Supplementary information (methods and results) for Vaginal Microbiome Diversity and Preterm Birth: Results of a Nested Case Control Study in Peru

Freida BLOSTEIN MPH, Bizu GELAYE PhD, Sixto E. SANCHEZ MD, MPH, Michelle A. WILLIAMS PhD, Betsy FOXMAN PhD

Supplementary Methods

Reads were in a mixed orientation file, therefore the reads were separated into forward and reverse reads using qiime 1's extract_barcodes.py script. Reads were then demultiplexed using the idemp program. After demultiplexing, adapters were trimmed using DADA2's filterandTrim commands "trimLeft" option. Runs were thereafter processed using DADA2. After examining quality plots for forward and reverse reads, reads were trimmed at the 240th and 244th nucleotide position respectively, or at the first instance of a quality score <= to 11. The following parameters were used for filtering reads: a maximum of 2 expected errors for both forward and reverse reads and no ambiguous bases.

Of a total 8811003 reads, 490020 were lost to filtering and trimming (5.56% lost).

Errors were learned (independently for each run) using 1 million bases. The forward and reverse reads were denoised and then forward and reverse reads were merged together. Post denoising and merging, 8190868 reads remained (1.56% lost).

At this point, the runs were merged together, and chimeras were removed using DADA2s removeBimeraDenovo commmand and the consensus method. 60.37% of the ASVs were determined to be chimeric; removing chimeras resulted in a loss of 4.85% of reads.

Subsequently taxonomic names were assigned to the genus level using the Silva v132 database. Species names were assigned using DADA2s addSpecies command and the Silva v132 database. Although species level assignment is inexact using the 16S rRNA gene, we felt that species level assignment was informative for vaginal communities, where different species within the same genus (i.e. *Lactobacillus*) have different epidemiological significance.

After initial processing using DADA2 the average read count per sample was 56887. A histogram of the read counts in samples is presented in Figure 1.

Technical replicates were examined visually for differences. Overall, the duplicates closely matched each other in terms of the distribution of ASVs' relative abundance, except for one sample which was submitted in triplicate. For this sample, two of three replicates closely matched while the third did not. The triplicate which did not match the other two was therefore excluded from analysis, and the other two replicates kept in. All other replicates were similar to each other and counts from replicates were summed (Figure 2).

After removing the two mock communities and summing the duplicates, 125 samples remained (1 sample per individual in the study). No samples had less than 1000 reads, so all 125 samples were kept for downstream analysis.

Next we filtered out taxa. DADA2 reported an original 1453 ASVs. We removed any taxa that were not bacterial or that were identified as chloroplasts or mitochondria (129 of 1453 ASVs). At this point we collapsed all ASVs named to the same species into a single ASV - if an ASV was not identified to the species level, it remained classified as it's own ASV. This resulted in a loss of 71 ASVs. For the diversity analysis, this was the extent of filtering that we did.

For analyses involving the Dirichilet multinomial modeling and individual taxa testing, we performed two additional filters.

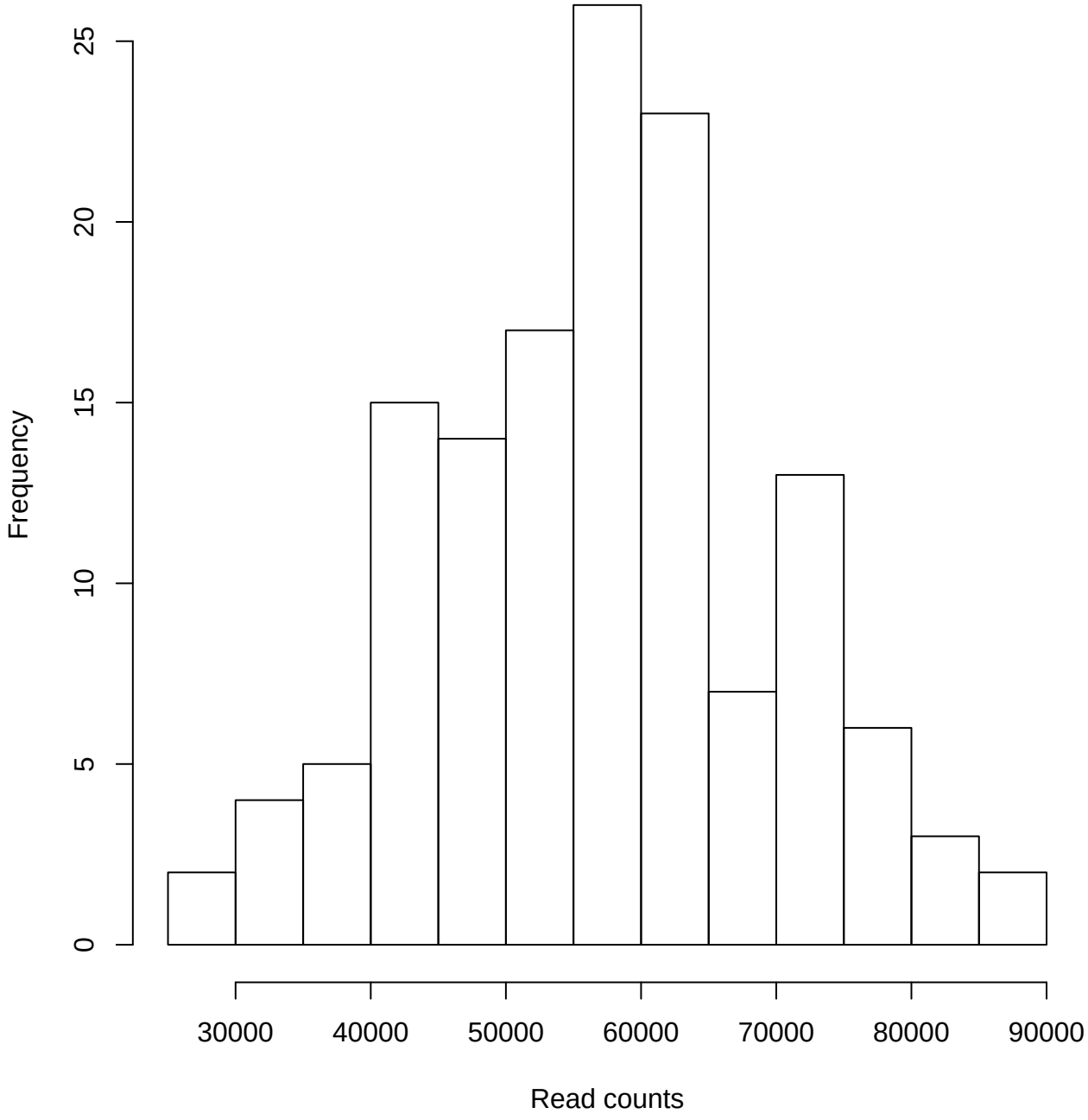

Figure 1: Histogram of read counts in all samples (including mocks and duplicates)

