Citizen-science map of the vaginal microbiome

Sarah Lebeer (sarah.lebeer@uantwerpen.be)
UAntwerpen https://orcid.org/0000-0002-9400-6918

Sarah Ahannach
UAntwerpen

Stijn Wittouck
UAntwerpen

Thies Gehrmann
UAntwerpen

Tom Eilers
UAntwerpen

Eline Oerlemans
UAntwerpen

Sandra Condori
UAntwerpen

Jelle Dillen
University of Antwerp

Irina Spacova
UAntwerpen https://orcid.org/0000-0003-0562-7489

Leonore Vander Donck
UAntwerpen

Caroline Masquiller
UAntwerpen

Peter Bron
University of Antwerp

Wannes Van Beeck
UAntwerpen

Charlotte De Backer
UAntwerpen

Gil Donders
UAntwerpen

Veronique Verhoeven
veronique.verhoeven@uantwerpen.be

Biological Sciences - Article

Keywords: Citizen science, vaginal microbiome, lactobacilli, large-scale remote sampling, population cohort, lifestyle impact

Posted Date: February 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1350465/v1

Citizen-science map of the vaginal microbiome

Sarah Lebeer[1,*,$], Sarah Ahannach[1,$], Stijn Wittouck[1,$], Thies Gehrmann[1,$], Tom Eilers[1], Eline Oerlemans[1], Sandra Condori[1], Jelle Dillen[1], Irina Spacova[1], Leonore Vander Donck[1], Caroline Masquillier[2], Peter A. Bron[1], Wannes Van Beeck[1], Charlotte De Backer[3], Gilbert Donders[4,5,6,°], Veronique Verhoeven[7,°]

*corresponding author

$shared first authors

°shared responsible clinicians

Affiliations

[1]Department of Bioscience Engineering, Research Group Environmental Ecology and Applied Microbiology, University of Antwerp, Groenenborgerlaan 171, 2020 Antwerp, Belgium

[2]Department of Sociology, Center for Population, Family and Health, University of Antwerp, Sint-Jacobstraat 2, 2000 Antwerp, Belgium

[3]Department Communication Sciences, University of Antwerp, Sint-Jacobstraat 2, 2000 Antwerp, Belgium

[4]Department of Obstetrics and Gynaecology, University Hospital Antwerp, Drie Eikenstraat 655, 2650 Edegem, Belgium

[5]Regional Hospital Heilig Hart, Kliniekstraat 45, 3300 Tienen, Belgium

[6]Femicare, Clinical Research for Women, Gasthuismolenstraat 33, 3300 Tienen, Belgium

[7]Department of Family Medicine and Population health (FAMPOP), University of Antwerp, Doornstraat 331, 2610 Antwerp, Belgium

Keywords

Citizen science / vaginal microbiome / lactobacilli / large-scale remote sampling / population cohort / lifestyle impact

Abstract

The vaginal microbiome is crucial for women's health and reproduction, but its ecology and determinants in the general population are still unclear. This lack of a reference framework hampers much-needed innovations in diagnostics and therapeutics. Here, we remotely mapped the vaginal microbiome of 3,345 women in Western Europe via a citizen-science approach. More than 75% of the vaginal samples were dominated by *Lactobacillus* taxa, but not in discrete community state types. Compositional correlation network analysis validated with public data pointed at six main modules of interacting microbes: a *Lactobacillus crispatus-*, *Lactobacillus iners-, Gardnerella-*, *Prevotella-, Anaerococcus-,* and gut-derived module. In the first module, *Limosilactobacillus* taxa were functionally connected to *L. crispatus* and *Lactobacillus jensenii*. This module was positively associated with the luteal phase of the menstrual cycle and negatively with the number of vaginal complaints, while the *Gardnerella*-module was associated with discharge and increasing age. Contraceptives with oestrogen correlated with higher levels of the *L. crispatus*- and less of the *Gardnerella*-module, with the opposite found for a hormonal intrauterine device or having multiple partners. Mothers had lower relative abundance of the *L. crispatus*-module and more *Bifidobacterium, Lactobacillus gasseri* and *Streptococcus*. Other covariates such as BMI, menstrual pads and cups, smoking and dietary habits were also associated with the microbial constellation. These findings suggest that lifestyle interventions have potential to improve vaginal health when combined with dedicated therapies.

Introduction

The vaginal microbiome plays a central role in women's health and reproduction, but detailed knowledge about its general ecology and the host-side determinants of its composition is lacking. For more than a century, the vagina has been considered a rather simple ecosystem characterized by a low diversity and a high abundance of lactic acid-producing bacteria[1]. In 1892, Döderlein and colleagues described a gram-positive bacterium, as the key bacterium in the vagina[2]. Since then, it has been well established that *Lactobacillus* taxa are the most dominant bacteria in female populations from European and Asian[3–5]. The dominance of these lactobacilli in the vagina is linked to health: when disrupted by an overgrowth of anaerobic bacteria such as *Gardnerella vaginalis* during bacterial vaginosis (BV), or because of inflammation during aerobic vaginitis (AV) or pelvic inflammatory disease (PID)[6,7], an increased susceptibility to conditions such as sexually transmitted diseases[8–10] and adverse reproductive outcomes[11,12] is observed. In 2020, the taxonomy of the family *Lactobacillaceae* was significantly revised[13]. This was an important taxonomic update, as it revealed that the typical vaginal species all belong to the same genus: the *Lactobacillus* genus *strictu sensu*. In addition, the update highlighted the evolutionary distances to other lactobacilli such as *Lacticaseibacillus rhamnosus*, *Lactiplantibacillus plantarum* and *Limosilactobacillus reuteri* that are commonly studied as gut probiotics[13].

With the advent of amplicon sequencing, the vaginal microbiome has been generally described based on five vaginal community state types (CSTs)[3]. *L. crispatus* is dominant in CST I, *L. gasseri* in CST II, *L. iners* in CST III and *L. jensenii* in CST V. CST IV is not dominated by *Lactobacillus*, but rather a mix of more facultative or strict anaerobes such as *Gardnerella, Atopobium, Prevotella,* and *Finegoldia*[14]. This CST IV is found in asymptomatic women but is more associated with dysbiosis and problems such as BV. The recent VALENCIA (VAginaL community state typE Nearest CentroId clAssifier) study proposed thirteen CSTs, based on meta-analysis of 1,976 women from different study cohorts, with particularly extra subdivisions for this CST IV[15]. The CST framework has been very useful to simplify high-dimensional microbial community datasets and facilitate statistical analyses. However, it is currently unclear how well the vaginal CSTs reflect the inherent biology.

