Check for updates

RESEARCH ARTICLE

REVISED Genome wide association study of vaginal microbiota genetic diversity in French women

[version 2; peer review: 2 approved]

Samuel Alizon[1,2], Claire Bernat[2,3], Vanina Boué[2], Sophie Grasset[2], Soraya Groc[2,4], Tsukushi Kamiya[1], Massilva Rahmoun[2], Christian Selinger[2,5], Nicolas Tessandier[2], Marine Bonneau[6], Vincent Foulongne[4], Christelle Graf[6], Jacques Reynes[7], Michel Segondy[4], Vincent Tribout[7], Jacques Ravel[8], Nathalie Boulle[4], Carmen Lia Murall[9], Vincent Pedergnana[2], Jean-François Deleuze[10]

[1]CIRB, INSERM, Collège de France, Université PSL, Centre National de la Recherche Scientifique (CNRS), Paris, Île-de-France, France
[2]MIVEGEC, IRD, Université de Montpellier, Centre National de la Recherche Scientifique, Montpellier, Occitanie, France
[3]Institut de Génomique Fonctionnelle, Universite de Montpellier, Montpellier, Occitanie, France
[4]PCCEI, Inserm, University Montpellier, Montpellier, Occitanie, France
[5]Swiss Tropical and Public Health Institute, Basel, Switzerland
[6]Department of Obstetrics and Gynaecology, Centre Hospitalier Universitaire de Montpellier, Montpellier, France
[7]Department of Infectious and Tropical Diseases, Centre Hospitalier Universitaire de Montpellier, Montpellier, France
[8]Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, Maryland, USA
[9]National Microbiology Laboratory (NML), Montreal Public Health Agency of Canada (PHAC), Winnipeg, Canada
[10]Centre National de Recherche en Génomique Humaine (CNRGH),, Université Paris-Saclay, Évry-Courcouronnes, France

v2 **First published:** 15 Jul 2025, **5**:186
https://doi.org/10.12688/openreseurope.20462.1
Latest published: 11 Mar 2026, **5**:186
https://doi.org/10.12688/openreseurope.20462.2

Abstract

Background

The composition of the vaginal microbiota is known to be highly structured into five main community state types (CSTs) that are found in all human populations. Several associations between perceived ethnicity and the type of community have been reported but analyses of human genetic data, especially genome wide association studies (GWAS), remain limited and mostly rely on phenotypic traits rather than microbial DNA data.

Methods

Analysing genotyping data from in 168 women from the PAPCLEAR cohort study in France, we perform a GWAS looking for human

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 11 Mar 2026	✓ view	✓ view
version 1 15 Jul 2025	? view	? view

1. **Antonio Charlys da Costa**, University of São Paulo, São Paulo, Brazil
Tania Regina Tozetto-Mendoza, Universidade de Sao Paulo Instituto de Medicina Tropical de Sao Paulo, São Paulo, Brazil

genetic polymorphisms associated with vaginal microbiota community composition. For the latter, we use Simpson diversity and community state type (CST) as summary statistics to summarise 16S RNA metabarcoding data.

Results

We show that inverse Simpson diversity is the trait related to the vaginal microbiota that is best explained by the human genome. Furthermore, we identify several genomic regions associated with variations in this trait and show that the covariates associated with vaginal microbiota composition do not correlate with these genetic variants.

Conclusion

This is one of the first GWAS to use microbial genetic data instead of symptoms to characterise the vaginal microbiota. However, it remains limited because of the size of our cohort and our results call for more powered studies in terms of participants and genome coverage.

Plain Language Summary

In all the populations sampled across the world, the composition of the vaginal microbiota falls into five main community state types (CST), four of which are dominated by a single species of Lactobacillus bacteria. Furthermore, the relative frequency of these CSTs consistently varies according to covariates related to age or lifestyle. In particular, several studies report that CST frequency vary with self-reported ethnicity. However, to date, very few genome wide association studies (GWAS) have been performed on the diversity of the vaginal microbiota to identify genetic variants that could explain the observed variation.

Building on the PAPCLEAR cohort, which was set up in Montpellier (France) to study HPV infections and the vaginal microbiota, we perform a GWAS using the Shannon diversity of the vaginal microbiota as our response trait. We identify two genomic regions associated with microbiota diversity variations.

This is one of the first GWAS to use microbial genetic data instead of symptoms to characterise the vaginal microbiota and our results call for more powered studies in terms of participants and genome coverage.

Keywords

vaginal microbiota, human genetics, GWAS, SNP, variants, diversity, CST

2. **Luisa W. Hugerth**, Uppsala University, Uppsala, Sweden

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the European Research Council (ERC) gateway.

This article is included in the Horizon 2020 gateway.

Corresponding author: Samuel Alizon (samuel.alizon@cnrs.fr)

Author roles: Alizon S: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Bernat C**: Resources; **Boué V**: Resources; **Grasset S**: Resources; **Groc S**: Resources; **Kamiya T**: Data Curation; **Rahmoun M**: Resources; **Selinger C**: Data Curation; **Tessandier N**: Data Curation; **Bonneau M**: Resources; **Foulongne V**: Resources, Supervision; **Graf C**: Resources; **Reynes J**: Resources, Supervision; **Segondy M**: Resources, Supervision; **Tribout V**: Resources, Supervision; **Ravel J**: Resources, Supervision; **Boulle N**: Resources, Supervision; **Murall CL**: Conceptualization, Data Curation, Funding Acquisition; **Pedergnana V**: Conceptualization, Methodology, Writing – Review & Editing; **Deleuze JF**: Funding Acquisition, Resources, Supervision

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 648963, to SA).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Alizon S *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Alizon S, Bernat C, Boué V *et al.* **Genome wide association study of vaginal microbiota genetic diversity in French women [version 2; peer review: 2 approved]** Open Research Europe 2026, **5**:186
https://doi.org/10.12688/openreseurope.20462.2

First published: 15 Jul 2025, **5**:186 https://doi.org/10.12688/openreseurope.20462.1

REVISED **Amendments from Version 1**

As detailed in the Supplementary files describing the response to the reviewers, in this revised version, we redid the whole analysis and performed some major structural changes to the manuscript.

First, and most importantly, following the (very insightful) suggestion from a reviewer, we used Simpson diversity instead of Shannon diversity as a summary statistics for vaginal microbiota diversity. This led to much stronger results and resulted in the updating of most of the results section (especially Figure 3 and 4, as well as Table 3, 4, and 5).

Second, thanks to the accurate reviewing of the other reviewer, we noticed an issue with our metadata table. We therefore checked all the data, which led to the correction of one patient CST and minor adjustments for diversity metrics. These updates were of course included in the new analysis and led to the updating of Figure 2.

Third, we now explicitly discuss the fraction of the variance that is explained by human genetics with the addition of the new Table 2.

Finally, we restructured the whole text to follow the suggestions of the reviewers. The Introduction should now avoid repetitions. Furthermore, the Methods and the Discussion better stress the difficulty to work with microbiota as a response trait for GWAS and make the underlying assumptions more explicit. More generally, we also give more details and references about the methods used.

Any further responses from the reviewers can be found at the end of the article

Introduction

The vaginal microbiota has been studied since the end of the 19th century, which is consistent with the massive impact it is now known to have on women's susceptibility to sexually-transmitted infections,[1] fertility,[2] and general well-being.[3] The vaginal microbiota is also highly structured into five main community state types, or CSTs,[4] which tend to be relatively stable over several months.[5,6] Its composition has also been long reported to be associated with perceived ethnicity.[4,7–9]

Contrarily to other human microbiota,[10] only a few studies, which are detailed below, have attempted to identify potential genetic factors associated with the observed variation in human vaginal microbiota. This is surprising since all studies conducted around the globe detect the same five main CSTs. Four of these communities are dominated, sometimes exclusively, by a single species of *Lactobacillus* (*L. crispatus* for CST-1, *L. jenseni* for CST-2, *Lactobacillus iners* for CST-3, and *Lactobacillus gasseri* for CST-5). Conversely, the fifth one, CST-4, is a diverse assemblage of anaerobic bacteria from genera such as *Gardnerella*, *Prevotella*, or *Fannyhessea.* It is also often associated with bacterial vaginosis (BV) and other health issues.[24,4,1,28] The relative proportion of each CST can vary widely between populations, although CST-1, CST-3, and CST-4 are always the most frequent. In North America, CST-4 was found to be four times more common in women who identify themselves as 'Black' or 'Hispanic' than in those who identify themselves as 'White'.[4] These differences could have behavioural origins. For example, vaginal douching is known to be strongly associated with CST-4 and this practice varies across populations.[15] However, to date, we do not know to what extent this difference is associated with human genetics. One of the motivations to identify potential single nucleotide polymorphisms (SNPs) associated with CST-4, also sometimes referred to as 'molecular bacterial vaginosis',[16] is that half of the time they are not associated with symptoms, which could be consistent with some women being able to better tolerate these CSTs than others.

