TITLE: Sub-communities of the vaginal ecosystem in pregnant and non-pregnant women

Authors & affiliations:

Laura Symul[1,11], Pratheepa Jeganathan[2,11], Elizabeth K. Costello[3,11], Michael France[4,5,11], Seth M. Bloom[6,7,8,11], Jacques Ravel[4,5,11], Douglas S. Kwon[6,7,8,11], David A. Relman[3,9,10,11], Susan Holmes[1,11]

1. Department of Statistics, Stanford University, 390 Jane Stanford Way, Stanford, CA 94305, USA
2. Department of Mathematics and Statistics, McMaster University, 1280 Main Street, West Hamilton, Ontario L8S 4K1, Canada
3. Department of Medicine, Stanford University School of Medicine, 291 Campus Drive, Stanford, CA 94305 USA
4. Institute for Genome Sciences, University of Maryland School of Medicine, 670 W. Baltimore Street, Baltimore, MD 21201, USA
5. Department of Microbiology and Immunology, University of Maryland School of Medicine, 685 West Baltimore Street, HSF-I Suite 380, Baltimore, MD 21201, USA
6. Division of Infectious Diseases, Massachusetts General Hospital, 55 Fruit Street, Boston MA 02114, USA
7. Harvard Medical School, 25 Shattuck St, Boston, MA 02115, USA
8. Ragon Institute of MGH, MIT and Harvard, 400 Technology Square, Cambridge MA 02139, USA
9. Department of Microbiology & Immunology, Stanford University School of Medicine, 299 Campus Drive, Stanford, CA 94305, USA
10. Infectious Diseases Section, Veterans Affairs Palo Alto Health Care System, 3801 Miranda Avenue, Palo Alto, CA 94304, Palo Alto, CA 94304, USA
11. The Vaginal Microbiome Research Consortium (VMRC)

Corresponding author(s): Susan Holmes susan@stat.stanford.edu

Author contribution:

SH, DR, JR designed the study.

SH, LS, PJ, DR conceived and designed the analyses.

DR, JR, DK, SB, EC, MF collected and annotated the data.

LS, PJ performed the analysis.

LS, SH, EC, DR wrote the manuscript draft.

All authors contributed to the final version of the manuscript.

Competing Interest Statement: J.R. is the cofounder of LUCA Biologics, a biotechnology company focusing on translating microbiome research into live biotherapeutics drugs for women's health. All remaining authors have no disclosures to declare.

Abstract:

Diverse and non-*Lactobacillus*-dominated vaginal microbial communities are associated with adverse health outcomes such as preterm birth and acquisition of sexually transmitted infections. Despite the importance of recognizing and understanding the key risk-associated features of these communities, their heterogeneous structure and properties remain ill-defined. Clustering approaches have been commonly used to characterize vaginal communities, but they lack sensitivity and robustness in resolving community substructures and revealing transitions between potential sub-communities. We used a more highly resolved approach based on mixed membership topic models with multi-domain longitudinal data from cohorts of pregnant and non-pregnant subjects to identify several non-*Lactobacillus*-dominated sub-communities common to women regardless of reproductive status. These sub-communities correlated with clusters of metabolites. In non-pregnant subjects, we identified a few sub-communities that were more common during menses but did not predict an increased likelihood of non-*Lactobacillus*-dominated communities during the rest of the menstrual cycle. The menstrual cycle was a strong driver of transitions between sub-communities and was correlated with changes in levels of cytokines, for example, elevated TNF-α concentrations at the time of ovulation, and metabolites, for example, elevated kynurenine concentrations during menses. In pregnant women, some metabolite clusters were predictive of changes in vaginal microbiota structure. Overall, our results show that the vaginal community substructure is shaped by the menstrual cycle and that specific sets of metabolites are associated with community instability during pregnancy.

INTRODUCTION

Several critical aspects of women's health are associated with the composition of the vaginal microbiota (1–3). When associated with positive health outcomes, the vaginal microbiota is dominated by *Lactobacillus* species (3). A paucity of *Lactobacillus* and a diverse array of strict and facultative anaerobes, however, are associated with negative health outcomes such as preterm birth (4, 5) and susceptibility to sexually transmitted infections (6–9), including HIV (10–12). Some *Lactobacillus* species, such as *L. crispatus* or *L. gasseri*, are more robust to invasion or take-over by non-*Lactobacillus* species and create greater vaginal ecosystem stability during and outside of pregnancy (13–15). Other *Lactobacillus* species, such as *L. iners*, are more frequently associated with non-optimal communities (13–15). In some individuals, the vaginal microbiota composition may change several times over the course of a few months or even a few days (4, 13, 14, 16). In non-pregnant menstruating women, menses are associated with drastic changes in microbiota composition (13). In a cross-sectional study, recent sexual intercourse was associated with the presence of specific bacteria such as *Corynebacterium* (17). These associations suggest that changes in women's physiology, the interactions with the penile microbiome, and/or other factors may perturb the vaginal microbiota (3). However, the underlying mechanisms driving changes, and predictors of change in microbiota structure are still not identified, especially at times of the menstrual cycle other than menses. It is unclear why the vaginal microbiota of some individuals easily transitions back to an optimal state from a non-optimal one, while other microbiotas do not (4, 13, 14, 18).

In this study, we pursued three objectives. First, we sought to deepen our understanding of the structure of non-optimal vaginal microbiotas. To address this objective, we used mixed membership statistical methods to explore whether distinct sub-communities could be identified in a robust manner. Second, we evaluated the potential associations between reproductive status of the host (gestational age and phase of the menstrual cycle) and variation in features of the vaginal ecosystem, defined here as the set of bacteria, microbial and host metabolites, and host cytokines present in the vagina. Third, we sought to describe the temporal dynamics of the vaginal microbiota and identify metabolites or cytokines predictive of changes in vaginal microbiota structure.

Our first objective was motivated by the fact that non-optimal vaginal microbiotas are highly heterogeneous among and across individuals (4, 13, 14). Clustering approaches are commonly used to define community structure and have led to the adoption of the concept of community state types (CST) (19, 20). However, while an important dimensionality reduction tool for these complex datasets, this categorization tends to oversimplify community structure in non-optimal microbiota (often called the Community State Type IV), and importantly, hides transition states between clusters. For example, the vaginal microbiota of two individuals could switch from one cluster (CST) to another and be described by the same sequence of CSTs, but one individual may experience a progressive shift in her vaginal microbial structure (*e.g.,* the proportion of *Gardnerella* could be slowly decreasing over time) while the other may experience an abrupt change of her vaginal microbiota (*e.g., Gardnerella* completely disappearing over a few days). These two situations may be driven by different mechanisms and have different health implications. Moreover, while clustering approaches can identify sets of species which frequently co-occur, they are not well suited to identify subsets of species which may have similar functions but that are not frequently found together. Consequently, to address our first objective, we turned to mixed membership models, and in particular topic models, to describe vaginal microbiota structure.

Mixed membership models have been developed to infer population structures from multilocus genotype data (21). More recently, topic models were described by Blei et al. (22) as Latent Dirichlet Allocation (LDA) in natural language research (NLP), and have been adopted for analyzing microbiota with the goal of identifying bacterial sub-communities (23). LDA models documents as a mixture of topics, and each topic is identified by a particular distribution of word frequencies (22). In the case of microbiota, documents are equivalent to biological samples, words are equivalent to bacterial species or strains, and topics can be viewed as bacterial sub-communities. Concretely, the structure of the community found in each sample is summarized by the proportion of each topic, and topics are characterized by the prevalence of each species or strains. For example, a sample could be described as 70% one topic and 30% another topic. This means that the species subsumed in the first topic account for 70% of the sample, while the species in the second topic account for the remaining 30%. Some species can be found in several topics (*e.g.,* a species can co-exist within two distinct sub-communities). Topics may be composed of a few species or strains (sparse topics) or include a larger number. Here, we have applied topic analysis to vaginal samples from two longitudinal cohorts of women in order to identify specific sub-communities, and report their temporal dynamics and the transitions between them. The first cohort was composed of pregnant subjects who provided weekly samples throughout their pregnancy. The second cohort was composed of non-pregnant subjects of reproductive age who provided daily samples for ten weeks.

Our second and third objectives, which were to identify factors driving changes in microbiota structure across reproductive status (*i.e.,* in pregnant and non-pregnant individuals), required multi-domain data (data quantifying distinct biological components) acquired longitudinally from both menstruating and pregnant women. While several studies have collected longitudinal vaginal microbiota data (4, 5, 14, 18, 24, 25), relatively few of them also quantified metabolite concentration (13, 26) or cytokine levels (27–29), and none collected samples in both pregnant and non-pregnant subjects or accounted for potential menstrual cycle effects on the vaginal environment. In this study, both metabolites and cytokines were quantified longitudinally (five samples per subject) in forty subjects from each cohort (40 pregnant and 40 non-pregnant participants). The data integration from 16S rRNA gene sequencing, metabolites, and cytokines quantification allowed us to characterize the relationships among these different kinds of data and to describe the vaginal ecosystem throughout the menstrual cycle and pregnancy. Finally, we investigated the associations between metabolite clusters and stability of the vaginal microbiota composition.

RESULTS

Topic analysis reveals shared and unique bacterial sub-communities across pregnant and non-pregnant women.

Topic models were fit to the count data of 16S rRNA amplicon sequence variants (ASVs) in 2,179 vaginal samples collected weekly from 135 pregnant individuals enrolled at two sites and 2,281 vaginal samples collected daily from 40 non-pregnant individuals enrolled at three sites (Methods). Demographic characteristics are described in Table S1.1 and CST time-series are displayed in Fig S2.1-3. Distinct strain ASVs may share the same taxonomic label. Of particular interest, the three most abundant "*Gardnerella* ASVs" in the pregnancy samples have been found to be differentially associated with preterm birth risk (4) and may correspond to specific strains of *Gardnerella* (30). To facilitate linkage with the published literature, we added a G1/G2/G3 label to these ASVs (see Suppl Mat for details). Our Supplementary Material shows that these ASVs, even though they differ by only one nucleotide, match the 16S rRNA gene sequence from the genomes of cultivated *Gardnerella* isolates (Fig S1.2).

Topic analysis requires choosing K, the number of topics, for the modeling of the provided count data. This number can be estimated using cross-validation or, as recently proposed (31), by performing topic alignment across models with different resolutions (*i.e.,* with different K). This approach offers the advantage of providing diagnostic scores to characterize individual topics and evaluate deviations from LDA assumptions. In pregnant subjects, the alignment suggested that 10 topics provided the best compromise between dimension reduction and accurate modeling of the ASV counts, while in non-pregnant subjects, 11 topics were best suited (Methods, Suppl Mat).

