Supporting Information

Ravel et al. 10.1073/pnas.1002611107

SI Materials and Methods

Taxonomic Classification of 16S rRNA Gene Sequences. Each processed 16S rRNA gene sequence was classified at a genus level using the Ribosomal Database Project (RDP) Naïve Bayesian Classifier (1). These assignments were refined using the following algorithm. If the median RDP score of all 16S rRNA gene sequences assigned to a genus was less than 0.5 threshold, the median RDP score for the next-higher taxonomic level (family) was calculated. If it was above the RDP score threshold then the Operational Taxonomic Unit (OTU) name "FamilyName_idx" was assigned to all of the reads, where idx was some integer index. For each taxonomic level, the algorithm was applied in an iterative manner until the RDP score was above 0.5. For example, the median RDP score of reads assigned to the genus *Ignavigranum* was 0.17, and the median RDP score at the family (Aerococcaceae) level was 0.36, whereas the median RDP score for the order (Lactobacillales) was (0.73). Thus we assigned the OTU name Lactobacillales_5 to all reads originally assigned to the genus *Ignavigranum*. The OTU has index 5, because it was the fifth OTU created that belonged to order Lactobacillales.

Often, taxonomic assignments were made that resulted in no more than two sequence reads assigned to a genus with a mean RDP score of less than 0.9, and the read(s) came from only one sample. These were considered to be low-quality assignments, and reads assigned to this genus were excluded from the community structure analysis. This eliminated a total of 70 low-quality genus assignments.

Overall, 70% of all sequence reads generated in this study were taxonomically assigned to the genus *Lactobacillus*. The median RDP *Lactobacillus* reads score was 0.94. Approximately 92% of the reads assigned to the genus *Lactobacillus* by the RDP classifier had a score of 0.8 or higher, whereas only 0.3% had a score less than 0.5.

Species-Level Taxonomic Assignment of *Lactobacillus* 16S rRNA Gene Sequences. Because of the short read lengths obtained by 454 pyrosequencing, the classification of 16S rRNA sequences using phylogenetic approaches is typically limited to the genus level (1). However, in studies of vaginal microbiota it is essential to classify *Lactobacillus* at the species level to differentiate the four species of *Lactobacillus* sp. that distinguish kinds of vaginal communities, namely *L. crispatus*, *L. iners*, *L. jensenii*, and *L. gasserii* (2).

We have developed an algorithm to achieve accurate and rapid species-level assignments from short V2 16S rRNA genes sequences generated by 454 pyrosequencing. Full-length 16S sequences of *Lactobacillus* species were assembled from the RDP database (3) and trimmed to 240 base pairs (nucleotide positions 98–338). Sequences containing ambiguous base calls have been removed from the dataset. Only species with at least 10 representative sequences were retained in the dataset. This resulted in a total of 3,108 sequences representing 42 *Lactobacillus* species.

Using the software HMMER version 1.8.5, hidden Markov models (HMM) were built for each of the 42 known species of the genus *Lactobacillus*, including those previously found in the vagina. Each V2 region of 16S rRNA gene sequences assigned to the genus *Lactobacillus* was aligned to all species-level HMM models. A read was assigned to the i-th HMM model if the highest HMM alignment score came from the i-th HMM model and that score was at least as high as the lowest score of the sequences used to build the i-th model.

The sequence reads not assigned to any HMM model were classified as OTUs within the genus *Lactobacillus* using the DBSCAN clustering algorithm (4) on a 3D projection of unclassified reads using HMM scores. More precisely, to each unclassified read r we assigned a 5-tuple [Li(r), Lc(r), Lj(r), Lg(r), Lv(r)], where Li(r), Lc(r), Lj(r), Lg(r), and Lv(r) are HMM scores for r obtained by aligning r to the HMM models of the five most abundant species in the dataset: *L. iners*, *L. crispatus*, *L. jensenii*, *L. gasseri*, and *L. vaginalis*, respectively. The assignment

$$r \rightarrow [Li(r),\ Lc(r),\ Lj(r),\ Lg(r),\ Lv(r)]$$

induces an embedding of the unclassified reads into a five-dimensional Euclidean space. We used principal component analysis (PCA) to project these points from the five-dimensional space to a 3D space using the first three PCA components.