To our knowledge, at least four Genome Wide Association Study (GWAS) studies have been conducted to identify the genetic bases for the observed variations in vaginal microbiota composition. All are recent and tend to have limitations. One of the first studies dates from 2020 and involved 359 pregnant women in China to perform a GWAS where the response trait was the presence or absence of specific bacterial species based on 16S RNA data.[13] Despite its statistical power and use of microbial DNA data, a limitation of this study is that pregnancy is known to affect vaginal microbiota composition with an increased presence of lactobacilli.[29] Another exception comes from a 2021 study in Kenya that performed a GWAS on 171 women,[12] using as a response trait the presence or absence of one of three bacterial species, the Shannon diversity, or the CST. In 2024, an analysis was performed in the USA on 686 women living with and at risk for HIV infection, the trait of interest being the presence or absence of bacterial vaginosis (BV).[11] However, this analysis was focused on 627 SNPs across 41 genes important in mucosal defense identified from a previous GWAS. Finally, again in 2024, a study performed a GWAS on 12,815 women living in Estonia and identified SNPs associated with recurrent vaginitis.[14] Overall, besides their rarity, these earlier studies illustrate the challenge represented by the analysis of microbiota data in human genomics association studies. Most of these rely on symptoms and few consider the genetic diversity of the bacterial population; even though the low species richness of the human vaginal microbiota makes it an ideal system for designing relevant summary statistics.

In this study, we perform a GWAS on 168 women from the PAPCLEAR cohort, which was implemented to analyse human papillomavirus infections,[17] using longitudinal vaginal microbiota metabarcoding data (16S) as a response trait. This work is of limited statistical power but stands out in terms of the response traits considered and of the study population.

Methods

Data generation

189 participants were enrolled in the PAPCLEAR study, which aimed to understand the human papillomavirus kinetics in the vaginal environment.[18]

DNA was extracted using the QIAamp® DNA mini kit (QIAGEN Inc.) from PBMCs extracted from 10 mL of circulating blood following standard protocol for body fluids. Eleven samples were lost during the extraction process, and DNA could not be retrieved.

We then genotyped 168 participants using an Illumina Global Screening Assay Multi-Disease (GSA-MD) genotyping assay, which targets approximately 800,000 SNPs in the genome.

For each of the participants, we used vaginal microbiota 16S metabarcoding data described in an earlier study.[5]

Data curation

In order to perform a GWAS, we had to transform the vaginal microbiota data into a single summary statistic. Given the uneven number of samples for each participant and given that prior studies show that a single sample predicts with good accuracy the vaginal microbiota trajectory over the following 18 months,[6] we averaged all the visits to obtain a mean dataset for the relative abundance of 372 bacterial species identified from the 16S sequence data using the SpeciateIt software package.[30] We analysed this dataset for community state type (CST) composition using the VALENCIA nearest centroid classifier.[19] Intuitively, this algorithm uses a large reference dataset (3,160 taxonomic profiles from 1975 women) to compute a distance to reference CSTs for any sample based on the relative frequency of up to 199 bacterial taxa. From this, we could obtain the distance to each centroid, as well as the CST assignment. We used this data to compute two traits: a binary CST assignment (CST-4 *vs.* non-CST-4) and a quantitative shortest distance to CST-4.

Using the R diversity function from the vegan package with the mean relative abundance data, we computed we computed Shannon and Simpson diversity indexes, which are given by the following formula, with p_i denoting the relative frequency of species i:

$$D_{sh} = -\sum_{i \in s} p_i \ln(p_i)$$

$$D_{si} = \frac{1}{\sum_{i \in S} p_i^2}$$

While Shannon's entropy (D_{sh}) is widely used in microbiome research, Simpson's diversity (D_{si}) places more weight on evenness and is, therefore, more sensitive to changes in the low-richness communities.

Genome Wide Association Study

We performed the association study itself using plink2, using the linear regression with the --covar option to use the first ten MDS components as covariates. Only the results assuming an additive genetic effect were kept, and SNPs with p-values higher than 0.01 were ignored for further analyses.

SNP imputation

In order to enrich our dataset, we imputed genomic positions not directly covered by the GSA-MD assay using IMPUTE2.[20] For this, we downloaded the reference genome v.38 and exported each chromosome from the main file using plink v.2.0.[21]

Quality control

We performed classical steps for quality control, as summarised in Ref 22, namely, we:

- removed variants that were too rare or participants with too many missing variants with a 2% threshold;
- removed variants with a minority allele frequency with a 5% threshold;
- checked for a Hardy-Weinberg equilibrium of the variants;
- removed individuals with more than 2 standard deviations in terms of heterozygosity;
- checked for cryptic relatedness, assuming a π threshold value of 0.2.

Using the 1000 Genomes database reference, we also investigated the shared ancestry between our participants. For this, we performed a multidimensional scaling (MDS) clustering, pooling our PAPCLEAR dataset with that from the 1000 Genomes.

The output of this quality control is available in the supplementary HTML file at https://doi.org/10.57745/LGZANK

Heritability estimation

To help identify the most suitable phenotypic trait for GWAS investigation, we estimated the variance explained by all the SNPs using GCTA,[23] which implements a restricted maximum likelihood (REML) approach.

Results

Cohort description

The PAPCLEAR study enrolled women aged from 18 to 25 years old, living in the area of Montpellier (France), and who were primarily university students (119/138, 86%) at the time of the study. The main characteristics of the study population have been reported in earlier studies and some are shown in Table 1 for the samples analysed here, with a stratification in terms of vaginal microbiota composition. Most of the associations observed are consistent with earlier studies since women with a CST-4, i.e. *Lactobacillus*-poor vaginal microbiota, report lower condom usage from the male partner, higher lifetime number of sexual partners, and identify themselves less as 'caucasian'.[24] Their body mass index (BMI) and age are also higher, and they have more often been pregnant. In terms of CST differences, the most frequent is CST-1, followed by CST-3, and then CST-4 (which is *Lactobacillus*-poor). As expected, Shannon and Simpson diversity indexes are much lower in the *Lactobacillus*-dominated microbiotas.[4]

Table 1. Cohort profile stratified by vaginal microbiota composition. Numerical variables are centred and scaled. The p-value indicated the outcome of a Fisher exact test for categorical variables and a Kruskal-Wallis rank sum test for continuous variables (the interquartile range is shown in brackets). Significant differences are in italic.

	CST-4	*Lactobacillus*	p-value
Number of participants	19	149	
female affinity (%)	4 (21.1)	26 (17.6)	0.751
male affinity (%)	18 (94.7)	145 (99.3)	0.218
smoker (%)	10 (76.9)	61 (70.9)	0.754
regular sport practice (%)	7 (36.8)	77 (51.7)	0.330
chlamydia infection (%)	0 (0.0)	6 (5.2)	1.000
menstrual cup user (%)	3 (15.8)	24 (16.1)	1.000
vaginal products user (%)	8 (42.1)	69 (46.3)	0.810
tampon user (%)	8 (42.1)	45 (30.2)	0.304
previous pregnancy (%)	3 (15.8)	5 (3.4)	0.049
vaginal douching (%)	1 (5.3)	3 (2.0)	0.384
male partner using condom (%)	4 (21.1)	84 (56.4)	0.006
identifies as Caucasian (%)	11 (57.9)	125 (83.9)	0.012

Table 1. *Continued*

	CST-4	*Lactobacillus*	p-value
alcohol consumption	−0.07 [−0.52, 0.39]	−0.07 [−0.52, 0.84]	0.583
body mass index	0.11 [−0.28, 1.23]	−0.35 [−0.73, 0.38]	0.030
age at inclusion	0.23 [0.23, 0.72]	−0.27 [−0.77, 0.72]	0.123
age at first menstruation	−0.18 [−1.11, 0.19]	0.19 [−0.55, 0.93]	0.244
lifetime number of partners	0.30 [−0.38, 1.45]	−0.38 [−0.67, 0.30]	0.020
microbiota Shannon diversity	1.67 [1.01, 2.11]	0.46 [0.22, 0.76]	<0.001
microbiota Simpson diversity	3.46 [2.02, 5.14]	1.30 [1.08, 1.63]	<0.001
CST (%)			<0.001
I.A	0 (0.0)	60 (40.3)	
I.B	0 (0.0)	24 (16.1)	
II	0 (0.0)	2 (1.3)	
III.A	0 (0.0)	39 (26.2)	
III.B	0 (0.0)	19 (12.8)	
IV.A	3 (15.8)	0 (0.0)	
IV.B	15 (78.9)	0 (0.0)	
IV.C	1 (5.3)	0 (0.0)	
V	0 (0.0)	5 (3.4)	

The genetic composition of the large majority of the study population corresponds to the individuals from European ancestry in the 1000 genomes database (Figure 1). Given the limited size of the sample, we chose not to remove the participants far away from the European-ancestry cluster but used the MDS coordinates as covariates in the association study as a control.