Our analysis showed that several bacterial sub-communities (*i.e.,* topics) were shared across pregnant and non-pregnant women (Fig 1a). Besides the topics dominated by Lactobacillus spp., several topics without *Lactobacillus* showed strong agreement between cohorts and were robustly identified (Method, Suppl Mat). For example, pairs of ASVs could be identified in three topics from both pregnant and non-pregnant women. Specifically, in topics E and F, ASV 15 labeled as *Gardnerella piotii* ("G3") was frequently observed with ASV 5 - *Gardnerella leopoldii/swidzinski* ("G1") - but much less so with ASV7 *Gardnerella vaginalis* ("G2") which co-occurred with ASV10 (*Atopobium vaginae*) in topic G. These two ASVs (G2 and Atopobium) constituted a robust topic in both pregnant and non-pregnant women and showed an ability to be a dominant topic (*i.e.,* making up almost 100% of the sample composition), especially in pregnant individuals (Fig S4.4-5). Topics E highlighted the observation that ASV 8 (*Megasphaera lornae*) was more frequently observed together with ASV 9 (*Ca*. Lachnocurva vaginae, "BVAB1") than with other non-*Lactobacillus* species (Fig S4.6-7). Time series displays with these pairs can be found in the Supplementary Material (Fig S4.4-11).

The topic alignment (Fig 1b-c, S3.6,9,12,15,16) shows that the topics dominated by *Lactobacillus* were found at low resolution (*i.e,* for low K values) and were coherent topics with a median coherence score of 0.85 in K = 8 to 10 (pregnancy) and of 0.74 in K = 6 to 11 (non-pregnancy). The "*Atopobium* & G2" topic (topic G) also emerged at low resolution (K = 5 (pregnancy), 9 (non-pregnancy)) and remained coherent as the number of topics increased (median coherence scores of 0.58 (P) and 0.94 (NP)). Other non-*Lactobacillus* topics showed a weaker alignment as K increased, but two groups of non-*Lactobacillus* topics could be identified at low K across both groups. Specifically, at K = 4 (both pregnant and non-pregnant participants), two topics were identified as "parents" of distinct topic groups (Fig 1b,c). One group was composed of topics containing *Gardnerella*, *Megasphaera*, *Ca*. Lachnocurva and specific *Prevotella* ASVs. The other group is composed of topics with *Finegoldia*, *Streptococcus*, *Enterococcus*, and *Staphylococcus* ASVs. Interestingly, two of those topics were largely composed of species associated with menses in non-pregnant women. These two topics (and their corresponding ASVs) are much less abundant in pregnant subjects (Fig 1a, Suppl Mat). Topic analyses repeated on the samples from both cohorts yielded similar results (Suppl Mat). Examples of time series displays showing the topic composition of two participants are shown in Fig 2. These examples also indicate how metabolites and cytokines concentrations varied over time by displaying the metabolites or cytokine clusters for each sample.

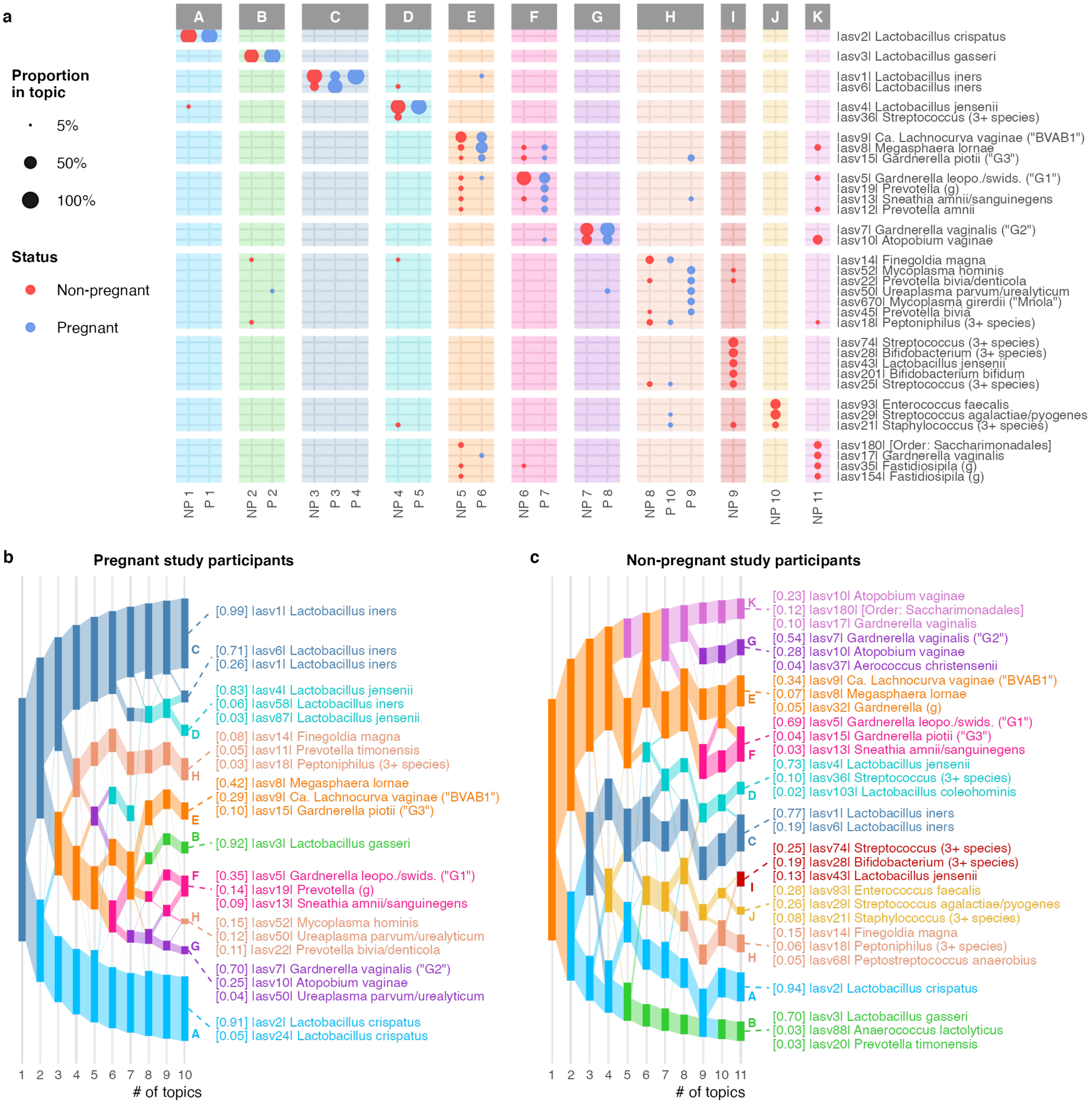

Figure 1: Topic analysis on pregnant and non-pregnant women samples reveals shared and unique sub-communities. (a) Composition of topics identified with samples of pregnant and non-pregnant subjects considered separately. The size of the dots is proportional to the LDA β parameters estimated by the topic model. These parameters model the proportion of a specific ASV in that topic. Dot color indicates whether this topic was identified in samples of pregnant or non-pregnant participants. To improve the readability of the figure, ASVs were filtered such that only ASVs that accounted for at least 5% of the topic composition were displayed. Topics were grouped (horizontal panels labeled from A to K, with background color matching topics in lower panels) by similarity based on the Jensen-Shannon divergence between topic composition. ASVs were ordered within topics by prevalence. **(b & c)** Alignment of topics (rectangles) identified on samples of pregnant subjects **(b)** and of non-pregnant subjects **(c),** as the number of topics (x-axis) was varied from 1 to 10 or 11 (optimal number of topics for samples of pregnant or non-pregnant subjects respectively). The alignment of topics revealed by analyses using K from 1 to 17 topics can be found in the Supplementary Material and used to justify the number of topics chosen here. Rectangles' height is scaled according to the total proportion of the corresponding topic in all samples: taller topics were more frequently found across samples than were smaller topics. The numbers in brackets in front of each ASV indicate the proportion of that ASV in the topic.

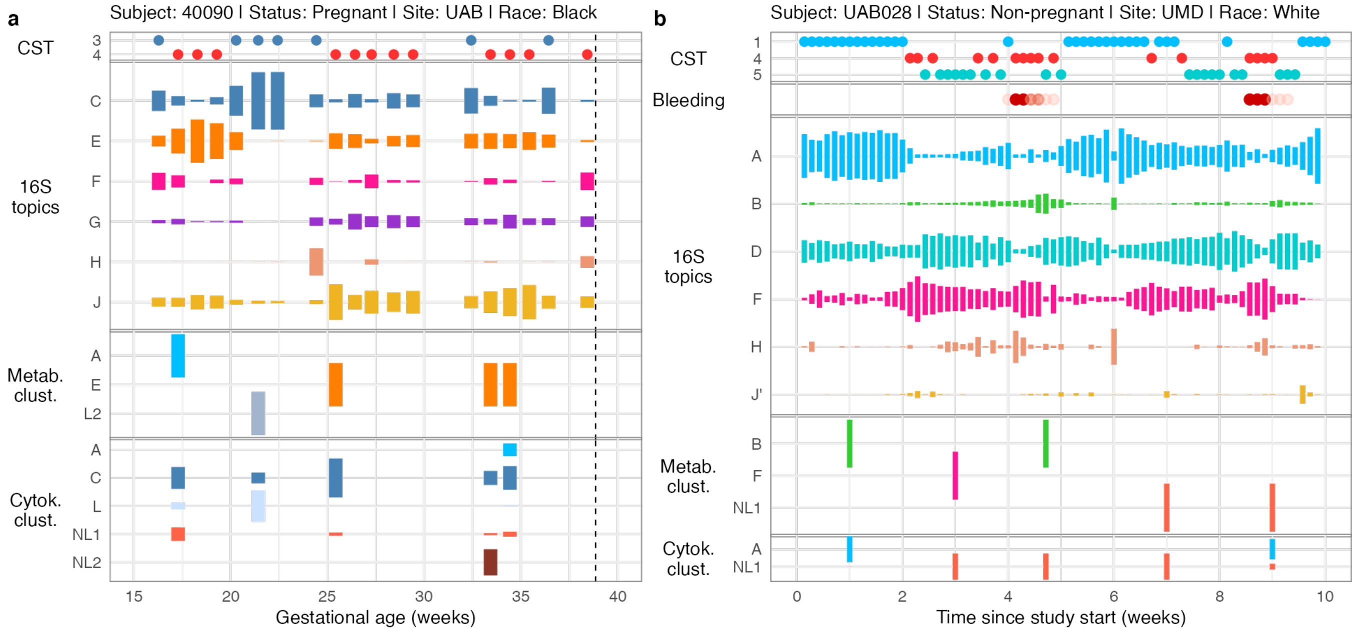

Figure 2: Examples of time-series displays of data from two subjects

(a & b) Time-series displays of the samples from two participants (one pregnant and one non-pregnant). From the top: CSTs as defined in refs (14, 19) are shown for each sample with 16S rRNA data along with self-reported bleeding for the non-pregnant participant, 16S rRNA topic composition is shown just below CST or bleeding data, while at the bottom, clusters are shown for metabolites and cytokines. For the 16S rRNA topics, the height of the bar is proportional to the topic proportion in each sample. For each participant, topics were included if they accounted for at least 10% of sample composition in at least one sample of that participant. The height of bars showing the metabolite or cytokine clusters is proportional to the inverse log10 probability of the sample belonging to each cluster (a higher bar means that the sample was closer to the cluster mean). Similarly, clusters were filtered such that their relative transformed probability was higher than 10%.

