The structure and diversity of strain-level variation in vaginal bacteria

Brett A Tortelli
Amanda L Lewis
Justin C Fay

Follow this and additional works at: https://digitalcommons.wustl.edu/oa_4

Part of the Medicine and Health Sciences Commons

Please let us know how this document benefits you.
The structure and diversity of strain-level variation in vaginal bacteria

Brett A. Tortelli¹,², Amanda L. Lewis³,⁴† and Justin C. Fay⁵,*

Abstract
The vaginal microbiome plays an important role in human health and species of vaginal bacteria have been associated with reproductive disease. Strain-level variation is also thought to be important, but the diversity, structure and evolutionary history of vaginal strains is not as well characterized. We developed and validated an approach to measure strain variation from metagenomic data based on SNPs within the core genomes for six species of vaginal bacteria: Gardnerella vaginalis, Lactobacillus crispatus, Lactobacillus iners, Lactobacillus jensenii, Lactobacillus gasseri and Atopobium vaginae. Despite inhabiting the same environment, strain diversity and structure varies across species. All species except L. iners are characterized by multiple distinct groups of strains. Even so, strain diversity is lower in the Lactobacillus species, consistent with a more recent colonization of the human vaginal microbiome. Both strain diversity and the frequency of multi-strain samples is related to species-level diversity of the microbiome in which they occur, suggesting similar ecological factors influencing diversity within the vaginal niche. We conclude that the structure of strain-level variation provides both the motivation and means of testing whether strain-level differences contribute to the function and health consequences of the vaginal microbiome.

DATA SUMMARY
All vaginal metagenomic sequence data generated for this project can be found on the Sequence Read Archive under BioProject PRJNA639592.

INTRODUCTION
A diverse range of microbial communities have been found to be associated with human anatomical sites, including the skin, gastrointestinal tract, oral cavity and vagina [1]. Surveys of these microbial communities have demonstrated significant differences between anatomical sites but also variation among individuals [1, 2]. Inter-individual variation in microbial communities has in many instances been associated with a variety of host factors, including human health and disease, e.g. obesity and inflammatory bowel disease [3], leading to continued investigation of the implications of microbial variation.

Inter-individual variation has largely been explored by means of characterizing differences in species presence or relative abundance. However, prevalent species not only show differences in relative abundance but also exhibit appreciable strain-level variation [1, 2]. Individual strains may be unique to a person’s microbiome and bacterial strains of a species isolated from the same individual have been noted to be more similar to each other than strains isolated from different individuals [2, 4]. When examined, strain-level variation is characterized by functional differences, prominent among these are differences in metabolic potential and antibiotic resistance [2, 5, 6]. This suggests that strain-level variation may contribute to...
For certain vaginal bacterial species, functional, phenotypic differences in personal microbiomes observed between individuals. However, knowing the extent to which strain-level differences translate to functionally distinct strains remains an open and important question. Currently, most comparisons of microbial communities utilize operational taxonomic units as means of grouping similar strains together and differentiating them from other groups.

Strain differences and their relationships define the population structure of a species. Population structure is relevant for both grouping strains but also making inferences about their history. In the absence of recombination, strains continually diversify, but those lineages that are most successful will expand and others will be lost. Eventually such lineages can converge in function and even establish new species. Under the ecological species concept, two species cannot stably coexist unless they differ in their niche [7]. However, population structure can also be established by limited migration, in which case subpopulations may have the same functions in the community but diverge (neutrally) in their genome [8]. Although distinguishing functional populations from neutral populations is difficult, population structure remains an important component of describing groups of strains with shared functional differences or shared population history. As each human may carry or enable the formation of unique microbial strains, the characterization of population structures and their determinants is important to addressing the role of strain-level variation in the human microbiome [9].

Among human microbial communities, the vaginal microbiome differs in its community composition. Both 16S ribosomal profiling and metagenomic community profiling have indicated that the vaginal microbiome often exhibits lower community diversity when compared to other anatomical sites, frequently being dominated by a single species [1, 2]. The composition of the bacterial community is often described in terms of five common community types [10]. Four of the five community types are dominated by a single Lactobacillus species (Lactobacillus crispatus, Lactobacillus iners, Lactobacillus jensenii or Lactobacillus gasseri). The fifth community type is characterized by a lack of Lactobacillus dominance and often exhibits higher community diversity. This diverse community has been correlated with a high vaginal pH (>4.5) and bacterial vaginosis (BV) [10], a dysbiosis associated with the overgrowth of anaerobic bacteria including Gardnerella vaginalis and Atopobium vaginae. The prevalence of these vaginal communities vary by self-reported race/ethnicity [10, 11] and have been associated with reproductive health [12]. While community-type classification offers a convenient method for categorizing the overall composition of the vaginal microbiome, the significance of strain-level variation is of increasing interest.

For certain vaginal bacterial species, functional, phenotypic and genomic differences have been described among isolated strains. An example of this is the classification of G. vaginalis into distinct phylogenetic clades (groups) through genomic approaches such as gene ontology and genome-wide SNP analysis [13–15]. Characterization of individual strains have shown functional differences (including sialidase activity) between groups with phenotypic consequences [16–19]. Such functional differences may explain why some groups but not others have been associated with BV [20–22]. While less is known about other vaginal bacterial species, genomic analysis of L. crispatus and L. iners strains has provided some insight into the population structure of these species [4, 23, 24]. An examination of 41 strains found that L. crispatus may be comprised of two closely related groups [4], but identification of phenotypic differences between these groups is lacking. Additionally, the population structure of L. iners appears to lack strain groupings, but rather each strain appears to be distinct [4, 24]. These assessments of strain-level variation have focused on isolated strains, and assessments of strain-level variation within the vaginal microbiome have been limited [4]. The use of variable regions of the 16S gene to define genovarants has been used by some as a proxy for strain diversity [4, 25, 26]. However, the use of 16S genovarants to explore strain level variation and associations with health is limited by the resolution of genovarants and their correspondence to phylogeny.