To better understand the ecology and function of vaginal lactobacilli and other microbiome members and to better design diagnostic and therapeutic options for vaginally associated diseases, more reference datasets are also necessary. So far, female populations in North America, Scandinavia and South-Africa have been mainly characterized[3,15–18], while there seems to be a vast knowledge gap on the vaginal microbiome in other populations. Moreover, other valuable information can come from human-animal comparisons[19]. Humans appear to be the only animals with a vagina mostly dominated by *Lactobacillus* taxa under healthy conditions[20,21]. This unique phenomenon is at present not yet well understood, but the typical hormonal fluctuations throughout the menstrual cycle, particularly estrogen[14]; the glycogen accumulated in the vaginal epithelial cells[22]; the typical human diet since agriculture was introduced[23]; and the strong antimicrobial capacity of lactobacilli that protect the limited offspring of humans from infections[20] have all been suggested to play a role. A detailed mapping of lifestyle and personal characteristics in relation to the vaginal microbiome can aid to better understand the unique build-up of the human vaginal microbiome.

In this citizen science-based self-sampling study, we mapped the vaginal microbiome in a large cohort in Belgium, with a particular focus on the prevalence and abundance of key taxa of the lactobacilli, and their association with life-course and lifestyle factors. Two self-collected vaginal swabs were donated by 3,345 women ranging from 18 to 98 years old: one for 16S rRNA amplicon sequencing and one for culturing and metabolic analyses. The project was named 'Isala', after Isala Van Diest (1842-1916), honoring the very first female doctor in Belgium.

Results

Citizen Science-based study cohort. The call for participation was launched in Belgium (Western Europe) in March 2020. Within ten days, 6,007 participants registered on the website and registrations were closed (https://isala.be/en/). A total of 4,682 of the original registrants completed the questionnaire with an average completion time of 49 minutes and received the self-sampling kit (Figure 1A-B). The sole exclusion criteria were pregnancy and being younger than 18 years. Of the participants that filled in the questionnaire, 3,345 provided two vaginal swabs, allowing microbiome, culturomics and metabolomics analyses (Figure 1C). The mean age and body mass indexes (BMIs) of the included participants were 31.8 +- 9.5 years and 24.3 +- 4.6 kg/m^2, respectively (Figure 1D).

The call was directed towards the general female population outside a clinical setting. Indeed, 69.7 % of the women did not report a single vaginal health symptom at the time of sampling based on the questionnaire data (Table S3). 18.3% had one self-reported vaginal symptom, ranging from redness, dryness, odor, increased and/or discoloration of discharge, pain during intercourse, itching, swelling, burning, to urinary infection. Only 7% and 2.6% reported two, or three symptoms respectively. Nevertheless, more than 50% and 70% of the participants answered to have at least once experienced a fungal infection or bladder infection, respectively, which are prevalences in agreement with previous studies[24,25].

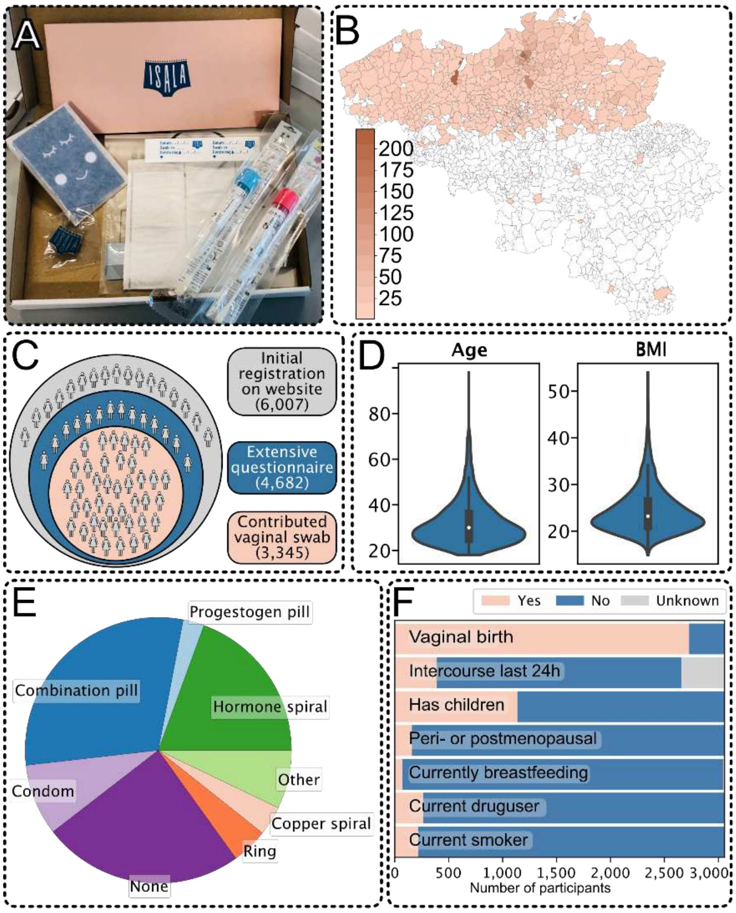


Figure 1 – Characterization of the Isala study cohort and key physiological, behavioral, lifestyle and environmental factors of the participating women. (A) The self-sampling kit sent to the participants via the national postal service. (B) Geographical overview of the participants that sent in samples for this project, by overlaying their zip codes on a map of the Flemish region and some cities from the Wallonia region of Belgium. Darker colors represent higher numbers of participants with that specific zip code. (C) An overview of the population cohort that registered within ten days after the first announcement, with their different citizen-science roles to the Isala project: minimal involvement by expressing online interest as potential donor via website and answering five questions on age, pregnancy, contraceptive use, country of living until three years and zip code (gray), partial involvement by filling out the extensive questionnaire (blue) and full involvement as donors and 24h follow-up questionnaire (pink). The distribution of a selection of the questionnaire variables: (D) age and BMI, (E) reported contraceptive use of the whole cohort, (F) a subset of the binary variables.

5.2% of the Isala participants were menopausal. 30.2% used a combined oral contraceptive pill, 19.9% a hormonal intrauterine device, 13.1% condoms, 3.7% a copper intrauterine device and 2.5% a progestogen-only pill (Figure 1E). Other forms of contraception (implant, cup, periodic celibacy, sterilization of the participant and/or partner, etc.) were less frequent at 5.7% combined (Table S3). About four out of ten (39.2%) of all women had ever been pregnant. 16.0% reported sexual intercourse within 24h before sampling. 9.1% identified themselves as a smoker, while 8.6% reported drug use (Figure 1F). As expected, age was significantly correlated with BMI, previous pregnancy, having kids and menopause (Figure S1). 4.8% of the participants were not born in Belgium, and 10.0% identified with a culture besides the Belgian one. Ethnicity or race, as previously collected as metadata in US vaginal microbiome studies (Caucasian, African-American, Asian, Hispanic)[26] was not explicitly questioned, since considered not relevant to the Belgian population with its diverse ethnography[27]. 163 participants (5.4%) reported to be part of families below the national poverty threshold, calculated based on the total family income and number of dependents[28].