The DBSCAN algorithm requires two parameters: ε and the minimum number of points required to form a cluster (minPts). Two points p and q are in the same precluster if there is a chain of points $P = x1, \ldots, xn = q$, such that the distance between each pair of consecutive points xi, xi+1 is at most ε. A precluster forms a cluster if there are at least minPts number of elements in it. If the precluster has fewer than minPts elements, then its elements are labeled as noise. The choice of ε DBSCAN parameter was done by visual investigation of clustering quality on the 3D PCA projection of unclassified reads. The minimum number of points required to form a cluster was set to 10. Each identified cluster formed a new *Lactobacillus* OTU, and a new HMM was built for that OTU. A total of four new *Lactobacillus* sp. OTUs were identified in the dataset.

DBSCAN was implemented in C++ language using ANN library (http://www.cs.umd.edu/~mount/ANN/) for nearest neighbor searching in Euclidean spaces of various dimensions.

Validation of the HMM-Based Species Assignments for *Lactobacillus* sp. To validate the HMM-based speciation algorithm we used the dataset of Zhou et al. (2), who sequenced the 8–926 region of 1,892 cloned 16S rRNA genes of *Lactobacillus* sp. from the vagina and assigned them to species of *Lactobacillus* using phylogenetic algorithms. A total of six species of *Lactobacillus* was identified. As shown in Table S8, the algorithm developed for the present study was able to correctly classify each species with 98.69–100% accuracy.

Statistical Methods. ***Community clustering analyses.*** The clustering of communities based on community composition and abundance in Fig. 1 (main text) and Fig. S1 were done using complete linkage hierarchical clustering with five clusters using the R package (2). Two singleton clusters were identified and omitted from the figures that were generated using a modified version of the heatmap routine in the R package.

Shannon diversity analyses. The following equation was used to calculate the Shannon diversity index of a community as previously described (5):

$$H(p) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

where p_i is the proportion of the i-th member of the community, and n is the number of all community members. Shannon diversity is a nonnegative function that reflects the entropy of the probability mass function {pi}. It is zero for a community with only one species and attains its maximum value of $\log_2(n)$ when all taxa of

a community are equally abundant p1= p2 = . . . = pn = 1/n. The Shannon diversity indices for each of the 394 samples analyzed in this study are shown in Fig. 1 (main text) and Fig. S1.

PCA of vaginal microbial communities. The 3D projection of all communities shown in Fig. 4 (main text) was generated by PCA using the prcomp routine in the R package (6) on a dataset consisting of the percentage abundances of taxa in each community (Table S4). The three principle components explained 82% of the variance. To obtain a more symmetrical tetrahedron shape of community states, 100 copies of the median point of states dominated at the 95% level by *L. jensenii* were added to the dataset.

A gradient coloring scheme was selected to show the relationships between the different communities.

Correlation profiles of taxa and Nugent scores. Spearman correlation coefficients. We evaluated potential pairwise interactions between all 282 bacterial phylotypes of the vaginal communities analyzed in this study using Spearman's rank correlation coefficients between relative abundances of phylotype pairs. Spearman correlation coefficients assess how well the relationship between the relative abundances of two phylotypes can be described using a monotonic function. If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when the relative abundance of each phylotype is a perfect monotone function of the other. We tested whether an observed value of the correlation coefficient is significantly different from zero by estimating the probability that the correlation coefficient would be greater than or equal to the observed value, given the null hypothesis, by using a permutation test. The *P* value was set to 0.001 to alleviate the problem of multiple testing. Correlation coefficients that were not significant at the above significance level were set to 0.

Correlations between phylotype relative abundances and Nugent scores were calculated using Spearman's rank correlation coefficients. Nugent scores were classified as low (score 0–3), medium (score 4–6), and high (score 7–10) and treated as ordered categorical variables. Correlations that were not significant at the 0.001 level were set to 0.

All correlation analyses were performed in the R package (6).

Correlation profiles. A profile of relative abundance correlation coefficients of a given phylotype is the vector of Pearson correlation coefficients of the relative abundances of the phylotype with the relative abundances of all other phylotypes. In this study we refer to this as a "correlation profile". For example, if P_1, . . ., P_282 is a list of phylotypes with P_1 = *L. iners*, then the correlation profile of *L. iners* is the set of numbers

$$\mathrm{cor}(\mathrm{P\_1},\ \mathrm{P\_1}),\ \mathrm{cor}(\mathrm{P\_1},\ \mathrm{P\_2}),\ \ldots,\ \mathrm{cor}(\mathrm{P\_1},\ \mathrm{P\_282}),$$

where cor(P_i, P_j) is the Pearson correlation coefficient between phylotype P_i and P_j.