A critical factor in evaluating strain-level variation is how it is measured. Early studies employed multilocus sequence typing (MLST) [27], but recombination and horizontal gene transfer (HGT) can cause results to differ depending on the loci employed. Strain-level variation has also been examined using gene ontology or copy-number-variation (CNV)
analysis [4, 6], which has the advantage that many CNVs are functionally important. However, CNV can be hard to detect in low-coverage samples and HGT can stimulate CNV [28]. Genome-wide SNP analysis has also been used [2, 4, 5], but limited reference genomes for some species and variation in genome content present challenges [29]. Furthermore, widely divergent species may have limited core genomes and alignment methods for such divergent species present significant difficulties [30]. Additionally, it can be challenging to distinguish between strains with mixed ancestry from multi-strain samples.

The goal of the present study was to define and compare the population structure of common vaginal bacteria and identify patterns of strain-level variation among vaginal microbiomes. We developed and validated a genome-wide SNP analysis based on available reference genomes. We applied this approach to metagenomic data from vaginal samples and found that diversity present among the vaginal samples was well represented by the available reference strains. We found species-specific differences in strain variation and structure, identifying clear groupings within most of the species. Although our power was limited, no strong associations between strain and host factors were identified. Together, our results provide insight into how vaginal microbiome community types developed over the course of human history and lay the groundwork for assessing the importance of strain-level variation in the vaginal microbiome and human health.

METHODS

Metagenomic sequencing of vaginal samples

We obtained 197 cervicovaginal swabs from 195 pregnant women: 25 cervical swabs and 142 vaginal swabs (collectively referred to as vaginal samples) through the Global Alliance to Prevent Prematurity and Stillbirth (GAPPS) biobank and 30 vaginal swabs from the Women and Infants Health Consortium (WHISC) at Washington University in St. Louis (IRB no. 201610121). When selecting samples from the GAPPS biobank, efforts were made to: (1) select all available specimens from women who delivered preterm (<37 weeks of gestational age), (2) increase the representation of specimens from women of non-White race/ethnicity among the cohort, and (3) balance samples across all three trimesters. We augmented the samples selected from GAPPS with samples obtained from women currently enrolled in other studies with WHISC. Patient data including gestational age at time of swab collection, gestational age at the time of birth, birthweight, maternal age and race/ethnicity were obtained from GAPPS and WHISC. Women who delivered prior to 37 weeks of gestational age were considered preterm and represented both spontaneous and indicated preterm delivery. To extract genomic DNA, frozen vaginal swabs were eluted in 250 µl of an enzyme solution containing 0.5 mg ml⁻¹ lysozyme, 150 units ml⁻¹ mutanolysin, 12 units ml⁻¹ lysisostaphin, 0.025 units ml⁻¹ zymolase in 0.05 M potassium phosphate buffer (pH 7.5) and incubated for 1 h at 37 °C. A ZR Fungal/Bacterial DNA MicroPrep Kit (Zymo Research) was used to extract and purify genomic DNA from swab elutions. Metagenomic sequencing libraries were prepared with a Nextera DNA Sample Prep Kit (Illumina) using 5 ng of genomic DNA and a small volume protocol [31]. PCR was performed using KAPA Hi-Fr HotStart ReadyMix (KAPA Biosystems) and libraries were purified with Ampure XP magnetic beads (Beckman Coulter). Libraries were pooled and sequenced on an Illumina NextSeq platform (75 cycles).

Sequence processing and classification of the vaginal microbiome

Sequence reads from our vaginal samples were trimmed and quality filtered using fastqtrim (version 0.9.7) to remove reads less than 50 base pairs in length and trim read ends where quality scores drop below 10. Reads were then aligned to the human genome with Bowtie2 [32] (version 2.3.4) and human reads were discarded. Metagenetic data from 128 vaginal specimens collected as part of the Human Microbiome Project [1] were obtained from NCBI’s Sequence Read Archive (SRP002163). Data were filtered to remove reads less than 50 base pairs in length and remove human reads comprised of Ns using fastq-mcf (version 1.04.803). Taxonomic profiling was performed on non-human reads using MetaPhlAn2 [33] (version 2.6.0). Each microbiome was classified into community types based on the dominant Lactobacillus species present, defined as 50% relative abundance or greater and referred to as, ‘L. crispatus-dominant’, ‘L. iners-dominant’, ‘L. gasseri-dominant’ or ‘L. jensenii-dominant’. Communities without a single Lactobacillus species reaching 50% were referred to as ‘diverse’, as most communities without an abundant Lactobacillus species have high species diversity [10]. Read data for all vaginal samples were aligned with BWA and Stampy as described below.