Dominance of *Lactobacillus* taxa. 3,345 fully involved Isala donors delivered vaginal samples between July and October 2020, of which 3,196 (96.6%) passed quality control based on estimated DNA concentrations. The high-quality samples totaled over 82 million high-quality V4 16S rRNA read pairs, ranging from 2,126 to 376,242 read pairs per sample with an average of 25,909. Read pairs were merged and denoised into a total of 4,972 unique Amplicon Sequence Variants (ASVs). Short-read 16S rRNA gene sequencing studies generally do not allow for species-level identification[29]. This also applies to many vaginal species: for example, the species *L. jensenii* and *Lactobacillus mulieris* both occur in the vagina, but cannot be discriminated using 16S rRNA gene regions[30]. To be able to analyze the data at the functionally interpretable genus level, but still be able to discriminate between the “big four” vaginal *Lactobacillus* species, the *Lactobacillus* genus was divided into subgenera based on a high-quality core genome phylogeny (Figure 2C-D and Figure S2). This resulted in nine subgenera, four of which are known to be associated with the vagina: the *L. crispatus* group, *L. iners* group, *L. jensenii* group and *L. gasseri* group. To validate this subgenus-level classification approach, shotgun metagenomic sequencing was done for a subset of samples (n = 18, Figure

2C). For the four subgenera containing the four typical vaginal *Lactobacillus* species, the relative abundance correlations between the methods were remarkably large (Figure S3).

For each sample, the dominant (sub)genus was then determined as the (sub)genus with the largest relatively abundance over 30%. Employing these criteria, the *L. crispatus* group (163 ASVs) dominated the largest number of samples (43.2% of the participants), followed by the *L. iners* group (120 ASVs) (27.7%) and *Gardnerella* (49 ASVs) (9.8%). Several smaller dominant taxa also occurred, namely the *L. jensenii* group (54 ASVs) (3.5%)*, Prevotella* (421 ASVs) (3.4%), the *L. gasseri* group (56 ASVs) (3.2%), *Bifidobacterium* (18 ASVs) (1.8%) and *Streptococcus* (52 ASVs) (1.2%) (Figure 2A-B).

Because of the citizen-science nature of the project, the personal vaginal microbiome profiles were communicated to the participants before the submission of this manuscript (Figure S4 and https://isala.be/en/results/). Participants received information about the top eight taxa in the dataset accompanied by information for non-microbiology experts (Figure S5). A feedback questionnaire (n = 2,000) showed that 83% of participants who received their results, perceived them as easy to interpret and 99.6% of participants would volunteer again in future Isala endeavors.

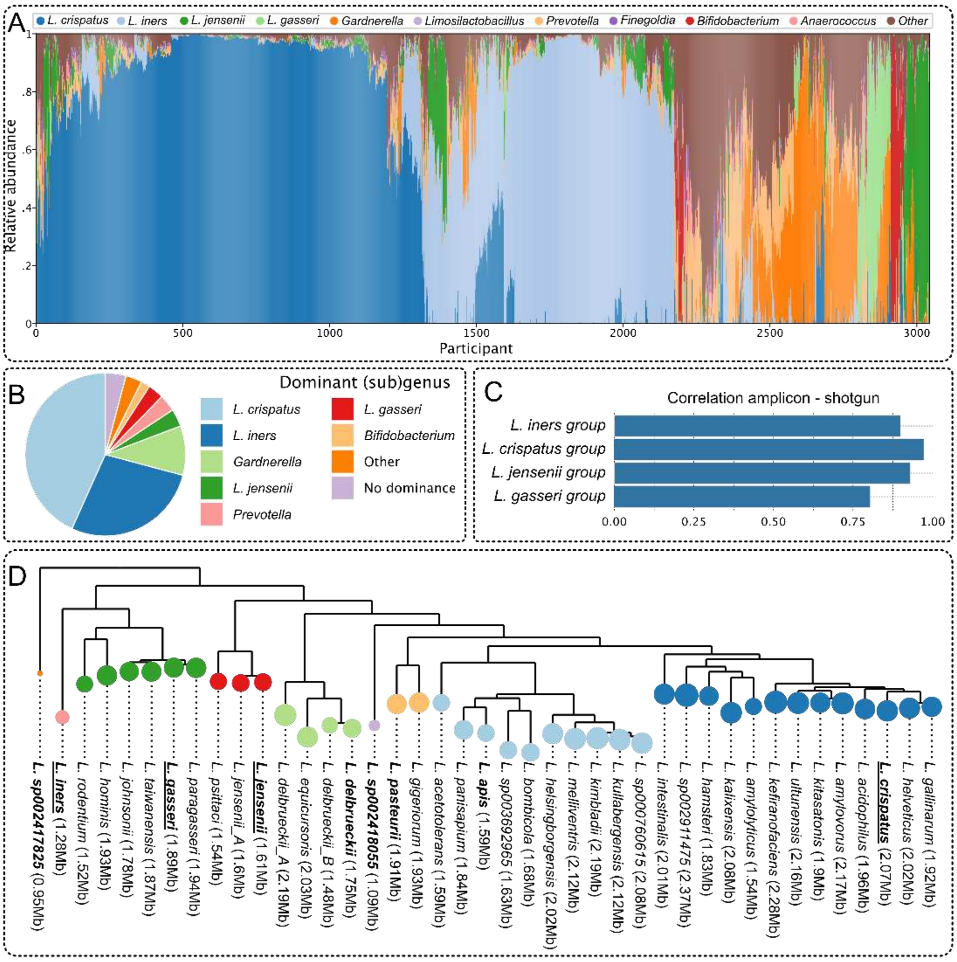


Figure 2 - Overview of the most abundant taxa in the vaginal microbiome of the Isala cohort, with particular focus on the *Lactobacillus* taxa. (A) Stacked bar chart describing the microbiome composition of all participants in the study in terms of the 10 most abundant taxa. (B) Occurrence of the most dominant taxa in the vaginal microbiome of the Isala cohort based on the highest taxonomic resolution possible with our available data. Dominance was defined as the most abundant taxon that constituted at least 30% of the profile. "Other" refers to the number of samples where a different (sub)genus was dominant from the seven that are shown; "no dominance" refers to the number of samples where not a single (sub)genus reached at least 30% abundance. (C) Validation of the 16S amplicon sequencing pipeline, including classification to *Lactobacillus* subgenera, with shotgun sequencing data (n = 18). For the "big four" *Lactobacillus* subgenera, the spearman correlations between their relative abundances in the amplicon and shotgun samples are shown. (D) Maximum-likelihood phylogeny of species of the genus *Lactobacillus* inferred from the amino acid sequences of 100 single-copy core genes. Colors indicate the nine custom-defined subgenera used in this study. Bold tip labels indicate representative species of the subgenera. Species names were taken from the Genome Taxonomy Database[31], which splits species that are very diverse, yielding e.g., *L. delbrueckii_A* and *L. jensenii_A,* the latter recently identified as *L. mulieris*[32]. The size of the circles reflects the genome size of representative genomes of the species (with the average genome size also put between brackets).