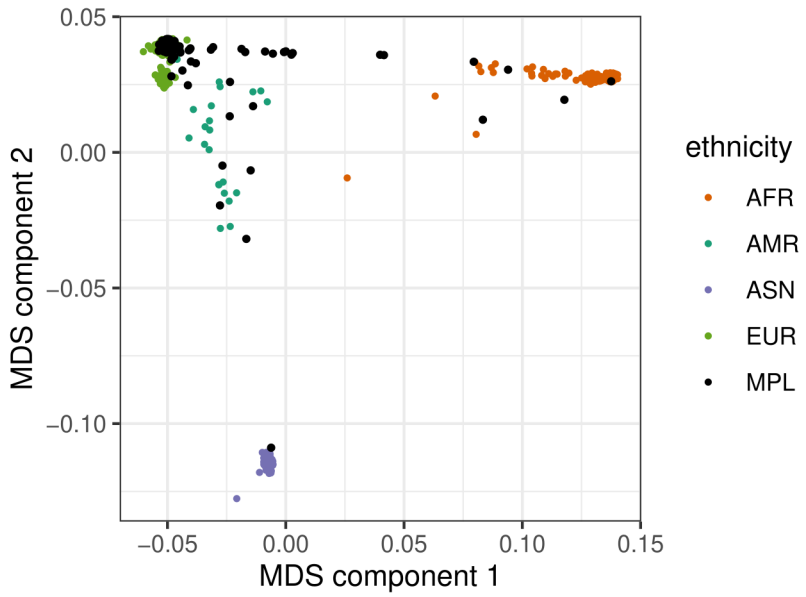

Figure 1. Multidimensional scaling (MDS) plot of population stratification. The four colours show the main ancestries for the 1000 Genomes database. The PAPCLEAR study participants from Montpellier ('MPL') are shown in black.

Genetic control over the traits

When considering the explanatory power of all the SNPs on of a trait of interest, the signal was limited when using the binary trait (CST-4 *vs.* non-CST-4) with an estimated genetic variance explained by all the SNPs of the genome of 0.09 (Table 2). For the Shannon diversity, which is highly correlated with the distance to CST-4 (Figure 2), this genetic variance was 0.24. For the Simpson diversity, it was even higher with a value of 1.80. Therefore, we focus on the latter trait in the following. Note that GCTA found little environmental variance, which is likely due to the small size of our dataset.

Each of the dots corresponds to one of the participants. The line shows the output of the linear model, showing the negative correlation between Simpson diversity and distance to CST-4. Colours show the different CSTs (CST-4 is poor in lactobacilli).

Association study

The genome wide association study, assuming an additive model and using the ten major MDS components as covariates, identified several specific genomic regions as being significantly associated with vaginal microbiota Simpson diversity, assuming a classical $5 \times 10–8$ threshold for p-value significance (Figure 3).

The most significant association was with several SNPs located in chromosome 10 and in the gene coding the Coiled-coil domain containing 3 (CCDC3) protein (Table 3). This secretory is expressed in vascular endothelial cells and has no

Table 2. Variance explained by the whole genome for microbiota-related traits. V(G) stands for genetic variance, V(e) for environmental variance, h2 is the fraction of the total variance explained by the genetic variance, log(L) is the log likelihood and P-value the p-value associated with the final estimate for the restricted maximum likelihood (REML) model (see Ref. 23 for details).

Source	CST-4 vs. non-CST-4	Simpson diversity	Shannon diversity
V(G)	0.09	1.80	0.29
V(e)	0.0	0.000003	0.0
h2	0.99	0.99	0.99
log(L)	109.50	−63.95	3.60
p-value	0.0013	0.0008	0.015

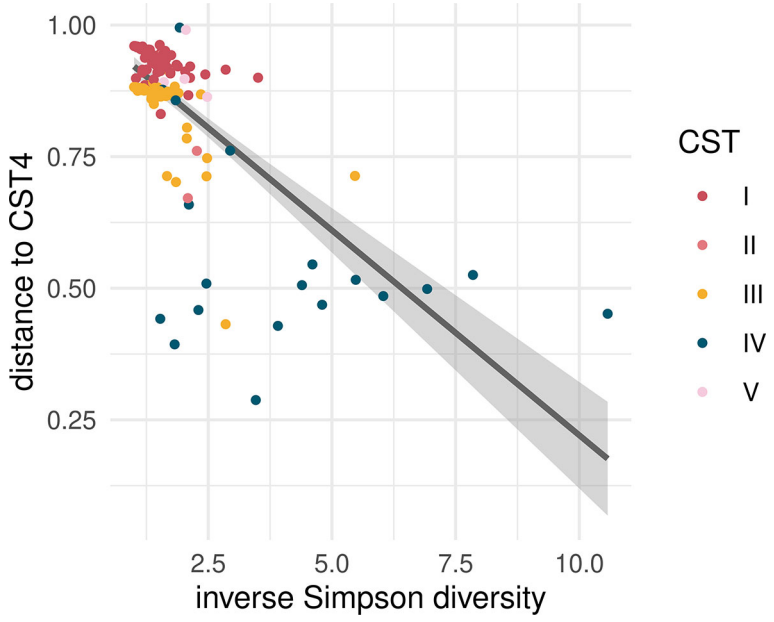

Figure 2. Distribution of the phenotypic traits of interest.

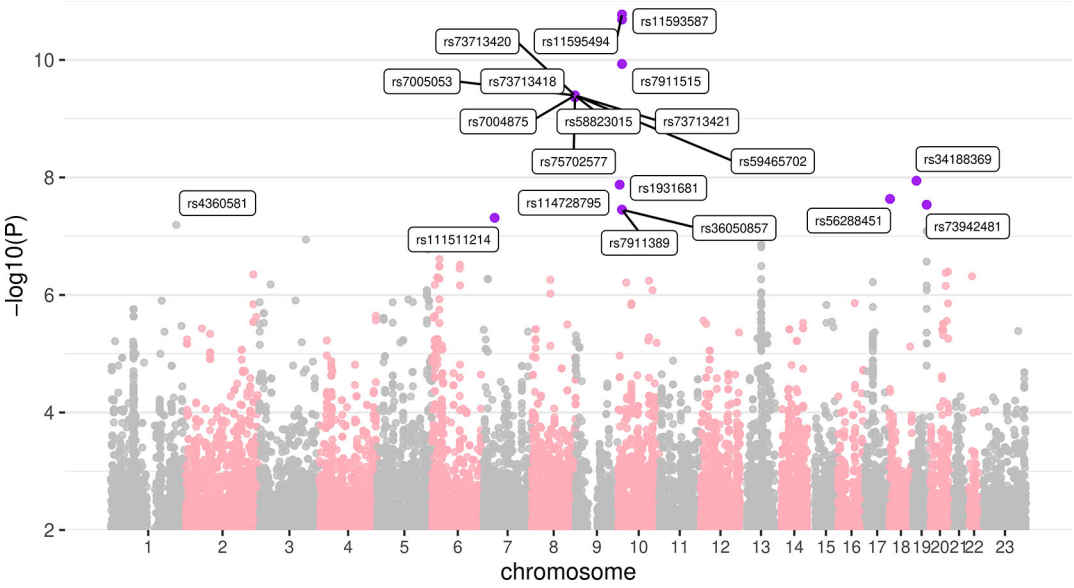

Figure 3. Manhattan plot showing associations with vaginal microbiota Shannon diversity. Ancestry is taken into account in the regression via the MDS components. Colours indicate SNPs or regions of interest, which are also labelled. The X chromosome is labelled as '23'.

reported associations with microbiota. The next two most strongly associated regions were located in chromosomes 8 and 19, and far from any human gene. A fourth region of interest was located in a non-coding area of chromosome 10, close to the region coding for the Aldo-keto reductase family 1 member C4 (AKR1C4) protein. The final three significant SNPs are located in chromosomes 18, 19, and 7, and in the coding regions of Elastin microfibril interfacer 2 (EMILIN2), Paraneoplastic antigen-like protein 8B (PNMA8B), and Endonuclease/exonuclease/phosphatase family domain containing 1 (EEPD1) genes. EMILIN2 is an extracellular matrix constituent involved in angiogenesis but with no reported link with microbiota. Finally, another SNP was nearly significant and located in a non-coding region of chromosome 1, close to the gene coding for the Estrogen-related receptor gamma (ESRRG) protein.

A closer look at the distribution of these SNPs in the dataset revealed a potential confounding effect from one participant. Indeed, as shown in the top panels of Figure 4, except for rs1931681, this person is homozygote for the minority allele and her microbiota has one of the highest diversity values. Furthermore, as illustrated by a linear model stratified by variant for rs11595494 (Table 4), the participant reported using vaginal douching, a practice known to be strongly associated with CST-4, i.e. *Lactobacillus*-poor vaginal microbiota, but rare in our cohort (Table 1). Incidentally, note that we did not detect any other significant difference in this analysis.

