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], Douglas S. Kwon[6,7,8,11], Jacques Ravel[4,5,11], 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:**

Susan Holmes, Department of Statistics, Stanford University, 390 Jane Stanford Way, Stanford, CA 94305, USA

Email: susan@stat.stanford.edu

Authors' 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 highly resolved approach based on mixed membership topic models with multi-domain longitudinal data from cohorts of pregnant and non-pregnant study participants to identify several non-*Lactobacillus*-dominated sub-communities common to both cohorts regardless of reproductive status. These sub-communities correlated with clusters in metabolite concentrations. In non-pregnant individuals, we identified non-*Lactobacillus*-dominated sub-communities that were more common during menses but did not predict an increased likelihood of other 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 and metabolites, such as elevated TNF-α concentrations at the time of ovulation, and elevated kynurenine concentrations during menses, respectively. In pregnant individuals, some metabolite clusters were predictive of temporal instability in vaginal microbiota structure. Overall, our analyses based on mixed membership models revealed new substructures of the vaginal ecosystems which may have potentially important clinical or biological associations.

Main Text

Introduction

Several critical aspects of women's health are linked to the structure of the vaginal microbiota (1–3). Vaginal microbiota dominated by beneficial *Lactobacillus* species are associated with positive health outcomes (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 robustly resistant to invasion or replacement 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 structure 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 structure (13). A cross-sectional study showed that recent sexual intercourse was associated with the presence of specific bacteria such as *Corynebacterium* (17). These associations suggest that changes in female reproductive tract 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 during phases of the menstrual cycle other than menses. It is unclear why the vaginal microbiota of some individuals easily transitions back to an optimal state (dominated by beneficial *Lactobacillus* species) from a non-optimal one, while other microbiota do not (4, 13, 14, 18).

In this study, we pursued three objectives. First, we sought to deepen our understanding of the fine structure of non-optimal vaginal microbiota. 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 potential associations between reproductive status of the host (gestational age or phase of the menstrual cycle) and variation in features of the vaginal ecosystem, defined here as the set of bacteria, 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 microbiota are highly heterogeneous within and between 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 this categorization serves as an important dimensionality reduction tool for these complex datasets, it tends to oversimplify community structure in non-optimal microbiota (often labeled CST IV), and importantly, hides transition states between clusters. For example, the vaginal microbiota of two individuals could switch from one CST to another and be described by the same sequence of CSTs, but one individual might have experienced a progressive shift in her vaginal microbial structure (*e.g.,* the proportion of *Gardnerella* might have been slowly decreasing over time) while the other might have experienced an abrupt change (*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 that frequently co-occur, they are not well suited to identify subsets of species that 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.

Topic models, developed to infer population structures (21) and formally described as Latent Dirichlet Allocation (LDA) (22), have been adopted for analyzing microbiota to identify bacterial sub-communities (23). Microbial samples are modeled as a mixture of topics (sub-communities), and each topic is characterized by a particular distribution of bacterial species or strains. For example, if a sample were described as 70% topic 1 and 30% topic 2 this would mean that the species subsumed in topic 1 accounted for 70% of the sample, while the species in topic 2 accounted 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 microbiota datasets from two longitudinal cohorts of women to identify specific sub-communities and report their temporal dynamics and the transitions between them. The first cohort was composed of pregnant participants who provided weekly samples throughout pregnancy. The second cohort was composed of non-pregnant participants of reproductive age who provided daily samples for ten weeks.

Our second and third objectives, which were to identify ecosystem features associated with reproductive status (*i.e.,* in pregnant and non-pregnant individuals) and predictive of changes in microbiota structure, 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 also quantified metabolite concentration (13, 26) or cytokine levels (27–29), and none collected samples in both pregnant and non-pregnant individuals or accounted for potential menstrual cycle effects on the vaginal environment. In this study, metabolites and cytokines were quantified longitudinally from five samples per participant in each of forty participants from each cohort (40 pregnant and 40 non-pregnant participants). The integration of microbiota structure, metabolite, and cytokine data 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 and cytokine concentrations and stability of vaginal microbiota composition.

Results

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

We analyzed data from 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, Table S1 for demographic data). Topic models were fit to the count data of 16S rRNA amplicon sequence variants (ASVs) in these samples (Fig S1-3 for CST time series). To facilitate linkage with the published literature, “G1/G2/G3” labels were added to the three most abundant *Gardnerella* ASVs as these were previously found to be differentially associated with preterm birth risk (4) and may correspond to specific species of *Gardnerella* (30) (SI, Fig S4).

Topic analysis requires choosing K, the number of topics, for the modeling of the provided count data. K can be estimated using cross-validation or, as recently proposed (30), by performing topic alignment across models with different resolutions (*i.e.,* with different K). This latter approach offers the advantage of providing diagnostic scores to characterize individual topics and evaluate deviations from LDA assumptions. In pregnant participants, the alignment suggested that 10 topics provided the best compromise between dimension reduction and accurate modeling of ASV counts, while in non-pregnant participants, 11 topics were best suited (Methods, SI).