The resulting sequence of numbers

$$\mathrm{cor}(\mathrm{P\_1},\ \mathrm{P\_1}),\ \mathrm{cor}(\mathrm{P\_1},\ \mathrm{P\_2}),\ \ldots,\ \mathrm{cor}(\mathrm{P\_1},\ \mathrm{P\_282}),$$

can be considered as a point in 282 dimensional space with one correlation coefficient per dimension. Phylotypes with similar correlation profiles correspond to clusters of points in this 282 dimensional community state space.

Correlograms. A standard method of displaying correlations between a small number of variables utilizes a matrix of scatter plots as shown in Fig. S2*A*. But for a larger number of variables, scatter plots are not legible, and this is why we resorted to the use of heatmap-based correlograms as shown in Fig. 3 (main text) and Fig. S2*B*. Phylotypes with the strongest positive or negative association with Nugent scores (Fig. S2*C* and Table S7) were used to construct Fig. 3.

Fig. S2*B* is a graphical representation of the correlation profiles for the 50 taxa with the largest cumulative Spearman's correlation coefficient (Table S7). As in Fig. 3, the rows and the columns of the correlation matrix of Fig. S2*B* have been reordered so that phylotypes with similar correlation profiles are placed near each other. The reordering was accomplished using complete linkage hierarchical clustering.

1. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267.
2. Zhou X, et al. (2007) Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. *ISME J* 1:121–133.
3. Cole JR, et al. (2009) The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* 37(Database issue):D141–D145.
4. Ester M, Kriegel H-P, Sander J, Xu A (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, eds Simoudis E, Han J, Fayyad UM (AAAI Press, Menlo Park, CA), pp 226–231.
5. Bent SJ, Forney LJ (2008) The tragedy of the uncommon: Understanding limitations in the analysis of microbial diversity. *ISME J* 2:689–695.
6. R Development Core Team (2008) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).