Description of our clinical cohort

For the 195 women in our study, we obtain clinical and demographic data. Data on self-reported race/ethnicity showed most (96%) reported White, Black, Hispanic or Asian. The remaining women (4%) reported either American Indian/Alaskan Native, multiple races or their race/ethnicity was unknown. Maternal age (years), gestational age at sample collection (days), birthweight (grams) and gestational age at delivery (days) was also collected. Preterm delivery was defined as delivery prior to 37 weeks. Sixty-nine (35%) women had L. crispatus-dominant microbiomes, 53 (27%) had L. iners-dominant microbiomes, 9 (5%) had L. jensenii-dominant microbiomes, 9 (5%) had L. gasseri-dominant microbiomes and 55 (28%) had diverse microbiomes. A summary of clinical and demographic data can be found in Table S1 (Supplementary Data Sheet 1 available in the online version of this article). We noted a higher prevalence of L. crispatus-dominant microbiomes among White (42%) than Hispanic (32%) or Black (20%) women; a higher prevalence of L. iners-dominant microbiomes among Hispanic (39%) and Black (35%) than White women (19%); and a higher prevalence of diverse microbiomes among Black (45%) than White (25%) or Hispanic (25%) women (Table S2).
Reference strain analysis and validation

As reference for the vaginal samples and to identify the core genome we obtained genome assemblies for reference strains from NCBI for the six bacterial species of interest. A total of 101 *G. vaginalis*, 60 *L. crispatus*, 21 *L. iners*, 18 *L. jensenii*, 31 *L. gasseri* and 5 *A. vaginae* assemblies were obtained (Table S3). Two *G. vaginalis* strains were not included in our analysis: UMB0388, which mapped extensively to other genomes, suggesting the assembly was not a pure isolate; and 6420LIT, which had a particularly small genome when compared to all other *G. vaginalis* genomes. ART-MountRainer [34] (version 2.5.8) was used to generate simulated Illumina data (75 base-pair reads, NextSeq500 platform v2) at 20× coverage for each assembly.

Simulated read data were aligned to a concatenated reference database containing a representative assembly for each species (Table S4). A concatenated database was used in order to eliminate reads with low mapping quality due to equivalent mappings to multiple species. Alignments were performed with Stampy [30] (version 1.0.32) using the BWA-facilitated option and with an expected divergence of 0.05. We used this divergence parameter since higher rates of divergence (0.10) in some cases decreased the size of our filtered dataset, presumably due to higher rates of reads mapping to multiple reference sequences. Alignment data was filtered to remove reads with a mapping quality score of less than 10 using Samtools [35] (version 1.9).

We identified SNPs among all reference strains for a species and removed all variants not present within the core genome. The core genome was defined as all sites that had coverage for all reference strains in each species, and represented 12% (*G. vaginalis*), 47% (*L. crispatus*), 82% (*L. iners*), 72% (*L. jensenii*), 72% (*L. gasseri*) and 24% (*A. vaginae*) of the genome. SNPs were called using GATK UnifiedGenotyper [36, 37] (version 3.6.0). Variant calls were filtered using GATK to remove variants that met the following criteria: QD<50, FS>60.0, MQ<40.0, MQRankSum <-12.5, and ReadPosRankSum<-8.0. A small number of variants were removed due to cross-species read mapping. These sites were called based on reads from one species mapping to an incorrect reference genome, e.g. variant calls in the *L. crispatus* genome based on simulated reads from an *L. iners* reference assembly. Variant selection and removal was completed using VCFtools [38] (version 0.1.14). Among the reference strains, variant sites represented 15% (*G. vaginalis*), 3% (*L. crispatus*), 3% (*L. iners*), 2% (*L. jensenii*), 4% (*L. gasseri*) and 17% (*A. vaginae*) of the core genome.

To determine whether the choice of reference genome for read mapping affected the relationship between reference strains we aligned all simulated reference-strain data to a second set of alternative reference genomes (Table S4) and genotyped variants as described above. Variant sites with more than two alleles were then removed from both the original and alternate call set using VCFtools [38] (version 0.1.14) and genotypes were extracted using a custom script. A Euclidean distance matrix of strains for each species and each mapping reference was generated and compared by a Mantel test in R (version 3.5.1). The distance matrices for each mapping reference were found to be significantly correlated (Pearson’s correlation coefficient>0.9, P<0.05) for each species, indicating that the choice of mapping reference genome did not alter the relationships among strains. Based on this finding we utilized a single mapping reference set (Primary Reference Set) for all analyses.

Alignment and SNP calling for vaginal samples

Sequence reads from vaginal samples obtained for this study (N=197) and HMP samples (N=128) were aligned to the reference genomes as described above. SNPs were independently called from alignment files for all 325 vaginal samples and 234 reference strains and filtered using GATK as described above. SNPs outside of the core genome and genotype calls with less than 4× coverage were removed with VCFtools. We removed vaginal samples with >50% missing sites, and subsequently variant sites with >10% missing genotypes across all samples. The resulting SNP dataset had a sample size of 668, with 234 from reference strains and 434 from vaginal samples (some vaginal samples had sufficient genotype calls in multiple species). Most samples (574/668) had an average read coverage over 10× at genotyped sites, and the average coverage was 53× in the vaginal samples (Table S5).

Population structure

Population structure was inferred using both principal component analysis (PCA) and ADMIXTURE [39] and compared to phylogenetic trees. ADMIXTURE assumes free recombination among sites via sexual reproduction but PCA does not infer subpopulations and admixture proportions [39]. Our rationale was that both methods can handle genetic exchange between populations and should thus complement one another if they produce similar results. Additionally, because recombination can disrupt phylogenetic relationships, we employed the results of ADMIXTURE/PCA to eliminate recombinant samples from our phylogenetic analysis (next section).