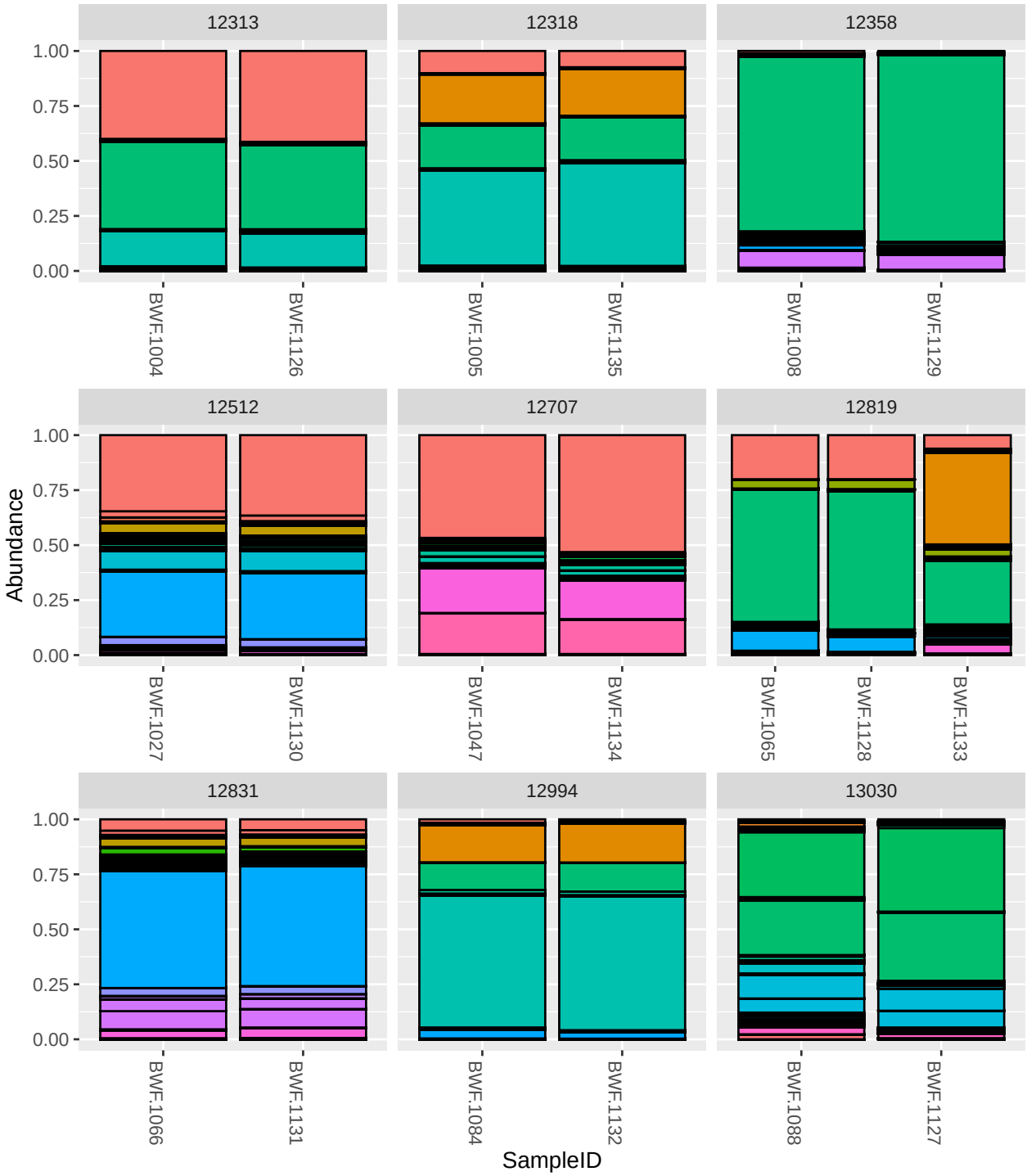

Figure 2: Visual examination of duplicates

We removed phyla that occured in less than or equal to 1 sample on average. This removed the phyla Acidobacteria, Armatimonadetes, Chloroflexi, Cyanobacteria, Gemmatimonadetes, Planctomycetes, Verrucomicrobia, resulting in a loss of 45 ASVs.

Some studies will filter ASVs that occur in a low percentage of the total samples. However, in our study, some ASVs occured in a single sample but composed a substantial portion of the reads in that sample. Therefore, we instead filtered ASVs which represented less than 00.5% of the sample's relative abundance in every sample. This resulted in filtering an additional 1007 ASVs, for a total of 201 ASVs for the Dirichilet multinomial modeling and testing of individual taxa in Aldex2.

Supplementary Results

Supplementary Table 1 compares the sample and swab characteristics by gestational age at vaginal swab. Supplementary Table 2 compares sample and swab characteristics of women who go on to have preterm vs term births, stratified by gestational age at swab (swabbed before 12 weeks gestation vs swabbed at or after 12 weeks gestation, up to 16 weeks gestation).

Table 1: Summary descriptives table by groups of 'Gestational age at swab'

	$<$ 12 weeks N=69	$\geq$ 12 to 16 weeks N=55	p.overall
Preterm status:			0.791
No	54 (78.3%)	45 (81.8%)	
Yes	15 (21.7%)	10 (18.2%)	
Community state type:			0.221
Diverse	30 (43.5%)	31 (56.4%)	
*Lactobacillus ASV2 dominated**	23 (33.3%)	11 (20.0%)	
L. iners dominated	16 (23.2%)	13 (23.6%)	
Shannon	1.22 (0.69)	1.40 (0.74)	0.160
Maternal age	29.6 (6.22)	26.9 (6.39)	0.019
Maternal age (categories):			0.169
18 to 19	2 (2.90%)	4 (7.27%)	
20 to 29	34 (49.3%)	35 (63.6%)	
30 to 34	17 (24.6%)	8 (14.5%)	
35 and older	16 (23.2%)	8 (14.5%)	
Education:			0.788
>12th grade	28 (40.6%)	25 (45.5%)	
7th to 12th grade	37 (53.6%)	28 (50.9%)	
$\leq$ 6th grade	4 (5.80%)	2 (3.64%)	
Mestizo:			0.734
No	18 (26.1%)	12 (21.8%)	
Yes	51 (73.9%)	43 (78.2%)	
Married:			0.667
No	13 (18.8%)	13 (23.6%)	
Yes	56 (81.2%)	42 (76.4%)	
Employment:			0.747
No	27 (39.1%)	24 (43.6%)	
Yes	42 (60.9%)	31 (56.4%)	
Trouble paying for basics:			0.377
No	31 (44.9%)	30 (54.5%)	
Yes	38 (55.1%)	25 (45.5%)	
gadelivery	38.1 (2.31)	38.0 (2.33)	0.803
Planned pregnancy:			0.147
No	32 (47.1%)	34 (61.8%)	
Yes	36 (52.9%)	21 (38.2%)	
Early pregnancy BMI:			0.202
<18.5	1 (1.45%)	1 (1.82%)	
18.5-24.9	27 (39.1%)	31 (56.4%)	
25-29.9	24 (34.8%)	16 (29.1%)	
$\geq$ 30	16 (23.2%)	6 (10.9%)	
Missing	1 (1.45%)	1 (1.82%)	
Nulliparous:			0.212
No	36 (52.9%)	22 (40.0%)	
Yes	32 (47.1%)	33 (60.0%)	
Bacterial vaginosis (Hay-Ison criteria):			0.660
I	10 (14.5%)	10 (18.2%)	
Missing	2 (2.90%)	1 (1.82%)	
N	42 (60.9%)	28 (50.9%)	
VB	15 (21.7%)	16 (29.1%)	

* The second CST's dominating organism – labeled ASV2 - was identified either *L. acidophilus* or *L. crispatus* by a BLAST search.

Table 2: Summary descriptive tables by preterm birth stratified by gestational age