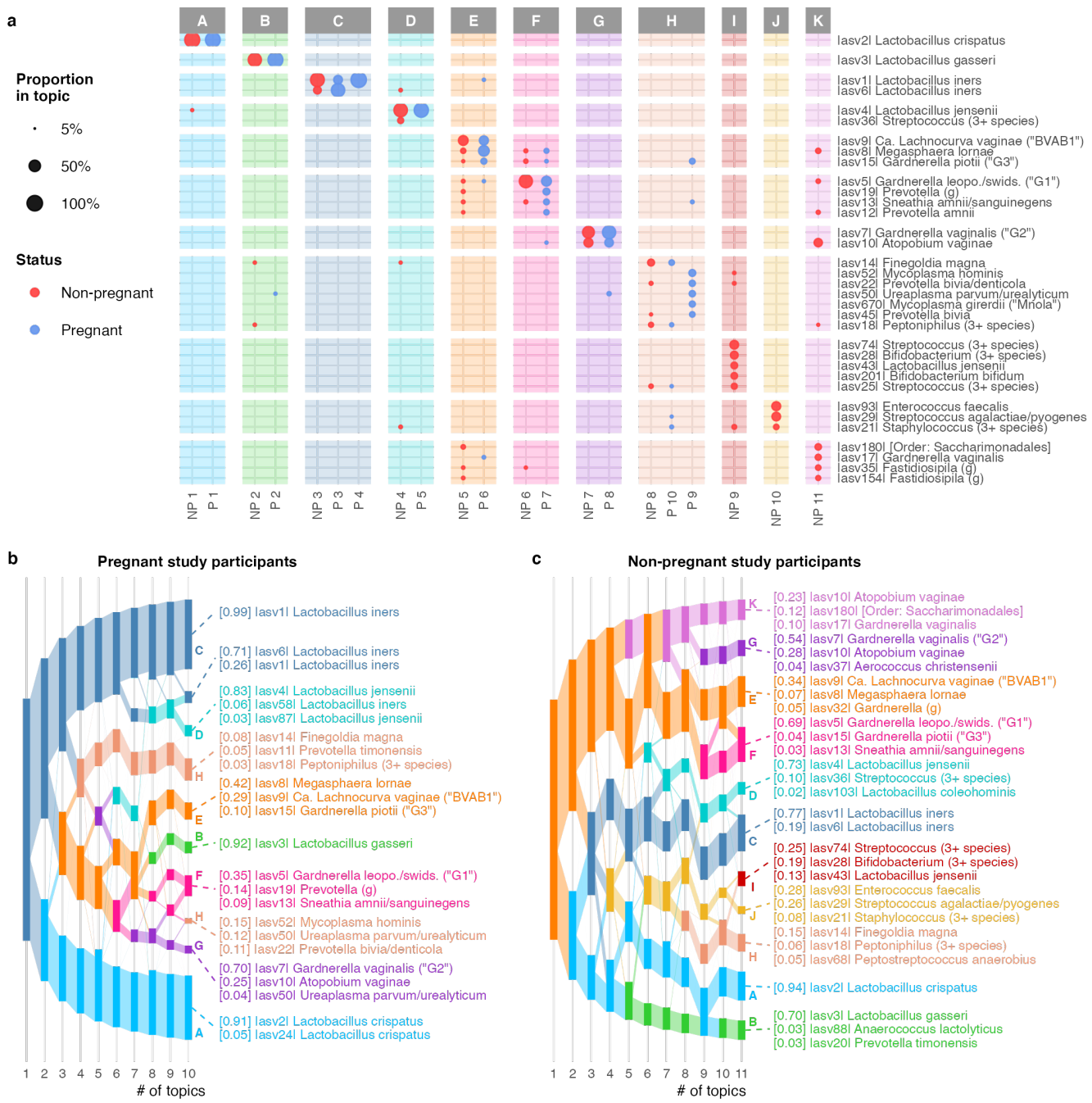

Figure 1. Topic analysis on samples from pregnant and non-pregnant women reveals shared and unique sub-communities. **(a)** Composition of topics identified with samples of pregnant (P) and non-pregnant (NP) participants (topic models fitted on the samples of the two groups separately). Dots size indicates the proportion of a specific ASV in each topic. Dot color indicates whether the topic was identified in samples of pregnant (blue) or non-pregnant (red) participants. To improve the readability of the figure, ASVs were displayed if they accounted for at least 5% of the topic composition. Topics were grouped (A to K) by similarity based on the Jensen-Shannon divergence between topic composition using hierarchical clustering (SI). Topic background color matches topic colors in panels (b) and (c). ASVs (rows) were ordered within topics by prevalence. **(b & c)** Alignment of topics (rectangles) identified separately with samples of pregnant participants **(b)** and non-pregnant participants **(c),** as the number of topics (x-axis) was varied from 1 to 10 or 11, the optimal number of topics for samples of pregnant or non-pregnant participants respectively (SI). The height of the rectangles is scaled according to the total proportion of the corresponding topic in all samples: taller rectangles represent topics more frequently found among all samples than those represented by shorter rectangles. The numbers in brackets in front of each ASV indicate the proportion of that ASV in the topic.

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, SI). 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 together with ASV 5 - *Gardnerella leopoldii /swidzinskii* (“G1”) - but much less so with ASV 7, *Gardnerella vaginalis* (“G2”) which co-occurred with ASV10 (*Atopobium vaginae*) in topic G. These two ASVs (“G2” and *A. vaginae*) constituted a robust topic in both pregnant

and non-pregnant women and showed an ability to dominate (*i.e.,* explain almost 100% of the sample composition), especially in pregnant individuals (Fig S5-6). Topic 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 S7-8).

Topic alignments (Fig 1b-c) constructed from the pregnant and non-pregnant women's samples provided insights into the coherence of the identified topics and relationships between them. Overall, we observed several similarities when comparing the alignment structures in both groups. Of particular interest, two topics at K = 4 were identified as "parents" of two distinct topic sets similar across groups. The first set of topics contained topic E, F, G, and – in non-pregnant participants - topic K, in which *Gardnerella*, *Megasphaera*, *Ca*. Lachnocurva and specific *Prevotella* ASVs had a high prevalence. The second set contained topic H, and, in non-pregnant participants, topics I and J. In this second set, *Finegoldia*, *Streptococcus*, *Enterococcus*, and *Staphylococcus* ASVs were highly prevalent. Topic alignments repeated on technical replicates of selected samples from both cohorts yielded similar results (SI, Fig S9). The distribution of topic proportions per CST highlights that most samples are a mixture of topics (Fig S10-11). Examples of time series displays showing the topic composition of two participants are shown in Fig 2 (similar displays for all participants are available in the SI). These examples also show how metabolites and cytokines concentrations varied over time.