The core-genome variants were filtered to select for biallelic sites with VCFtools. Heatmaps of the hierarchically clustered variants and samples were generated in R using the heatmap function. PCA was performed on the variant data in R with the package ‘FactoMineR’. All principal components (PCs) explaining>5% variance were assessed for associations with host factors (see Statistical Analysis). For *L. iners* where no PC explained>5% of the variance, we assessed the PC that explained the most variance.

VCF files containing core-genome SNPs for reference strains and vaginal samples were converted to PLINK format with PLINK [40] (version 1.9). ADMIXTURE [39] (version 1.3.0) was then used to identify populations and infer ancestry. The number of populations (groups) was estimated based on the cross-validation (CV) error for the number of groups [K1...10]. The estimated number of groups (K) was identified as the point at which the CV error plateaued to a minimum.
Reference strains and samples with <90% of ancestry estimated to be derived from a single group were classified as mixed ancestry (Table S5).

For multi-strain *G. vaginalis* and *L. crispatus* samples we estimated subpopulation (group) abundance within vaginal samples using the relative allele depth at group-specific SNPs. Group-specific SNPs were defined as those with an allele frequency of 80% or more in one group but 20% or less in all other groups. Based on this designation we identified 22 (group 1), 210 (group 2), 19 (group 3) and 178 (group 4) SNPs out of 4884 SNPs in the 88 *G. vaginalis* reference strains with less than 10% mixed ancestry. For *L. crispatus* we identified 604 (group 1), 55 (group 2) and 75 (group 3) SNPs out of 4469 SNPs. Using these SNPs we extracted the allele depth supporting each nominally heterozygous genotype. The abundance of each group in each mixed sample was estimated by the average proportion of allele depth for each group-specific SNP.

**Phylogenetic analysis**

Reference strains without mixed ancestry were used for phylogenetic analysis. Fourfold degenerate synonymous SNPs were selected using SNPeff [41] (version 4.3T) and SNPsift [42] (version 4.0). The SNPs for each sample were concatenated into FASTA format with VCF-kit [43]. The number of fourfold synonymous sites surveyed was determined by identifying all fourfold synonymous sites within the core genome using the same filters as described above except both variant and non-variant sites were retained. A distance matrix of pairwise differences in fourfold synonymous SNPs/fourfold synonymous sites surveyed (4π) was used to generate a neighbour-joining (NJ) tree in R with the packages ‘ape’ and ‘phangorn’. We calculated Watterson's estimator of diversity [44] with the formula: \( \theta_w = \frac{S}{\bar{a}_n} \), where \( S \)=the number of SNPs, \( \bar{a}_n = \sum_{i=1}^{n-1} \frac{1}{i} \) and \( n \) is the sample size.

In parallel we identified zerofold degenerate non-synonymous SNPs and zerofold non-synonymous sites surveyed. The average pairwise difference in zerofold non-synonymous variant sites/zerofold non-synonymous sites surveyed (0π) were determined. Tajima's D was calculated as previously described [45].

For each species we estimated the average time to the most recent common ancestor in generations as \( t = d/(2\mu) \), where \( \mu \) is the mutation rate and \( d \) is the average or maximum pairwise distance between strains at synonymous sites. We used a bacterial mutation rate of \( 2 \times 10^{-10} \) mutations per base pair per replication from *Escherichia coli* [46]. We used an *in vitro* doubling rate of *G. vaginalis* (7.1 h) [47] to estimate a replication rate of approximately 3.38 generations per day for all of the species and convert time in generations to time in years.

**Statistical analysis**

All PCs explaining >5% variance were assessed for associations with host factors. For *L. iners* where no PC explained >5% of the variance, we assessed the PC that explained the most variance. Kruskal–Wallis and Spearman rank correlation tests were used to test for associations between PCs and host factors as appropriate. Due to multiple comparisons, a *P*-value below 0.001 was considered significant for associations with PCs and a *P*-value below 0.05 was considered significant for associations with microbiome community type. Statistical analysis was conducted in R.

**RESULTS**

To characterize the diversity and structure of variation within common vaginal bacterial species, we generated metagenomic data from 197 vaginal swabs. A median of 94% of metagenomic reads per sample mapped to the human genome (range 50–99%). After removing human reads, a median of \( 8.43 \times 10^5 \) reads remained per sample, providing adequate coverage of the microbial genome. An analysis of the composition of the microbiome using MetaPhlAn2 [33] indicated compositions similar to those described in prior studies with *Lactobacillus*-dominant and diverse community types: 71 (36.0%) were *L. crispatus*-dominant, 9 (4.6%) were *L. jensenii*-dominant, 53 (26.9%) were *L. iners*-dominant, 55 (27.9%) were diverse and 9 (4.6%) were *L. gasseri*-dominant (Table S1). The prevalence of these community types differed by self-reported race/ethnicity (Table S2).

Using MetaPhlAn2 community composition data, we identified species for strain analysis that were well represented among our vaginal samples. Six bacterial species (*L. crispatus, L. iners, L. jensenii, L. gasseri, G. vaginalis* and *A. vaginae*) showed a minimum relative abundance of 10% in at least ten samples. When metagenomic data was mapped to a set of reference genomes (Table S4), most of the samples (60%) showed more than 4x coverage to one or more of the bacterial species. This indicated that many of the samples had sufficient shotgun metagenomic data to identify SNPs and examine strain-level variation. To increase the number of metagenomic samples for strain analysis, we included data from an additional 128 vaginal samples collected as part of the Human Microbiome Project (HMP) [1]. A summary of the number of metagenomic samples included in the strain analysis is presented in Table 1.