Our analysis controls for shared ancestry by using the principal components of the MDS clustering but we performed additional analyses to investigate the robustness of our results. First, we redid the previous analysis without the aforementioned participant. Only one of the regions remained significant with an SNP located in chromosome 10 close to the AKR1C4 gene (model "Simpson2" in Table 2). Furthermore, we performed a multivariate linear regression using the Simpson diversity as a response variable, and, as explanatory variables, the main variants at seven genomic positions, the covariates associated with vaginal microbiota composition (in Table 1), and the first ten components of the MDS. As shown in Table 5, the only variables significantly associated with Simpson diversity were the SNPs and the addition of the behavioural covariates only marginally improved the percentage of the variance explained (the adjusted R-squared increased only from 0.63 to 0.65).

For completeness, we also performed GWAS using the Shannon diversity index or the CST composition (CST-4 vs. non-CST-4). In both models, a non-coding region from chromosome 17 was found to be significantly associated with the trait.

Table 3. Characteristics of the SNPs significantly associated with vaginal microbiota Simpson diversity. ID is the reference SNP cluster ID, CHR is the chromosome number, BP is the position in the chromosome using genome reference GRCh37, REF is the reference allele, MUT is the alternative allele, MAF is the minority allele frequency in the dataset, EUR freq. and AFR freq. Are the MAF in European and African populations based on the Allele Frequency Aggregator (ALFA) project, Obs. is the number of observations of the position in our dataset, Model is the analysis used (CST is when using CST-4 vs. non-CST-4 as a response trait and Simpson2 is when removing one key participant from the dataset), Beta is the regression coefficient of the model, SE its standard error, and p-val. Its asymptotic p-value. The non-significant value, i.e. with a p-value above 5·10–8 is in italic. 'coding' indicates any gene the SNP may be located in and 'nearby' indicates any potential gene of interest within less than 6,000 base.

ID	CHR	BP	REF	MUT	MAF	EUR freq.	AFR freq.	Obs.	Model	Beta	SE	p-val.	gene	nearby
rs11595494	10	12,992,365	G	C	0.052	0.050	0.233	163	Simpson	1.80	0.25	1.69e-11	CCDC3	---
rs73713420	8	142,685,979	A	G	0.052	0.020	0.305	164	Simpson	1.91	0.29	4.03e-10	---	---
rs34188369	19	13,293,164	G	A	0.058	0.036	0.253	164	Simpson	1.64	0.27	1.14e-08	---	---
rs1931681	10	5,263,915	G	T	0.053	0.98	0.78	162	Simpson	1.54	0.26	1.33e-08	---	AKR1C4
rs56288451	18	2,909,700	C	T	0.055	0.047	0.133	164	Simpson	1.54	0.26	2.32e-08	EMILIN2	---
rs73942481	19	46,993,058	A	G	0.064	0.054	0.105	164	Simpson	1.43	0.24	2.92e-08	PNMA8B	---
rs111511214	7	36,267,254	G	A	0.053	0.034	0.316	161	Simpson	1.50	0.26	4.87e-08	EEPD1	---
rs4360581	1	216,671,992	C	T	0.049	0.048	0.072	164	Simpson	1.42	0.25	*6.41e-08*	---	ESRRG
rs8081213	17	25,297,867	A	C	0.056	0.069	0.182	161	CST	0.70	0.11	1.32e-08	---	---
rs113602804	10	5,265,452	T	C	0.052	0.004	0.005	163	Simpson2	1.33	0.22	1.70e-08	---	AKR1C4
rs8081213	17	25297867	A	C	0.056	0.069	0.182	161	Shannon	0.70	0.12	2.20E-08	---	---

Open Research Europe 2026, 5:186 Last updated: 19 MAR 2026

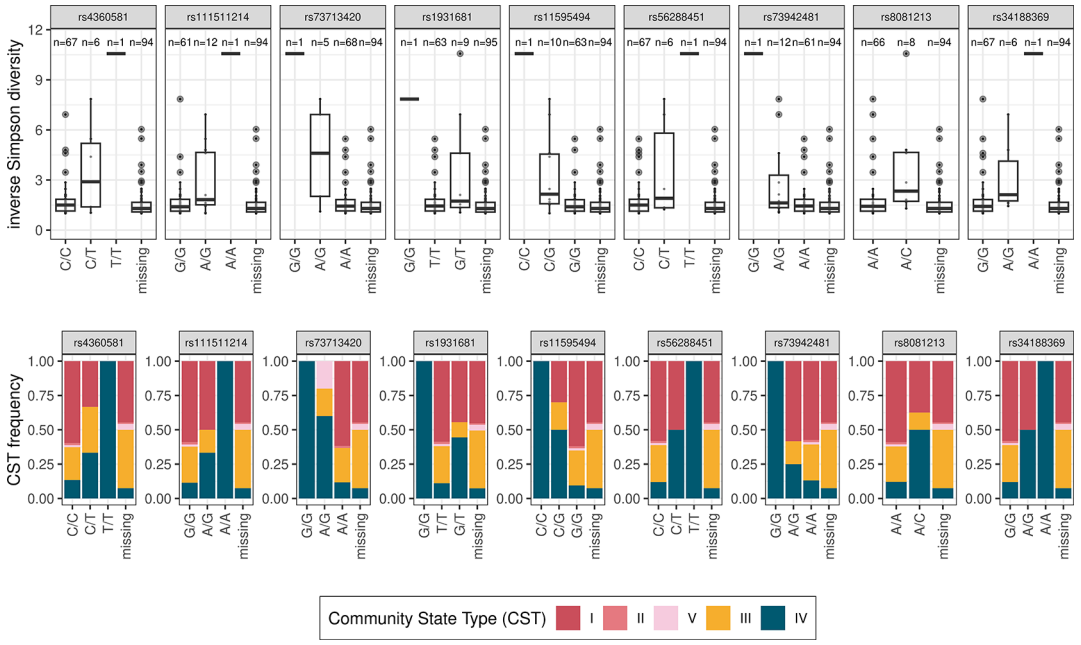

Figure 4. Vaginal microbiota Simpson diversity (top) and community state type (bottom) stratified by genetic variant. On the top panels, boxes show the medians, quartiles, and quantiles. Dots show the outliers. For some individuals, the coverage for the genomic position is missing. Numbers above the boxes indicate group sample sizes.

Table 4. Cohort profile stratified by mutation at rs11595494. See Table 1 for details.

	C/C	C/T	T/T	p
number of participants	67	6	1	
female affinity (%)	12 (17.9)	1 (16.7)	0 (0.0)	1.000
male affinity (%)	67 (100.0)	6 (100.0)	1 (100.0)	NA
smoker (%)	26 (63.4)	6 (100.0)	0 (NaN)	0.157
regular sport practice (%)	36 (53.7)	2 (33.3)	0 (0.0)	0.305
menstrual cup user (%)	15 (22.4)	0 (0.0)	0 (0.0)	0.469
vaginal products user (%)	30 (44.8)	3 (50.0)	0 (0.0)	1.000
tampon user (%)	18 (26.9)	1 (16.7)	0 (0.0)	1.000
previous pregnancy (%)	3 (4.5)	1 (16.7)	0 (0.0)	0.334
vaginal douching (%)	*1 (1.5)*	*0 (0.0)*	*1 (100.0)*	*0.033*
male partner using condom (%)	34 (50.7)	3 (50.0)	0 (0.0)	1.000
alcohol consumption	−0.07 [−0.52, 0.61]	−0.52 [−0.52, −0.18]	−1.43 [−1.43, −1.43]	0.235
BMI	−0.25 [−0.57, 0.67]	0.33 [−0.43, 1.15]	−0.64 [−0.64, −0.64]	0.385
age at inclusion	0.72 [−0.27, 0.72]	0.72 [0.35, 0.72]	−1.76 [−1.76, −1.76]	0.209
age at first mentruations	0.19 [−0.55, 0.93]	−0.18 [−1.11, 0.19]	0.19 [0.19, 0.19]	0.482
number of sexual partners	−0.19 [−0.62, 0.30]	0.73 [−0.19, 1.79]	0.49 [0.49, 0.49]	0.238

Table 5. Linear model with socio-demographic covariates. 'ref.' indicates the reference value for categorical variables, beta is the effect term, and Ci stands for confidence interval. MDS PC indicates the principal components of the multidimensional scaling clustering (see the main text). The adjusted R-squared of the model is 0.65 (and 0.63 without the behavioural covariates).