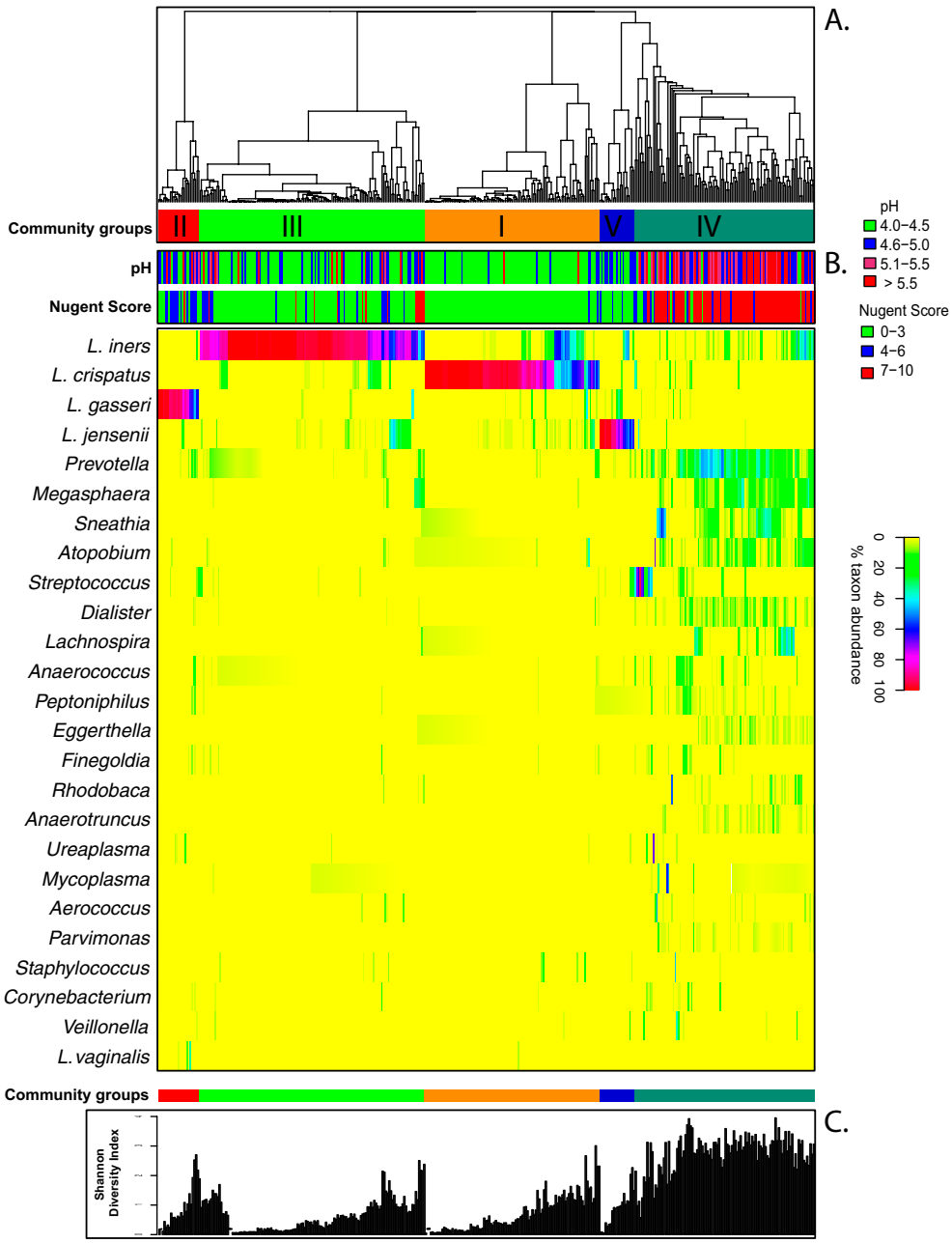

Fig. S1. Heatmap of percentage abundance of microbial taxa found in the vaginal microbial communities of 394 reproductive-age women. (*A*) Complete linkage clustering of samples based on species composition and abundance in communities. (*B*) Nugent scores and pH measurements for each of the 394 samples. (*C*) Shannon diversity indices calculated for each of the 394 vaginal communities (two singletons were excluded).

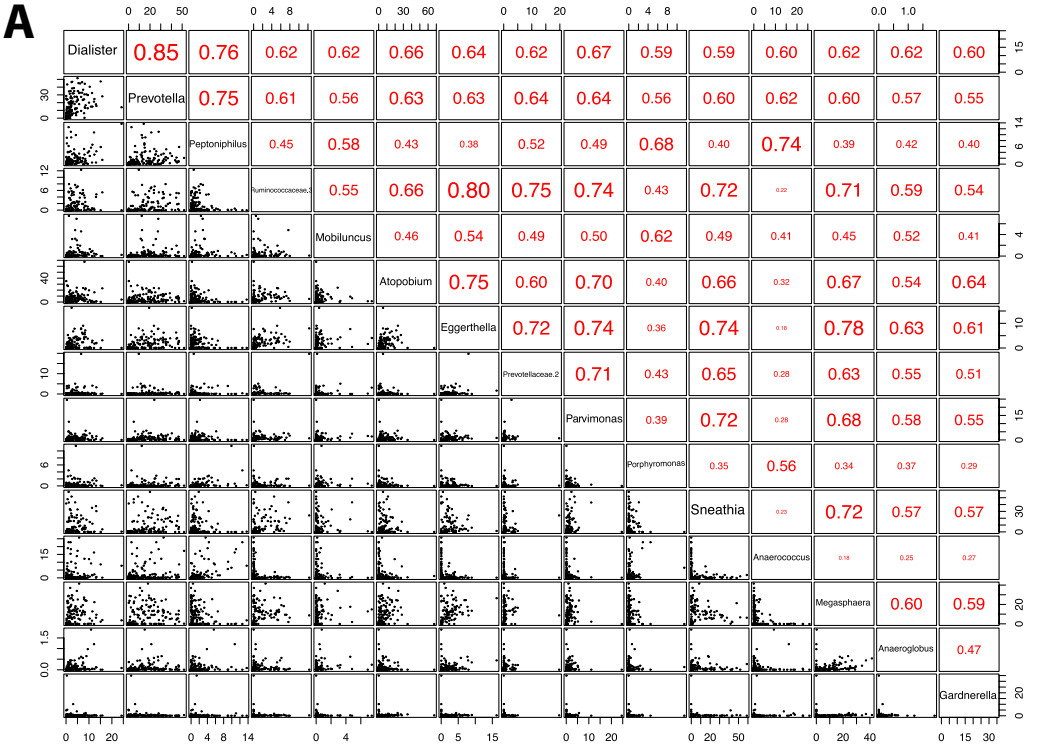

Fig. S2. (Continued)