Vaginal community structure. To enable a detailed map of the different constellations of the vaginal microbiota in our cohort, samples were embedded in a two-dimensional t-SNE space[33]. t-SNE projects a high dimensional space into a low dimensional space while aiming to preserve inter-sample distances, placing higher weight on smaller distances to preserve sample neighborhoods. This allows a better global representation of the diversity compared to other commonly used approaches such as PCoA plots[33]. This t-SNE plot was annotated with the two most dominant taxa per sample (Figure 3A-B). Several high-density regions were observed in this two-dimensional representation that broadly corresponded to the five previously described CSTs[3], but these high-density regions were connected by intermediate regions (Figure 3A). A clear example was provided by the *L. crispatus* and *L. iners* high-density regions, which were connected by samples with *L. crispatus* and *L. iners* as the two most abundant taxa. This was the case for 454 samples, of which 22% contained *L. crispatus* and *L. iners* in near-equal proportions (Figure 3C, gray dashed enclosure). This observation suggests that the previously described CSTs are not distinct possibilities in vaginal community composition. This is especially apparent when visualizing the samples based on the second most dominant (sub)genus (Figure 3B) and the relative abundance of the top (sub)genus (Figure 3C). Intermediate regions can be observed in which at least two subgenera are co-dominant, with the same patterns observed in the datasets aggregated in the VALENCIA study[15] (Figure 3D-G). As in the Isala data, samples dominated by *L. iners* and *L. crispatus* at near equal abundances were also observed here.

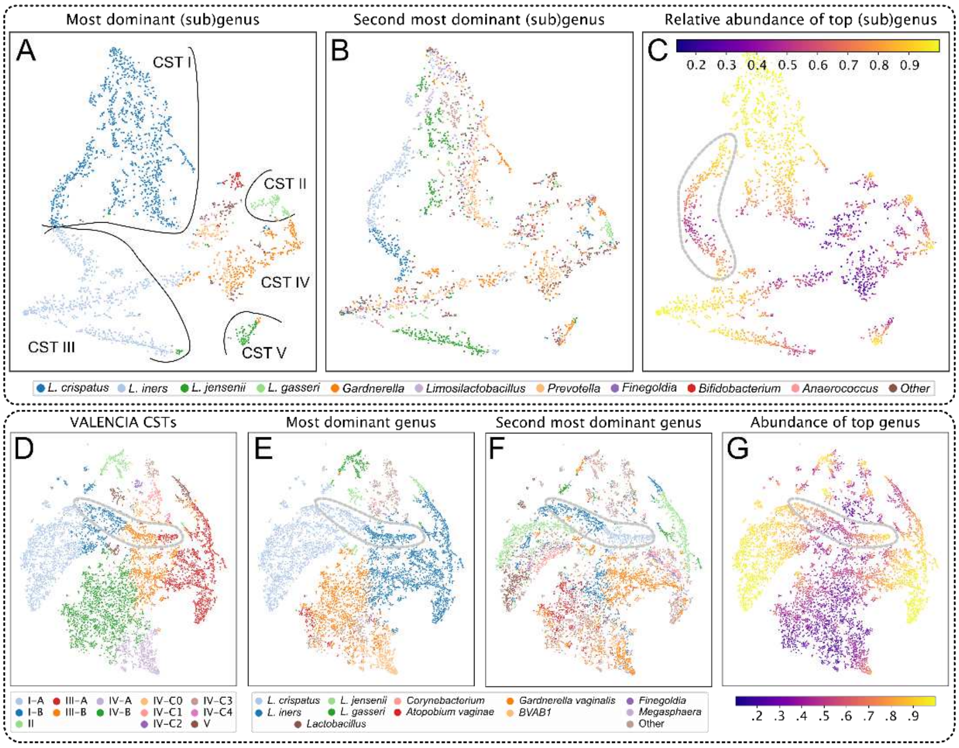


Figure 3 – Vaginal microbiome structure of the Isala cohort. (A) t-SNE plot of microbiome samples in the Isala study. Embedding colored by the most abundant (sub)genus. Broad community state types (CSTs) are delineated with black lines, except CST IV, which is composed of the remaining samples. (B) Samples are colored by the second-most abundant (sub)genus. (C) Samples are colored by the largest relative abundance level in each sample. (D) Structure of the vaginal microbiome of the VALENCIA public dataset. A t-SNE plot of all microbiome samples of the VALENCIA dataset (multi-temporal samples per participant included), colored by the 13 CSTs presented in that paper. CST I—*L. crispatus* dominated (A high relative abundance, B lower relative abundance), CST II—*L. gasseri* dominated, CST III—*L. iners* dominated (A high relative abundance, B lower relative abundance), and CST V—*L. jensenii* dominated. CST IV-A - *Candidatus Lachnocurva vaginae* (BVAB1) with some *G. vaginalis.* CST IV-B - *G. vaginalis* with low relative abundance of *Ca. L. vaginae*. CST IV-C0 - *Prevotella*, CST IV-C1—*Streptococcus,* CST IV-C2—*Enterococcus* dominated, CST IV-C3—*Bifidobacterium* dominated, and CST IV-C4—*Staphylococcus* dominated. Samples of the VALENCIA dataset colored by (E) the most dominant genus, (F) the second most dominant genus, (G) and by the largest relative abundance level in each sample. The branching point between *L. crispatus* dominated and *L. iners* dominated samples is indicated with a grey line. Of note, BVAB1 corresponds to genus EU728721_g in the Isala dataset, where it only occurred in 1.4% of the participants (not visualized in panel A-B because not in top 10).