	Beta	95% CI	p-value
rs4360581 (ref. C/C)			
C/T	1.2	0.39, 2.0	0.004
missing	−0.10	−1.8, 1.6	>0.9
T/T	7.6	5.3, 9.8	<0.001
Rs111511214 (ref. A/A)			
A/G	0.53	−0.09, 1.1	0.095
rs73713420 (ref. A/A)			
A/G	1.1	0.14, 2.0	0.025
rs1931681 (ref. G/G)			
G/T	−3.0	−5.1, −0.98	0.004
T/T	−2.9	−5.0, −0.84	0.006
missing	−2.6	−5.3, 0.07	0.056
rs11595494 (ref. C/C)			
C/G	0.66	0.05, 1.3	0.034
rs56288451 (ref. C/C)			
C/T	0.70	−0.06, 1.5	0.072
rs73942481 (ref. A/A)			
A/G	0.12	−0.52, 0.76	0.7
rs8081213 (ref. A/A)			
A/C	0.60	−0.16, 1.4	0.12
rs34188369 (ref. A/A)			
A/G	1.2	0.42, 2.0	0.003
MDS PC1	2.3	−0.79, 5.5	0.14
MDS PC2	0.59	−5.4, 6.6	0.8
MDS PC3	9.8	−1.8, 21	0.10
MDS PC4	−2.6	−10, 5.3	0.5
MDS PC5	−6.4	−43, 30	0.7
MDS PC6	−13	−32, 6.6	0.2
MDS PC7	8.1	−7.1, 23	0.3
MDS PC8	−16	−42, 9.4	0.2
MDS PC9	−32	−67, 3.0	0.072
MDS PC10	−6.2	−41, 28	0.7
body mass index	0.12	−0.01, 0.25	0.078
regular sport practice (ref. no)	−0.14	−0.40, 0.12	0.3
age at first menstruation	0.05	−0.08, 0.18	0.5
condom use (ref. no)	−0.12	−0.37, 0.13	0.3
previous pregnancy (ref. no)	−0.32	−0.99, 0.35	0.3
age at inclusion	−0.01	−0.15, 0.13	>0.9

Discussion

Few studies have explored associations between human genetics and the vaginal microbiota. Most of these rely on symptoms rather than genetic microbiological data. Furthermore, few studies have been conducted on women of European ancestry.

Building on data collected during the PAPCLEAR study,[5] we performed a GWAS using the Simpson diversity index of the vaginal microbiota as our trait of interest. Our microarray allowed us to identify 320,452 variants, which we enriched to 2,927,125 variants using imputation methods, and the analysis revealed several SNPs of interest, most of which were in or close from coding regions of the human genome. However, the mechanistic links between the genes involved and the vaginal microbiome are limited except for the two cases where the SNPs were close from the gene coding region. The first one, AKR1C4, has been shown to be associated with inflammatory bowel disease.[31] This protein is also expected to strongly interact with proteins involved in the degradation of testosterone and progesterone (SDR5A1 and SDR5A2) according to the STRING prediction algorithm.[32] The second one, ESRRG is involved in estrogen signaling and interacts with other hormone receptor proteins. It has also been shown to be involved in bacterial infections by *Helicobacter pylori*.[33]

When using Shannon diversity or a binary trait (CST-4 vs. non-CST-4) as the response variable in our model, we identified another genomic area that was not associated with Simpson diversity. We do expect Shannon diversity and the CST stratification to be more alike than the Simpson diversity, and the SNPs identified with the former were in non-coding regions, which makes it difficult to provide more mechanistic interpretations.

An additional issue is that most of the SNPs significantly associated with Simpson diversity were conditioned by the inclusion in the dataset of a specific participant. We decided to report these results because our analysis takes into account genetic ancestry and because reporting these SNPs makes it possible for future, more powered, studies to confirm or refute these associations. Note also that the SNPs close from the AKR1C4 gene remained significantly associated with Simpson diversity when the participant was removed from the dataset.

In addition to the number of participants, our study is also limited in terms of genome coverage. The variants we find do not seem to be correlated with covariates classically involved in vaginal microbiota differences and a deeper sequencing of the human genome could identify positions that were not sufficiently covered in the assay used.

Performing a GWAS on microbiota raises several challenges by definition. A first issue has to do with the temporal variability. In the case of the vaginal microbiota, earlier studies show that the communities are highly stable on a weekly basis[5] and that a single time point single sample predicts with good accuracy the vaginal microbiota trajectory over the following 18 months.[6] To further buffer potential variations, we averaged all the observations of a single participant to define the observed trait. Another possibility could have been to analyse the temporal variability of the vaginal microbiota. However, this would have been problematic because the number of observations per participant was highly variable, some being followed for less than two months and others for more than two years.[5] A second issue has to do with the trait used to describe the microbiota. The high level of clustering of vaginal microbiota samples offers several options to address this problem. One of them is to focus on the most abundant species, e.g. using the Shannon diversity index or the presence/absence of lactobacilli, whereas the other is to put more importance on evenness in species abundance, e.g. using the Simpson diversity index. We show that the latter yields much more variance across samples and is associated with more genomic regions. With more homogeneous follow-ups between participants and a larger cohort size, a possibility could have been to perform a hierarchical clustering on the follow-ups themselves to include a third category of women who alternate between CSTs.[5,6] Nevertheless, compared to other microbiota, the temporal stability and strong structure of the vaginal microbiota make it an ideal study system to investigate the coevolution between hosts and their microbiota.

Ethics

The PAPCLEAR study has been approved by the Comitee of Persons Protection (Comité de Protection des Personnes, CPP) Sud Méditerranée I on 25 April 2016 (reference number 2016-A00712–49); by the Consultative Comitee on Information processing in the context of research in health (Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé, reference number 16.504); by the National Commission on Informatics and Freedom (Commission Nationale Informatique et Libertés, reference number MMS/ABD/AR1612278, decision number DR-2016-488), by the National Agency of Security of Drugs and Health Products (Agence Nationale de Sécurité du Médicament et des Produits de Santé, reference 20160072000007), and is registered at ClinicalTrials.gov under the ID NCT02946346. All participants provided written informed consent to participate in the study.

Data availability

The output of the GWAS for Shannon diversity, the metadata, and the output of the quality control analysis can be accessed on the *Recherche Data Gouv* data repository at https://doi.org/10.57745/LGZANK, along with the scripts used to generate the figures and table in the manuscript.[26] It is also available on the Open Science Framework DOI 10.17605/OSF.IO/MZTWX.[27]

Creative Commons Attribution 4.0 International Public License.

Acknowledgements

The authors acknowledge the ISO 9001 certified IRD i-Trop HPC (member of the South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this article (https://bioinfo.ird.fr and https://www.southgreen.fr). We are extremely grateful to one of the reviewers for their suggestion to use Simpson diversity instead of Shannon diversity as a response trait.