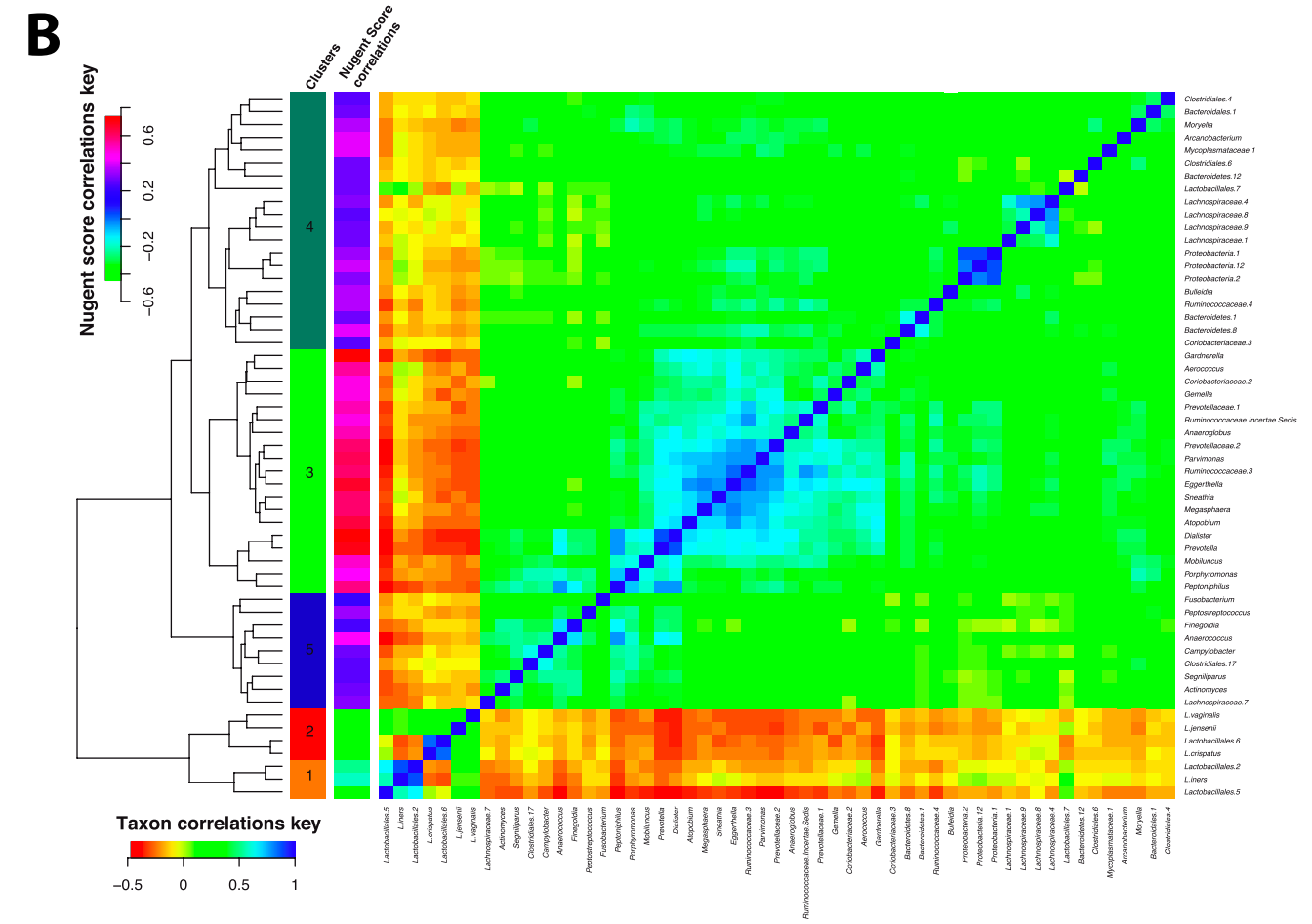

Fig. S2. (Continued)