A significant challenge to SNP identification from mixed metagenomic samples is being able to reliably call variants for the correct species. This is complicated by HGT, close relationships among the *Lactobacillus* species and variation in genome content. To address these issues, we generated simulated metagenomic read data from publicly available reference genomes (Table S3) for each of the bacterial species (99 *G. vaginalis*, 60 *L. crispatus*, 21 *L. iners*, 18 *L. jensenii*, 31 *L. gasseri* and 5 *A. vaginae*). We aligned the simulated metagenomic reads to the reference set and found low mean misalignment rates for each species: 0.011 (*G. vaginalis*), 0.020 (*L. crispatus*), 0.003 (*L. iners*), 0.018 (*L. jensenii*), 0.017 (*L. gasseri*) and 0.092 (*A. vaginae*). While infrequent, misalignment did result in a small number of SNPs being called to the wrong species (e.g. variant calls in the *L. crispatus* genome based on simulated reads from an *L. iners* reference assembly). These invalid SNPs...
were excluded from our analysis. Next we identified the core-genome SNPs for each species, based on sites represented among all reference strains. Within the core genome, we identified thousands of SNPs for each of the species (Table 2).

To evaluate strain variation among vaginal samples, we independently called SNPs among all vaginal samples and reference strains within core genomic regions. When the data were hierarchically clustered based on SNP profiles, clusters of strains were present for all of the species except *L. iners*. Notably, many of the vaginal samples contained numerous genotype calls with both alleles present (nominally heterozygotes) (Table S5), indicating the presence of multiple strains in a single sample (Fig. 1). We conservatively defined samples as having multiple strains present if more than 10% of the SNPs were called heterozygous. Using this definition, we identified a high proportion of multi-strain samples for *G. vaginalis* (63%) and *A. vaginae* (77%) and a lower proportion (<22%) for the *Lactobacillus* species (Table 1).

To compare strain diversity among vaginal samples and isolated reference strains we performed PCA of the core-genome SNPs. Plotting vaginal samples and reference strains by PCs revealed that most reference strains formed clusters and the reference strains represent much of the variation observed among the vaginal samples (Fig. 2).

PCA can distinguish different strain groups but it does not identify strains of mixed ancestry, which can occur through conjugation, transduction and transformation. To examine the structure of strain diversity within each species we used ADMIXTURE [39] to classify samples into groups (subpopulations) and identify samples with mixed ancestry to multiple groups. This analysis identified multiple groups for most of the vaginal species we studied: four *G. vaginalis*, three *L. crispatus*, three *L. gasseri* and two each for *L. jensenii* and *A. vaginae*. We did not identify any population structure for *L. iners*. These groups closely correspond to the clusters observed by PCA (Fig. 2).

Consistent with the high proportion of multi-strain samples, many of the vaginal samples were inferred to be mixtures of groups (Table S5). While ADMIXTURE is unable to distinguish vaginal samples with distinct strains from multiple groups, from samples with a single strain of mixed ancestry, it can identify reference strains with mixed ancestry. Among our reference panel, mixed ancestry was present but uncommon, representing 11 of 99 (11.1%) *G. vaginalis*, 8 of 60 (13.3%) *L. crispatus*, 0 of 18 (0.0%) *L. jensenii*, 1 of 31 (3.2%) *L. gasseri*, and 0 of 5 (0.0%) *A. vaginae* strains (Table S5).

Among the vaginal samples inferred to have mixed ancestry, many were multi-strain samples. In these samples, the relative abundance of reads supporting each allele should be indicative of strain frequency in a sample. We thus used allele-specific read counts of population-specific SNPs to quantify relative abundance of each group in mixed vaginal samples. This was done for *G. vaginalis* and *L. crispatus*, which both have well-defined groups based on

Table 1. Metagenomic vaginal samples included in strain analysis for each species

<table>
<thead>
<tr>
<th></th>
<th><em>G. vaginalis</em></th>
<th><em>L. crispatus</em></th>
<th><em>L. iners</em></th>
<th><em>L. jensenii</em></th>
<th><em>L. gasseri</em></th>
<th><em>A. vaginae</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total samples</td>
<td>67 (100)</td>
<td>137 (100)</td>
<td>115 (100)</td>
<td>53 (100)</td>
<td>32 (100)</td>
<td>30 (100)</td>
</tr>
<tr>
<td>From this study</td>
<td>42 (62.7)</td>
<td>71 (51.8)</td>
<td>70 (60.9)</td>
<td>20 (37.7)</td>
<td>16 (50.0)</td>
<td>16 (53.3)</td>
</tr>
<tr>
<td>From HMP</td>
<td>25 (37.3)</td>
<td>66 (48.2)</td>
<td>45 (39.1)</td>
<td>33 (62.3)</td>
<td>16 (50.0)</td>
<td>14 (46.7)</td>
</tr>
<tr>
<td>Multi-strain samples</td>
<td>42 (62.7)</td>
<td>29 (21.2)</td>
<td>22 (19.1)</td>
<td>1 (1.9)</td>
<td>3 (9.4)</td>
<td>23 (76.7)</td>
</tr>
</tbody>
</table>

Values are n (%)