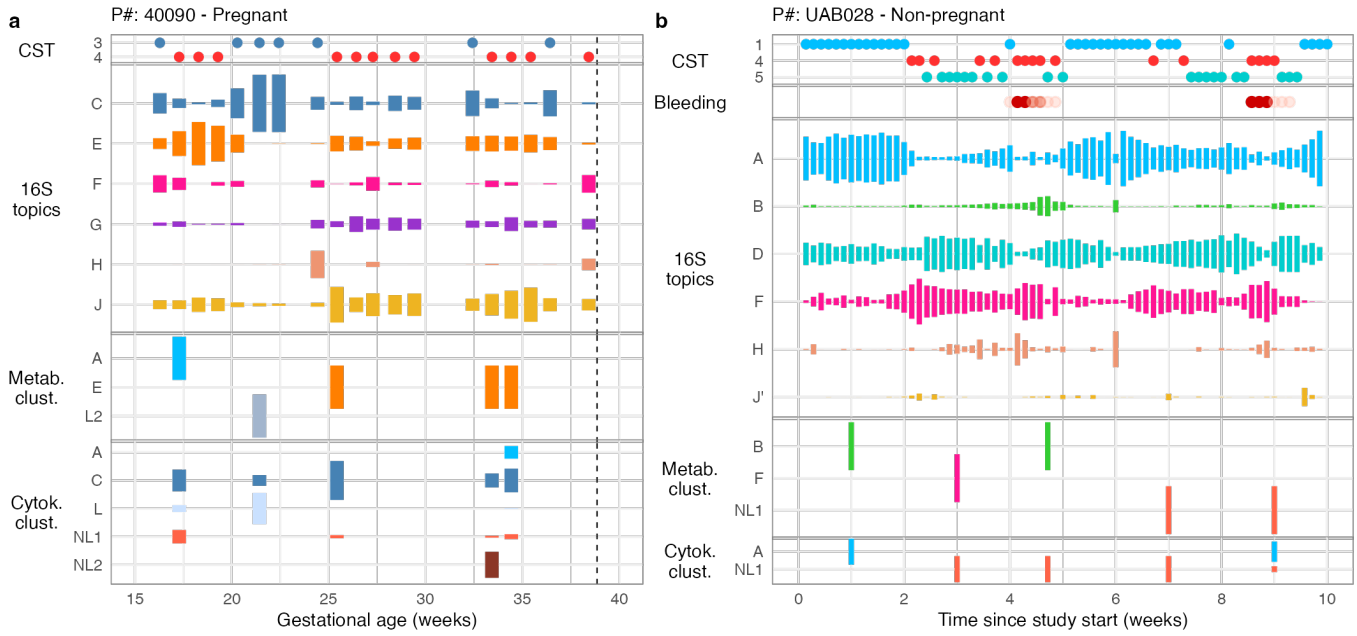

Figure 2. Examples of time-series displays of data from two participants
Time-series displays of data from one pregnant participant (participant 40090, panel a) and one non-pregnant participant (participant UAB028, panel b). From the top, for each sample (x-axis): CSTs as defined in refs (14, 19), reported bleeding (just for the non-pregnant participant), topics based on 16S rRNA data, metabolite clusters, and cytokine clusters. 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 from that participant. For the metabolite or cytokine clusters, the height of bars is proportional to the inverse log10 probability of the sample belonging to that 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 (see Methods for sample selection). Using a mixture model (Methods) to identify clusters in metabolite and cytokine concentrations, we found that metabolite clusters, and less so, cytokine clusters, were closely related to microbial

structure (Fig 3). The UMAP (Uniform Manifold Approximation and Projection) embeddings of cytokine and metabolite concentrations further showed that the menstrual cycle, and in particular menstruation, had a strong effect on both metabolite and cytokine concentrations (Fig 3).