Metabolite concentrations are strongly correlated with microbial sub-communities.

Metabolite and cytokine concentrations were available at a lower temporal resolution than ASVs (Methods, Fig 2). Their concentrations were measured in 5 samples per participant in 40 pregnant and 40 non-pregnant women. Using a mixture model (Methods) to identify clusters in metabolite and cytokine concentrations, we found that metabolite clusters, but less so for cytokines clusters, were closely related to the microbial composition (Fig 3). The UMAPs (Uniform Manifold Approximation and Projection) of cytokine and metabolite concentrations further showed that the menstrual cycle, and in particular menstruation, had a strong effect on both metabolites and cytokines concentrations (Fig 3).

Vaginal concentrations of cytokines were, in general, more strongly associated with the menstrual timing than with microbial structure: cytokine clusters were weakly aligned with 16S rRNA topics (Fig 3d) and only a few cytokines were associated with the proportion of non-*Lactobacillus* species in a sample when adjusting for menstrual bleeding (Fig 3a, Fig S10.8-9). IP10, MIG, and IL8 were positively associated with *L. iners* and *L. jensenii* (but not with *L. crispatus* or *L. gasseri*) and negatively associated with several topics dominated by non-*Lactobacillus* species (Fig S10.12-13). One topic, not strongly dominated by a specific ASV, but composed of one *Finegoldia* ASV and over ten *Corynebacterium* and seven *Prevotella* ASVs, was associated with several cytokines, including IL1α, IL1β, IL8, MIG, and IP10 (Fig S10.12-13). Pregnancy status was also not associated with differences in cytokine concentration, except for MIP3α, IL1α and IL1β, which were mildly elevated levels in samples collected before 5 months of gestation (Fig S10.10).

Metabolite concentrations were strongly associated with microbiota structure. Both in pregnant and non-pregnant women, most metabolites (58% in P, 48% in NP, 34% in both) were associated with the proportion of *Lactobacillus* species in a sample (Fig 3b-c, Suppl. Mat.) when adjusting for gestational age (pregnant) or bleeding (non-pregnant). Further, there was a strong correlation between metabolite clusters and 16S rRNA topics (Fig 3e-f), indicating that metabolite concentrations are strongly associated with specific sub-communities of the vaginal microbiota. In pregnant women, metabolites were associated with gestational age, as hinted by the mild gradients on the UMAP (earlier pregnancy stages in the center on the left panel of Fig 3b). Nine metabolites were significantly associated with gestational age and using cross-validated sparse logistic regression, gestational age could be predicted with an R2 of 0.78 based on the concentration of 37 metabolites (Suppl. Mat., Fig S9.13). In non-pregnant women, menses was associated with the metabolite concentrations as most samples collected during menstruation were located in the same area of the UMAP (Fig 3c).

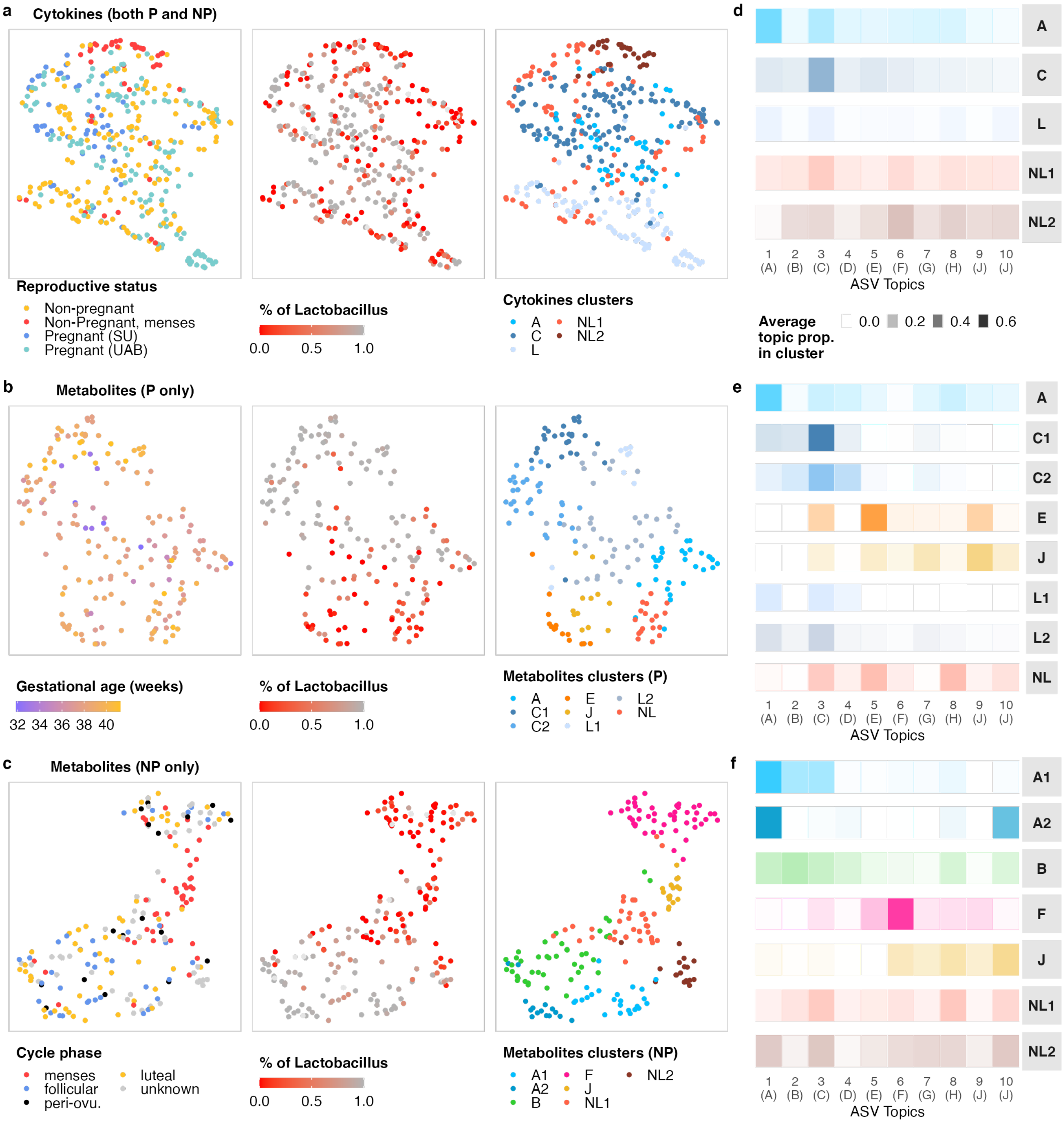

Figure 3: Integration of metabolite and cytokine clusters with microbial sub-communities

(a-c) UMAP (Uniform Manifold Approximation and Projection) of cytokines (a) and metabolites (b-c) concentrations (each dot is a sample). Because metabolite samples from pregnant and non-pregnant women were processed slightly differently, batch effects prevented a combined analysis of samples from the two groups. The same UMAP was reproduced three times with different colors for each modality. Samples were colored by reproductive status, gestational age, or cycle phase on the left panels; by the proportion of Lactobacillus species in each sample in the center panels; and by their respective clusters on the right. **(d-f)** Color intensity indicates the proportion of each 16S rRNA sub-community (topic from analysis on all P and NP samples, horizontally) in each cluster (vertically). Color hues match clusters in corresponding panels on the left.

The menstrual cycle shapes the vaginal ecosystem.

Among the 40 non-pregnant subjects, 33 had reported vaginal bleeding patterns that allowed for the identification of at least one menstrual cycle within the ten study weeks (see Methods for menstrual cycle identification from bleeding reports). For 24 subjects, two menstrual cycles could be identified reliably. Given that the luteal phase (after ovulation) is known to vary less in duration than the follicular phase (before ovulation) (32, 33), cycles were standardized starting from 18 days before to 7 days after the first day of menses (Fig 4a, Methods). Ovulation was assumed to occur around 2 weeks before the first day of menses based on the average luteal phase duration (Methods).

Consistent with past results (19), the vaginal pH of *Lactobacillus*-dominated samples was lower (4.4, 90% 4.0-5.3) than that of non-*Lactobacillus*-dominated samples (5, 90% 4.0-5.8). The pH remained stable throughout the cycle (*Lactobacillus*-dominated: 4.3, 90% 4.0-5.3; non-*Lactobacillus* dominated: 4.9, 90% 4.0-5.5), except during menses when it increased by about 0.5 units in both *Lactobacillus*-dominated samples (4.7, 90% 4.0-5.8) and non-*Lactobacillus*-dominated samples (5.4, 90% 4.4-7.0) (Fig 4b).

Half of the measured cytokines showed concentration changes associated with the menstrual cycle (Fig S10.2-4, Table S10.1). Most cytokines had elevated levels during menses (Fig S10.1,5). However, a few cytokines (e.g., INFγ), increased around the predicted time of ovulation or, like MIP3α, exhibited a gradual increase throughout the luteal phase (Fig 4c). The abundance of about 45 metabolites (out of 355) was associated with the menstrual cycle (Fig S9.10-11). While the largest changes were often found around the time of menses (Fig S9.10-11), several metabolites also exhibited changes during specific phases of the menstrual cycle, such as following ovulation (e.g., isoleucine), or in the mid- or late-luteal phase (e.g., aspartate) (Fig 4d, Suppl Fig S9.10-11).