	< 12 weeks			≥ 12 to 16 weeks		
	No N=54	Yes N=15	p.overall	No N=45	Yes N=10	p.overall
Community state type:			0.365			0.094
Diverse	21 (38.9%)	9 (60.0%)		28 (62.2%)	3 (30.0%)	
*Lactobacillus ASV2 dominated**	19 (35.2%)	4 (26.7%)		7 (15.6%)	4 (40.0%)	
L. iners dominated	14 (25.9%)	2 (13.3%)		10 (22.2%)	3 (30.0%)	
Gestational age at swab:			.			.
< 12 weeks	54 (100%)	15 (100%)		0 (0.00%)	0 (0.00%)	
≥ 12 to 16 weeks	0 (0.00%)	0 (0.00%)		45 (100%)	10 (100%)	
Shannon	1.20 (0.69)	1.28 (0.73)	0.707	1.49 (0.72)	1.01 (0.72)	0.076
Maternal age	29.4 (5.87)	30.3 (7.55)	0.653	26.2 (6.03)	29.8 (7.50)	0.184
Maternal age (categories):			0.583			0.687
18 to 19	1 (1.85%)	1 (6.67%)		4 (8.89%)	0 (0.00%)	
20 to 29	28 (51.9%)	6 (40.0%)		29 (64.4%)	6 (60.0%)	
30 to 34	13 (24.1%)	4 (26.7%)		6 (13.3%)	2 (20.0%)	
35 and older	12 (22.2%)	4 (26.7%)		6 (13.3%)	2 (20.0%)	
Education:			0.562			0.018
>12th grade	23 (42.6%)	5 (33.3%)		24 (53.3%)	1 (10.0%)	
7th to 12th grade	27 (50.0%)	10 (66.7%)		20 (44.4%)	8 (80.0%)	
≤ 6th grade	4 (7.41%)	0 (0.00%)		1 (2.22%)	1 (10.0%)	
Mestizo:			1.000			0.096
No	14 (25.9%)	4 (26.7%)		12 (26.7%)	0 (0.00%)	
Yes	40 (74.1%)	11 (73.3%)		33 (73.3%)	10 (100%)	
Married:			0.270			1.000
No	12 (22.2%)	1 (6.67%)		11 (24.4%)	2 (20.0%)	
Yes	42 (77.8%)	14 (93.3%)		34 (75.6%)	8 (80.0%)	
Employment:			0.330			0.486
No	19 (35.2%)	8 (53.3%)		21 (46.7%)	3 (30.0%)	
Yes	35 (64.8%)	7 (46.7%)		24 (53.3%)	7 (70.0%)	
Trouble paying for basics:			0.655			0.158
No	23 (42.6%)	8 (53.3%)		27 (60.0%)	3 (30.0%)	
Yes	31 (57.4%)	7 (46.7%)		18 (40.0%)	7 (70.0%)	
Planned pregnancy:			0.398			0.725
No	23 (43.4%)	9 (60.0%)		27 (60.0%)	7 (70.0%)	
Yes	30 (56.6%)	6 (40.0%)		18 (40.0%)	3 (30.0%)	
Early pregnancy BMI:			0.320			0.066
<18.5	0 (0.00%)	1 (6.67%)		0 (0.00%)	1 (10.0%)	
18.5-24.9	22 (40.7%)	5 (33.3%)		25 (55.6%)	6 (60.0%)	
25-29.9	20 (37.0%)	4 (26.7%)		14 (31.1%)	2 (20.0%)	
≥ 30	11 (20.4%)	5 (33.3%)		6 (13.3%)	0 (0.00%)	
Missing	1 (1.85%)	0 (0.00%)		0 (0.00%)	1 (10.0%)	
Nulliparous:			0.796			0.498
No	29 (54.7%)	7 (46.7%)		17 (37.8%)	5 (50.0%)	
Yes	24 (45.3%)	8 (53.3%)		28 (62.2%)	5 (50.0%)	
Bacterial vaginosis (Hay-Ison):			0.421			1.000
I	6 (11.1%)	4 (26.7%)		8 (17.8%)	2 (20.0%)	
Missing	2 (3.70%)	0 (0.00%)		1 (2.22%)	0 (0.00%)	
N	33 (61.1%)	9 (60.0%)		23 (51.1%)	5 (50.0%)	
VB	13 (24.1%)	2 (13.3%)		13 (28.9%)	3 (30.0%)	

* The second CST's dominating organism (labeled ASV2) was identified as either *L. acidophilus* or *L. crispatus* by a BLAST search

Figure 3 displays the same information as in Figure 1 of the main results section limited to only Mestizo women to examine the possibility that the observed differences in mean Shannon diversity by term status in those sampled at or after 12 weeks' gestation is due to the uneven distribution of ethnicities in the samples collected at or after 12 weeks' gestation. The same pattern holds when restricted to Mestizo participants only (i.e. lower Shannon diversity among women with preterm births than term births among those sampled at or after 12 week's gestation.)

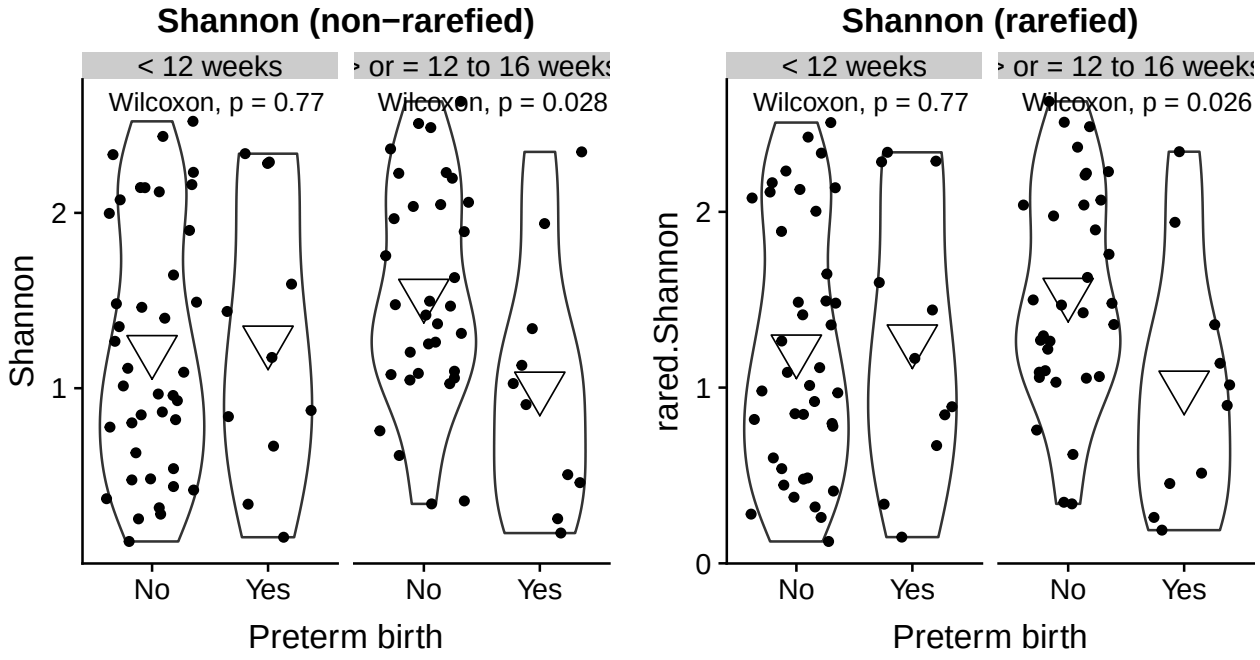

Figure 3: Alpha diversity violins by preterm birth status and trimester of sampling limited to only Mestizo women in a sample of Peruvian women (N=95)

Figure 4 displays the results of running an NMDS ordination on the Bray-curtis distances, colored by various characteristics of interest (i.e., preterm birth status and community state type).

Figure 5 and 6 compare Shannon diversity measures in Mestizo vs non-Mestizo and parous vs nulliparous women, respectively. Interesting, nulliparous women appear to be more diverse than parous women. Tables 3 and 4 examine sample and swab characteristics by community state type assignment, respectively when not stratified and when stratified by gestational age. BV result always correlates strongly with vaginal community state type assignment.

Table 3: Summary descriptives table by groups of 'Community state type'

	Diverse N=62	*Lactobacillus ASV2 dominated* N=34	*L. iners dominated* N=29	p-value
Preterm status:				0.811
No	50 (80.6%)	26 (76.5%)	24 (82.8%)	
Yes	12 (19.4%)	8 (23.5%)	5 (17.2%)	
Maternal age	28.0 (6.76)	29.2 (6.08)	27.9 (6.30)	0.633
Maternal age (categories):				0.524
18 to 19	4 (6.45%)	0 (0.00%)	3 (10.3%)	
20 to 29	36 (58.1%)	18 (52.9%)	15 (51.7%)	
30 to 34	10 (16.1%)	8 (23.5%)	7 (24.1%)	
35 and older	12 (19.4%)	8 (23.5%)	4 (13.8%)	
Education:				0.748
>12th grade	29 (46.8%)	12 (35.3%)	12 (41.4%)	
7th to 12th grade	30 (48.4%)	21 (61.8%)	15 (51.7%)	
$\leq$ 6th grade	3 (4.84%)	1 (2.94%)	2 (6.90%)	
Mestizo:				0.871
No	14 (22.6%)	8 (23.5%)	8 (27.6%)	
Yes	48 (77.4%)	26 (76.5%)	21 (72.4%)	
Married:				0.818
No	13 (21.0%)	6 (17.6%)	7 (24.1%)	
Yes	49 (79.0%)	28 (82.4%)	22 (75.9%)	