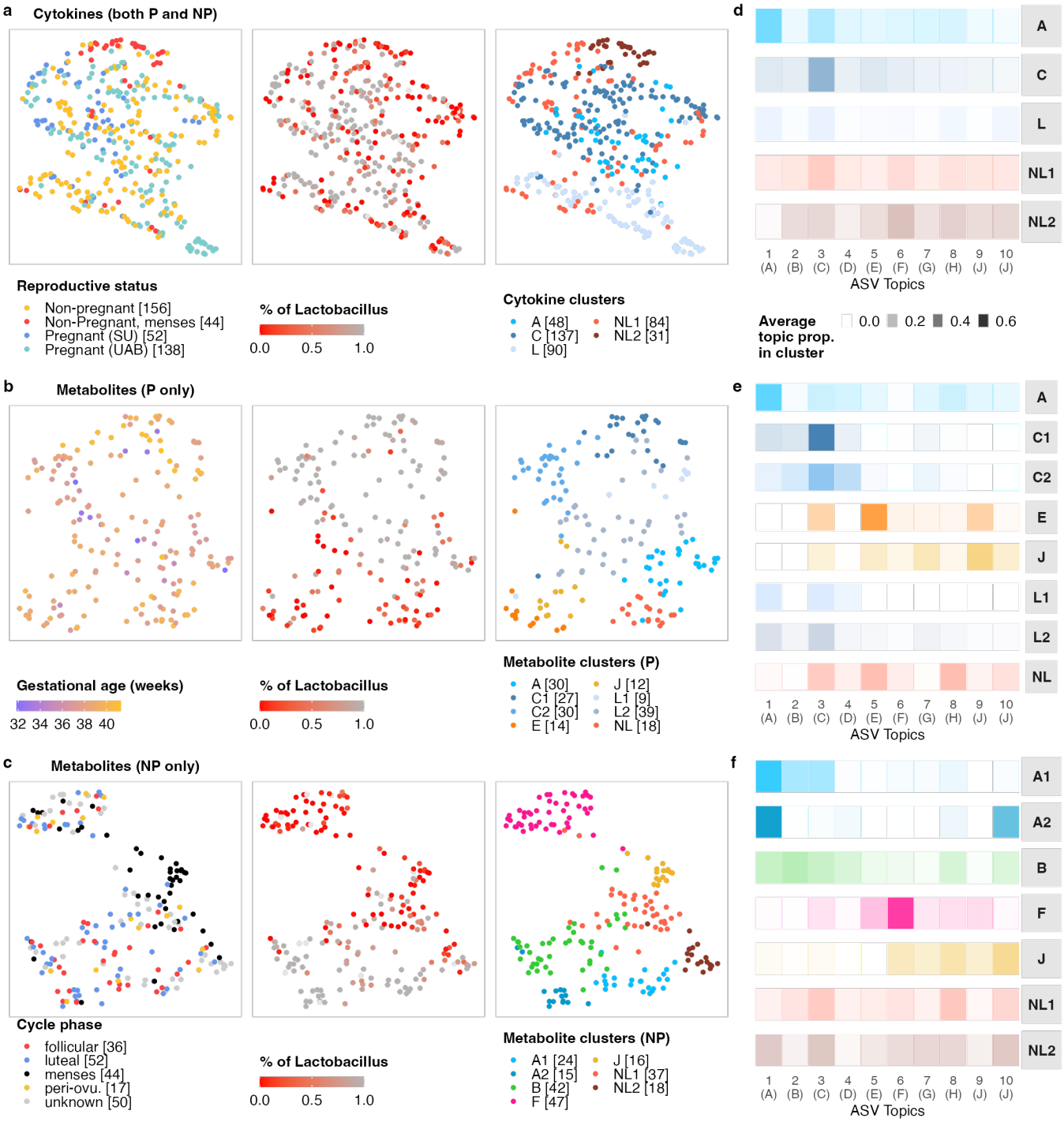

Figure 3. Integration of metabolite and cytokine clusters with microbial sub-communities
(a-c) UMAP (Uniform Manifold Approximation and Projection) of cytokine (a) and metabolite (b-c) concentrations (each dot is a sample). Because samples from pregnant and non-pregnant women were processed slightly differently for metabolite measurements, batch effects prevented a combined analysis of samples from the two groups. For each modality, the same UMAP was reproduced three times with different colors. Samples were colored by reproductive status, gestational age, or cycle phase in the left panels; by the proportion of *Lactobacillus* species in each sample in the center panels; and by their respective clusters on the right. Numbers in brackets indicate the number of samples in each category. **(d-f)** Color intensity indicates the proportion of each 16S rRNA sub-community (topic from analysis on all P and NP samples, displayed horizontally as columns) in each cluster (displayed vertically as rows). Color hues match cluster designations in corresponding panels on the left.

Vaginal concentrations of cytokines were, in general, more strongly associated with the menstrual cycle than with microbial community structure: cytokine clusters were weakly aligned with bacterial sub-communities (topics) (multinomial prediction accuracy of 0.4, Fig 3d, TableS2) and only a few cytokines (IP10, MIG, IL1β, IL8, IL12p70)

were associated with the proportion of non-*Lactobacillus* species in a sample when adjusting for menstrual bleeding (Fig 3a, Fig S12). 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 S13). One topic (H'), not strongly dominated by any 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 S13). Pregnancy status was also not associated with differences in cytokine concentration, except with MIP3α, IL1α, and IL1β, which were mildly elevated in samples collected before 5 months of gestation (Fig S14).

Metabolite concentrations were strongly associated with microbiota structure. Both in pregnant and non-pregnant women, most metabolites (58% in pregnant, 48% in non-pregnant, 34% in both) were associated with the proportion of *Lactobacillus* species in a sample (Fig 3b-c, SI) when adjusting for gestational age (in pregnant participants) or bleeding (in menstruating non-pregnant participants). Further, there was a strong correlation between metabolite clusters and bacterial sub-communities (topics) (multinomial prediction accuracy of 0.49 in pregnant participant samples and 0.56 in non-pregnant participant samples, Fig 3e-f, Table S2), indicating that metabolite concentrations are strongly associated with specific sub-communities of the vaginal microbiota. Because metabolite concentrations were quantified in separate batches for pregnant and non-pregnant individuals, metabolite clusters in these two groups (Fig 3e-f) are not directly comparable (Fig S15). In pregnant individuals, metabolites were associated with gestational age, as hinted by the mild gradients on the UMAP embeddings (earlier gestational age in the center on the left panel of Fig 3b). Nine metabolites were significantly associated with gestational age (4-hydroxyglutamate, hydroxypalmitoyl sphingomyelin (d18:1/16:0(OH)), 3-methylglutarate/2-methylglutarate, 3-hydroxybutyroylglycine, adipoylcarnitine (C6-DC), heptadecasphingosine (d17:1), quinolinate, glycerophosphoinositol*, prolylhydroxyproline, SI, Fig S16). In non-pregnant women, menses was associated with metabolite concentrations, as most samples collected during menstruation were located in the same area of the UMAP embeddings (Fig 3c).

The menstrual cycle shapes the vaginal ecosystem.

Among the 40 non-pregnant participants, 33 had reported vaginal bleeding patterns that allowed for the identification of at least one menstrual cycle within the ten study weeks (Methods). For 24 participants, 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) (31, 32), 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 *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 S17-18, Table S4). Most cytokines had elevated levels during menses (Fig S17). However, a few cytokines (e.g., INF-γ) 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 (Table S5, Fig S19). While the largest changes were often found around the time of menses (Fig S19), 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, Fig S19).