To assess the associations between the vaginal microbiota composition and the menstrual cycle, we used a longitudinal approach and compared the vaginal microbiota composition between two consecutive cycles of a given participant. The vaginal composition, characterized by the ASV rank (or relative abundance in the Suppl. Mat., Fig S7.7-9), showed a high correlation between two consecutive cycles (Fig 4e-f). We used this longitudinal (within-subject) approach because the absolute abundance of ASV cannot be established from 16S rRNA gene amplicon sequencing, so we could not test for their association with time in the menstrual cycle as we did for metabolites and cytokines. We also used a cross-sectional approach to evaluate the impact of the menstrual cycle on the probability of an ASV being present (*vs.* absent) on a given cycleday. We found a strong association for 85 ASVs (Fig S7.19), among which 52 (61%) had a higher probability to be present during menses compared to the rest of the cycle (Fig S7.19). A total of 17 (20%) ASVs, including four *Prevotella* and one *Atopobium* ASV, were associated with the late luteal phase (Fig 4g).

We found that menses-associated species (e.g., *Finegoldia* or *Pseudomonas*, see Table S7.1) were not associated with the presence of other non-*Lactobacillus* species, such as *Gardnerella* spp., at the other phases of the menstrual cycle (Fig S7.23). Participants whose pre-menses samples were dominated by *Lactobacillus* species were as likely to see an increase in non-*Lactobacillus* species during the menses (Suppl. Mat.). Similarly, the presence of these menses-specific species (or topics, see Suppl. Mat. for replicated analysis on topic proportions) was not associated with a higher risk of non-*Lactobacillus* species following menses. Finally, we found that some menses associated ASVs are also found around the time of ovulation in some individuals (Fig S7.22). Because most participants used a combination of pads and tampons throughout their menses, we did not find significant associations between topics and menstrual protection use (Table S7.2, Fig S7.27). Our analysis also confirmed that behaviors suspected to disrupt the vaginal microbiota, such as sexual intercourse or the use of vaginal gels or powder, were associated with changes in the vaginal microbiota. Indeed, we observed higher distributions of changes in microbiota composition for participants who more frequently reported these behaviors (Fig S7.26). However, this study did not have the statistical power to identify which behaviors were most disruptive.

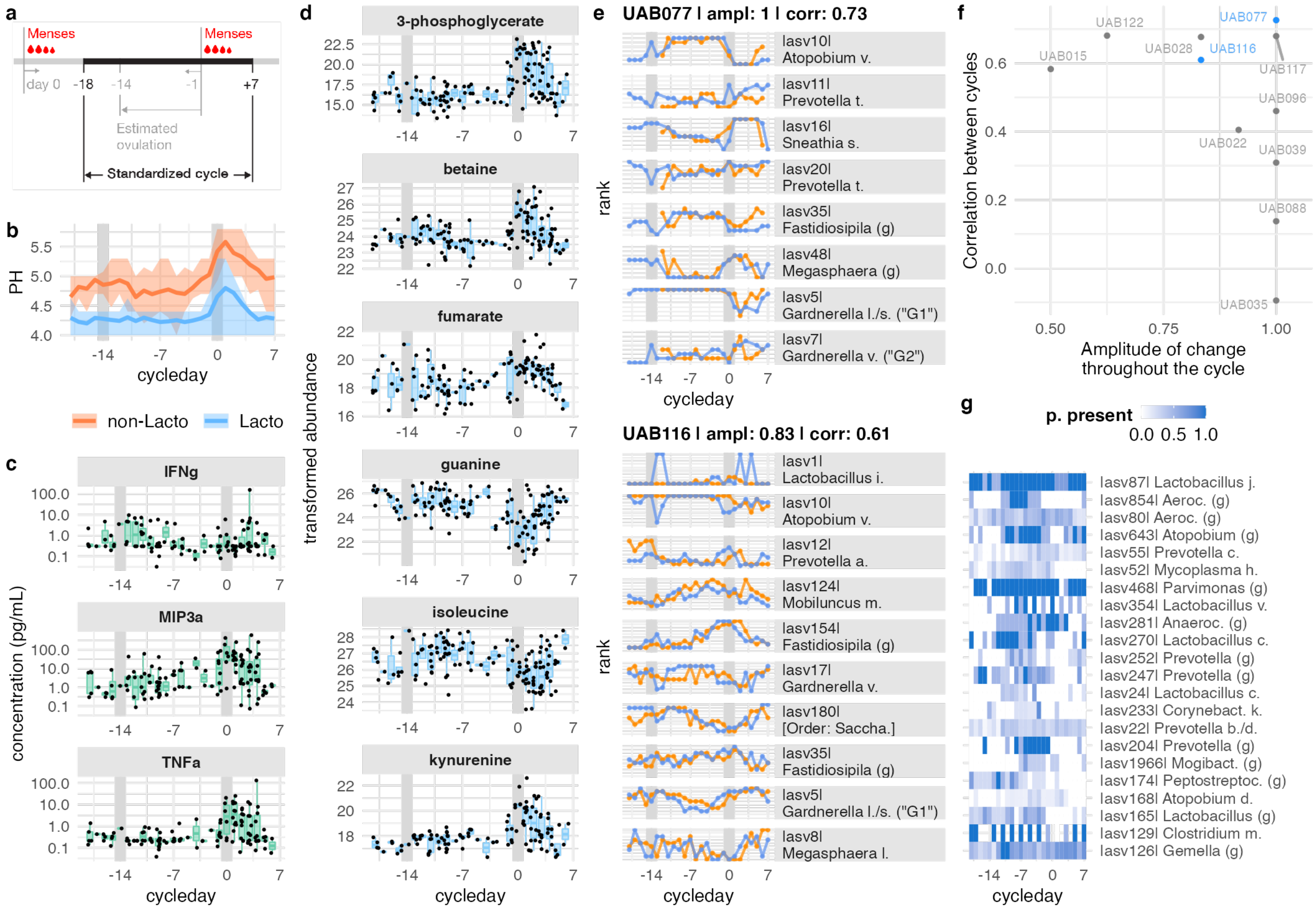

Figure 4: The menstrual cycle shapes the vaginal ecosystem.

(a) Schematic illustrating the definition of standardized cycles. **(b)** Vaginal pH throughout the menstrual cycle for *Lactobacillus* or non-*Lactobacillus*-dominated vaginal microbiotas. **(c)** Vaginal cytokine concentration (pg/mL) throughout the menstrual cycle for three cytokines associated with time in the menstrual cycle. Profiles of all measured cytokines are shown in Fig S10.5. **(d)** Vaginal metabolite relative concentration throughout the menstrual cycle for 6 metabolites showing a strong association with time in the menstrual cycle. Profiles of all metabolites with an association with time in the menstrual cycle are shown in Fig S9.10) **(e)** Vaginal microbiota structure of two subjects with data available for at least two full menstrual cycles. The first menstrual cycle is displayed in orange, the second in blue. These two subjects were selected to show the diversity of temporal profiles. The time series display shows ASV rank (y-axis, a higher rank value indicates that this ASV was more abundant than other ASVs) on each cycle day (x-axis). For each study participant, ASVs were included if they were part of a pool accounting for 98% of the total composition in a sample and if they were present in at least ten samples from that participant. **(f)** Scatter plot, in which each dot is a participant, the correlation in vaginal microbiota composition between two cycles of an individual is on the vertical (y-axis), and the relative amplitude of change throughout the cycle is on the horizontal axis (Methods). **(g)** Heatmap of the probabilities an ASVs is present on a given cycleday. ASVs whose presence probability was significantly higher in the luteal phase were selected for this panel. A similar heatmap with ASVs associated with other phases of the menstrual cycle is available in the Supplementary Material (Fig S7.19).

Metabolite clusters are associated with changes in vaginal microbiota composition.

To address the third objective of this study (*i.e.,* to identify drivers of changes in the composition of the vaginal microbiota), we relied on the multi-domain nature of our datasets. Specifically, we used the pregnant participants' samples to identify metabolites and/or cytokines that were predictive of the future composition of the vaginal microbiota. The samples of the non-pregnant participants were not used for this objective for two reasons. First, only a few non-pregnant subjects switched from a *Lactobacillus*-dominated state to a non-*Lactobacillus*-dominated state (or reversely) (Fig S1.4), and second, our results showed that the menstrual cycle is modulating the vaginal microenvironment. The combination of these two elements implied that non-pregnant women's data would not allow disentangling the effects of the menstrual cycle on the composition of the vaginal microbiota from those of specific metabolites or cytokines.

Intrigued by the fact that samples from different metabolite clusters (A, C1, C2, L1, L2, fig 3e) shared a similar topic (and ASV) composition, we investigated whether metabolite clusters were indicative of microbial stability. To do so, we compared the distributions of changes in topic composition in each metabolite cluster (Fig 5). We observed that samples of the A and L2 clusters were more likely than those of the C1, C2, and L1 clusters to be associated with changes in topic composition (Fig 5), despite all containing samples fully dominated by *L. crispatus* or *L. iners*. In other words, samples fully dominated by *L. crispatus* were more likely to still be dominated by *L. crispatus* in the previous or next sample if their metabolic state was best represented by clusters C1, C2, or L1.

In addition, we fitted sparse regressions on the pregnant subjects' samples to predict the relative abundance (proportion) of *Lactobacillus* species or the proportion of specific ASV topics at the next time-point (Suppl Mat). Three models were fitted to assess the predictive power of metabolites and cytokines. In the first model, metabolite abundances were included as input variables; in the second model, cytokine abundances were included; and in the third model, both metabolites and cytokines were included as input variables. Given that we observe a high correlation (0.76) between the *Lactobacillus* proportion (or topic proportions) from one sample to the next (approximately sampled a week apart), all models also included the current *Lactobacillus* (or topic) proportion as one of the input variables. When predicting the proportion of *Lactobacillus* species at the next time-point, the first and third models produced the most accurate predictions. The inclusion of cytokines abundance in the model did not improve the predictions (Fig S12.9-10). There was moderate overlap between the metabolites found to be predictive of the proportion of specific topics (Fig S12.21). Metabolites positively associated with increases in *L. crispatus* were often negatively associated with increases in *L. iners* and those positively associated with increases in topics of group E and F (*i.e.,* containing "G1" and "G3" *Gardnerella* ASVs) were negatively associated with increases in either *L. crispatus* or *L. iners* (Fig S12.21).