Table 2. SNP counts and nucleotide diversity measures

<table>
<thead>
<tr>
<th></th>
<th><em>G. vaginalis</em></th>
<th><em>L. crispatus</em></th>
<th><em>L. iners</em></th>
<th><em>L. jensenii</em></th>
<th><em>L. gasseri</em></th>
<th><em>A. vaginae</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Core-genome sites</td>
<td>2.01E+05</td>
<td>1.00E+06</td>
<td>1.06E+06</td>
<td>1.19E+06</td>
<td>1.41E+06</td>
<td>3.52E+05</td>
</tr>
<tr>
<td>Core-genome SNPs</td>
<td>3.12E+04</td>
<td>2.61E+04</td>
<td>3.24E+04</td>
<td>1.84E+04</td>
<td>6.01E+04</td>
<td>6.14E+04</td>
</tr>
<tr>
<td>Fourfold degenerate SNPs</td>
<td>1.04E+04</td>
<td>6.60E+03</td>
<td>1.10E+04</td>
<td>5.49E+03</td>
<td>1.85E+04</td>
<td>2.33E+04</td>
</tr>
<tr>
<td>Zerofold degenerate SNPs</td>
<td>7.33E+03</td>
<td>8.36E+03</td>
<td>7.08E+03</td>
<td>3.04E+03</td>
<td>1.23E+04</td>
<td>1.70E+04</td>
</tr>
<tr>
<td>Average π₄</td>
<td>0.047</td>
<td>0.011</td>
<td>0.031</td>
<td>0.019</td>
<td>0.033</td>
<td>0.190</td>
</tr>
<tr>
<td>Average π₀</td>
<td>0.003</td>
<td>0.002</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.032</td>
</tr>
<tr>
<td>Average π₀/π₄</td>
<td>0.063</td>
<td>0.172</td>
<td>0.100</td>
<td>0.068</td>
<td>0.077</td>
<td>0.167</td>
</tr>
<tr>
<td>θₜ</td>
<td>0.068</td>
<td>0.012</td>
<td>0.026</td>
<td>0.011</td>
<td>0.028</td>
<td>0.206</td>
</tr>
<tr>
<td>Tajima's D</td>
<td>−1.06</td>
<td>−0.273</td>
<td>0.806</td>
<td>2.65</td>
<td>0.749</td>
<td>−0.610</td>
</tr>
</tbody>
</table>
Fig. 1. Hierarchically clustered core-genome variants for reference and vaginal samples for (a) *G. vaginalis*, (b) *L. crispatus*, (c) *L. iners*, (d) *L. jensenii*, (e) *L. gasseri* and (f) *A. vaginae*. The reference allele is indicated by black, the alternate allele is indicated by red and the presence of both alleles is indicated by yellow, missing data is indicated by white. To the left of each heatmap is a bar indicating the source of the samples either vaginal or reference strain, and the group to which the strain for each sample was assigned using ADMIXTURE. Samples with mixed ancestry or multi-strain samples from multiple groups are identified as 'Mixed Ancestry'.
Fig. 2. PCA of SNP data for (a) \textit{G. vaginalis}, (b) \textit{L. crispatus}, (c) \textit{L. iners}, (d) \textit{L. jensenii}, (e) \textit{L. gasseri} and (f) \textit{A. vaginae}. Each vaginal sample and reference strain is shown by their PC1 and PC2 coordinates. Vaginal samples are indicated by blue points and reference strains are indicated by black points. Samples that were identified as multi-strain are represented by triangles and single-strain samples are represented by circles. Subpopulation groups were determined by ADMIXTURE analysis and ellipses drawn to show vaginal samples and reference strains belonging to a single group.
DISCUSSION

Strain-level variation is thought to be functional, motivating fine-scale measurement of strain variation and testing for its association with reproductive health. In this study we developed and validated a reference genome-based analysis of metagenomic vaginal samples to study the structure of strain level variation within and between individuals. We find reference genomes encompass the majority of strain-level vaginal samples, thereby providing a means interpreting strain-level variation and structure in vaginal samples. Despite occupying the same environment, we find differences in strain-level variation, multi-strain samples, population structure and strength of selection among the vaginal species. Below, we discuss these results in relation to prior studies of strain-level variation, and the ecology and evolutionary history of vaginal strains relevant to identifying functional differences among groups and their role in human health.

Strain level variation

Our analysis of core-genomic SNPs provides fine-scale measures of strain-level variation and captures known and new aspects of population structure present in vaginal species. Previous genomic studies encompassed only reference genomes, were limited to smaller sample sizes, or did not accommodate multi-strain or admixed samples (Table S10). We made use of combined metagenomic and reference genomes to (1) survey metagenomic SNPs in core-genomic regions and establish that most metagenomic variation is captured in reference genomes, and (2) identify groups or subpopulations within each species while accounting for a number of mixed ancestry genomes and numerous multi-strain samples. Mixed ancestry due to genetic exchange can confound phylogenetic analysis of strain-level variation [48], and multi-strain samples are difficult to resolve due to the challenges of accurate assemblies from metagenomic data [49, 50]. While our approach does not resolve multi-strain samples into individual lineages, we make use of multiple allele genotype calls to estimate relative abundance of groups within multi-strain samples.