References

1. van de Wijgert JHHM: **The vaginal microbiome and sexually transmitted infections are interlinked: consequences for treatment and prevention.** *PLoS Med.* 2017; **14**(12): e1002478. **Publisher Full Text**
2. Haahr T, Zacho J, Bräuner M, *et al.*: **Reproductive outcome of patients undergoing in vitro fertilisation treatment and diagnosed with Bacterial Vaginosis or abnormal vaginal microbiota: a systematic PRISMA review and meta-analysis.** *BJOG.* 2019; **126**(2): 200–207. **PubMed Abstract** | **Publisher Full Text**
3. Bilardi JE, Walker S, Temple-Smith M, *et al.*: **The burden of Bacterial Vaginosis: women's experience of the physical, emotional, sexual and social impact of living with recurrent Bacterial Vaginosis.** *PLoS One.* 2013; **8**(9): e74378. **Publisher Full Text**
4. Ravel J, Gajer P, Abdo Z, *et al.*: **Vaginal microbiome of reproductive-age women.** *Proc Natl Acad Sci U S A.* 2011; **108**(Suppl 1): 4680–4687. **Publisher Full Text**
5. Kamiya T, Tessandier N, Elie B, *et al.*: **Factors shaping vaginal microbiota long-term community dynamics in young adult women.** *Peer Community J.* 2025; **5**: pcjournal.527. **Publisher Full Text**
6. Tamarelle J, Thiébaut ACM, de Barbeyrac B, *et al.*: **Vaginal microbiota stability over 18 months in young student women in France.** *Eur J Clin Microbiol Infect Dis.* 2024; **43**(12): 2277–2292. **Publisher Full Text**
7. Borgdorff H, van der Veer C, van Houdt R, *et al.*: **The association between ethnicity and vaginal microbiota composition in Amsterdam, the Netherlands.** *PLoS One.* 2017; **12**(7): e0181135. **Publisher Full Text**
8. Fettweis JM, Paul Brooks J, Serrano MG, *et al.*: **Differences in vaginal microbiome in African American women versus women of European ancestry.** *Microbiology (Reading).* 2014; **160**(Pt 10): 2272–2282.
9. Zhou X, Brown CJ, Abdo Z, *et al.*: **Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women.** *ISME J.* 2007; **1**(2): 121–133. **PubMed Abstract** | **Publisher Full Text**
10. Kurilshikov A, Medina-Gomez C, Bacigalupe R, *et al.*: **Large-scale association analyses identify host factors influencing human gut microbiome composition.** *Nat Genet.* 2021; **53**(2): 156–165. **Publisher Full Text**
11. Murphy K, Shi Q, Hoover DR, *et al.*: **Genetic predictors for Bacterial Vaginosis in women living with and at risk for HIV infection.** *Am J Reprod Immunol.* 2024; **91**(5): e13845. **Publisher Full Text**
12. Mehta SD, Nannini DR, Otieno F, *et al.*: **Host genetic factors associated with vaginal microbiome composition in Kenyan women.** *mSystems.* 2020; **5**(4): e00502–e00520. **Publisher Full Text**
13. Fan W, Kan H, Liu HY, *et al.*: **Association between human genetic variants and the vaginal bacteriome of pregnant women.** *mSystems.* 2021; **6**(4): e0015821. Publisher: American Society for Microbiology. **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**
14. Mutli E, Mändar R, Koort K, *et al.*: **Genome-wide association study in Estonia reveals importance of vaginal epithelium associated genes in case of recurrent vaginitis.** *J Reprod Immunol.* 2024; **162**: 104216. **Publisher Full Text**
15. Brotman RM, Ghanem KG, Klebanoff MA, *et al.*: **The effect of vaginal douching cessation on Bacterial Vaginosis: a pilot study.** *Am J Obstet Gynecol.* 2008; **198**(6): 628.e1–628.e7. **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**
16. McKinnon LR, Achilles SL, Bradshaw CS, *et al.*: **The evolving facets of Bacterial Vaginosis: implications for HIV transmission.** *AIDS Res Hum Retroviruses.* 2019; **35**(3): 219–228. **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**
17. Tessandier N, Elie B, Boué V, *et al.*: **Viral and immune dynamics of genital human papillomavirus infections in young women with high temporal resolution.** *PLoS Biol.* 2025; **23**(1): e3002949. **Publisher Full Text**
18. Murall CL, Rahmoun M, Selinger C, *et al.*: **Natural history, dynamics, and ecology of human papillomaviruses in genital infections of young women: protocol of the PAPCLEAR cohort study.** *BMJ Open.* 2019; **9**(6): e025129. **Publisher Full Text**
19. France MT, Ma B, Gajer P, *et al.*: **VALENCIA: a nearest centroid classification method for vaginal microbial communities based on composition.** *Microbiome.* 2020; **8**(1): 166. **Publisher Full Text**
20. Howie BN, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.** *PLoS Genet.* 2009; **5**(6): e1000529. **Publisher Full Text**
21. Chang CC, Chow CC, Tellier LC, *et al.*: **Second-generation PLINK: rising to the challenge of larger and richer datasets.** *Gigascience.* 2015; **4**. **PubMed Abstract** | **Publisher Full Text**
22. Marees AT, de Kluiver H, Stringer S, *et al.*: **A tutorial on conducting genome-wide association studies: quality control and statistical analysis.** *Int J Methods Psychiatr Res.* 2018; **27**(2): e1608. **Publisher Full Text**
23. Yang J, Lee SH, Goddard ME, *et al.*: **GCTA: a tool for Genome-wide Complex Trait Analysis.** *Am J Hum Genet.* 2011; **88**(1): 76–82. **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**
24. Ma B, Forney LJ, Ravel J: **Vaginal microbiome: rethinking health and disease.** *Annu Rev Microbiol.* 2012; **66**(1): 371–389. **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**
25. van Ravenswaaij-Arts CM, Hefner M, Blake K, *et al.*: **CHD7 disorder.** Adam MP, Feldman J, Mirzaa GM, *et al.*, editors. *GeneReviews®.* Seattle (WA): University of Washington, Seattle; 1993.
26. Alizon S: **Supplementary files to 'genome wide association study of vaginal microbiota genetic diversity in French women'. Recherche Data Gouv, UNF: 6:9XQ2mHB7nzkGG4v4e1dGxQ==.**

2025.
Reference Source

27. Alizon S: **Supplementary files to "genome wide association study of vaginal microbiota genetic diversity in French women."** *Open Science Framework.*
Publisher Full Text

28. France M, Alizadeh M, Brown S, *et al.*: **Towards a deeper understanding of the vaginal microbiota.** *Nat Microbiol.* 2022; **7**(3): 367–378.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

29. Romero R, Hassan S, Gajer P, *et al.*: **The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women.** *Microbiome.* 2014; **2**(1): 4.
Publisher Full Text

30. Holm JB, Gajer P, Ravel J: **SpeciateIT and vSpeciateDB: novel, fast, and accurate per sequence 16S rRNA gene taxonomic classification of vaginal microbiota.** *BMC Bioinformatics.* 2024; **25**: 313.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

31. Deris Zayeri Z, Parsi A, Shahrabi S, *et al.*: **Epigenetic and metabolic reprogramming in inflammatory bowel diseases: diagnostic and prognostic biomarkers in colorectal cancer.** *Cancer Cell Int.* 2023; **23**: 264.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

32. Szklarczyk D, Kirsch R, Koutrouli M, *et al.*: **The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest.** *Nucleic Acids Res.* 2023; **51**: D638–D646.
Publisher Full Text

33. Kang M-H, Eyun S, Park Y-Y: **Estrogen-related receptor-gamma influences Helicobacter pylori infection by regulating TFF1 in gastric cancer.** *Biochem Biophys Res Commun.* 2021; **563**: 15–22.
PubMed Abstract | **Publisher Full Text**

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 19 March 2026

https://doi.org/10.21956/openreseurope.24881.r71043

© **2026 Hugerth L.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Luisa W. Hugerth**
Uppsala University, Uppsala, Sweden

The authors have thoroughly addressed all my previous concerns.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Microbiome; Women's health

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 19 March 2026

https://doi.org/10.21956/openreseurope.24881.r71042

© **2026 da Costa A et al.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Antonio Charlys da Costa**
University of São Paulo, São Paulo, Brazil
Tania Regina Tozetto-Mendoza
Virology Lab, Universidade de Sao Paulo Instituto de Medicina Tropical de Sao Paulo, São Paulo, State of São Paulo, Brazil

I consider the article ready for indexing

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Metagenomics and virus discovery

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 01 September 2025

https://doi.org/10.21956/openreseurope.22142.r57586

© 2025 Hugerth L. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Luisa W. Hugerth
Uppsala University, Uppsala, Sweden

Alizon and colleagues set out to identify human genetic loci associated with the vaginal microbiome. This is interesting for two reasons. Firstly, the vaginal microbiome has been previously associated with many adverse obstetric and gynecologic outcomes. While in some cases there are clear mechanistic evidence for it, it is equally possible that the microbiome is (partly) a biomarker for an internal underlying process, e.g. connected to the immune system or to mucosal integrity. Secondly, abundant data from around the globe shows a stratification of the vaginal microbiome in regards to ethnicity, as well as diverging risk patterns for some microbes in different populations. For all these reasons, the present research has microbial ecological and medical interest.

However, defining a "microbiome" as a phenotype is a complicated endeavour, and could be improved in the current work. Additionally, the treatment of confounders is not entirely clear, as detailed below:

1. Data availability and methods.
https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/LGZANK contains important summary data, but I wouldn't call it raw data. At least for the microbiome, raw fastq reads (depleted of human DNA) should be submitted to ENA, together with appropriate patient-level data.

Related to this, even if the 16S data is previously published, please provide minimal information on it in the methods, namely: hypervariable region amplified, platform used for clustering or denoising (Deblur, DADA2, Unoise, other?) and database+algorithm used for taxonomic assignment.

Additionally, the provided link http://www.doi.org/10.17605/OSF.IO/MZTWX redirects me to the baseline page https://osf.io/, so I'm not sure what should be in it.

2. Demographics and other confounders.
Since the data in table 1 is centralized and standardized, I have no clue whether the participants included are school-age, reproductive age or postmenopausal. This should be stated more clearly.

Additionally, it would be important to provide the median and range for the number of samples collected for each participant, since this may impact the averaging procedure. Related to this, there is strong shaping of the microbiome in relation to the menstrual cycle (see e.g. https://pubmed.ncbi.nlm.nih.gov/39160615/). Is there any information on the participants cycle (assuming they are all pre-menopausal?) Is averaging all samples the best approach? Or would it make more sense to look for "mid-cycle" samples (and then average these)? At any rate, more information on the menstrual cycle would be needed for data interpretation.

Furthermore, parity is another very strong predictor of the vaginal microbiome, and there is a significant imbalance in gravidity at least for one locus. Excluding the 8 women that have previously been pregnant (as well as any partipants aged 45+) may give a cleaner signal.

To be clear, I would love to see a similar study as yours focusing on middle-aged women, but from what I can tell from your data, you're not powered for it; and therefore I would advise on focusing on the nulliparous 18-35-years-old and really getting that signal out.

3. Defining dysbiosis.
This is of course a very complicated endeavour, and I appreciate that you have attempted a variety of methods, focusing the GWAS on the most promising. Still, there are some things that could be improved in this area, as follows:

VALENCIA centroids are based on a number of pre-defined taxa. How well did your taxa match VALENCIA's? Were species not matching VALENCIA excluded (and the data re-normalized) or were they kept? Describe the procedure a bit more clearly in the methods.

VALENCIA also has a coding quirk that, if no CST are of high confidence, defaults to CST-I. Were there any cutoffs for CST assignment? How did you handle low-confidence assignments? Please specify in the methods. Would it make sense to have them as putative CST-IV, in the sense of "not dominated by any of the 4 cannonical Lacto species"?

Conversely, were there any samples dominated by a non-canonical Lacto, such as L. vaginalis or Limosilactobacillus spp? How were these handled? VALENCIA would likely not be able to assign them with any confidence, but they strike me as definitely not CST-IV.