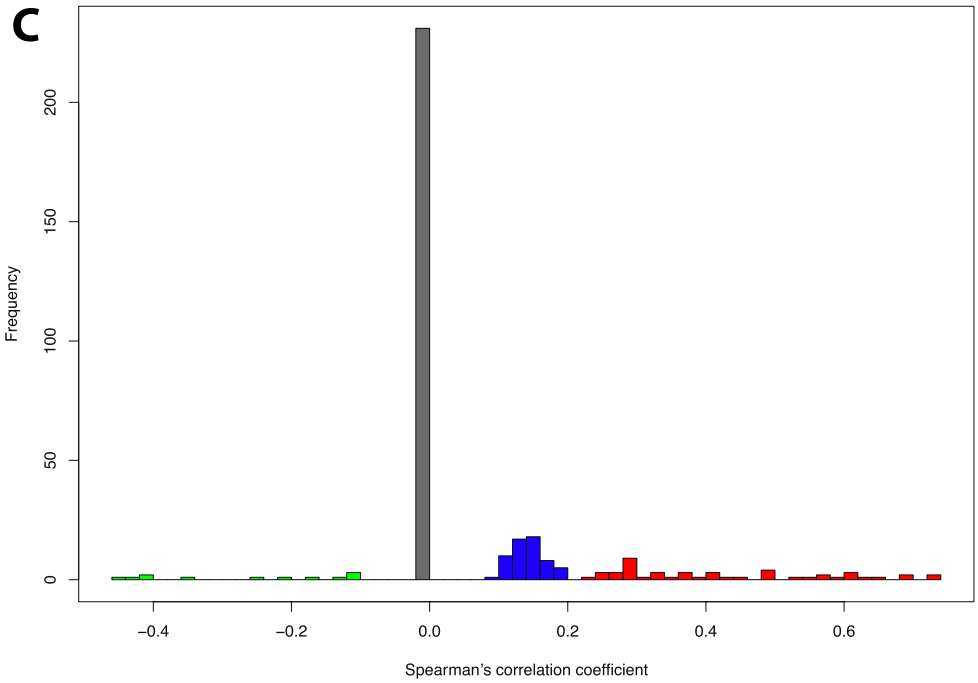

Fig. S2. (*A*) Taxa correlation plots of 15 taxa with highest cumulative Spearman's correlation coefficients found in the vagina microbiota. A Spearman correlation coefficient positive value indicates positive correlation, and a negative value indicates a negative correlation. In these correlation plots, each taxa is compared with all 14 remaining taxa graphically (lower half of the correlation plot) and statistically (upper half of the correlation plot). In the graphical comparison, two taxa are plotted against each other in each subplot, with one taxa's abundance on the *x* axis and another taxa's abundance on the *y* axis. In the statistical comparison, the significance of the correlation was tested using a two-tailed *t* test (i.e., the significance of the difference of Spearman coefficient from 0) at $P = 0.05$. The font sizes of correlation coefficients are proportional to their values. In this figure all correlations are significant at the 5% level. This matrix correlogram is appropriate to demonstrate a correlation between 10 and 20 most-abundant taxa or those with the strongest correlation coefficient. However, this representation is inappropriate to establish correlation between a large number of taxa or groups of taxa. (*B*) Correlogram of the 50 taxa with the largest cumulative Spearman's rank correlations. In this correlation plots, each taxa is compared with all 49 remaining taxa, and the Spearman's correlation coefficient is displayed using shades of red to represent negative correlations and shades of green to blue to represent positive correlations. The taxa were clustered using complete hierarchical clustering with seven clusters using the profiles of 49 correlation coefficients for each taxon. The combination of clustering and heatmap display is amenable to discovery of strong correlations between groups of organisms. (*C*) Histogram of Spearman's correlation coefficients reflecting the relative abundance of a taxon and Nugent score category (low, intermediate, and high). The colors of bars indicate four groups of taxa: green, taxa negatively correlated with Nugent score (high relative abundance, low Nugent score); gray, taxa not correlated with Nugent score; blue, taxa showing weak positive correlation with Nugent score; red, taxa showing medium to strong positive correlation with Nugent score. The 60 taxa comprising the green and red groups were selected and used to construct the correlogram shown in Fig. 3 (main text).