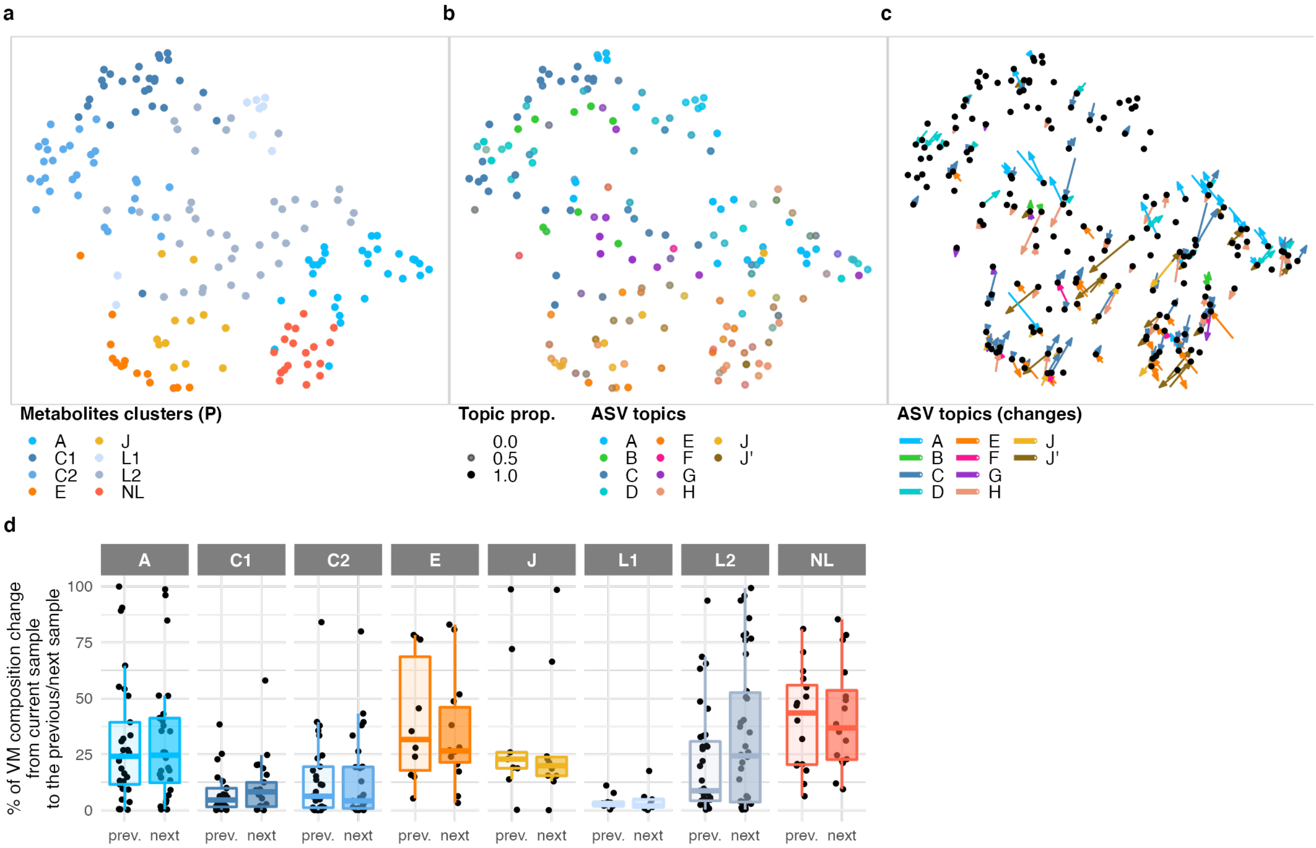

Figure 5: Metabolite clusters are associated with instability in the vaginal microbiota.

(a-c) UMAP (dots) of the metabolite concentrations in samples from pregnant women (same as in Fig3b). Each dot is representing a sample collected at a specific time-point. **(a)** Same as in Fig 3b. **(b)** Same UMAP where each dot is colored by 16S topic proportion. Transparency is used to reflect topic proportion such that dots with bright colors matching the legends represent samples dominated by a specific topic while dots with mixed colors represent samples composed of several topics. **(c)** Arrows show changes in topic proportion from the previous or the next time-point. Incoming arrows show the difference in topic proportion from the previous time-point to the current time-point. Outgoing arrows show the difference in topic proportion from the next time-point to the current time-point. The angle of the arrow is related to the topic, with arrows for topics dominated by Lactobacillus on top of the dots and arrows for topics dominated by non-Lactobacillus topics on the bottom. The length of the arrow is proportional to the difference in topic proportion. For example, a dot with two large light sky-blue arrows indicates that both the previous and following time-points were dominated by topic A, but not the current one. **(d)** Distribution of changes in vaginal microbiome composition from one sample to the next (right boxes in each panel) or from the previous sample to the current sample (left boxes in each panel) per metabolite cluster in the current sample (horizontal panels).

DISCUSSION

In this study, samples from both pregnant and non-pregnant participants were used to identify specific bacterial sub-communities within non-*Lactobacillus* dominated samples. We studied the impact of pregnancy and the menstrual cycle on the vaginal ecosystem, as quantified by vaginal microbes, cytokines, and metabolites. Applying topic analysis separately on samples from pregnant and non-pregnant women, we identified several non-*Lactobacillus* sub-communities (*i.e.,* topics) that presented a substantial overlap across the two groups. Within these sub-communities, we identified several pairs of bacterial ASVs that appear to often co-exist. For example, specific *Gardnerella* ASVs (G1 and G3) co-occur more frequently than others (G2 and G3) and the most common *Atopobium* ASV appears to form a dominating and robust sub-community with G2 in 42% of the subjects for which *Atopobium* is present. It remains to be experimentally tested if these bacteria are symbiotic and if specific conditions, such as a given hormonal milieu, are favorable to their emergence.

The main difference between the approach used here (topic analysis) and clustering approaches traditionally used to identify sub-groups in vaginal microbiota is that topics enable mixed membership models allowing samples to be associated with several topics (sub-communities) in different proportions. When comparing the sub-communities identified by our topic analysis to the community state types (CSTs), we observe a good overall correspondence. The first robust topics emerging at low K (the number of topics) correspond roughly to the 5 previously identified CSTs (19) with the exception that *L. jensenii* does not become its own topic until more topics are allowed. This could be because *L. jensenii* is less abundant than other *Lactobacillus* species and is often co-present with *L. iners*. Two groups of non-*Lactobacillus*-dominated topics emerge at low K, with topic composition roughly matching the description of the two sub-types CST IV-A and IV-B (34). These two topics further split into more topics which also share some overlap with the more recently introduced CST IV sub-types (20). However, given the differences in taxonomic resolution (species *vs* ASVs), the comparison between these clusters and our topics is only approximative. The clustering approach in (20) provides three main subtypes (A, B, and C) for the non-*Lactobacillus* samples. One observation is that the separation between clusters A and B and cluster C seems to match one of the early splits in our non-*Lactobacillus*-dominated topics. Another observation is that, in (20), cluster C is further broken down into 5 categories. In contrast, topic models allow the species dominating each of these categories to remain grouped in a single or in a few topics. That is because, in the case where two species are found interchangeably (but not simultaneously) with a specific group of other species, these two species will be found in the same topic. Indeed, topic models allow for synonyms, which may reflect potential functional equivalence in the microbiota context.

We found that clusters fitted on metabolites concentrations correlated well with the microbiota sub-communities. However, there was a weaker association between cytokine concentrations and the microbiota composition.
While several studies had already shown that menses were associated with disruptions of the vaginal microbiota (13, 35, 36), the vaginal ecosystem, defined as the set of microbes, metabolites, and cytokines, had not yet been characterized throughout the whole menstrual cycle, and longitudinal correlations between cycles had not been investigated. Overall, our findings indicate that the menstrual cycle has a substantial impact on the vaginal ecosystem. Our results suggest that hormonal fluctuations occurring throughout the cycle are associated with changes in the concentration of some metabolites and cytokines, and in the composition of the vaginal microbiota. In most cases, it is difficult to assess whether the cycle has a direct effect on each domain or only on one of them. For example, the cycle may affect the microbiota composition, which in turn alters the other domains. While our

cytokine and metabolite data are too sparse to disentangle these effects outside of menses, we observe that the shifts in cytokine concentrations during menstruation are independent of the microbiota composition.

Almost all subjects with two menstrual cycles of data show a high between-cycle correlation in their vaginal microbiome variations. However, some species reach their maximal relative abundance at different menstrual cycle phases across individuals. These differences could simply be due to our inability to detect absolute levels (ASV relative levels or rank prevent a reliable identification of the peak phase of each ASV). Besides this technical limitation, these inter-individual differences could be due to (i) inter-individual differences in menstrual timing (for example, one subject might have a 10-day luteal phase while another one might have a 14-day luteal phase), (ii) to inter-individual differences in hormonal levels (or the rates of change), (iii) to the set of species present in each subject and how each of these species might respond differently to the menstrual cycle while competing for resources. Future clinical studies should include the measurement of hormonal levels and use methods to reliably estimate ovulation timing.

Similarly, additional data would be necessary to understand the substantial changes in the composition of the vaginal microbiota during menses and why some of the menses-associated ASVs, such as specific ASVs assigned to *Finegoldia*, *Corynebacterium,* and *Sneathia sanguinegens*, are also found around ovulation. Several hypotheses might explain these observations. The first one is that the presence of these menses-associated ASVs could be associated with menstrual protections, such as pads, which are also used by some women when they experience vaginal discharge between menses (37). Here, we only observed minor differences in topic composition in samples in which participants reported using pads and/or tampons. Another hypothesis is that these ASVs could be part of the uterine (or cervical) microbiota and be found in the vagina as fluid (blood or mucus) flows through the cervix. Unfortunately, the resolution of past works on the microbiota of the upper genital tracts does not offer much opportunity to support or reject this hypothesis (38–40). These bacteria could also be hormone-dependent and thrive when estradiol levels drop, which happens both during menses and around ovulation (and after delivery). Alternatively, these bacteria may not be estradiol-dependent and remain when estradiol-dependent bacteria disappear. Hormonal measurements at a high temporal resolution would be necessary to address this hypothesis. Finally, uterine blood may bring nutrients or changes in the biophysical and biochemical environment that are favorable for the growth of these species. Their presence at ovulation would be explained by light ovulation bleeding experienced by some individuals (41). This hypothesis is not supported by the data presented here since none of our subjects reported bleeding around the time of ovulation. Interestingly, the presence or levels of these menses associated ASVs is not associated with a higher probability of having non-*Lactobacillus* bacteria at other times in the cycle. In fact, many participants with *L. crispatus*-dominated vaginal microbiota show a high increase in these ASVs during menses. These species are also rarely found in the pregnant participants' vaginal microbiota.