Table 3: Summary descriptives table by groups of 'Community state type' *(continued)*

	Diverse	*Lactobacillus ASV2 dominated**	*L. iners dominated*	*p-value*
Employment:				0.196
No	30 (48.4%)	10 (29.4%)	12 (41.4%)	
Yes	32 (51.6%)	24 (70.6%)	17 (58.6%)	
Trouble paying for basics:				0.384
No	30 (48.4%)	14 (41.2%)	17 (58.6%)	
Yes	32 (51.6%)	20 (58.8%)	12 (41.4%)	
Planned pregnancy:				0.658
No	34 (55.7%)	16 (47.1%)	16 (57.1%)	
Yes	27 (44.3%)	18 (52.9%)	12 (42.9%)	
Early pregnancy BMI:				0.723
<18.5	1 (1.61%)	0 (0.00%)	1 (3.45%)	
18.5-24.9	30 (48.4%)	14 (41.2%)	14 (48.3%)	
25-29.9	20 (32.3%)	12 (35.3%)	9 (31.0%)	
$\geq$ 30	11 (17.7%)	6 (17.6%)	5 (17.2%)	
Missing	0 (0.00%)	2 (5.88%)	0 (0.00%)	
Nulliparous:				0.107
No	24 (39.3%)	21 (61.8%)	13 (44.8%)	
Yes	37 (60.7%)	13 (38.2%)	16 (55.2%)	
Bacterial vaginosis (Hay-Ison criteria):				<0.001
I	14 (22.6%)	1 (2.94%)	6 (20.7%)	
Missing	1 (1.61%)	1 (2.94%)	1 (3.45%)	
N	18 (29.0%)	31 (91.2%)	21 (72.4%)	
VB	29 (46.8%)	1 (2.94%)	1 (3.45%)	

* The second CST's dominating organism (labeled ASV2) was identified as either *L. acidophilus* or *L. crispatus* by a BLAST search

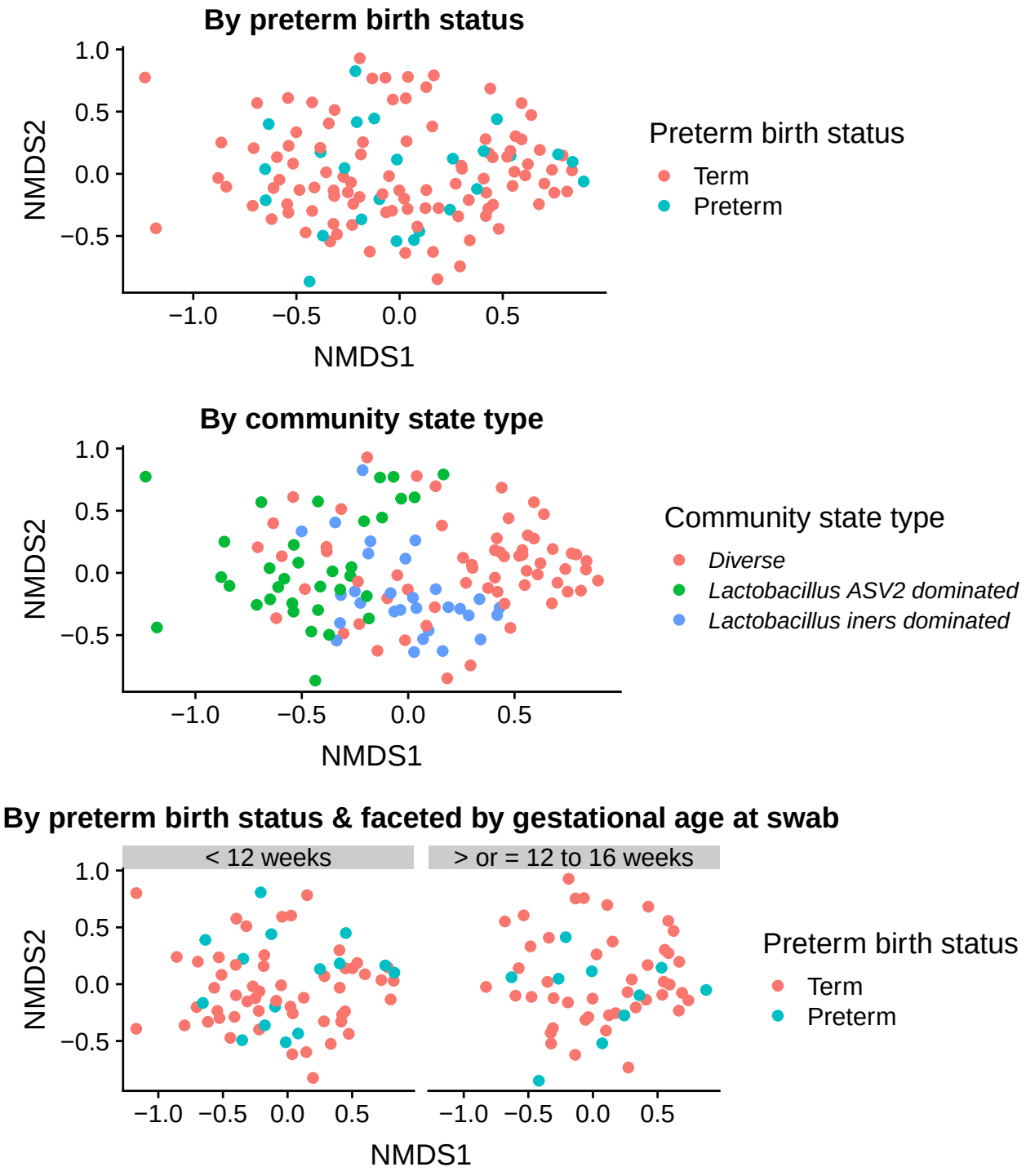

Figure 4: Bray-curtis distance based ordination plots (NMDS)