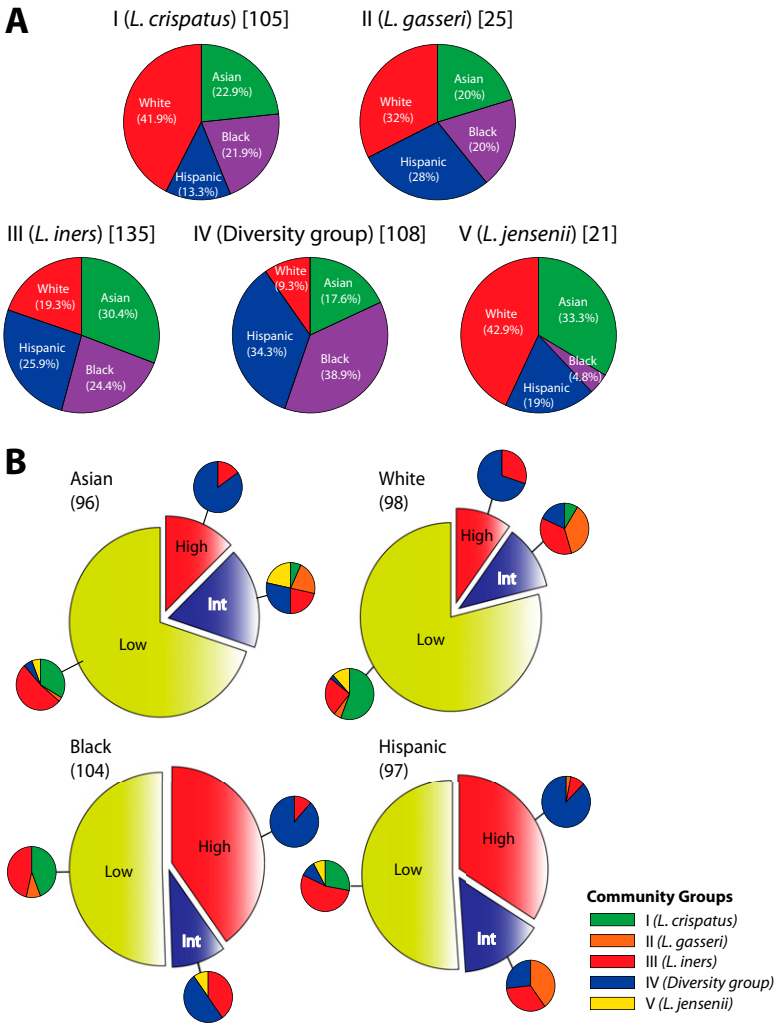

Fig. S3. (*A*) Contribution of ethnicity to each of the five vaginal community groups, expressed as percentage. Sectors are colored according to ethnicity and labeled accordingly. The percentage represents the proportion of subjects of each ethnicity divided by the total number of subjects assigned to a community group (indicated in square brackets). The dominant species for each community group is indicated in parentheses. (*B*) Apportionment of Nugent score categories within ethnic groups and the proportion of each of the five community groups in each Nugent score category (0–3 = low, 4–6 = medium, 7–10 = high). The total number of subjects in each ethnic group is indicated in parentheses. Community groups are colored according to the following: community group I (*L. crispatus*): green; community group II (*L. gasseri*): orange; community group III (*L. iners*): red; community group IV (diversity group); community group V (*L. jensenii*): yellow.