About 45 metabolites show an association with a specific phase of the menstrual cycle. Some of the metabolites, which we found to have a strong association with the menstrual cycle, include components such as kynurenine or isoleucine. Kynurenine is a tryptophan catabolite via a pathway involving IDO1-mediated degradation and is known to play a role in blood vessel dilatation during inflammatory events (42). The elevated levels of kynurenine during menses found in our study are thus consistent with these roles and with past studies showing differential levels in kynurenine serum and urinary levels through the cycle (43, 44). Isoleucine is a branched-chain amino acid, and as such, plays important metabolic functions (45). In our vaginal samples, isoleucine levels are found to be highest in the luteal phase and lowest during menses. Interestingly, serum levels of isoleucine show opposite trends (46). Over 50% of measured metabolites show an association with the proportion of *Lactobacillus*.

Cytokines mostly show an association with the menstrual cycle (half of them have a significant association) but not with pregnancy. Only a few of them, such as IP-10, show associations with the proportion of *Lactobacillus* or specific topics. Previous studies in non-pregnant subjects found similar variations in cytokine levels throughout the menstrual cycle (27, 28). Specifically, this study confirms the positive association between IP-10 and the proportion of *Lactobacillus* species found in (27). Another study (12) based on a large cohort of young South African women found associations between vaginal cytokine levels and microbiota composition. Specifically, IL-1α, TNF-α, and IL-8 were higher in women whose vaginal microbiota was dominated by non-*Lactobacillus* species (12). However, about half of the participants in that study were on an injectable progestin contraceptive and samples were not collected during menses, preventing the adjustments for time in the menstrual cycle. Here, after adjusting for bleeding in non-pregnant subjects, we find the same association for TNF-α in pregnant but not in non-pregnant subjects. As in (12) IL-8 is found here to be mildly positively associate with L. iners compared to L. crispatus but, in contrast to (12), IL-8 is not found to be associated with the proportion of non-Lactobacillus species , in contrast is found with an opposite association than in (12) in non-pregnant subjects. In addition, IL-1β shows a strong negative association with the abundance of *Lactobacillus crispatus* species in our dataset, which was also previously found elevated during and after BV events (27, 47, 48).

When investigating if specific metabolites or cytokines were associated with changes in the vaginal microbiota in pregnant subjects, we found that two metabolite clusters with a similar ASV composition were differentially associated with temporal changes in the ASV composition. The differences in metabolite concentrations between these two clusters could be explained by different transcriptional activities of apparently similar microbiota composition. It was indeed recently suggested that differential transcriptional activities were associated with changes in metagenomic composition (49) and these changes in activity could be reflected in the metabolic differences observed here. We also used supervised learning to identify metabolites or cytokines associated with the composition of the vaginal microbiome at the next time point in pregnant subjects and found that metabolites were better predictor of changes than cytokines. These results suggest that beyond the strong association between microbial composition and metabolite concentrations, the metabolic state is also associated with local stability in the microbial composition. This highlight the important role of metabolomics for vaginal health as specific metabolites were also found to be predictive of spontaneous preterm births (26).

The main strength of this study is the richness of the dataset as several modalities (microbiota, cytokines and metabolites) were collected longitudinally in cohorts of pregnant and non-pregnant women. Another strength is the use of high-resolution statistical methods (e.g., topic analysis) to address the study objectives. The longitudinal nature of the dataset allowed us to study the effect of time in the menstrual cycle and to overcome the limitations inherent to 16S rRNA sequencing in quantifying absolute level changes. Our data integration across several domains allowed for a detailed description and understanding of the changes in the vaginal microenvironment in pregnant and non-pregnant women. However, the lack of hormonal measurements, of biomarkers indicative of ovulation, and of mucus quality may have hindered our ability to uncover potential associations with ovulation, hormonal fluctuations, and mucus characteristics. For example, cervical mucus, whose production, and biochemical properties depend on steroid hormones and undergo drastic changes around ovulation, may play an important role in regulating the growth of specific vaginal bacteria. Inversely, the presence of specific bacterial species might affect the ability of cervical epithelial cells to produce mucus and maintain mucosal barrier integrity. The data collected in this study did not allow us to investigate these interactions. While we observe that sexual activity is sometimes associated with changes in the vaginal microbiota, measurements of partners' microbiome, which have been shown to be relevant for vaginal microbiome composition or bacterial vaginosis (50–53), were not included here, preventing the study of potentially related associations. Finally, a larger number of subjects in the non-pregnant cohort may have allowed us to confirm the associations between metabolites and changes in the composition of the vaginal microbiota that we find in the pregnant subjects, while a higher temporal resolution for the metabolites and cytokines data would have increased our statistical power.

Conclusions

Topic analysis revealed bacterial sub-communities (topics) shared across pregnant and non-pregnant women. Compared to clustering approaches traditionally used to describe microbial composition, topics provide an expanded characterization of the heterogeneity of the previously described community state type IV (CST IV) and a high-resolution view of transitions between communities. We found that metabolite concentrations highly correlated with the make-up of microbial sub-communities and that the menstrual cycle had a strong impact on the vaginal ecosystem, defined here as the composition of the vaginal bacteria, metabolites, and cytokines. In-vitro studies will provide further functional insights into the identified sub-communities, their ecological network, and their effect on the vaginal epithelium. Future clinical studies should include hormonal measurements to allow for a deeper understanding of the impact of the hormonal milieu on the vaginal microbiota. These studies could then inform whether certain phases of the cycle are more favorable to the growth of *Lactobacillus*, and if windows of time in the cycle would be better suited for interventions targeting vaginal dysbiosis. Finally, specific metabolic states and a few cytokines were found to be associated with changes in microbial composition. In vitro confirmation of these associations could improve the development of products supporting the growth of *Lactobacillus* species and the restoration of an optimal vaginal microbiota.

MATERIAL AND METHODS

Cohorts and sample collection

Daily samples from non-pregnant women. For this study, the samples from two cohorts with identical recruitment criteria and sample collection methods. The samples from 10 participants from a larger multi-site cohort recruited at Emory (EM) and University of Maryland, Baltimore (AYAC) were used along with samples from 30 participants recruited at the University of Alabama Birmingham (UAB). Each participant self-collected daily vaginal swabs for 10 weeks, resulting in a maximum of 10 x 7 = 70 samples per individual. For further detail about recruitment criteria and sample collection, see (16).

Weekly samples from pregnant women. We used the samples from both cohorts presented in (4). 39 pregnant subjects were recruited at Stanford University (SU) and 96 pregnant subjects were recruited at the University of Alabama Birmingham (UAB). Participants recruited at UAB received intra-muscular injections of progesterone throughout pregnancy. Participants were enrolled from the 4th month of their pregnancy (earliest enrollment at week 8, latest at week 22) and vaginal swabs were collected weekly (approximately) until delivery.

Selection of the “VMRC samples”. All samples described above were used for 16sRNA sequencing. Metabolites and cytokine concentrations were quantified on a subset of these samples. Specifically, 5 samples from 40 pregnant and 40 non-pregnant women were selected and labeled “VMRC samples”. The 40 pregnant women were selected out of the initial pool of 135 participants as a representative sample of the larger pool based on their Jensen-Shannon stability index, the CST composition, and gestational age at delivery. 5 samples were then selected so that they were equally distributed in time throughout pregnancy. The 5 samples from the 40 non-pregnant women were similarly selected so that samples were distanced by approximately 2 weeks.

Ethics

All participants provided written informed consent. Ethical approval was obtained from the Institutional Review Boards of Stanford University, the University of Alabama, Birmingham, and the University of Maryland. All research was conducted in compliance with relevant guidelines and regulations.

Vaginal microbiome sequencing

Daily samples from the 10 non-pregnant participants recruited at EM and AYAC. the V1V4 regions of the 16S rRNA gene were amplified then sequenced using 454 sequencer, as described in (16).

Daily samples from the 30 non-pregnant participants recruited at UAB (1534 samples). The V3V4 regions of the 16S rRNA gene were amplified then sequenced with Illumina HiSeq/MiSeq

Weekly samples from pregnant participants of both cohorts (SU and UAB) (2179 samples): Raw sequence data from pregnant participants samples were generated and processed as described in (4). In brief, genomic DNA was extracted from vaginal samples using a PowerSoil DNA isolation kit (MO BIO Laboratories). Barcoded primers 515F/806R (54) were used to amplify the V4 variable region of the 16S rRNA gene from each sample. Pooled amplicons were sequenced on Illumina HiSeq/MiSeq instruments at the Roy J. Carver Biotechnology Center, University of Illinois, Urbana-Champaign.

VMRC samples (393 samples) from the 40 pregnant and 40 non-pregnant participants were re-sequenced using the same protocol as for the pregnant participants’ samples.

Demultiplexed raw sequence data from the Illumina HiSeq/MiSeq (samples of pregnant subjects, of the 30 UAB non-pregnant participants, and VMRC samples) were resolved to amplicon sequence variants (ASVs) as described in the DADA2 Workflow for Big Data (https://benjjneb.github.io/dada2/bigdata.html) (55).

Taxonomic assignment. Automated taxonomic calls were made using DADA2’s implementation of the RDP naive Bayesian classifier (56) and a Silva reference database (version 132) (57). The assignment of sequences of the most abundant ASVs were refined and standardized by BLASTing against NCBI RefSeq type strains. This is the case for *Lactobacillus, Candidatus* Lachnocurva vaginae (previously referred to as BVAB1), *Gardnerella*, and *Megasphaera lornae* species-level assignments, following recently published work on these species (58, 59). *Gardnerella* ASVs were tagged as G1, G2, or G3 sensu (4) based on exact matching.

ASV harmonization across assay. While most analyses were performed the data from each assay separately, we were interested to also fit the topic model on the combined assays. However, given that two different sequencing technologies were used for the samples from the pregnant and the non-pregnant subjects, the sequenced region of the 16S rRNA gene differed between assays but shorter regions were all included within the longer regions. ASV sequences were thus matched across assays. We used exact matches (*i.e.,* a shorter sequence had to be fully and exactly included into a longer sequence). If several longer sequences exactly matched a given shorter sequence, the counts of the corresponding longer ASVs were aggregated. The taxonomic assignment of the shorter sequence was used for the matched ASVs.