Inference of population structure and admixture can depend on the methods and data used. Our dataset includes some samples (94/668) with low (4–10×) coverage. Although these samples may have higher rates of genotyping error, they were intermingled with high coverage samples in each subpopulation and represented a small fraction of admixed strains (17/117). Thus, while the branch tips leading to low coverage samples may be slightly longer due to genotyping error, the exclusion of these samples would not affect the population structure of strain-level variation, Our analyses used PCA, ADMIXTURE and phylogenetics to identify population structure. While our use of ADMIXTURE violates the programme’s assumption of free recombination via sexual reproduction, the results were consistent with PCA. Additionally, population groups were further supported by phylogenetic groups after removing mixed ancestry strains. Thus, our results support a consistent picture of population structure with genetic exchange.
**Fig. 3.** Neighbour-joining trees of reference strains created from synonymous sites for (a) *G. vaginalis*, (b) *L. crispatus*, (c) *L. iners*, (d) *L. jensenii*, (e) *L. gasseri*, (f) *A. vaginae*. Branch lengths represent pairwise differences per site surveyed. Groups were determined by ADMIXTURE analysis. Select bootstrap values for nodes separating groups are shown. Bootstrap values indicate the number of supporting iterations out of 100 as calculated by resampling with replacement.
With the exception of strains showing mixed ancestry, the structure of bacterial strain diversity is largely consistent with prior studies (Table S10). We find no population structure of L. iners, consistent with previous genomic analyses that showed a highly conserved genome among L. iners strains with little difference in gene content [4, 24]. A lack of population structure does not convey a lack of strain diversity. Indeed, most L. iners strains identified appeared to be unique and average nucleotide diversity among L. iners strains was greater than that seen among L. crispatus strains (Table 1). The three L. gasseri groups we identified correspond to two previously defined groups with distinct gene content [51]. Notably, recent studies [52, 53] suggest our L. gasseri group 3 represents the closely related L. paragasseri. The three L. crispatus groups correlate with two previously reported groups described in a genomic analysis of 41 strains [4]. Notably, most vaginal samples harboured group 2 or group 3 strains, while reference isolates from avian hosts were common in the more diverse group 1. This suggests that group 1 strain colonization of the human vagina may be rare.

The four groups of G. vaginalis that we found encompass and are largely consistent with prior groups (Table S11) [4, 13–15]. However, a number of these previously defined groups correspond to strains we find to have mixed ancestry. Ahmed et al. proposed the division of the species into four groups after a phylogenetic analysis of the core genome of 17 isolated strains [13]. Subsequent studies of the cpn60 gene [17, 22, 54] as well as our strain group assignments are consistent with those described by Ahmed et al. [13]. One exception is that strains 1400E and 55152, which were assigned to group 1 but our analysis suggested were of mixed ancestry (mostly groups 1 and 2). However, assemblies may give the appearance of mixed ancestry if unknowingly generated using a mixture of two or more strains. More recent studies [4, 14, 15] have expanded the number of strains as well as the number of groups (Table S10). However, many of these new groups are comprised of strains our analysis indicates are of mixed ancestry. The placement of strains with mixed ancestry into a separate group is not incorrect; such groups may be functionally distinct. While our approach to inferring population structure does not place mixed ancestry strains into separate groups, our results provide insight into the historical origin of these mixed ancestry groups.

Mixed ancestry strains likely derive from genetic exchange between population groups. While we did not investigate the mechanism of exchange, G. vaginalis encodes predicted competence-promoting proteins, prophages and transposable elements. Moreover, prior studies have provided evidence of HGT [13, 55–57]. A recent analysis of G. vaginalis core and accessory genomes within and between populations suggests that population structure is maintained through barriers to genetic exchange [57]. Patterns of HGT appear to differ between population groups, with some groups displaying more evidence of genetic exchange than others. Interestingly, this analysis suggests that intergroup HGT may have been more common in the remote past and indicates that at least some genes (e.g. vaginolysin) are freely exchanged within and between groups [57].

Ecological diversity of the vaginal microbiome

Vaginal microbial diversity has been correlated with reproductive health and determinants of this diversity are of significant clinical interest. Species diversity within the vaginal niche is determined through ecological interactions within that environment. A key correlate of species diversity is vaginal pH. A vaginal pH less than 4.5 is thought of as healthy and is associated with low diversity, Lactobacillus-dominated communities [10, 58]. It is believed that through the production of lactic acid, Lactobacillus species are able to outcompete other vaginal bacteria and dominate that niche [58]. Among Lactobacillus-dominated microbiomes, multiple Lactobacillus species may be present but a single Lactobacillus species usually dominates [10]. This suggests that these species may occupy very similar niches within the vagina. According to ecological theory, multiple species cannot occupy the same niche indefinitely and one species will eventually outcompete the others [59]. Conversely, a more neutral pH correlates with greater diversity and an abundance of BV-associated anaerobes including G. vaginalis and A. vaginae [10]. These polymicrobial communities support multiple species, which may be explained by the theory of resource partitioning in which competing species utilize different subsets of resources to occupy niche divisions within an environment [60]. However, such a co-occurrence could also result from spatial structure within the vagina [61].

We find that patterns of strain-level variation mimic those of species-level diversity. We observed greater strain-level diversity within G. vaginalis and A. vaginae, which are found in more species-diverse communities. Furthermore, most samples with G. vaginalis (62.7%) and A. vaginae (76.7%) harbour multiple strains from different groups, while multi-strain samples among lactobacilli are much less common (2–21%). The high frequency of multi-strain samples is consistent with prior studies of G. vaginalis [4, 20–22] and A. vaginae [62]. The co-occurrence of different groups can be explained by ecotype theory, which suggests that different strains of the same species may occupy the same niche if they function as different ecological species (ecotypes), exploiting different resources [7, 18].