CST-IV is, as you know, highly polymorphic. Indeed, you seem to have 15 individuals with CST IV-B, 3 CST IV-A and 1 CST IV-C. Is it possible that you would get a cleaner GWAS signal by focusing on CST IV-B, and perhaps excluding the other 4 individuals as neither case nor control?

Relatedly, how come a participant with CST-IVC is placed in table 1 as "Lactobacillus-dominated"? How was Lacto-dominance determined, if not through CST? Please specify in the methods.

If all of this cleaning doesn't improve the CST-IV signal (or if you decide not to pursue it), it may still be productive to improve the diversity signal. While Shannon's entropy is widely used in microbiome research, Simpson's diversity (which places more weight on evenness) is, in my experience, more sensitive to changes in the low-richness vaginal microbiome. Alternatively, you may find it productive to break down the global diversity measure into its components of richness and evenness. There could be quite different processes controlling richness (the establishment of new species, even at a very small number) and evenness (the possibility for one species to thoroughly dominate the vaginal microbiome).

All in all, this is a very interesting paper. Small adjustments described above could really make the conclusions much more convincing and harder to dismiss by competing teams.

References

1. Hugerth L, Krog M, Vomstein K, Du J, et al.: Defining Vaginal Community Dynamics: daily microbiome transitions, the role of menstruation, bacteriophages, and bacterial genes. *Microbiome*. 2024; **12** (1). Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Microbiome; Women's health

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 27 Jan 2026

Samuel Alizon

Alizon and colleagues set out to identify human genetic loci associated with the vaginal microbiome. This is interesting for two reasons. Firstly, the vaginal microbiome has been previously associated with many adverse obstetric and gynecologic outcomes. While in some cases there are clear mechanistic evidence for it, it is equally possible that the microbiome is (partly) a biomarker for an internal underlying process, e.g. connected to the immune system or to mucosal integrity. Secondly, abundant data from around the globe shows a stratification of the vaginal microbiome in regards to ethnicity, as well as diverging risk patterns for some microbes in different populations. For all these reasons, the present research has microbial ecological and medical interest.
However, defining a "microbiome" as a phenotype is a complicated endeavour, and could be improved in the current work. Additionally, the treatment of confounders is not entirely clear, as detailed below:
Reply: Thank for the positive appreciation of our work! We completely agree that finding an appropriate summary statistics to characterise the microbiote is challenging.
_____ *Q 2.1*
Data availability and methods.
https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/LGZANK contains important summary data, but I wouldn't call it raw data. At least for the microbiome, raw fastq reads (depleted of human DNA) should be submitted to ENA, together with appropriate patient-level data.
Related to this, even if the 16S data is previously published, please provide minimal information on it in the methods, namely: hypervariable region amplified, platform used for clustering or denoising (Deblur, DADA2, Unoise, other?) and database+algorithm used for taxonomic assignment.
Additionally, the provided link http://www.doi.org/10.17605/OSF.IO/MZTWX redirects me to the baseline page https://osf.io/, so I'm not sure what should be in it.
Reply: We are still in the process of recovering the raw data for the upload in ENA. The process is taking some time because half of the files were archived. This also partly explains the delay in resubmitting. Hopefully this can be solved in the next weeks, but we preferred to resubmit to avoid wasting even more time. We added details about the generation of the 16S data in the methods from our previous publication. We also now mention summary data instead of raw data to avoid ambiguities. Regarding the link, this is normal: the Open Science Foundation creates a DOI associated with the dataset (this was asked by Open Research Europe because they do not recognise the data.gouv.fr data warehouse).
_____ *Q 2.2*
Demographics and other confounders. Since the data in table 1 is centralized and standardized, I have no clue whether the participants included are school-age, reproductive age or postmenopausal. This should be stated more clearly.
Additionally, it would be important to provide the median and range for the number of samples collected for each participant, since this may impact the averaging procedure. Related to this, there is strong shaping of the microbiome in relation to the menstrual cycle (see e.g. https://pubmed.ncbi.nlm.nih.gov/39160615/). Is there any information on the participants cycle (assuming they are all pre-menopausal?) Is averaging all samples the best approach? Or would it make more sense to look for "mid-cycle" samples (and then average these)? At any rate, more information on the menstrual cycle would be needed for data interpretation.
Furthermore, parity is another very strong predictor of the vaginal microbiome, and there is a significant imbalance in gravidity at least for one locus. Excluding the 8 women that have

previously been pregnant (as well as any partipants aged 45+) may give a cleaner signal.
To be clear, I would love to see a similar study as yours focusing on middle-aged women, but from what I can tell from your data, you're not powered for it; and therefore I would advise on focusing on the nulliparous 18-35-years-old and really getting that signal out.
Reply: Thank you for pointing out the fact that the normalisation of the values make them difficult to interpret. We now present more clearly the demographics. About the multiple samples per participants, we refer to the figure from our previous publication. Information about the menstrual cycle was not available and, as we now explain in further details, we pooled the data from each participant. Regarding the 8 participants who reported earlier pregnancies, we indeed find a significant difference in terms of vaginal microbiota composition with slightly less Lactobacillus-rich compositions, although if we correct for multiple hypothesis testing it becomes non significant. We hesitated to perform the GWAS on a restricted dataset but decided not to for two reasons. First, thanks to your suggestion to focus on other diversity metrics, we now have a much stronger signal. Second, there are other factors that may bias our populations of women from 18 to 25, and some have even stronger impact on vaginal microbiota (e.g. the lifetime number of sexual partners). If we attempt to correct for all of these, we will be left with a tiny cohort. Therefore, we decided to maintain our post hoc analysis to check whether these factors differ significantly when stratifying our populations with respect to the allele found at positions of interest.

_____ *Q 2.3*

Defining dysbiosis. This is of course a very complicated endeavour, and I appreciate that you have attempted a variety of methods, focusing the GWAS on the most promising. Still, there are some things that could be improved in this area, as follows:
VALENCIA centroids are based on a number of pre-defined taxa. How well did your taxa match VALENCIA's? Were species not matching VALENCIA excluded (and the data re-normalized) or were they kept? Describe the procedure a bit more clearly in the methods.
VALENCIA also has a coding quirk that, if no CST are of high confidence, defaults to CST-I. Were there any cutoffs for CST assignment? How did you handle low-confidence assignments? Please specify in the methods. Would it make sense to have them as putative CST-IV, in the sense of "not dominated by any of the 4 cannonical Lacto species"?
Conversely, were there any samples dominated by a non-canonical Lacto, such as L. vaginalis or Limosilactobacillus spp? How were these handled? VALENCIA would likely not be able to assign them with any confidence, but they strike me as definitely not CST-IV.
CST-IV is, as you know, highly polymorphic. Indeed, you seem to have 15 individuals with CST IV-B, 3 CST IV-A and 1 CST IV-C. Is it possible that you would get a cleaner GWAS signal by focusing on CST IV-B, and perhaps excluding the other 4 individuals as neither case nor control?
Reply: We now present the VALENCIA algorithm in further details but it is based on 13,160 taxonomic profiles from 1975 women, meaning that the external database is quite large. Regarding the matching to the database, we can investigate the distance to the CSTs (although these represent "idealised" communities and not actual samples). About the assignment, we are not aware of any 'default' assignment to CST-1. For example, one of the samples that seems difficult to attribute to any CST is labelled as a CST5. Other GWAS have considered species diversity in detail (e.g. in China). However, we are unsure that this can be done by correctly controling for multiple hypothesis testing. Furthermore, the size of our cohort strongly limits our possibility to go beyond illustrative results. Finally, we wish to stress that the discussion on the CSTs is somehow secondary since our main results do not involve these. We merely use them to show that an asset of the vaginal microbiota is that

there is high correlation between the metric used (although, as you intuited, Simpson diversity yields more signal).

Relatedly, how come a participant with CST-IVC is placed in table 1 as "Lactobacillus-dominated"? How was Lacto-dominance determined, if not through CST? Please specify in the methods.
Reply: Thank you so much for your careful reading!!! This actually led us to realise that there was a column swap in the data. We reran the whole averaging script and triple checked the outputs. In the end, besides this line, there were only a few very minor differences in diversity metrics that did not change the results.

If all of this cleaning doesn't improve the CST-IV signal (or if you decide not to pursue it), it may still be productive to improve the diversity signal. While Shannon's entropy is widely used in microbiome research, Simpson's diversity (which places more weight on evenness) is, in my experience, more sensitive to changes in the low-richness vaginal microbiome. Alternatively, you may find it productive to break down the global diversity measure into its components of richness and evenness. There could be quite different processes controlling richness (the establishment of new species, even at a very small number) and evenness (the possibility for one species to thoroughly dominate the vaginal microbiome).