The correlation between taxa abundances was investigated with SparCC, considering the compositionality of the relative abundance data[34]. Six main modules of intercorrelated taxa were determined (Figure 4). The first module contained the *L. crispatus group*, *L. jensenii group,* and *Limosilactobacillus*. Correlations between the taxa in this module were weakly positive (r = 0.18 – 0.40). A second module was assigned to a group of taxa that included

Gardnerella, Sneathia, *Atopobium* and *Aerococcus* (*Gardnerella* module, r= 0.11-0.5). A third module contained the relatively strongly correlated *Anaerococcus*, *Peptoniphilus* and *Finegoldia* taxa (*Anaerococcus* module, r= 0.1-0.71), together with some more weakly correlated taxa such as *Staphylococcus.* A fourth module was composed of *Prevotella* and *Dialister* (r=0.78), which jointly correlated positively with both the *Gardnerella* and *Anaerococcus* modules, while the latter two were negatively correlated with each other. A fifth module was composed of taxa associated with the gut, including *Ruminococcus, Bacteroides,* and *Subdoligranulum* (Gut module, r=0.16-0.28). Interestingly, the Gut module was positively correlated with the *L. crispatus* module. Finally, the sixth main module constituted the *L. iners* group and the genus *Ureoplasma*. A few taxa did not show any strong correlations with other taxa, notably *Bifidobacterium, Streptococcus* and the *L. gasseri* group. Yet, when computing SparCC correlations in the VALENCIA dataset, we identified a striking concordance with the modules identified in the Isala dataset (Figures S6 and S7). In both datasets, the *L. crispatus* module showed moderately negative correlations to the taxa in the *Gardnerella*, *Anaerococcus* and *Prevotella* modules (-0.22, -0.15, and -0.27 respectively), which is in line with the previously documented inhibitory capacity of *L. crispatus*-dominated communities against these potential vaginal pathobionts [35,36].

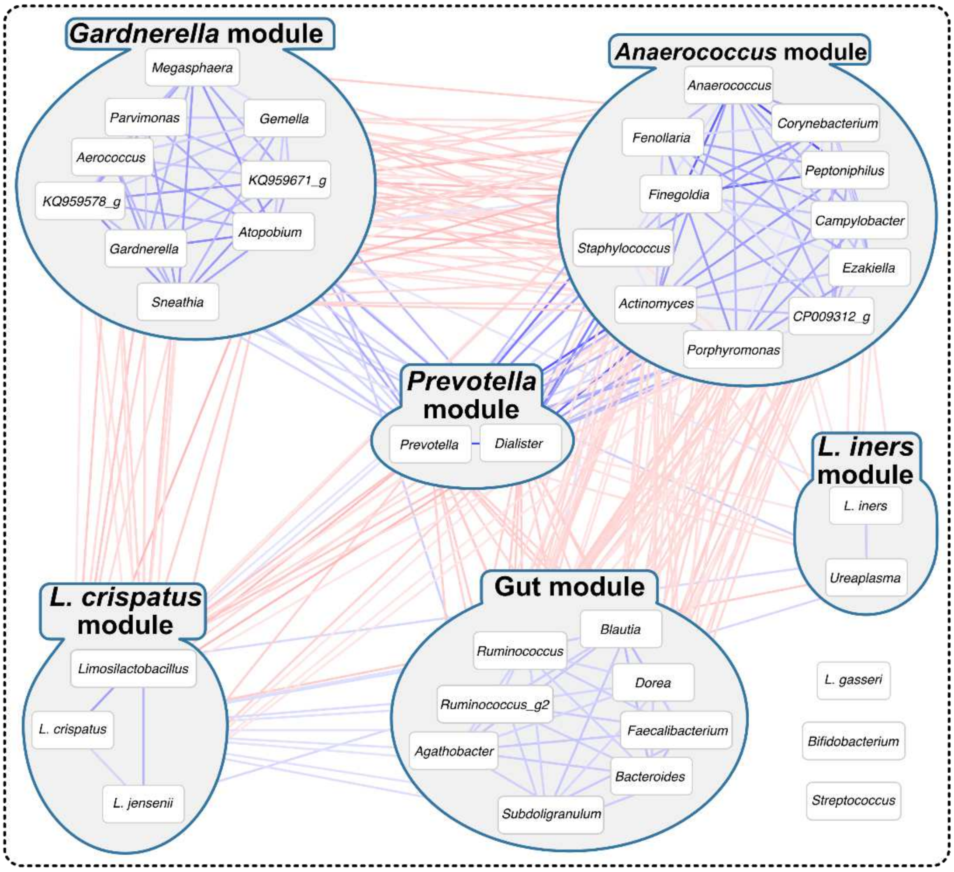


Figure 4 – Six main modules of interacting microbes as defined by a compositional correlation analysis. Modules are enclosed in gray. Positive correlations in blue, negative correlations in red. Thickness of the line indicates the strength of the correlation. Exact correlations are given in Figures S6 and S7.

Our analysis also pointed at a strong correlation between the genus *Limosilactobacillus* and both the *L. crispatus* and *L. jensenii* groups (which were also positively correlated with each other). *Limosilactobacillus* taxa did not show a high average relative abundance (0.4%) in our dataset, but had a surprisingly high prevalence of 47.8% (Figure 2A and Table S1). Based on a case-by-case ASV sequence comparison with a 16S reference database, we could assign the ASVs classified as *Limosilactobacillus* to one of three groups within the genus: a *Lactobacillus reuteri* group, the species *Limosilactobacillus coleohominis* and the species *Limosilactobacillus fermentum*. The *L. reuteri* group contained the species *Limosilactobacillus reuteri*, *Limosilactobacillus vaginalis* and five other species that are not known to occur in the human vagina. We found a prevalence of 43.7% for the *L. reuteri* group, 11.5% for *L.*

coleohominis and 4.1% for *L. fermentum*. In addition to our large dataset with amplicon-sequenced samples, we also inspected the 264 vaginal metagenomes of the VIRGO metastudy for the presence of *Limosilactobacillus* species. The most prevalent species were *L. coleohominis* (25%), *L. vaginalis* (20%) and *L. fermentum* (1%)[37]. *L. fermentum* was most frequently cultured from a subset of 592 vaginal swabs, with even more isolates obtained than for *L. crispatus* and *L. jensenii* based on standard growth conditions for lactobacilli (Table S1). Overall, culture of the vaginal lactobacilli was cumbersome under the standard conditions and remains to be further optimized.

Impact of host covariates on the vaginal microbiome. We then analyzed the association of personal data with key features of the vaginal microbiome (Figure 5). As an alternative to reducing the dimensionality of the microbial community data through a classification into CSTs, alpha and beta-diversity metrices, twelve individual (sub)genera of interest and eigentaxa (see Methods) of the four largest modules of intercorrelated taxa were selected for association testing. The functional relevance of this latter approach was confirmed by the association observed between change in discharge and an increase in the *Gardnerella*-module, but not with specific taxa. Similarly, a lower relative abundance level of the *L. crispatus*-module was associated with an increased number of vaginal complaints specifically. Considering age had the largest effects, the data were also adjusted for this parameter.