The co-occurrence of differentiated groups within vaginal communities is important for understanding group associations with health. The presence of multiple groups of G. vaginalis has been correlated with BV [20, 22]. We find that the frequency of group 4 is negatively correlated with the frequency of all other groups in mixed vaginal samples, potentially indicating that it competitively excludes these groups. In contrast, groups 1, 2 and 3 co-occur but only 1 and 3 are negatively correlated. This is particularly interesting as the co-occurrence of multiple groups of G. vaginalis has been correlated with BV [20, 22]. Additionally, prior studies have failed to show an associated between group 4 and BV [20, 22], which may indicate that group 4 strains are less pathogenic. These findings suggest that mixed group communities may confound G. vaginalis group associations with vaginal health and should be accounted for in future models.
While a low pH may enable Lactobacillus species to exclude high pH species from the vaginal niche, this does not explain why multiple Lactobacillus strains are not observed more frequently in the same sample. If Lactobacillus groups represented distinct ecotypes, one would expect groups to co-occur as observed with G. vaginalis and A. vaginae. One potential explanation is that there has not yet been enough time for Lactobacillus strains to diversify and evolve resource partitioning strategies.

Evolutionary origins of strain level diversity

Strain-level diversity is indicative of the species’ demographic history, including past changes in population size, population structure and migration between host microbiomes or other environments. Some insight into the evolutionary origins of the vaginal microbiome may be gleaned by comparing it to microbiome composition of other primates. While vaginal microbial signatures of non-human primates are unique to each species, community compositions more closely resemble the diverse structure associated with G. vaginalis and A. vaginae [63–65]. Only humans are dominated by Lactobacillus species [63]. While Lactobacillus species are closely associated with food and agriculture [66], the species that dominate the human vagina have reduced genome sizes when compared to other Lactobacillus species suggesting adaptation to the host environment [67].

Among the species studied here, A. vaginae and G. vaginalis exhibited the greatest strain divergence, which is consistent with large, long-term populations of G. vaginalis and A. vaginae in humans as part of an ancestral state. The Lactobacillus species showed much less diversity, which may reflect a smaller historic population size or more recent colonization during human history. Among the Lactobacillus species, L. iners and L. gasseri are the most divergent. The diversity of L. iners may reflect a larger historic population size and earlier association with human vaginal microorganisms compared to other Lactobacillus species. This idea is consistent with it being the most prevalent (most frequently detected) of the Lactobacillus species [10]. L. gasseri strain divergence is also greater than the other Lactobacillus species, but this is partly caused by strong divisions between groups, which could predate colonization of the human vaginal microbiome. L. jensenii has low diversity and a large positive Tajima’s D, consistent with a recent bottleneck, potentially related to colonization of the vaginal niche. Together, these differences in strain divergence support the hypothesis that the vaginal microbiome of ancestral humans was more similar to modern diverse communities, and that the emergence of Lactobacillus species is a more recent event in the evolution of the human vaginal microbiome.

The relative differences in divergence prompted us to examine the timescale over which strain divergence may have occurred and align this timescale with the migration of human populations out of Africa. Using an experimental estimate of mutation rate from E. coli [47] and an in vitro estimate of G. vaginalis replication rate [48], we estimated average time to most recent common ancestor using pairwise divergence between the two most distantly related strains of a species. This analysis gave a divergence time of 800 (A. vaginae), 180 (G. vaginalis), 144 (L. gasseri), 72 (L. jensenii), 71 (L. iners) and 31 (L. crispatus) thousand years. However, these divergence times could be off by a factor of two or more since they depend on a general estimate for replication time and mutation rate that likely differ from the true values at the species and possibly even the strain level. As such, these numbers should be interpreted with caution. Our estimates indicated A. vaginae and G. vaginalis groups likely diverged prior to the migration of modern humans out of Africa [68]. Of the two most commonly found Lactobacillus species, our estimates suggest that L. iners diverged around the time when modern humans were beginning to disperse out of Africa [68], while L. crispatus diverged after the time that it is believed modern humans settled Europe [65]. These observations seem to parallel earlier findings from many research groups showing that vaginal microbiomes with an abundance of G. vaginalis and/or L. iners are more common in women of African descent, whereas an abundance of L. crispatus within the vaginal microbiome is more commonly found in women of European descent [10, 11, 66].

CONCLUSION

Our results show that most species are characterized by multiple distinct groups of strains, and that strain diversity and the frequency of multi-strain samples is related to species-level diversity of the microbiome in which they occur. Future work will need to uncover the ecological variables that impact variation within and between communities at the strain level, and the historical genomic and functional differentiation that led to extant population structure. Doing so will not only help resolve the role of strain variation in the vaginal microbiome as it relates to reproductive disease, but could also provide insight into the establishment and subsequent changes in community composition as a function of important gynecologic and obstetric events including: sexual development, the menstrual cycle, pregnancy and menopause [69].

Funding information
This study was supported the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the NIH: F30HD094435, a pilot research grant from the Centre for Women’s Infectious Disease Research at Washington University School of Medicine and the March of Dimes Prematurity Research Centre at Washington University. The sponsors had no role in the study design, the collection, analysis or interpretation of data, or the decision to submit the article for publication. The content of this study is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions
Conflicts of interest

The authors declare that there are no conflicts of interest.

Ethical statement

Vaginal specimens were collected under appropriate IRB approval with informed consent and all human subject data was deidentified. Human metagenomic data were removed prior to analysis as described in the methods.

References

40. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575.


