin, a peptide possessing potent antibiotic activity against G. vaginalis and other Gram-positive pathogens (36). Although some evidence suggests that L. iners may contribute to adverse health outcomes (37), it would be consistent with our findings, and previous observations that the L. iners-dominated CST3 is relatively unstable (11, 32), if L. iners were itself harmless but promoted or at least did not inhibit the growth of other harmful microbes.

Population Effects. The differences between the Stanford and UAB cohorts strongly suggest that PTB–microbiota associations are population-dependent. This is especially likely here because of the multiple differences between these cohorts: Women in the UAB cohort are predominantly African American, had a prior history of PTB, and were treated with intramuscular 17α-hydroxyprogesterone caproate, an antiinflammatory agent, during pregnancy.

Women of black race have a different range of normal vaginal community compositions when not pregnant (15). A previous study of a cohort of predominantly African-American women found no significant association between Gardnerella or Lactobacillus and PTB (12). These findings are consistent with a higher normal frequency of L. iners and Gardnerella, and a lower susceptibility to Gardnerella-associated adverse health outcomes, in women of black race. Interestingly, the PTB-associated G2 variant was of nearly equal frequency in the Stanford and UAB cohorts, but the unassociated G1 and G3 variants were more than twice as frequent in UAB. Additionally, the less frequent L. gasseri and L. jensenii species were significantly associated with PTB in the UAB cohort but not in the Stanford cohort.

It is also possible that PTB–microbiota associations do not reach significance in the UAB cohort because of a lower fraction of PTB that is microbiota-associated. All women in the UAB cohort had a prior history of PTB. PTB has multiple causes, including factors unrelated to the microbiota such as vascular disorders (1). If nonmicrobiota-associated causes of PTB are more likely to persist between pregnancies, then PTB in women with a prior history will be enriched for nonmicrobiota causation. The use of a progestrone with antiinflammatory properties in the UAB cohort might have reduced potential inflammatory effects from the vaginal microbiota.

Genus-level associations were inconsistent between the UAB and Stanford cohorts, but our higher resolution analysis of Lactobacillus species revealed that PTB is significantly associated with a lower frequency of L. crispatus in both cohorts. Better discrimination of critical taxa may reveal consistencies across populations that are obscured when taxa are grouped at higher levels (38).

Comparison with Kindinger et al. A recent study also examined the relationship between the vaginal microbiota and PTB in a similarly sized cohort (n = 161) of predominantly Caucasian and Asian women with a prior history of PTB (37). Using a CST-based analysis, they corroborated an association between increased gestational frequency of L. crispatus and reduced risk for PTB. They also found a weaker or absent association between the vaginal microbiota and PTB in women of black race. However, their study differed from ours in finding a significant positive association between L. iners-dominated communities (CST3) and PTB, and no significant association between diverse BV-like communities (CST4) and PTB. We believe the differences in our findings may reflect methodological differences in our 16S rRNA gene sequencing protocols and analysis methods, rather than biological differences (SI Discussion). Kindinger et al. used a V1-V3 primer set that is sensitive to Lacticoccus, but insensitive to Gardnerella. Classification of BV-like communities (CST4) as L. iners-dominated communities (CST3) based on amplicon data depleted of Gardnerella sequences could create an association between CST3 and PTB that would otherwise be absent.

We believe these two studies have strengthened the evidence for a role of the vaginal microbiota in the pathophysiology of PTB, and point toward L. iners and G. vaginalis as taxa of particular importance for further study. Other taxa that often cooccur with G. vaginalis, such as the genera Prevotella, Dialister, Mobiluncus, and Atopobium, also merit additional investigation.

Strengths and Weaknesses. The strengths of our study include accurate quantification of total gestational exposure to each taxon from our dense longitudinal sampling regimen; high-resolution bioinformatics that discriminated health-relevant differences at the species and subspecies level; a hypothesis-driven approach based on previously proposed associations; and characterization of two very different cohorts with the same methodology. The weaknesses of our study include an inability to distinguish between the effects of race, progestrone treatment, and prior history of PTB because of their confounding across the two study cohorts; heterogeneity in the type of PTB considered (e.g., spontaneous or medically indicated); and incomplete use of the time-series aspects of the data in our analysis.