Phylotype level data for all samples of non-pregnant participants. To harmonize the data from all non-pregnant participants, the V3V4 (HiSeq/MiSeq) and V1V3 (454 sequencer) 16S count data were aggregated at the species level (or at a higher taxonomic level when taxonomic data was not available at the species level) resulting in a table of dimension 225 species x 2281 samples (approximately 40 subjects x 70 samples). This table was used to correlate metabolite and cytokines concentrations with relative abundance of Lactobacillus species. Topic analysis was also performed on this data and presented in the Supplementary Material.

Metabolite concentration quantification

Untargeted metabolomics was performed on the 400 VMRC samples by ultra-high-performance liquid chromatography/tandem mass spectrometry (Metabolon, Inc.). Metabolite identification was performed at Metabolon based on an internally validated compound library and results were expressed in relative concentrations, following the same protocol as in (60). Samples were sent in two groups to Metabolon. All samples from the non-pregnant women and 80 samples from the pregnant women were sent in a first batch. The 120 remaining samples from the pregnant women were sent in a second batch. We observed a mild batch effect between the samples from the pregnant and non-pregnant women, likely due to slightly different collection and storage procedures, and between the samples of the first and second batch in pregnant women's samples. While it was not possible to remove the batch effect between the samples from pregnant and non-pregnant women because of potential confounding from the biological state, we could re-align samples from the two batches in pregnant subjects after the variance stabilizing transformation (see below) because there were no other variables found to be different in these two batches (well-mixed populations).

Data transformation. We transformed the raw metabolite relative concentrations using a variance stabilizing method (61). Raw data included the concentrations of 855 (P) and 853 (NP) metabolites. However, the abundance of 465 (P) and 517 (NP) metabolites was missing in more than 50% of the samples. We removed these metabolites from the analysis. The remaining 390 (P) and 336 (NP) metabolites were still missing in at least one sample for most of them. Missing metabolites might be missing because their abundance is lower than the detection limits or because the overall quality of a sample was lower (the proportion of missing metabolites varies widely from one sample to another and from one batch to another). 21 (P) and 1 (NP) samples with more than 60% missing metabolites were excluded for the rest of the analysis. Metabolite distributions were re-aligned between the two batches of pregnant women's samples by subtracting the median differences between the two batches.

Imputation. For some analyses, such as dimension reduction of the metabolite space (e.g., UMAP) or future VM composition prediction, missing metabolite abundances were imputed using a mixture between a fraction (0.95) of the metabolite-specific lowest observed value and the imputed value using a k-nearest neighbor (KNN) imputation method. The mixture was based on the quality of the sample. Missing values in samples of high quality (with few missing metabolites) had a higher weight on the imputation to a fraction of the lowest value for that metabolite. Missing values in samples of low quality had a higher weight on the KNN imputation. For the KNN imputation, the Euclidean distance between samples was computed based on the values of co-present metabolites in sample pairs. Then, for each sample, the value of missing metabolites is imputed as the weighted mean of the values of the k=5 closest samples for which this metabolite was quantified. The weight (weighted mean) is inversely proportional to the distance between the considered sample and its k nearest neighbors.

Cytokine concentration quantification

20 cytokines were quantified in the 400 VMRC samples using a Luminex-based assay with a custom kit of 20 analytes (IFNγ, IL-1a, IL-1b, IL-4, IL-5, IL-6, IL-8, IL-10, IL-12p70, IL-13, IL-17, IL-21, IL-23, IP-10, ITAC, MIG, MIP-1a, MIP-1b, MIP-3a, and TNFα) following the same protocol as in (12). The assay was run on a Luminex FLEXMAP 3D instrument. For measurements that were below the limit of quantification for a given cytokine, values were imputed at half the lower limit of quantification (LLOQ / 2). For measurements that were above the limit of quantification for a given cytokine, values were imputed as equal to the upper limit of quantification (ULOQ). Values reported here represent medians of two technical replicates. The medians were calculated after imputation in one or both replicates (if necessary) as described above. Missing cytokine values represent technical failures of the assay for that analyte.

Data transformation and imputation. Raw cytokine abundances were log-transformed. Raw data included the abundance of 20 cytokines. Most of the cytokines could be quantified and only 14/8000 data points were missing. These missing data points were imputed using the same KNN imputation method as for the metabolite KNN imputation.

Data integration into a multi-assay experiment (MAE) object

All analyses were done in R (62). Specific packages used for the analyses are referred to in the next sections. The raw datasets were loaded and minimally processed before being formatted into SummarizeExperiment objects of the SummarizedExperiment bioconductor package (63), then combined into a single S4 object using the MultiAssayExperiment bioconductor package (64).

Identifying bacterial sub-communities using topic analysis

Microbial communities were estimated based on LDA (latent Dirichlet allocation) (22, 23). LDA models were fitted to the data for K (the number of topics) = 1 to 20 using the R package “topicmodels” (65). Models were fitted on the count data of each 16S rRNA assay separately and on the combined samples of the ASV-resolution assays.

Identifying robust topics across K (the number of topics). Topics were aligned across K using the topic alignment method described in (31). To identify robust topics across K, we used the alignment summary scores for topic coherence as defined in the same reference.

Matching topics across cohorts. Topics identified in different cohorts were matched using hierarchical clustering on the combined topics, with the Jensen-Shannon divergence used as the distance metric between topics.

Clustering of metabolite and cytokine concentrations

Samples were clustered using a parameterized finite Gaussian mixture model as implemented in the mclust R package (66). Specifically, Gaussian mixture models were fitted on the imputed transformed concentration and the model with the highest BIC was selected, determining the shape, orientation, and number of clusters. For each sample, the most likely cluster was reported as well as the probability of belonging to each cluster.

Identification of phases of the menstrual cycle

Menstrual cycles were identified from bleeding flows reported by participants on a scale from 0 (none) to 3 (heavy). Specifically, a hidden semi-Markov model was specified to account for empirically observed distribution of cycle length and bleeding patterns across the menstrual cycle (67). Data of participants who reported too few days with bleeding (*i.e.*, less than 3/70 study days) or too many (*i.e.,* more than 30/70 study days) were excluded from the menstrual cycle analyses. Once cycles were identified, cycle days were numbered forward and backward from the first day of the period. In order to align the two major menstrual events, i.e. ovulation and menses, across participant and given that the luteal phase has been well documented to vary less than the luteal phase (33), cycles were standardized starting from day -18 (*i.e.* 18 days before the start of the next cycle) and ending on day +7 (i.e. 7 days after the first day of the menses). This definition ensures that the standardized cycles would include the days leading to ovulation, estimated to happen around days -12 to -14 (33) and allows for the best possible alignment of the two major menstrual events (ovulation and menses) in the absence of hormonal and/or ovulation markers.

Testing for differential abundance throughout the menstrual cycle

To identify metabolites, cytokines or ASVs with differential abundance (metabolites or cytokines) or differential probabilities of being present at specific phases of the menstrual cycle, a linear model (for abundances) or logistic regression (presence/absence probabilities) was fit onto 5 degrees of freedom circular spline. Analysis of deviance was used to report p-values of the F-statistics and corrected for multiple testing using the Benjamini-Hochberg method.

Correlation in VM composition between two consecutive cycles

To evaluate how the menstrual cycle affects the vaginal microbiome, we first identified the most abundant species for each individual with at least two complete consecutive cycles. Then, the Spearman (rank) or Pearson (relative abundance) correlation between the ASV counts of two consecutive standardized cycles were computed.

Predicting changes in the composition of the vaginal microbiota in pregnant women

To identify metabolites or cytokines predictive of changes in the vaginal microbiota of pregnant women, sparse logistic regression (lasso) was used to predict the *Lactobacillus* proportion at the next time point. To increase the robustness of the model, sparse regressions were fitted ten times on sub-samples of the available data (70% of samples were randomly considered for each fit). The coefficients of the 15 first non-zero features (metabolites or cytokines) were identified for each fit. In parallel, a random forest model was fitted similarly so that the most important features identified by the random forest models could be compared to those identified using sparse regression. The glmnet and randomForest R packages were used for these predictions.

Availability of data and materials

The sequence data are available in SRA under BioProject accession numbers PRJNA208535 (samples beginning with UAB) and PRJNA575586 (samples beginning with AYAC and EM) and, for the samples of the pregnant subject, on the NCBI Sequence Read Archive, https://www.ncbi.nlm.nih.gov/sra (accession no. SRP115697). The supplementary material contains the R code that enables the reproduction of the analyses.

ACKNOWLEDGMENTS

The authors thank Anna Robaczewska for generating the amplicon libraries, and Nomfuneko Mafunda, Brooke Spencer, and Leah Froehle for running the cytokine quantification assays.
They also thank Dr K. Sankaran and Dr. J. Fukuyama for discussion on the visualization of hierarchical topic analysis and all other members of the VMRC consortium for fruitful discussions and interactions.

Funding: This work was supported by the Bill and Melinda Gates Foundation grant OPP1189205-2019 awarded to Dr J. Ravel (PI, University of Maryland) and Dr D. Relman (PD, Stanford University). S.M.B. was supported in part by a grant from the Harvard University Center for AIDS Research (CFAR), an NIH funded program (P30 AI060354), which is supported by the following NIH Co-Funding and Participating Institutes and Centers: NIAID, NCI, NICHD, NIDCR, NHLBI, NIDA, NIMH, NIA, NIDDK, NINR, NIMHD, FIC, and OAR. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

1. R. M. Brotman, Vaginal microbiome and sexually transmitted infections: an epidemiologic perspective. *J. Clin. Invest.* **121**, 4610–4617 (2011).

2. J. A. Gilbert, *et al.*, Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).

3. S. J. Kroon, J. Ravel, W. M. Huston, Cervicovaginal microbiota, women's health, and reproductive outcomes. *Fertil. Steril.* **110**, 327–336 (2018).

4. B. J. Callahan, *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.* **114**, 9966–9971 (2017).

5. M. A. Elovitz, *et al.*, Cervicovaginal microbiota and local immune response modulate the risk of spontaneous preterm delivery. *Nat. Commun.* **10**, 1305 (2019).

6. R. B. Ness, *et al.*, Bacterial Vaginosis (BV) and the Risk of Incident Gonococcal or Chlamydial Genital Infection in a Predominantly Black Population. *Sex. Transm. Dis.* **32**, 413–417 (2005).

7. J. E. Allsworth, J. F. Peipert, Severity of bacterial vaginosis and the risk of sexually transmitted infection. *Am. J. Obstet. Gynecol.* **205**, 113.e1-113.e6 (2011).