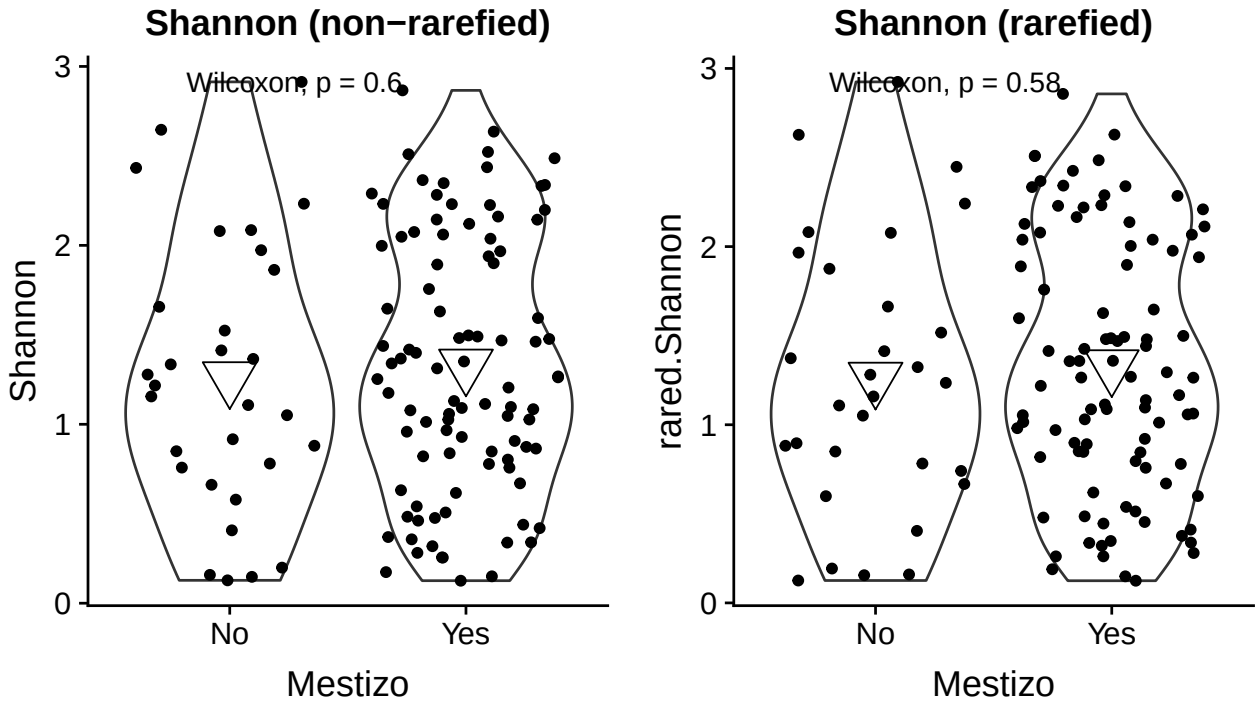

Figure 5: Shannon by Mestizo status

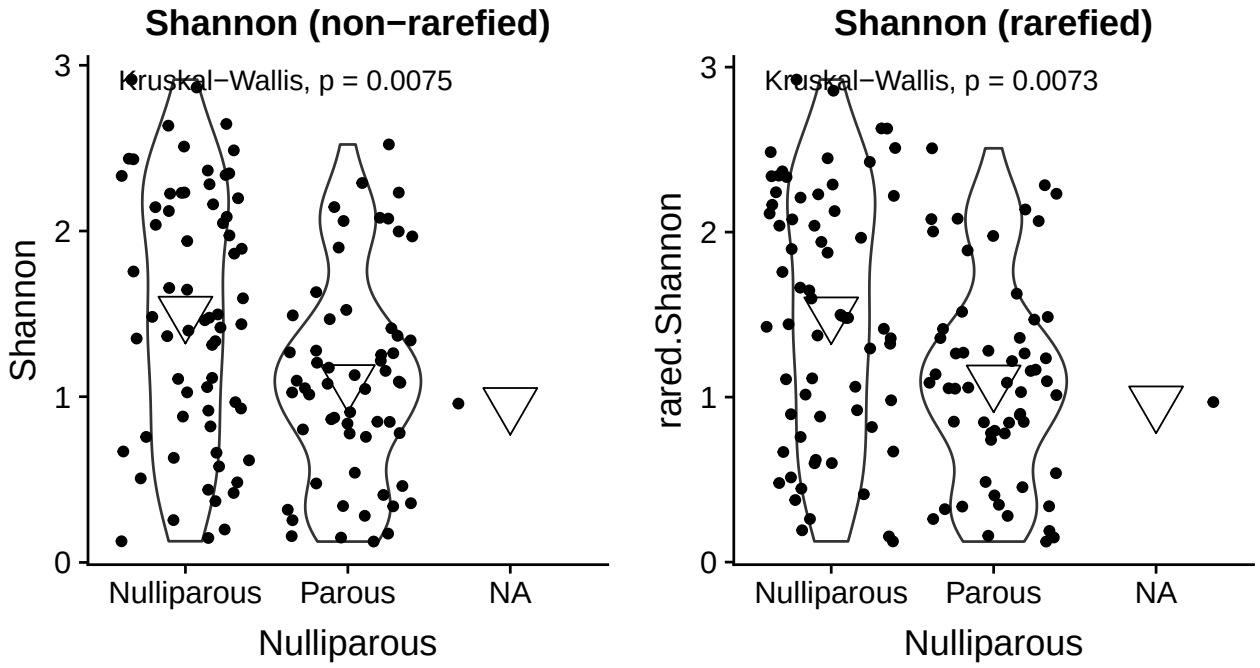

Figure 6: Shannon by Nulliparity

Table 4: Summary descriptive tables by Community state type stratified by gestational age

	< 12 weeks				≥ 12 to 16 weeks			
	Diverse N=30	*Lactobacillus ASV2 dominated** N=23	*L. iners dominated* N=16	*p-value*	*Diverse* N=31	*Lactobacillus ASV2 dominated* N=11	*L. iners dominated* N=13	*p-value*
Preterm status:				0.365				0.094
No	21 (70.0%)	19 (82.6%)	14 (87.5%)		28 (90.3%)	7 (63.6%)	10 (76.9%)	
Yes	9 (30.0%)	4 (17.4%)	2 (12.5%)		3 (9.68%)	4 (36.4%)	3 (23.1%)	
Maternal age	29.8 (7.13)	29.8 (5.77)	28.8 (5.26)	0.835	26.5 (5.95)	27.9 (6.79)	26.8 (7.46)	0.835
Maternal age (categories):				0.834				0.880
18 to 19	1 (3.33%)	0 (0.00%)	1 (6.25%)		2 (6.45%)	0 (0.00%)	2 (15.4%)	
20 to 29	15 (50.0%)	11 (47.8%)	8 (50.0%)		21 (67.7%)	7 (63.6%)	7 (53.8%)	
30 to 34	6 (20.0%)	6 (26.1%)	5 (31.2%)		4 (12.9%)	2 (18.2%)	2 (15.4%)	
35 and older	8 (26.7%)	6 (26.1%)	2 (12.5%)		4 (12.9%)	2 (18.2%)	2 (15.4%)	
Education:				0.964				0.701
>12th grade	13 (43.3%)	8 (34.8%)	7 (43.8%)		16 (51.6%)	4 (36.4%)	5 (38.5%)	
7th to 12th grade	15 (50.0%)	14 (60.9%)	8 (50.0%)		14 (45.2%)	7 (63.6%)	7 (53.8%)	
≤ 6th grade	2 (6.67%)	1 (4.35%)	1 (6.25%)		1 (3.23%)	0 (0.00%)	1 (7.69%)	
Mestizo:				0.938				0.550
No	7 (23.3%)	7 (30.4%)	4 (25.0%)		7 (22.6%)	1 (9.09%)	4 (30.8%)	
Yes	23 (76.7%)	16 (69.6%)	12 (75.0%)		24 (77.4%)	10 (90.9%)	9 (69.2%)	
Married:				0.625				1.000
No	6 (20.0%)	3 (13.0%)	4 (25.0%)		7 (22.6%)	3 (27.3%)	3 (23.1%)	
Yes	24 (80.0%)	20 (87.0%)	12 (75.0%)		24 (77.4%)	8 (72.7%)	10 (76.9%)	
Employment:				0.481				0.498
No	14 (46.7%)	7 (30.4%)	6 (37.5%)		15 (48.4%)	3 (27.3%)	6 (46.2%)	
Yes	16 (53.3%)	16 (69.6%)	10 (62.5%)		16 (51.6%)	8 (72.7%)	7 (53.8%)	
Trouble paying for basics:				0.225				0.998
No	13 (43.3%)	8 (34.8%)	10 (62.5%)		17 (54.8%)	6 (54.5%)	7 (53.8%)	
Yes	17 (56.7%)	15 (65.2%)	6 (37.5%)		14 (45.2%)	5 (45.5%)	6 (46.2%)	
Planned pregnancy:				0.200				0.013
No	16 (53.3%)	12 (52.2%)	4 (26.7%)		18 (58.1%)	4 (36.4%)	12 (92.3%)	
Yes	14 (46.7%)	11 (47.8%)	11 (73.3%)		13 (41.9%)	7 (63.6%)	1 (7.69%)	
Early pregnancy BMI:				0.862				0.124
<18.5	1 (3.33%)	0 (0.00%)	0 (0.00%)		0 (0.00%)	0 (0.00%)	1 (7.69%)	
18.5-24.9	11 (36.7%)	11 (47.8%)	5 (31.2%)		19 (61.3%)	3 (27.3%)	9 (69.2%)	
25-29.9	10 (33.3%)	7 (30.4%)	7 (43.8%)		9 (29.0%)	5 (45.5%)	2 (15.4%)	
≥ 30	8 (26.7%)	4 (17.4%)	4 (25.0%)		3 (9.68%)	2 (18.2%)	1 (7.69%)	
Missing	0 (0.00%)	1 (4.35%)	0 (0.00%)		0 (0.00%)	1 (9.09%)	0 (0.00%)	
Nulliparous:				0.335				0.397
No	14 (48.3%)	15 (65.2%)	7 (43.8%)		10 (32.3%)	6 (54.5%)	6 (46.2%)	
Yes	15 (51.7%)	8 (34.8%)	9 (56.2%)		21 (67.7%)	5 (45.5%)	7 (53.8%)	
Bacterial vaginosis (Hay-Ison criteria):				<0.001				0.003
I	6 (20.0%)	1 (4.35%)	3 (18.8%)		7 (22.6%)	0 (0.00%)	3 (23.1%)	
Missing	0 (0.00%)	1 (4.35%)	1 (6.25%)		1 (3.23%)	0 (0.00%)	0 (0.00%)	
N	9 (30.0%)	21 (91.3%)	12 (75.0%)		9 (29.0%)	10 (90.9%)	9 (69.2%)	
VB	15 (50.0%)	0 (0.00%)	0 (0.00%)		14 (45.2%)	1 (9.09%)	1 (7.69%)	