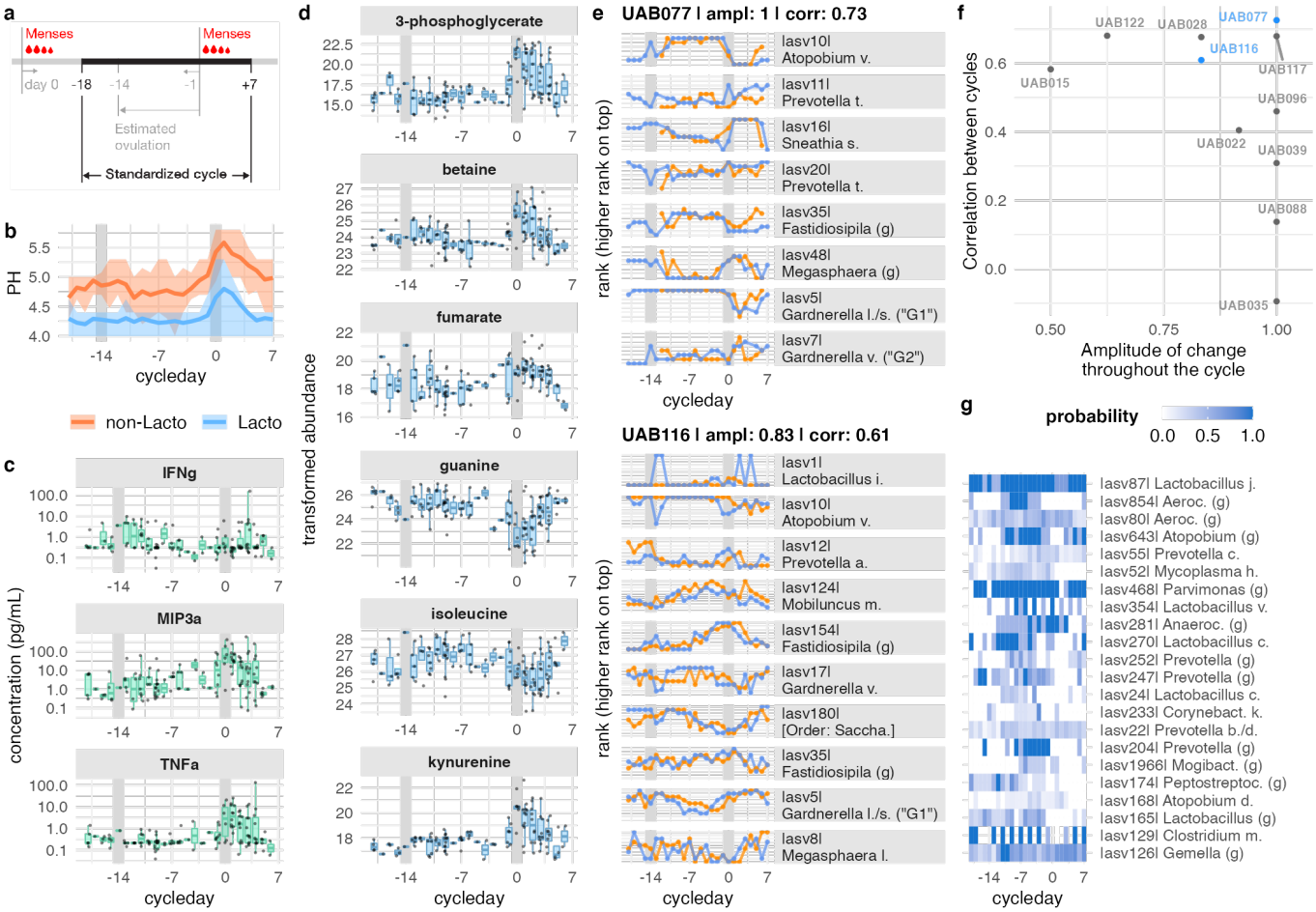

Figure 4. The menstrual cycle shapes the vaginal ecosystem.
(a) Schematic illustrating the features of standardized cycles. **(b)** Vaginal pH throughout the menstrual cycle for *Lactobacillus* or non-*Lactobacillus*-dominated vaginal microbiotas. **(c)** Vaginal cytokine concentrations (pg/mL) throughout the menstrual cycle for three cytokines correlated with the menstrual cycle. Profiles of all measured cytokines are shown in Fig S17-18 **(d)** Vaginal metabolite relative concentrations throughout the menstrual cycle for 6 metabolites showing a strong association with time in the menstrual cycle. Profiles of all metabolites associated with time in the menstrual cycle are shown in Fig S19 **(e)** Vaginal microbiota structure of two participants with data available for at least two full menstrual cycles. The first menstrual cycle is displayed in orange, the second in blue. These two participants 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, showing the relationships between correlation in vaginal microbiota structure between two cycles of same individual (y-axis), and the relative amplitude of change throughout the cycle (x-axis). **(g)** Heatmap of the probabilities that ASVs are present on a given cycle day. ASVs whose probability were 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 shown in Fig S21.

To assess possible associations between vaginal microbiota structure and menstrual cycle, we used a longitudinal approach and compared vaginal microbiota structure between two consecutive cycles for each participant with two menstrual cycles. Vaginal structure, characterized by ASV rank (or relative abundance), showed a high correlation between consecutive cycles (Fig 4e-f, Fig S20). This longitudinal (within-individual) approach was used because the absolute abundance of ASVs cannot be established from our 16S rRNA gene amplicon sequence data, so we could not test for their association with time in the menstrual cycle as we did for metabolites and cytokines. We still used a cross-sectional approach to evaluate the impact of the menstrual cycle on the probability of an ASV being detected on a given cycle day. We found a strong association for 91 ASVs (Fig S21), among which 55 (60%) had a higher probability of detection during menses compared to the rest of the cycle (Fig S22). A total of 17 (20%) ASVs, including four *Prevotella* ASVs and one *Atopobium* ASV, were associated with the late luteal phase (Fig 4g).

We tested whether the prevalence of menses-associated non-*Lactobacillus* species (such as *Finegoldia,* FigS23, Table S6) during the menses was associated with other non-*Lactobacillus* species, such as *Gardnerella* spp., at other phases of the menstrual cycle. We found no association (Fig S24, Table S7-8). Participants whose pre-menses samples were dominated by *Lactobacillus* species were as likely to see an increase in non-*Lactobacillus* species during menses (SI). Similarly, the presence of these menses-specific species (or topics H and J, Fig S24, Table S8) was not associated with a higher risk of non-*Lactobacillus* species following menses.

Because many participants used a combination of pads and tampons during their menses (Table S9), we did not find significant associations between topics and use of menstrual products (Fig S25). Our analysis also confirmed that behaviors suspected of disrupting the vaginal microbiota, such as sexual intercourse or the use of vaginal gels or powder, were associated with an increased risk of change in the vaginal microbiota. Indeed, we observed more changes in microbiota structure among participants who more frequently reported these behaviors (Fig S26). However, this study did not have the statistical power to identify which behaviors were most disruptive.