Reply: Again, thank you so much for your helpful comments! We followed your advice and reran all the analyses using Simpson diversity. As you will see in the revised version of the manuscript, we were amazed to see how this lit up the Manhattan plot, which many significant associations. One limitation is that most of the SNPs detected seem to rely on a single participant, whose ancestry differs from the majority of that of the population. When performing the analysis without this participant, only a single SNP remained. We discussed about which results to show in the main text and decided to keep the full dataset for two reasons. First, if we do not report it, future studies will not be able to test these candidate SNPs. Second, we corrected for shared ancestry in the GWAS using 10 components of the MDS clustering so, in statistically, this should be controlled for.

All in all, this is a very interesting paper. Small adjustments described above could really make the conclusions much more convincing and harder to dismiss by competing teams.
Thank you so much for your feedback and the insightful suggestion to look at Simpson diversity!

Competing Interests: NA

Reviewer Report 13 August 2025

https://doi.org/10.21956/openreseurope.22142.r57592

© 2025 da Costa A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Antonio Charlys da Costa
University of São Paulo, São Paulo, Brazil

This manuscript presents a significant and novel investigation into the genetic basis of vaginal microbiota diversity using GWAS approaches within a European cohort. The utilization of Shannon diversity from 16S metabarcoding data as a quantitative trait constitutes a novel approach that is substantiated by heritability estimates. However, the study's impact is constrained by its relatively modest sample size and the absence of genome-wide significant findings. The functional interpretation of the associated genomic regions remains speculative, and the averaging approach for longitudinal microbiota data may mask critical temporal variation influencing host-microbiota interactions. With moderate revisions aimed at clarifying the methodology, better addressing the limitations, and improving the presentation, this study would contribute to the emerging field of microbiome genetics. Subsequent, more extensive studies with more comprehensive genomic coverage and meticulous control of environmental confounders are necessary to validate and expand upon these findings.

The abstract could better emphasize the study's limitations, such as sample size and statistical power. It would be beneficial to explicitly state the novelty of using microbial genetic data rather than symptoms earlier in the abstract.

The extraneous background information from previous GWAS studies could be streamlined to reduce redundancy. In order to avoid the potential for overgeneralization, it would be advisable to enhance the referencing and clarification of certain statements pertaining to ethnicity and behavioral factors. A brief discussion of the challenges and opportunities inherent in the utilization of 16S metabarcoding data for the purpose of defining traits in the context of GWAS is warranted.

Further elaboration is necessary to clarify the averaging approach for longitudinal microbiota samples. The utilization of the mean as a metric has the potential to obfuscate the temporal dynamics inherent in the data. The rationale for imputation and its effect on downstream analyses could be expanded. There has been a paucity of discourse regarding the implementation of strategies for the control of potential confounders, including but not limited to sexual behavior, antibiotic utilization, and hormonal status.
The strongest associations do not reach genome-wide significance, a fact that merits greater discussion. The genomic regions identified are distant from known coding or regulatory regions, resulting in a speculative functional interpretation. Due to the modest sample size, the findings must be interpreted with caution and require replication in larger cohorts to ensure validity. The tables that summarize genotype-phenotype associations are not clearly formatted, and the clarity of these tables could be improved.
A more fruitful discussion would be one that integrates findings with those from previous GWAS or microbiome studies. A more critical examination is necessary to determine the impact of averaging temporal data on genetic associations. The potential influence of environmental or behavioral factors on vaginal microbiota diversity warrants further investigation. The recommendations for future research are somewhat generic; more specific recommendations would improve impact.

It has been noted that certain sections of the text contain dense jargon, which may be more effectively mitigated through the implementation of simpler language in certain sections. Minor typographical errors and inconsistencies in terminology (e.g., reference to CST4 vs. CST-4) occur

on occasion. The inclusion of more informative legends and clearer labeling would enhance the comprehensibility of the figures. It is recommended that the formatting of select tables be revised to enhance clarity, particularly those that present cohort characteristics.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Metagenomics and virus discovery

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 27 Jan 2026

Samuel Alizon

This manuscript presents a significant and novel investigation into the genetic basis of vaginal microbiota diversity using GWAS approaches within a European cohort. The utilization of Shannon diversity from 16S metabarcoding data as a quantitative trait constitutes a novel approach that is substantiated by heritability estimates. However, the study's impact is constrained by its relatively modest sample size and the absence of genome-wide significant findings. The functional interpretation of the associated genomic regions remains speculative, and the averaging approach for longitudinal microbiota data may mask critical temporal variation influencing host-microbiota interactions. With moderate revisions aimed at clarifying the methodology, better addressing the limitations, and improving the presentation, this study would contribute to the emerging field of microbiome genetics. Subsequent, more extensive studies with more comprehensive genomic coverage and meticulous control of environmental confounders are necessary to validate and expand upon these findings.
Reply: Thank you for this accurate summary of our work!

_____ *Q 1.1 The abstract could better emphasize the study's limitations, such as sample size and statistical power. It would be beneficial to explicitly state the novelty of using microbial genetic data rather than symptoms earlier in the abstract.*

Reply: We updated the abstract to mention these points. More precisely, we stress that earlier studies tend to use phenotypic traits as proxy instead of microbial DNA in the Background and we mention the size limitation explicitly in our conclusion.

_____ *Q 1.2 The extraneous background information from previous GWAS studies could be streamlined to reduce redundancy. In order to avoid the potential for overgeneralization, it would be advisable to enhance the referencing and clarification of certain statements pertaining to ethnicity and behavioral factors. A brief discussion of the challenges and opportunities inherent in the utilization of 16S metabarcoding data for the purpose of defining traits in the context of GWAS is warranted.*

Reply: We restructured the introduction to improve the reading. We now first present the general context. Then we grouped all the elements about the vaginal microbiota and its known association with self-reported ethnicity. Next, we discuss in the additional details what, to our knowledge, are the only four studies that have tried to identify human genetic bases for variations in vaginal microbiota composition. Finally, we introduce the present study and stress the challenge associated with using 16S metabarcoding data.

_____ *Q 1.3 Further elaboration is necessary to clarify the averaging approach for longitudinal microbiota samples. The utilization of the mean as a metric has the potential to obfuscate the temporal dynamics inherent in the data. The rationale for imputation and its effect on downstream analyses could be expanded. There has been a paucity of discourse regarding the implementation of strategies for the control of potential confounders, including but not limited to sexual behavior, antibiotic utilization, and hormonal status.*

Reply: Thank you for pointing this out since summarising microbiota data is indeed a delicate step in a GWAS. Building on the introduction, we now further explain why we decided to average the values from each participant. Ideally, we would have liked to perform a hierarchical clustering on the whole follow-up. Unfortunately, the number of samples per participant was highly variable. However, based on the study from Tamarelle et al. (2024), which we now cite explicitly in the methods, this would likely have led to similar results since they show that a single sample is highly informative for an 18-month trajectory. Regarding the confounders, we now stress that these are important because there are known associations with vaginal microbiota composition. We also explain how we chose the confounders analysed in the study.

_____ *Q 1.4 The strongest associations do not reach genome-wide significance, a fact that merits greater discussion. The genomic regions identified are distant from known coding or regulatory regions, resulting in a speculative functional interpretation. Due to the modest sample size, the findings must be interpreted with caution and require replication in larger cohorts to ensure validity. The tables that summarize genotype-phenotype associations are not clearly formatted, and the clarity of these tables could be improved.*

Reply: We now explicitly mention when some results are non-significant. We also rephrased the opening of the Discussion to stress all these limitations. However, note that thanks to the insightful suggestion from Reviewer #2 to use Simpson diversity, most of the SNPs we

find are now significant (although quite dependent on a single participant, which we stress as well). Regarding the genotype-phenotype associations tables, we now more clearly indicate what the contents correspond to. We also expanded the captions and no longer refer to Table 1.

_____ *Q 1.5 A more fruitful discussion would be one that integrates findings with those from previous GWAS or microbiome studies. A more critical examination is necessary to determine the impact of averaging temporal data on genetic associations. The potential influence of environmental or behavioral factors on vaginal microbiota diversity warrants further investigation. The recommendations for future research are somewhat generic; more specific recommendations would improve impact.*

Reply: We expanded on the discussion about the specificity of working with microbiota as a response trait in the introduction and, more thoroughly, in the discussion. About the effect of environmental/behavioural covariates, this is the case for most traits included in GWAS. The vaginal microbiota is actually interesting as a study subject because these covariates are well characterised (largely thanks to the high level of clustering of the microbiota), which makes it easy to control for them.

_____ *Q 1.6 It has been noted that certain sections of the text contain dense jargon, which may be more effectively mitigated through the implementation of simpler language in certain sections. Minor typographical errors and inconsistencies in terminology (e.g., reference to CST4 vs. CST-4) occur on occasion. The inclusion of more informative legends and clearer labeling would enhance the comprehensibility of the figures. It is recommended that the formatting of select tables be revised to enhance clarity, particularly those that present cohort characteristics.*

Reply: We went through the manuscript to try and remove the jargon as much as possible, to homogenise the notations, and clarify the captions. In particular, we mention as often as possible that CST-4 is lactobacillus-poor to facilitate the reading.

Competing Interests: NA