* The second CST's dominating organism (labeled ASV2) identified as either *L. acidophilus* or *L. crispatus* by a BLAST search

Table 5: Results of logistic regression for preterm birth status (odds ratios and 95 percent CI) in a sample of Peruvian women - sensitivity analysis: additional parameter for dispersion

	Model 1[a]	Model 2[b]	Model 3[c]
	OR (CI) [d]	OR (CI) [d]	OR (CI) [d]
	100 controls / 25 cases	99 controls / 25 cases	99 controls / 25 cases
Lactobacillus ASV2 dominated	1.28 (0.45, 3.54)	1.25 (0.42, 3.55)	1.27 (0.43, 3.63)
L. iners dominated	0.87 (0.25, 2.67)	0.85 (0.24, 2.63)	0.87 (0.25, 2.73)

Note:
Reference is Diverse community state type.
One individual missing parity status
[a] Unadjusted
[b] Adjusted for parity
[c] Adjusted for parity and Mestizo
[d] Profile-likelihood calculated confidence intervals
[e] The second CST's dominating organism – labeled ASV2 - was identified as being either *L. acidophilus* or *L. crispatus* by a BLAST search.

Table 6: Results of logistic regression for preterm birth status stratified by gestational age at vaginal swab (odds ratios and 95 percent CI) in a sample of Peruvian women - sensitivity analysis: additional parameter for dispersion

	< 12 weeks			> or = 12 weeks		
	Model 1[a]	Model 2[b]	Model 3[c]	Model 1[a]	Model 2[b]	Model 3[c]
	OR (CI)[c]	OR (CI)[c]	OR (CI)[c]	OR (CI)[c]	OR (CI)[c]	OR (CI)[c]
	54 controls / 15 cases	53 controls / 15 cases	53 controls / 15 cases	45 controls / 10 cases	45 controls / 10 cases	45 controls / 10 cases
Lactobacillus ASV2 dominated[e] *(reference = Diverse CST)*	0.49 (0.11, 1.83)	0.49 (0.11, 1.91)	0.49 (0.11, 1.92)	5.33 (0.92, 34.84)	5.06 (0.83, 34.75)	4.65 (0.92, 25.93)
L. iners dominated	0.33 (0.04, 1.59)	0.31 (0.04, 1.53)	0.31 (0.04, 1.54)	2.8 (0.43, 18.47)	2.71 (0.4, 18.52)	3.48 (0.6, 20.65)

Note:
Reference is Diverse community state type
[a] Unadjusted
[b] Adjusted for parity, one individual missing parity information
[c] Adjusted for parity and Mestizo, one individual missing parity information
[d] Profile-likelihood calculated confidence intervals
[e] The second CST's dominating organism – labeled ASV2 - was identified as being either *L. acidophilus* or *L. crispatus* by a BLAST search.

Results of multivariate model sensitivity analyses

We include a sensitivity analysis modeling an additional parameter for dispersion; however, as the data is not significantly overdispersed, the estimates from these models are very similar to those in the main analysis (Supplemental Tables 5 & 6)

As there were no non-Mestizo cases sampled at or after 12 weeks of gestation, the "Mestizo" predictor perfectly delineates cases and controls among women sampled at or after 12 weeks of gestation. Therefore, when numerically estimating confidence intervals for the Mestizo coefficient, fitted probabilities of 0 and 1 occurred, and our confidence intervals and effect estimates may not be reliable. However, when limiting the stratified sample to only Mestizo women, results of the models were similar to those controlling for Mestizo and parity and those controlling for parity alone (see Supplemental Table 7 and 8).

We also performed an effect modification analysis. In Table 9, we present the results of the multiplicative effect modification analysis performed by including an interaction term between bacterial community state type and gestational age at sampling. Due to our 3-level predictor (community state type), we had to run regressions on data subsets excluding individuals in one of the three community state types at a time in order to obtain relative excess risk due to interaction (RERI) statistics for an analysis of effect modification on the additive scale. Additionally, because measures of additive interaction can be inconsistent when using preventative factors, we have recoded the reference for the microbiome CSTs when calculating the RERI such that one of the *Lactobacillus* dominated community state types is always the reference. These results are presented in Tables 10 and 11.

We also conducted a taxa-wide assessment of association with preterm birth, both stratified and not stratified by gestational age. No taxa were significantly associated with preterm birth in stratified or unstratified analyses after correction for multiple testing. We show the two dominant *Lactobacillus* species and *Gardnerella vaginalis* here, although they were not significant, in Figure 8. Before correction for multiple testing, no taxa were significantly different between preterm and term births in the unstratified analysis, and two taxa were significant in the first trimester - these taxa are shown in Figure 9.