Metabolite clusters are associated with changes in vaginal microbiota structure.

Intrigued by the fact that samples with different metabolite structures (A, C1, C2, L1, L2, fig 3e) shared a similar topic (and ASV) composition, we investigated whether metabolite clusters were indicative of microbial community stability. We conducted this analysis using the pregnant participants' samples, excluding the samples from the non-pregnant participants to avoid confounding with menstrual cycle effects. We compared microbial community stability in each metabolite cluster identified in pregnant women's samples (Fig 5). Specifically, for each sample, we computed the percentage of the microbial composition that was different between that sample and the sample at the next or previous time-point. We observed that samples assigned to the A and L2 metabolite clusters were more likely to be associated with subsequent changes in topic composition than those assigned to the C1, C2, and L1 metabolite clusters (Fig 5). In particular, we observed that samples fully dominated by *L. crispatus* were more likely to be dominated by *L. crispatus* in the previous or next sample if their metabolic state was best represented by C1, C2, or L1 metabolite clusters than by the A or L2 clusters.

In addition, we fitted sparse regressions on the pregnant participants' samples to predict the relative abundance (proportion) of *Lactobacillus* species or the proportion of specific ASV topics at the next time-point (SI). We found that metabolite concentrations were better predictors than cytokine concentrations (Fig S27). 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 S28).

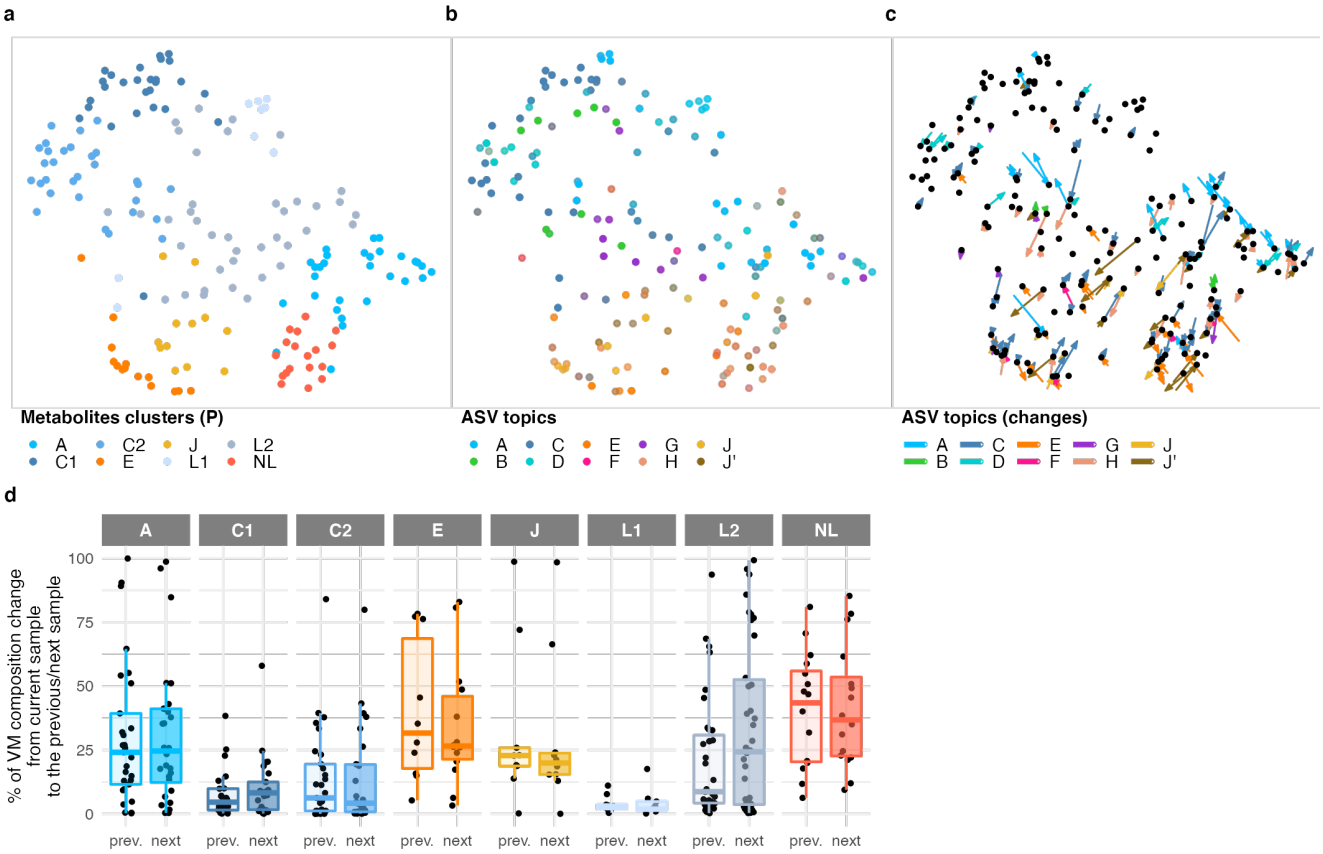

Figure 5. Metabolite clusters are associated with instability in the vaginal microbiota.
(a-c) UMAP of the metabolite concentrations in samples (dots) from pregnant women (same as in Fig 3b). **(a)** Same as in Fig 3b right-hand display. **(b)** UMAP of metabolite concentrations where samples are colored by 16S rRNA topic proportion. When samples are composed of several topics, dot colors are a mix of each individual topic colors. **(c)** Arrows show changes in topic proportion from the previous or the next time-point, which are collected about a week apart. 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. 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 (VM) composition from one sample to the next sample (right boxes in each panel) or from the previous sample to the current sample (left boxes in each panel) for each metabolite cluster (horizontal panels).