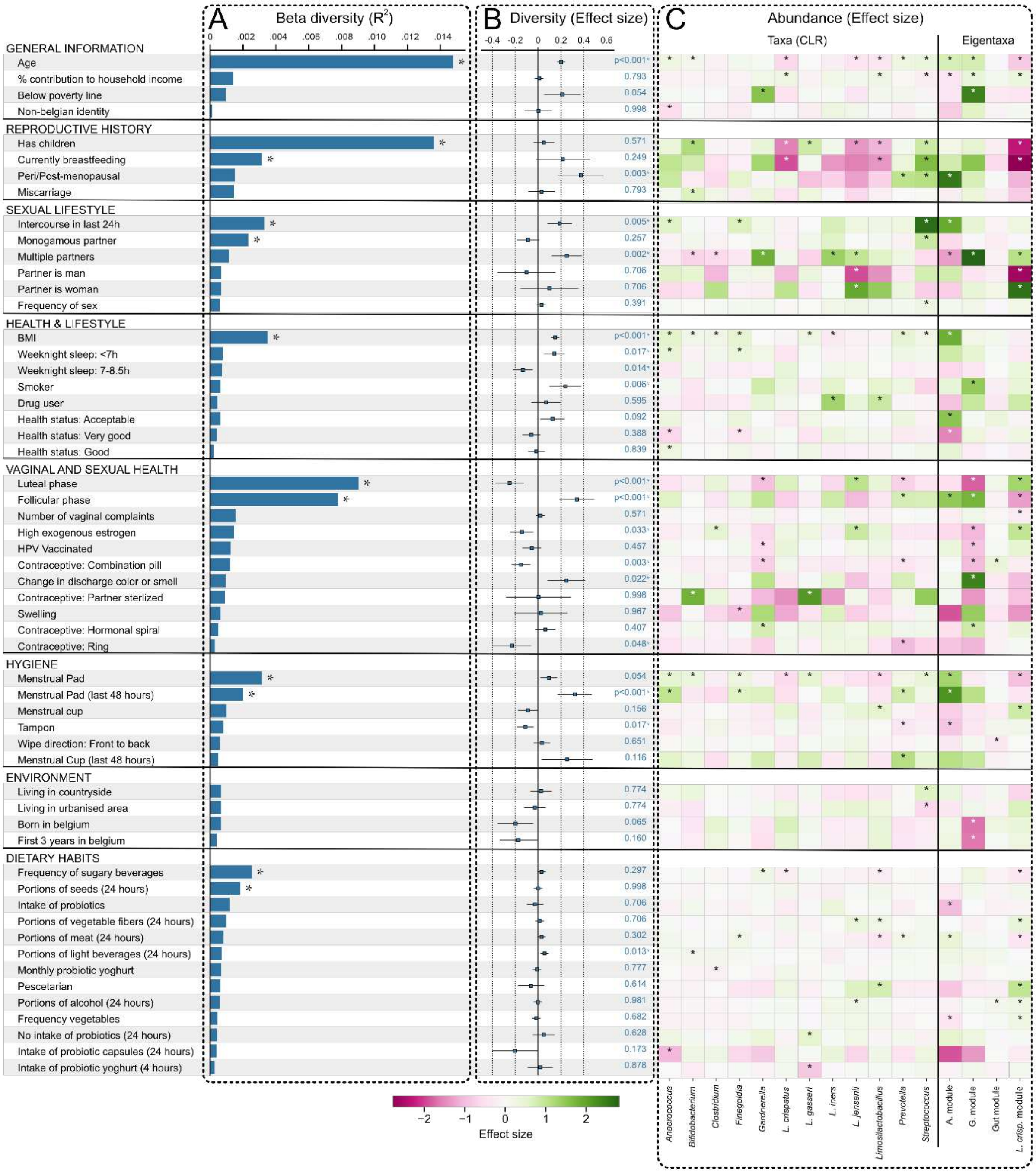


Figure 5 - Statistical analysis of the association of different personal, reproductive, lifestyle, health, hygiene, environmental and dietary factors with the vaginal microbiome space. Each panel displays effects on different levels of the microbiome: (A) the effect on the beta-diversity between the samples (Adonis test), (B) the effect on the alpha-diversity of the samples, (C) the effect on the abundances of specific taxa and on the eigentaxa of the modules discovered in the SparCC correlation analysis. The A and G modules refer to the *Anaerococcus* and *Gardnerella* modules, respectively. Asterisks represent significant associations (FDR adjusted and using a threshold of 0.05; white and black asterisks are merely for visualisation purposes). The number of samples for each question was almost the entire study (n = 3,043 participants). Due to missing data or specific comparisons, this can deviate, and detailed counts are provided in Table S3.

Besides age, having had children had the strongest association with beta-diversity, explaining 1.4% of the microbiome variation. It was significantly negatively associated with the abundance levels of *L. crispatu*s, *L. jensenii* and *Limosilactobacillus* (the *L. crispatus*-module), and positively with *Bifidobacterium, L. gasseri,* and *Streptococcus*. Breastfeeding at the time of sampling was correlated with beta-diversity, lower relative abundance of *L. crispatus* and *Limosilactobacillus* and higher levels of *Streptococcus*. Being “peri- or post-menopausal” did not show a significant association with the beta-diversity, but it was correlated with an increased alpha-diversity and levels of *Streptococcus, Prevotella* and the *Anaerococcus*-module. Having had intercourse in the last 24 hours was associated with a higher alpha diversity, and higher levels of *Anaerococcus, Finegoldia,* and in particular *Streptococcus.* We also investigated the associations of partnership with the vaginal microbiome. Compared to not being sexually active, having a monogamous relationship correlated with the beta-diversity and higher levels of *Streptococcus*, but no associations were noted for the alpha-diversity. However, having multiple partners was linked with a higher alpha-diversity and higher levels of the *Gardnerella*-module, but also higher levels of the *L. crispatus*-module, and less of the *Anaerococcus*-module*.* Having a male partner was associated with lower levels of *L. jensenii* and the *L. crispatus*-module, compared to having a female partner. The impact of the stage of the menstrual cycle was evaluated for pre-menopausal participants not taking any related hormonal contraceptives, with the follicular phase starting on the first day of menstruation and the luteal phase after ovulation (Figure S8). As expected, the follicular phase was associated with higher alpha-diversity, together with lower levels of the *L. crispatus*-module and higher levels of *Prevotella* and the *Gardnerella*- and *Anaerococcus*-modules, compared to the ovulation and luteal phase. The opposite was true for the luteal phase (compared to the ovulation and follicular phase). Combining the data for contraceptives with a high predicted exogenous estrogen level (combination pill, vaginal ring or patch) showed an association with an increase in the *L. crispatus*- module and less of the *Gardnerella*-module. The oral combination contraceptive pill, which disrupts the natural cycle and contains estrogen and progestin[23], correlated with lower alpha-diversity, lower relative abundances of *Prevotella* and *Gardnerella* but higher levels of the gut taxa module. Use of a ring contraceptive was linked to a significantly lower alpha-diversity and lower levels of *Prevotella*. Use of a hormonal intra-uterine device (containing only progestin) was associated with more of the *Gardnerella*-module. Having been vaccinated against HPV was linked to lower levels of the *Gardnerella*-module. Furthermore, we also observed associations for menstrual hygienic products, with a menstrual cup appearing more beneficial for the *L. crispatus*-module and pads being more associated with an increased alpha diversity. The menstrual pads also significantly reduced the *L. crispatus*-module and increased the *Anaerococcus*-module, especially when used in the last 48h. Wiping the vulva from front to back after a bathroom visit was associated with lower levels of the gut taxa module in the vagina.