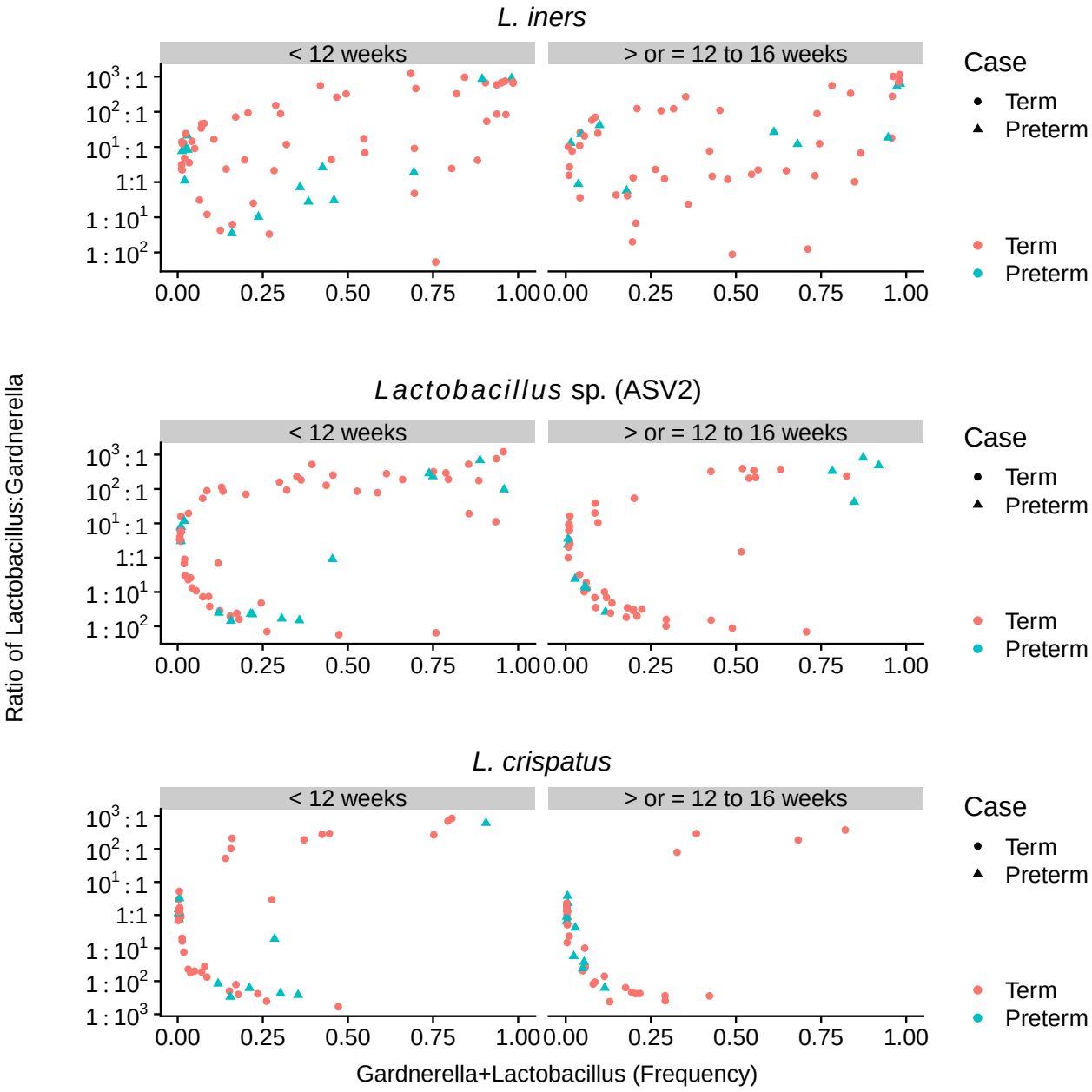

Figure 7: Exclusion of key taxa

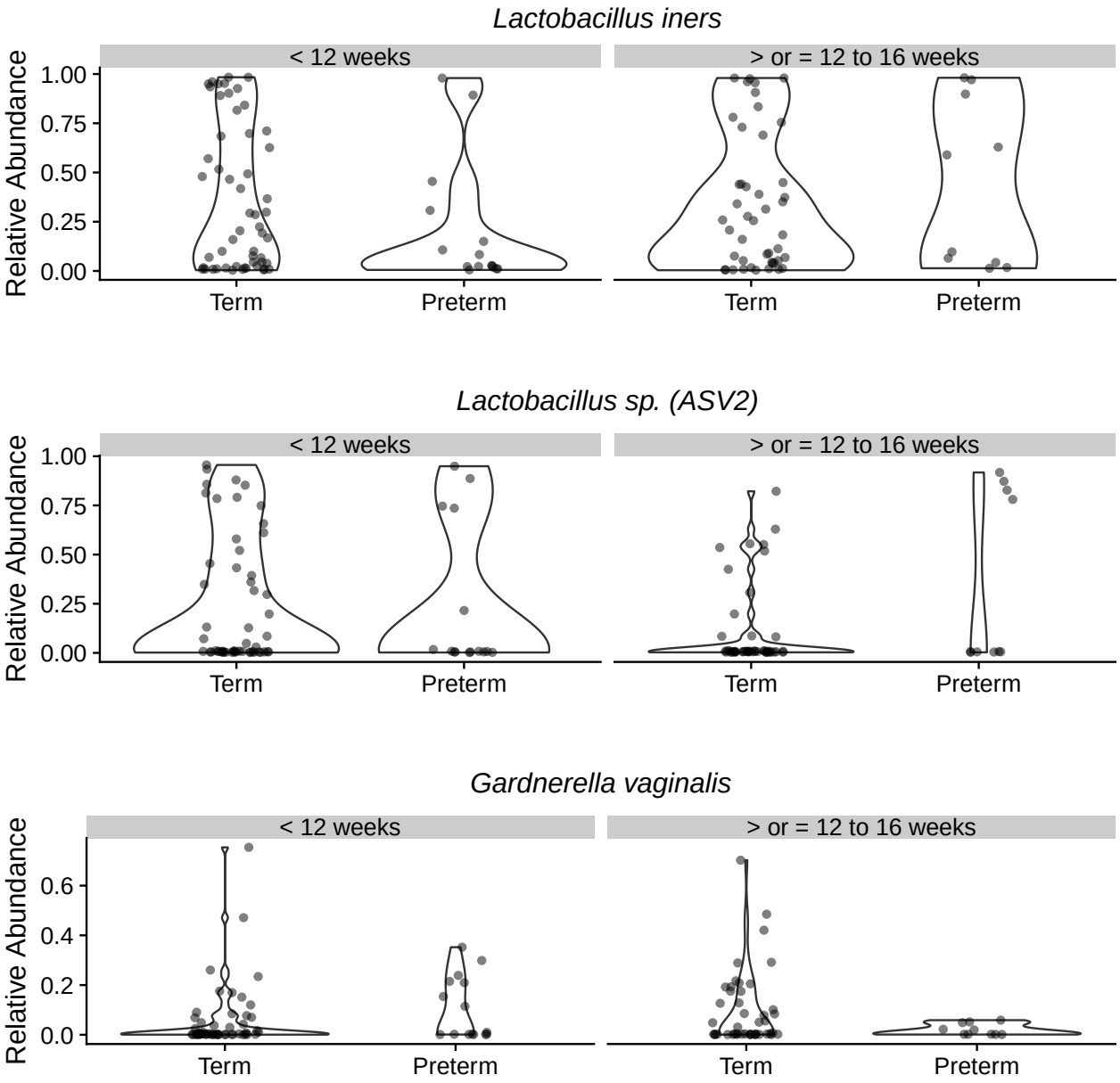

Figure 8: Aldex2 testing of individual taxa - hypothesized taxa of interest

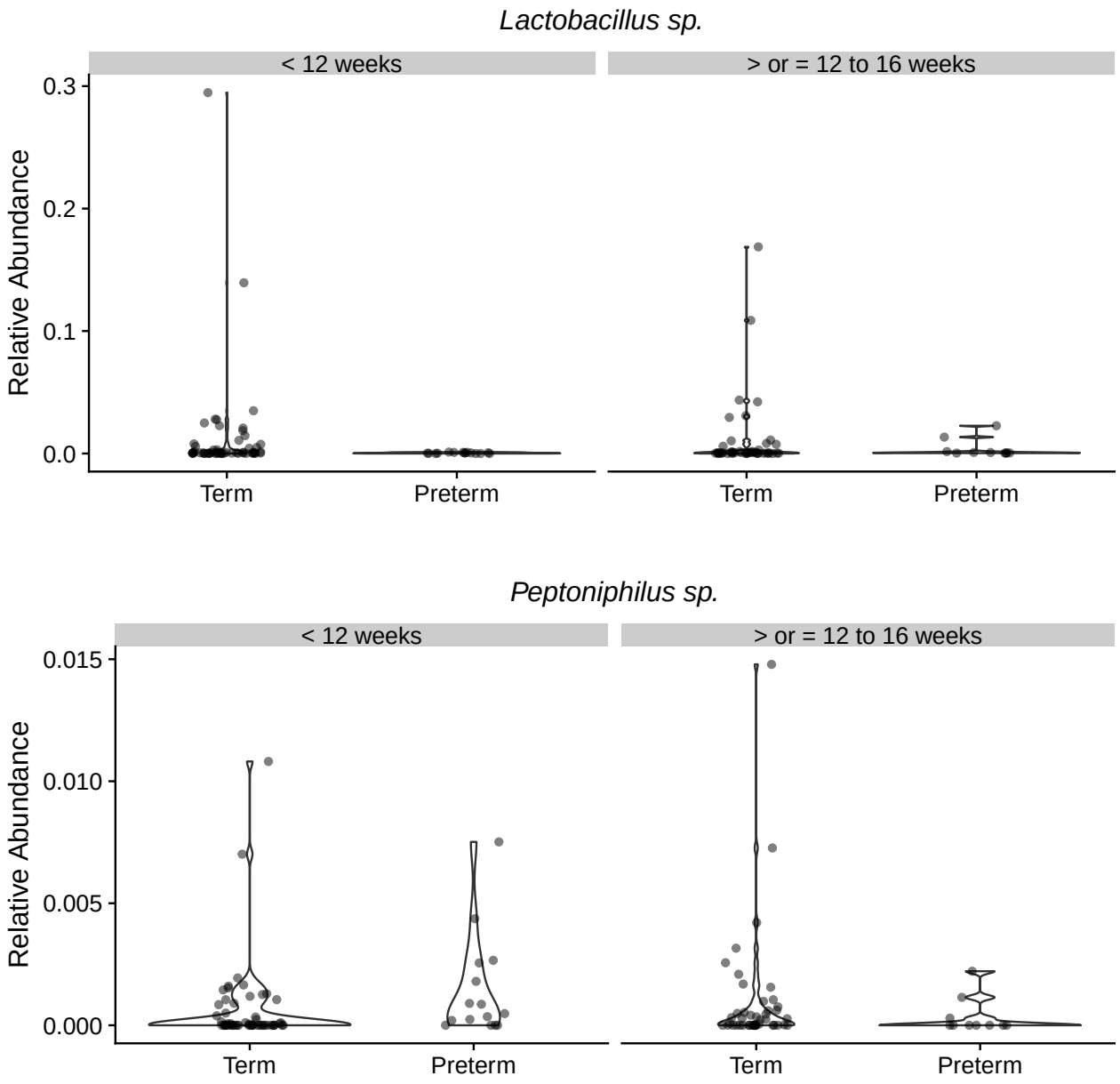

Figure 9: Aldex2 testing of individual taxa - taxa which were significant before correction for multiple testing