Discussion

In this study, we identified bacterial sub-communities within non-*Lactobacillus*-dominated samples from both pregnant and non-pregnant women. Using topic analysis, we found several non-*Lactobacillus* sub-communities (*i.e.,* topics) to be shared by the two groups, refining the structure of samples traditionally assigned to community state type (CST) IV (19). Within these sub-communities, we identified several pairs of co-existing bacterial ASVs to be typical of some sub-communities. For example, *Gardnerella* ASVs (G1 and G3) co-occurred more frequently than other *Gardnerella* ASVs (e.g., G2 and G3), and the most common *Atopobium* ASV defined a dominant, robust sub-community with G2 in 42% of the participants for which *Atopobium* was present. It remains to be seen if these bacteria are interacting with each other and if specific conditions, such as hormonal milieu, are favorable to their joint emergence.

The main difference between the approach used here (topic analysis) and clustering approaches traditionally used to identify sub-groups in the vaginal microbiota lies in the mixed membership nature of topic models, thereby allowing samples to be associated with multiple topics in different proportions. This property offers the advantage of revealing longitudinal transitions between sub-communities and the rate at which they occur, which is not possible with clustering approaches. In addition, topic models allow for synonyms, which may reflect potential functional equivalences in a microbial community context. Indeed, if two species are found interchangeably (but not simultaneously) with a specific combination of other species, these two species will be found in the same topic. In contrast, clustering approaches will tend to create two clusters, one containing each of these two species, artificially increasing the number of sub-communities. In fact, the comparison of the sub-communities identified here with the CST IV sub-types more recently introduced (20) suggests that topics I and J encapsulate the five CST IV-C subtypes.

Topic models used in this study are unsupervised methods. Like clustering, topic models identify dataset-specific features. This means that sub-communities identified on samples from a different cohort may differ from those identified in this study. However, since most sub-communities revealed by our analysis were shared across three distinct cohorts, encompassing both pregnant and non-pregnant individuals, we expect these sub-communities to be reproducibly observed in other populations. Other researchers applying topic models to samples from new cohorts can compare the sub-communities identified in these new cohorts against those identified here using the method we proposed for comparing sub-communities across reproductive status (Figure 1b).

Our study also quantified vaginal metabolite and cytokine concentrations in pregnant and non-pregnant individuals. When integrating data across domains (microbiota composition, metabolite concentrations, and cytokine concentrations), we found that clusters fitted on metabolite concentrations correlated well with microbiota sub-communities. There was a weaker association between cytokine concentrations and microbiota structure. In non-pregnant participants, samples collected during menses showed a distinct profile across domains.

Almost all non-pregnant participants with data available for two menstrual cycles showed a high between-cycle correlation in their vaginal microbiota variation. However, some species reached their maximal relative abundance at different menstrual cycle phases in different 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 participant might have a 10-day luteal phase while another one might have a 14-day luteal phase); (ii) inter-individual differences in hormone levels (or the rates of change in these levels); or (iii) the

set of species present in each individual 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 hormones and methods to estimate the timing of ovulation. Similarly, additional data would be necessary to understand if the substantial changes in vaginal microbiota composition during menses are driven by abrupt hormonal changes, the presence of blood, or the use of menstrual protections such as pads or tampons.

Over 50% of measured metabolites showed an association with the proportion of *Lactobacillus* and about 45 metabolites, such as kynurenine or isoleucine, showed an association with a specific phase of the menstrual cycle. 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 (33). The elevated levels of kynurenine during menses found in our study are thus consistent with these roles and with past studies showing varying levels of kynurenine in serum and urine through the cycle (34, 35). In our vaginal samples, isoleucine, a branched-chain amino acid with important metabolic functions (36), was found with the highest levels in the luteal phase and lowest during menses. Interestingly, serum levels of isoleucine show opposite trends (37).

Cytokines were primarily associated with the menstrual cycle (half of them had a significant association) but not with the stage of pregnancy. Previous studies in non-pregnant individuals found similar variations in cytokine levels throughout the menstrual cycle (27, 28). Regarding the associations between cytokine concentrations and vaginal microbiota structure, we found that most cytokines were not associated with the proportion of *Lactobacillus* or with microbiota topic composition. Only a few cytokines showed associations with the proportion of specific topics. IP-10 was positively associated with topic C (dominated by *L. iners*), which is consistent with previous findings (27), and negatively associated with topics E, F, and H (dominated by non-*Lactobacillus* ASVs). IL-1β also showed a strong negative association with the abundance of *L. crispatus* species in our dataset and was also previously found elevated during and after BV events (27, 38, 39). Our results showed limited overlap with those reported in a study of a large cohort of young South African women (12), but these differences could be explained by the fact that about half of the participants in that study were using an injectable progestin contraceptive and samples were not collected during menses, eliminating the possibility of adjustments for timing during the menstrual cycle.

When investigating temporal changes in the vaginal microbiota of pregnant participants, we found two metabolite clusters associated with increased instability in microbial composition. We also used supervised learning to identify specific metabolites or cytokines associated with the composition of the vaginal microbiome at the next time point and found that metabolites were better predictors of changes than cytokines. These results suggest that metabolic state is associated with both microbial composition and its stability. This highlights the important role of metabolomics for vaginal health as specific metabolites have also been found predictive of spontaneous preterm birth (26, 40).

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 risk-associated community state type IV (CST IV) and a high-resolution view of transitions between communities. We found that metabolite concentrations were strongly correlated with the composition 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 effects on the vaginal epithelium. Finally, specific metabolic states and a few cytokines were found to be associated with changes in microbial composition. Confirmation of these associations might facilitate 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 participants. For this study, the samples from two cohorts with identical recruitment criteria and sample collection methods were used. The samples from 10 participants from a larger multi-site cohort recruited at Emory (EM) and the University of Maryland Baltimore (AYAC) were used along with samples from 30 participants recruited at the University of Alabama, Birmingham (UAB) as part of the UMB-HMP study (16). These 40 participants were selected to represent women with stable *Lactobacillus* dominated microbiota, stable non-*Lactobacillus* dominated microbiota, and unstable microbiota (i.e., with samples dominated by *Lactobacillus* and others dominated by non-*Lactobacillus*). 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 individuals were recruited at Stanford University (SU) and 96 pregnant individuals were recruited at the University of Alabama, Birmingham (UAB). Pregnant participants from both cohorts were enrolled from the fourth month of their pregnancy (earliest enrollment at week 8, latest at week 22) and vaginal swabs were collected weekly (approximately) until delivery. There was an average of 16 samples per participant, and 2179 samples in total. The distributions of age, BMI, and race were significantly different between the two cohorts (Table S1). Participants recruited at UAB were part of a pool of individuals for which intramuscular injections of progesterone (17-OHPC) were indicated or recommended. UAB participants received that treatment throughout pregnancy. The treatment is intended to reduce the risk of preterm birth in pregnant women with a singleton pregnancy and who have a history of singleton spontaneous preterm birth.