Among the general health and lifestyle factors that were questioned, the largest effect was BMI, which was significantly associated with the beta-diversity, higher alpha-diversity, and higher levels of bacteria in the *Anaeroccocus*-module. Specific dietary components were also linked with the overall composition and diversity of the vaginal microbiome when adjusting for age. The consumption of sugary beverages was noticeably associated with beta-diversity, and with lower levels of the *L. crispatus* module, while the consumption of light beverages (marketed as diet, sugar-free, zero-calorie or low-calorie) in the last 24h was associated with a significantly higher alpha-diversity and higher levels of *Bifidobacterium*. A high portion of seed consumption was significantly associated with beta-diversity, but not with the specific taxa or modules that we examined. High frequency of vegetable consumption and its associated fibers, particularly in the last 24h, and being pescatarian were associated with a minor increase of *L. crispatus*-module. Ethanol consumption in the past 24h was associated with higher levels of the *L. crispatus*- and gut taxa module. Meat consumption was linked to lower levels of the *L. crispatus*-module, and higher levels of *Prevotella* and the *Anaerococcus*-module. Significantly lower levels of the *Anaerococcus* module taxa occurred when probiotic capsules were consumed in the last 24 hours. In contrast, consumption of probiotic yoghurts in the last 24h was associated with lower relative abundance of *L. gasseri*.

Additional lifestyle factors other than diet were also evaluated. Sleeping less than seven hours per weeknight corresponded to a significantly higher alpha-diversity and higher levels of *Anaerococcus* and *Finegoldia*, while sleeping between 7 and 8.5 hours corresponded to a lower alpha-diversity. In addition, smoking was associated with higher alpha-diversity, and higher levels of the *Gardnerella*-module. While taking drugs was not linked to the diversity of vaginal samples, it was linked to higher levels of *L. iners* and *Limosilactobacillus* . Income inequality within couples did not show a significant effect on the vaginal microbiome but being below the Belgian poverty threshold was linked to a higher alpha-diversity, and in particular, higher levels of *Gardnerella.* Being born in Belgium and living there for the first 3 years was associated with significantly lower levels of the *Gardnerella-module.* Furthermore, living in a more urbanized/polluted area (i.e., city center, village center, busy road, industrial zone) versus suburban/countryside environment (i.e., residential area, rural area, green zone/recreation zone) was associated with lower versus higher levels of *Streptococcus*.

All significant factors mentioned above could explain 8.01% of the variation in the vaginal microbiome, compared to 7.63% of the variations explained by covariates in a related study on the gut microbiome in the Belgian population[38].

Discussion

The Isala citizen science project on the vaginal microbiome was inspired by a strong need for a better understanding of the vaginal microbiome outside a clinical setting. The enthusiasm of participants willing to donate intimate samples is in line with the current trend of more women taking their health into their own hands. The fact that our study was fully remote had both advantages and limitations. No blood samples, clinical exams or host genetics data could be obtained, but the fully remote setting and large online questionnaire also provided us with unique opportunities to gain widespread access to samples and intimate data. Other inherent limitations of our study cohort were the slight bias towards a high socioeconomic status, like many other citizen science studies[39,40], and the fact that we had to rely on only one timepoint sampled per participant. On the other hand, the fact that intimate self-sampling could be done in the privacy of the home setting had a positive impact on the number of women willing to participate, resulting in a large, diverse set of samples with sufficient variation to study key parameters such as age, BMI, menstrual cycle, contraceptive use, menopausal status, obstetrical parameters, sexual and vaginal health, diet, income, and sleeping habits. The fact that the analysis of all samples was done within the same lab and a small timeframe minimized the technical variability. Taken together, this study set-up enabled us to obtain novel insights in the average vaginal microbiome constellation of this self-reported healthy Western European population.

The first key finding of this work was the high number of participants with a dominance of *Lactobacillus* in this Western-European population cohort: 75% of the women were dominated by *Lactobacillus* taxa, in particular by taxa belong to the *L. crispatus* and *L. iners* group, comparable to similar studies[3,20]. Subgenus or group level classification was preferred to better reflect the diversity in ASVs than generally reported. The *L. crispatus* group (163 ASVs) was detected in 43.2% of the participants. *L. iners* was dominant in 27.7% of the participants. As we and others have previously reviewed, *L. iners* has an ambiguous role in the vagina[41]. The fact that we found *L. iners* to be so prevalent in complaint-free women suggests that it is often probably rather a friend than a foe in healthy women. Yet, we observed a high diversity of ASVs for *L. iners* (120 unique ASVs), in line with previous suggestions of different clones of *L. iners* with distinct functional properties[42]. Similarly, *Gardnerella* was dominant in 9.8% of the Isala women, although it is often considered a pathobiont in the vagina. Yet, the association of *Gardnerella* with symptoms and disease appears to depend on the specific species and strains [43,44], the other members in the vaginal community[45] and the host[46]. This context- and taxon-dependent role of the vaginal bacteria highlights that it is important to capture the diversity of the vaginal ecosystem in the most biologically relevant way. From five[3] to thirteen CSTs[15] have been previously proposed. CSTs often confuse clinicians and researchers, as they have been mainly proposed for statistical and epidemiological purpose[15], and should not be interpreted as stable community state types. With t-SNE embedding analyses, we clearly showed that the vaginal microbiome space is a continuum, highlighting that CSTs should not be interpreted as the existence of fully discrete states of the vaginal microbiome, as is now also increasingly recognized[15,47,48]. For example, the two most abundant taxa, the *L. crispatus* and *L. iners* groups frequently co-occurred in varying and even equal proportions. As an alternative approach to maximally capture the diversity of the microbial space while still enabling the analysis of associations with as many metadata as possible, we introduced modules of taxa of interacting vaginal bacteria (with positive correlations within and mostly negative correlations between modules), for which we made eigentaxa for correlation analyses. The taxa-taxa correlations likely reflect relevant biologic phenomena including positive or negative microbial dependencies such as cross-feeding[49–51], inhibition via antimicrobial production[52] but also different immune or inflammation states of the host, where different "states" of the host enrich or restrict different bacteria[45]. The fact that we could validate the existence of these modules in another large independent dataset (VALENCIA) highlights their biological relevance and existence independent of our dataset, in contrast to CSTs obtained by hierarchical clustering which are more dataset dependent.