8. C. van der Veer, S. M. Bruisten, J. J. van der Helm, H. J. C. de Vries, R. van Houdt, The Cervicovaginal Microbiota in Women Notified for *Chlamydia trachomatis* Infection: A Case-Control Study at the Sexually Transmitted Infection Outpatient Clinic in Amsterdam, The Netherlands. *Clin. Infect. Dis.* **64**, 24–31 (2017).

9. J. Tamarelle, *et al.*, Vaginal microbiota composition and association with prevalent *Chlamydia trachomatis* infection: a cross-sectional study of young women attending a STI clinic in France. *Sex. Transm. Infect.* **94**, 616–618 (2018).

10. C. R. Cohen, *et al.*, Bacterial vaginosis and HIV seroprevalence among female commercial sex workers in Chiang Mai, Thailand: *AIDS* **9**, 1093–1098 (1995).

11. C. R. Cohen, *et al.*, Bacterial Vaginosis Associated with Increased Risk of Female-to-Male HIV-1 Transmission: A Prospective Cohort Analysis among African Couples. *PLoS Med.* **9**, e1001251 (2012).

12. C. Gosmann, *et al.*, Lactobacillus-Deficient Cervicovaginal Bacterial Communities Are Associated with Increased HIV Acquisition in Young South African Women. *Immunity* **46**, 29–37 (2017).

13. P. Gajer, *et al.*, Temporal Dynamics of the Human Vaginal Microbiota. 14 (2012).

14. D. B. DiGiulio, *et al.*, Temporal and spatial variation of the human microbiota during pregnancy. *Proc. Natl. Acad. Sci.* **112**, 11060–11065 (2015).

15. A. Munoz, *et al.*, Modeling the temporal dynamics of cervicovaginal microbiota identifies targets that may promote reproductive health. *Microbiome* **9**, 163 (2021).

16. J. Ravel, *et al.*, Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. *Microbiome* **1**, 29 (2013).

17. N. Noyes, K.-C. Cho, J. Ravel, L. J. Forney, Z. Abdo, Associations between sexual habits, menstrual hygiene practices, demographics and the vaginal microbiome as revealed by Bayesian network analysis. *PLOS ONE* **13**, e0191625 (2018).

18. H. Verstraelen, *et al.*, Longitudinal analysis of the vaginal microflora in pregnancy suggests that L. crispatus promotes the stability of the normal vaginal microflora and that L. gasseri and/or L. iners are more conducive to the occurrence of abnormal vaginal microflora. *BMC Microbiol.* **9**, 116 (2009).

19. J. Ravel, *et al.*, Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci.* **108**, 4680–4687 (2011).

20. M. T. France, *et al.*, VALENCIA: a nearest centroid classification method for vaginal microbial communities based on composition. *Microbiome* **8**, 166 (2020).

21. J. K. Pritchard, M. Stephens, P. Donnelly, Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**, 945–959 (2000).

22. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 30 (2013).

23. K. Sankaran, S. P. Holmes, Latent variable modeling for the microbiome. *Biostatistics* **20**, 599–614 (2019).

24. R. Romero, *et al.*, The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome* **2**, 4 (2014).

25. D. A. MacIntyre, *et al.,* The vaginal microbiome during pregnancy and the postpartum period in a European population. *Sci. Rep.* **5**, 8988 (2015).

26. W. F. Kindschuh, *et al.*, Preterm birth is associated with xenobiotics and predicted by the vaginal metabolome (2021) https:/doi.org/10.1101/2021.06.14.448190 (August 12, 2021).

27. V. Jespers, *et al.*, A longitudinal analysis of the vaginal microbiota and vaginal immune mediators in women from sub-Saharan Africa. *Sci. Rep.* **7**, 11974 (2017).

28. F. Bradley, *et al.*, The vaginal microbiome amplifies sex hormone-associated cyclic changes in cervicovaginal inflammation and epithelial barrier disruption. *Am. J. Reprod. Immunol.* **80**, e12863 (2018).

29. M. N. Anahtar, *et al.*, Cervicovaginal Bacteria Are a Major Modulator of Host Inflammatory Responses in the Female Genital Tract. *Immunity* **42**, 965–976 (2015).

30. D. S. A. Goltsman, *et al.*, Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. *Genome Res.* **28**, 1467–1480 (2018).

31. J. Fukuyama, K. Sankaran, L. Symul, Multiscale Analysis of Count Data through Topic Alignment. *ArXiv210905541 Stat* (2021) (September 15, 2021).

32. L. Symul, K. Wac, P. Hillard, M. Salathé, Assessment of menstrual health status and evolution through mobile apps for fertility awareness. *Npj Digit. Med.* **2**, 64 (2019).

33. K. M. Schmalenberger, *et al.*, How to study the menstrual cycle: Practical tools and recommendations. *Psychoneuroendocrinology* **123**, 104895 (2021).

34. P. Gajer, *et al.*, Temporal Dynamics of the Human Vaginal Microbiota. 14.

35. S. Srinivasan, *et al.*, Temporal Variability of Human Vaginal Bacteria and Relationship with Bacterial Vaginosis. *PLoS ONE* **5**, 8 (2010).

36. G. Lopes dos Santos Santiago, *et al.*, Longitudinal Study of the Dynamics of Vaginal Microflora during Two Consecutive Menstrual Cycles. *PLoS ONE* **6**, e28180 (2011).

37. S. Abraham, *et al.*, Menstrual Protection. Young Women’s Knowledge, Practice and Attitudes. *J. Psychosom. Obstet. Gynecol.* **4**, 229–236 (1985).

38. C. M. Mitchell, *et al.*, Colonization of the upper genital tract by vaginal bacterial species in nonpregnant women. *Am. J. Obstet. Gynecol.* **212**, 611.e1-611.e9 (2015).

39. C. Chen, *et al.*, The microbiota continuum along the female reproductive tract and its relation to uterine-related diseases. *Nat. Commun.* **8**, 875 (2017).

40. R. Koedooder, *et al.*, Identification and evaluation of the microbiome in the female and male reproductive tracts. *Hum. Reprod. Update* **25**, 298–325 (2019).

41. I. Fraser, H. Critchley, M. Broder, M. Munro, The FIGO Recommendations on Terminologies and Definitions for Normal and Abnormal Uterine Bleeding. *Semin. Reprod. Med.* **29**, 383–390 (2011).

42. Y. Wang, *et al.*, Kynurenine is an endothelium-derived relaxing factor produced during inflammation. *Nat. Med.* **16**, 279–285 (2010).

43. N. Hrboticky, L. A. Leiter, G. H. Anderson, Menstrual cycle effects on the metabolism of tryptophan loads. *Am. J. Clin. Nutr.* **50**, 46–52 (1989).

44. S. Brien, C. Martin, A. Bonner, Tryptophan Metabolism During the Menstrual Cycle. *Biol. Rhythm Res.* **28**, 391–403 (1997).

45. C. J. Lynch, S. H. Adams, Branched-chain amino acids in metabolic signalling and insulin resistance. *Nat. Rev. Endocrinol.* **10**, 723–736 (2014).

46. C. F. Draper, *et al.*, Menstrual cycle rhythmicity: metabolic patterns in healthy women. *Sci. Rep.* **8**, 14568 (2018).

47. S. Cauci, *et al.*, Interrelationships of interleukin-8 with interleukin-1b and neutrophils in vaginal ¯uid of healthy and bacterial vaginosis positive women. *Mol. Hum. Reprod.* **9**, 53–58 (2003).

48. K. Sturm-Ramirez, A. Gaye-Diallo, G. Eisen, S. Mboup, P. J. Kanki, High Levels of Tumor Necrosis Factor–a and Interleukin-1b in Bacterial Vaginosis May Increase Susceptibility to Human Immunodeficiency Virus. *J. Infect. Dis.* **182**, 476–73 (2000).

49. M. T. France, *et al.*, “Insight into the ecology of vaginal bacteria through integrative analyses of metagenomic and metatranscriptomic data” (Microbiology, 2021) https:/doi.org/10.1101/2021.06.17.448822 (November 17, 2021).

50. V. Jespers, *et al.*, The significance of Lactobacillus crispatus and L. vaginalis for vaginal health and the negative effect of recent sex: a cross-sectional descriptive study across groups of African women. *BMC Infect. Dis.* **15**, 115 (2015).

51. C. M. Liu, *et al.*, Penile Microbiota and Female Partner Bacterial Vaginosis in Rakai, Uganda. *mBio* **6** (2015).

52. S. D. Mehta, *et al.*, The Microbiome Composition of a Man’s Penis Predicts Incident Bacterial Vaginosis in His Female Sex Partner With High Accuracy. *Front. Cell. Infect. Microbiol.* **10**, 433 (2020).

53. S. D. Mehta, *et al.*, Vaginal and Penile Microbiome Associations With Herpes Simplex Virus Type 2 in Women and Their Male Sex Partners. *J. Infect. Dis.*, jiaa529 (2020).

54. W. Walters, *et al.*, Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. *mSystems* **1** (2016).

55. B. J. Callahan, *et al.*, DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

56. Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).

57. C. Quast, *et al.*, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).

58. J. B. Holm, *et al.*, Comparative Metagenome-Assembled Genome Analysis of “Candidatus Lachnocurva vaginae”, Formerly Known as Bacterial Vaginosis-Associated Bacterium−1 (BVAB1). *Front. Cell. Infect. Microbiol.* **10**, 117 (2020).

59. S. Srinivasan, *et al.*, Megasphaera lornae sp. nov., Megasphaera hutchinsoni sp. nov., and Megasphaera vaginalis sp. nov.: novel bacteria isolated from the female genital tract. *Int. J. Syst. Evol. Microbiol.* **71** (2019).

60. S. Srinivasan, *et al.*, Metabolic Signatures of Bacterial Vaginosis. *mBio* **6** (2015).

61. W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, M. Vingron, Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96–S104 (2002).

62. R Core Team, R: A Language and Environment for Statistical Computing.

63. M. Morgan, V. Obenchain, J. Hester, H. Pagès, SummarizedExperiment: SummarizedExperiment container (2020).

64. M. Ramos, *et al.*, Software for the Integration of Multiomics Experiments in Bioconductor. *Cancer Res.* **77**, e39–e42 (2017).

65. B. Grün, K. Hornik, **topicmodels** : An *R* Package for Fitting Topic Models. *J. Stat. Softw.* **40** (2011).

66. L. Scrucca, M. Fop, T. Murphy Brendan, A. Raftery E., mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* **8**, 289 (2016).

67. L. Symul, S. Holmes, Labeling self-tracked menstrual health records with hidden semi-Markov models. *IEEE J. Biomed. Health Inform.* (2021) https:/doi.org/10.1109/JBHI.2021.3110716 (September 15, 2021).