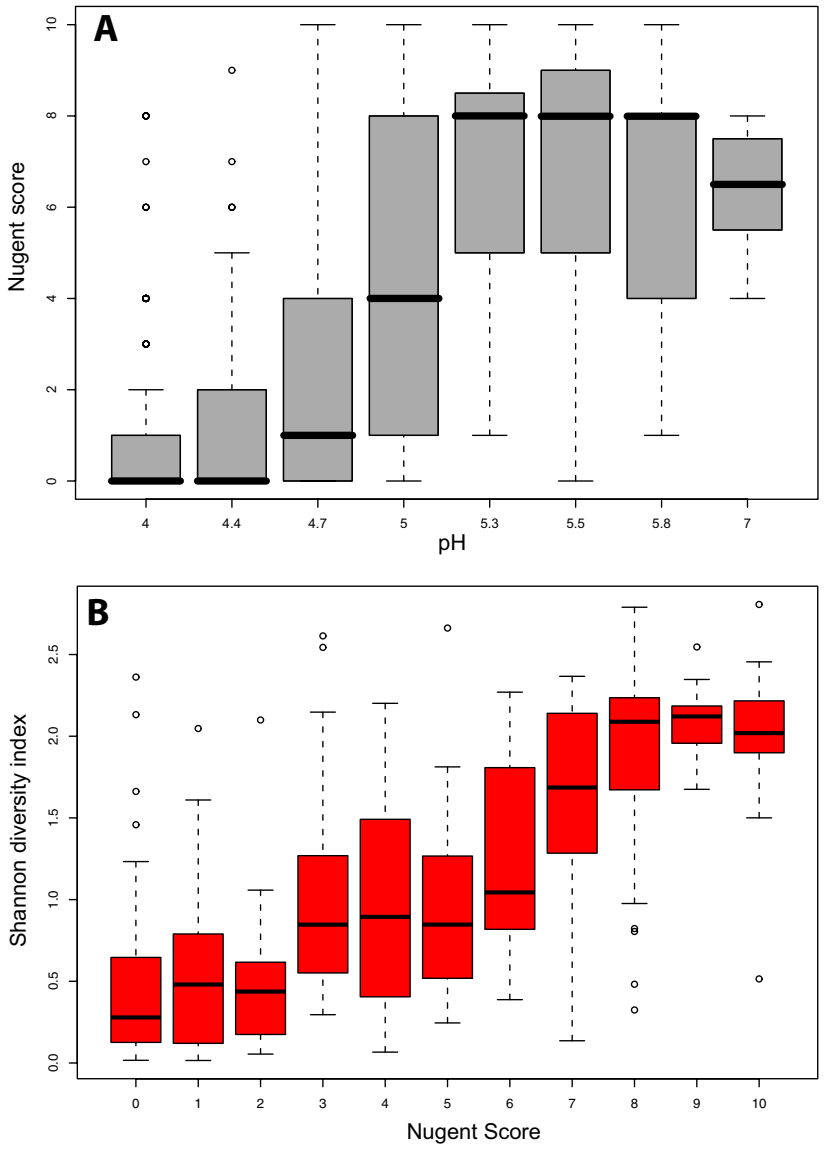

Fig. S4. Boxplots showing the positive correlation between (*A*) Nugent scores and vaginal pH, and (*B*) bacterial community diversity (Shannon diversity indices) and Nugent scores.

Table S1. Characteristics of reproductive-age women enrolled the cross-sectional analysis of the microbial species composition and abundance of the vaginal microbiota, Baltimore, MD, and Atlanta, GA.

Table S1 (XLSX)

Table S2. Barcoded PCR primers used for the amplification of 16S rRNA genes.

Table S2 (XLSX)

Table S3. Percentage of samples within a community group containing a taxon

Table S3 (XLS)

Taxa are sorted alphabetically.
[a]Defined as the total number of subject/samples in a community group.
[b]Total number of taxa detected within all of the samples of a community group.

Table S4. Taxonomic assignments and metadata for each sample analyzed in the study

Table S4 (XLSX)

[a]Self-described ethnic group;.
[b]Low = 1–3, intermediate = 4–6, and high = 7–10.
[c]Community group as defined in Fig. 1 (main text).
[d]Total number of high-quality 16S rRNA gene sequence reads; e RDP taxonomic assignments.

Table S5. Abundance of taxa in each community group

Table S5 (XLS)

The 16S rRNA gene sequences from all of the samples belonging to a community group were used to calculate the abundance of each taxon within that group.

Table S6. Percentage of samples within a community group containing a taxa

Table S6 (XLS)

Taxa are sorted by percentage within each community group. Each worksheet also contains the percentage of samples containing a taxa in the entire set of samples analyzed.

Table S7. Spearman's correlation coefficients between the relative abundance of a taxon and Nugent score category (low, intermediate and high).

Table S7 (XLSX)

Table S8. Validation of HMM-based algorithm for species-level classification of *Lactobacillus* sp.

Table S8 (XLSX)