Table 7: Results of logistic regression for preterm birth status (odds ratios and 95 percent CI) in a sample of Peruvian women - sensitivity analysis: restricted to Mestizo women

	Model 1[a]	Model 2[b]
	OR (CI)[c]	OR (CI)[c]
	74 controls / 21 cases	73 controls / 21 cases
Lactobacillus ASV2 dominated[d] (reference = Diverse CST)	1.01 (0.31, 3.08)	0.94 (0.28, 2.93)
L. iners dominated	0.79 (0.2, 2.7)	0.75 (0.19, 2.59)

Note:
Reference is Diverse community state type.
One individual missing parity status
[a] Unadjusted
[b] Adjusted for parity
[c] Profile-likelihood calculated confidence intervals
[d] The second CST's dominating organism – labeled ASV2 - was identified as being either *L. acidophilus* or *L. crispatus* by a BLAST search.

Table 8: Results of logistic regression for preterm birth status stratified by gestational age at vaginal swab (odds ratios and 95 percent CI) in a sample of Peruvian women - sensitivity analysis: restricted to Mestizo women

	< 12 weeks		> or = 12 weeks	
	Model 1[a]	Model 2[b]	Model 1[a]	Model 2[b]
	OR (CI)[c]	OR (CI)[c]	OR (CI)[c]	OR (CI)[c]
	40 controls / 11 cases	39 controls / 11 cases	40 controls / 11 cases	39 controls / 11 cases
Lactobacillus ASV2 dominated[d] (reference = Diverse CST)	0.27 (0.04, 1.3)	0.22 (0.03, 1.13)	4.67 (0.82, 29.96)	4.65 (0.81, 29.96)
L. iners dominated	0.17 (0.01, 1.13)	0.15 (0.01, 1.02)	3.5 (0.53, 23.81)	3.48 (0.52, 23.92)

Note:
Reference is Diverse community state type
[a] Unadjusted
[b] Adjusted for parity, one individual missing parity information
[c] Profile-likelihood calculated confidence intervals
[d] The second CST's dominating organism – labeled ASV2 - was identified as being either *L. acidophilus* or *L. crispatus* by a BLAST search.

Table 9: Results of logistic regression for preterm birth status (odds ratios and 95 percent CI) in a sample of Peruvian women - including interaction term between bacterial community state type and gestational age at sampling and testing for effect modification on multiplicative scale

		Community State Type (CST)						*ORs (95 percent CI) Lactobacillus ASV2 dominated CST within strata of gestational age*	*ORs (95 percent CI) for L. iners dominated CST within strata of gestational age*
		Diverse CST		*Lactobacillus ASV2 dominated CST*		*L. iners dominated CST*			
Gestational age	< 12 weeks' gestation	9/20	1.0	4/19	0.49,(0.11, 1.82); p=0.30	2/14	0.32,(0.04, 1.47); p=0.18	0.49,(0.11, 1.82); p=0.30	0.32,(0.04, 1.47); p=0.18
at vaginal swab	12 to 16 weeks' gestation	3/28	0.23,(0.05, 0.89); p=0.05	4/7	1.18, (0.27, 5.21); p=0.83	3/10	0.69, (0.15, 3.21); p=0.63	5.09, (0.89, 29.15); p=0.07	2.99, (0.5, 17.86); p=0.23
	Measure of interaction on multiplicative scale: Ratio of ORs (95 percent CI)				10.41,(1.21, 101.18); p=0.04		9.47,(0.86, 129.05); p=0.07		

Note:
ORs are adjusted for parity (nulliparous vs parous) and ethnicity (Mestizo vs not Mestizo)
All estimates presented in this table (including OR within strata of gestational age) were calculated using the regression formula $log(\hat{p}) = \beta_0 + \beta_1 MicrobialCST + \beta_2 I(GestationalAge \geq 12) + \beta_3 Mestizo + \beta_4 Parous + \beta_5 MicrobialCST * I(GestationalAge \geq 12)$.
Therefore estimates for ORs within strata of gestational age differ slightly from the ORs presented in the fully stratified analysis in Main Analysis, Table 3.

Table 10: Results of logistic regression for preterm birth status (odds ratios and 95 percent CI) in a sample of Peruvian women - including interaction term between bacterial community state type and gestational age at sampling and testing for effect modification on multiplicative and additive scales

Restricted to women not in L. iners dominated CST

		Community State Type (CST)				*ORs (95 percent CI) Diverse CST within strata of gestational age*
		Lactobacillus ASV2 dominated CST		*Diverse CST*		
Gestational age	< 12 weeks' gestation	4/19	1.0	9/20	2.08, (0.56, 8.92); p=0.29	2.08, (0.56, 8.92); p=0.29
at vaginal swab	12 to 16 weeks' gestation	4/7	2.43, (0.45, 13.38); p=0.29	3/28	0.49, (0.09, 2.57); p=0.39	0.20, (0.03, 1.15); p=0.07
	Measure of interaction on multiplicative scale: Ratio of ORs (95 percent CI)				0.10, (0.01, 0.83); p=0.04	
	Measure of interaction on additive scale: RERI (95 percent CI)				-3.02, (-8.64, 2.6); p=0.85	

Note:
ORs are adjusted for parity (nulliparous vs parous) and ethnicity (Mestizo vs not Mestizo)
Since measures of additive interaction can be inconsistent when using preventative factors we have recoded the reference for the CSTs, now the Diverse CST is the reference.
The models used to calculate the estimate presented in this table were run on a subsample excluding individuals assigned to the *L. iners* dominated CST

Table 11: Results of logistic regression for preterm birth status (odds ratios and 95 percent CI) in a sample of Peruvian women - including interaction term between bacterial community state type and gestational age at sampling and testing for effect modification on multiplicative and additive scales

Restricted to women not in Lactobacillus ASV2 dominated CST						
		Community State Type (CST)				
		L. iners dominated CST		*Diverse CST*		*ORs (95 percent CI) Diverse CST within strata of gestational age*
Gestational age at vaginal swab	< 12 weeks' gestation	2/14	1.0	9/20	3.24, (0.68, 23.87); p=0.18	3.24, (0.68, 23.87); p=0.18
	12 to 16 weeks' gestation	3/10	2.32, (0.32, 20.8); p=0.41	3/28	0.69, (0.1, 4.88); p=0.71	0.30, (0.05, 1.85); p=0.19
	Measure of interaction on multiplicative scale: Ratio of ORs (95 percent CI)				0.09, (0.01, 1.05); p=0.06	
	Measure of interaction on additive scale: RERI (95 percent CI)				-3.86, (-12.23, 4.51); p=0.82	

Note:
ORs are adjusted for parity (nulliparous vs parous) and ethnicity (Mestizo vs not Mestizo)
Since measures of additive interaction can be inconsistent when using preventative factors we have recoded the reference for the CSTs, now the Diverse CST is the reference.
The models used to calculate the estimate presented in this table were run on a subsample excluding individuals assigned to the *Lactobacillus* ASV2 dominated CST