Materials and Methods

Study Cohorts. Two study cohorts were enrolled prospectively from different locations within the United States. The Stanford cohort (Palo Alto, CA; 39 women, 9 of whom delivered preterm) was selected to enrich for PTB from an ongoing prospective study of women presenting to the obstetrical clinics of the Lucille Packard Children’s Hospital Stanford for prenatal clinical care; previous cohorts from this study provided samples that led to the identification of a vaginal microbiota signature of PTB (11). The underlying population is predominantly Caucasian and Asian and has low risk for PTB (<10%). The UAB cohort (Birmingham, AL; 96 women, 41 preterm) was enrolled prospectively from the pregnant women referred to UAB for prior history of PTB. The referral population is predominantly African American, and the analyzed cohort was representative of that population. The study was approved by an Administrative Panel for the Protection of Human Subjects (IRB, Institutional Review Board) of Stanford University (IRB protocol no. 21956) and by an IRB at UAB (protocol no. X121031002). All women provided written informed consent before completing an enrollment questionnaire and providing biological samples.

The UAB and Stanford cohorts were enrolled using the same study protocol, but because of the substantial differences in their demographic and clinical characteristics (Table S1), we treated them as separate case-control analyses. Study participants self-collected vaginal swabs weekly (897 swabs from 39 women in the Stanford cohort; 1,282 from 96 women in UAB) by sampling material from the midvaginal wall (Fig. S7). Controls were women who delivered at term; cases were women who delivered preterm, i.e., <37 wk of gestation (Table S2). See SI Materials and Methods for further details.

DNA Sequencing. The V4 hypervariable region of the 16S rRNA gene was PCR-amplified from genomic DNA extracted from vaginal swabs. Amplicons were sequenced on an Illumina HiSeq 2500 platform (2 × 250 bp), which yielded a median sequencing depth of ~200,000 reads per sample. See SI Materials and Methods for details.

Bioinformatics. DADA2 was used to infer the amplicon sequence variants (ASVs) present in each sample (39). Exact sequence variants provide a more accurate and reproducible description of amplicon-sequenced communities than is possible with OTUs defined at a constant level (97% or other) of sequence similarity (30). See SI Materials and Methods for details.
Statistical Analysis. Sequence variants were filtered before analysis if observed in fewer than three samples or with 100 reads study-wide ($10^{-7}$ frequency). Statistical testing was performed on untransformed frequencies and mean frequency of a taxon across all samples from the same gestation—in the R software environment (40). Based on prior hypotheses, associations between PTB and decreased frequencies of Lactobacillus, and increased frequencies of all other taxa, were tested by the one-sided Wilcoxon rank-sum method. Testing was performed separately for the Stanford and UAB cohorts, composite P values were calculated by Fisher’s method, and the FDR was controlled by the method of Benjamini and Hochberg (41).

MDS ordination plots based on the Bray–Curtis dissimilarity were generated with the phyloseq R package (42). Odds ratios and their confidence intervals were evaluated with the epitools R package (43). Figures were created with the ggplot2 R package (44).

Data Availability. Raw sequence data have been deposited at SRA under accession no. SRP115697. The processed sequence table and R scripts implementing the bioinformatics and analysis workflows are available from the Stanford Digital Repository (https://purl.stanford.edu/b681vm1809), and the R markdown file in Dataset S1.

ACKNOWLEDGMENTS. We thank study participants, as well as Cele Quaintance, Anna Robaczewska, March of Dimes Prematurity Research Center study coordinators, and nursing staff in the obstetrical clinics and labor and delivery unit of Lucille Packard Children’s Hospital. This research was supported by the March of Dimes Prematurity Research Center at Stanford University, and Stanford Child Health Research Institute, the Bill and Melinda Gates Foundation, the Gablian Fellowship (to S.P.H.), and the Thomas and C. Joan M. Merigan Endowment at Stanford University (to D.A.R.).