Selection of the "VMRC samples". All samples described above were used for 16S rRNA gene sequencing. Metabolites and cytokine concentrations were quantified on a subset of these samples. Specifically, 5 samples from 40 pregnant and each of 40 non-pregnant women were selected and labeled "VMRC samples". The 40 pregnant women were selected from 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. Five samples were then selected so that they were equally distributed in time throughout pregnancy. The five samples from the 40 non-pregnant women were similarly selected so that samples were separated by approximately 2 weeks.

Ethics

All participants provided written informed consent. Ethical approval was obtained from the Institutional Review Boards of Stanford University (IRB protocol no. 21956), the University of Alabama (protocol no. X121031002), Birmingham, Emory University and the University of Maryland Baltimore. 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 V1-V3 regions of the 16S rRNA gene were amplified and then sequenced using the 454 sequencing platform, as described in (16).

Daily samples from the 30 non-pregnant participants recruited at UAB (1534 samples). The V3-V4 regions of the 16S rRNA gene were amplified and then sequenced with the Illumina HiSeq/MiSeq platforms.

Weekly samples from pregnant participants of both cohorts (SU and UAB) (2179 samples): Raw sequence data from samples from pregnant participants 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 (42) were used to amplify the V4 variable region of the 16S rRNA gene from each sample. Pooled amplicons were sequenced on the Illumina HiSeq/MiSeq platforms 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 (technical replicas) using the same protocol as for the pregnant participants' samples.

Demultiplexed raw sequence data from Illumina HiSeq/MiSeq (samples of pregnant participants, 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) (43).

Taxonomic assignment. Automated taxonomic calls were made using DADA2's implementation of the RDP naive Bayesian classifier (44) and a Silva reference database (version 132) (45). The assignment of sequences of the most abundant ASVs were refined and standardized by using BLAST and 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 (46, 47). *Gardnerella* ASVs were tagged as G1, G2, or G3 sensu (4) based on exact matching of the ASV sequences. Tables with the taxonomic assignments are available (see data availability section).

ASV harmonization across sequencing assay. While most analyses were performed with the data from each 16S rRNA sequencing assay separately, we were also interested in fitting the topic model with data from the combined assays and in comparing analysis results across assays. Despite differences in the sequenced region of the 16S rRNA gene among different assays short regions were shared among the assays. ASV sequences were thus matched across assays. We used exact matches (*i.e.,* a shorter sequence had to be fully and exactly included in 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. Figure S29 shows the number of shared vs unique ASVs across the assays.

Phylotype level data for all samples of non-pregnant participants. To harmonize the data from all non-pregnant participants, the V3-V4 (HiSeq/MiSeq) and V1-V3 (454) 16S rRNA sequence count data were aggregated at the species level (or at a higher taxonomic level when taxonomic data were not available at the species level), resulting in a table of dimension 225 species x 2281 samples (approximately 40 participants x 70 samples). This table was used to correlate metabolite and cytokines concentrations with the relative abundance of *Lactobacillus* species. Topic analysis was also performed on these data and presented in the SI.

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 (48). 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 individuals after a 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 (49). 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. Despite this, measurements for most of the remaining 390 (P) and 336 (NP) metabolites were still missing in at least one sample. Metabolites might be missing because their abundance was lower than the detection limits or because the overall quality of a sample was lower. (The proportion of missing metabolites varied 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 was imputed as the weighted mean of the values of the k=5 closest samples for which this metabolite was quantified. The weight (weighted mean) was 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 performed in the R software environment (50). 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 (51), then combined into a single S4 object using the MultiAssayExperiment bioconductor package (52).

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” (53). 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 (30). 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 (54). Specifically, Gaussian mixture models were fitted on the imputed transformed (see above) concentrations 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 daily by participants on a scale from 0 (none) to 3 (heavy). A hidden semi-Markov model was specified to account for empirically observed distributions of cycle length and bleeding patterns across

the menstrual cycle, including spotting between menses (55). 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 (see Fig S30), 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 participants and given that the luteal phase has been well documented to vary less than the luteal phase (32), 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 (32) 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 fitted to circular splines parametrized with 5 degrees of freedom (R package "pbs"). 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 vaginal microbiota 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 individuals, 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 for samples from non-pregnant study participants are available in the NCBI Sequence Read Archive (SRA) under BioProject accession numbers PRJNA208535 (samples beginning with UAB) and PRJNA575586 (samples beginning with AYAC and EM). Sequence data from samples from pregnant study participants are available on the SRA (accession no. PRJNA393472). The raw data and R code enabling the reproduction of the analyses is available at . https://purl.stanford.edu/gp215vr4425. The code is also provided in the SI.

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 advice about 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). D.A.R. is supported by the Thomas M. and Joan C. Merigan Endowment at 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. J. Fukuyama, K. Sankaran, L. Symul, Multiscale Analysis of Count Data through Topic Alignment. *ArXiv210905541 Stat* (2021) (September 15, 2021).

31. 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).

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

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

34. 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).

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

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

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

38. 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).

39. 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).

40. P. Pruski, *et al.*, Direct on-swab metabolic profiling of vaginal microbiome host interactions during pregnancy and preterm birth. *Nat. Commun.* **12**, 5967 (2021).

41. 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).

42. 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).

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

44. 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).

45. 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).

46. 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).

47. 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).

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

49. 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).

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

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

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

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

54. 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).

55. 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).