The *L. crispatus*-module probably reflects the most common healthy homeostatic state, based on the known associations of these lactobacilli with vaginal health[53] and our own observations of a reduced abundance of this module with increased number of vaginal complaints versus its increase with increasing estrogen levels. Notably, the association between this module and vaginal complaints was lost with the individual taxa, showing the added value of implementing these modules. Another unprecedent finding for this module is the prevalence and possibly stabilizing capacity of *Limosilactobacillus*. This genus was shown to be highly prevalent, with occurrence in almost 50% of the women sampled, and showed to be easier to culture than the classic big four (i.e., *L. crispatus, L. iners, L. gasseri* and *L. jensenii*). Positive interactions between different taxa of lactic acid bacteria are very common in food fermentations where lactic acid bacteria dominate. In yoghurt, for instance, *Streptococcus thermophilus* and *Lactobacillus delbrueckii* subsp. *bulgaricus* exchange crucial metabolites, a process called protocooperation[49]. In kefir, it was recently shown that *Lactobacillus* *kefiranofaciens*, which dominates the kefir community, uses kefir grains to bind together all other microbes that it needs to survive[50]. Such mutualistic interactions have also been observed for related *Lactobacillus* taxa within vertebrate hosts. For example, in the rodent gastrointestinal tract, *Lactobacillus johnsonii* needs *L. reuteri* for biofilm formation[54]. It appears plausible that a similar interaction occurs in the vagina between species of the same two genera, where one or more *Limosilactobacillus* species support *L. crispatus* and *L. jensenii* as keystone taxa. Of note, one of the most widely used vaginal probiotics, *L. reuteri* RC-14, has been shown to have the capacity to prevent BV in women with HIV[55,56] and improve the BV cure rate with single dose of tinidazole[57]. Yet, in these previous studies, it is difficult to differentiate the effect of *L. reuteri* RC-14 from the other applied probiotic strain *Lacticaseibacillus rhamnosus* GR-1[57].

While the *L. crispatus* module contains presumed health-associated taxa, three of the modules contain taxa previously associated with dysbiosis: the *Gardnerella*-module consists mostly of taxa associated with BV[58,59], while the *Anaerococcus*- and *Prevotella* modules also contain taxa previously associated with BV[45,60], but also with more inflammatory host states such as AV[6,7,61], endometriosis[62] and PID[63]. The negative correlation between the *Gardnerella* and *Anaerococcus* modules is in line with the view that BV and other inflammatory states such as AV are different forms of dysbiosis with different underlying causes[7]. In this light, the positive correlation of the *Prevotella* module with both modules is harder to explain and requires further investigation. Interestingly, the number of different vaginal complaints reported by the participants was not significantly associated with any of the three modules containing taxa known to be dysbiosis-associated, but only with a reduction of *L. crispatus* module taxa. This suggests that the presence of these modules in itself is not sufficient for a dysbiotic state to develop; such a development would require an extra host-side factor such as a lack of immune control (such as sometimes thought for BV[45,64] or the development of an inflammatory state (such as observed in AV)[6,18]. For change in discharge, it is noticeable that we found a clear association with the *Gardnerella*-module, but not with the individual taxa, highlighting again the relevance of microbe-microbe interactions. Similarly, we interpret our observation of a gut taxa module by the existence of a gut-vagina axis, which is not only a source of potential urogenital pathogens but also of beneficial colonizers. For the latter, the positive correlation with the *L. crispatus* module is of particular interest.

Having established this update picture of the vaginal microbiome constellation and collecting a large dataset of personal data via questionnaires, allowed us to then perform an in-depth analysis of covariates. We could confirm previously found associations such as for BMI[65], the contraceptive pill[66] and smoking[67]. The fact that in our dataset especially estrogen-containing contraceptives had a positive association with the levels of the *L. crispatus*-module, and were also linked to less of the *Gardnerella*-module, is in a way reassuring, given the fact that it is so widely administered in Western Europe and completely abolishes the spontaneous menstrual cycle. A disruption of the vaginal microbiome does not seem a major side effect of the combination pill, although we and many Isala participants acknowledge the existence of other side effects, including impact on mood and libido[68–70] and increased risk for venous thromboembolism[71,72] , which are important to consider when choosing the personally most suitable contraceptive method. Notably, the association of a progestin-containing IUD and increased *Gardnerella*-module found here could be included in information provided to women choosing this contraceptive method. Our data are in line with clinical data that insertion of a hormonal IUD temporarily increases BV and over time increases *Candida* spp. colonization in the vagina[73], while systemic progestin-only contraceptives appear to have mixed effects on the vaginal microbiome[74].

The life event with the most significant impact on the vaginal microbiome was having children or having been pregnant, which correlated with an overall reduction in *L. crispatus, L. jensenii* and *Limosilactobacillus* levels and an increase in *Streptococcus, Bifidobacterium* and *L. gasseri* levels. A higher taxonomic resolution was not possible, but these three genera contain taxa beneficial to babies as initial colonizers of the oral cavity and gut of newborns[75]. It has been previously shown that most women experience a postdelivery disturbance in their vaginal microbiome, characterized by a decrease in *Lactobacillus* species and increase in diverse anaerobes which persisted for up to one year[76]. In our Isala dataset, it was surprising that we observed the signature of reduction in the *L. crispatus*-module and increase in *Streptococcus, Bifidobacterium* and *L. gasseri* in all women having biological children, independent of their age. This suggests that the impact of pregnancy could be long-lasting. We have at present no explanation for this phenomenon, although we do acknowledge we have a rather young cohort (average age 31.8 +- 9.5 years). Of note, breastfeeding women (who recently delivered) showed similar and even stronger associations for reduction in *L. crispatus* and increase in *Streptococcus*. Hormonal and associated sugar-level changes during pregnancy (including lower estrogens during breastfeeding), as well as the cervix shortening could all be involved and provide interesting aspects for further research. Moreover, the fact whether childbirth has taken place by vaginal or abdominal mode (C-section), the latter with or without preceding labor (i.e., secondary or primary C-section), may have played a major role, and remains to be elucidated in further studies.

Another intriguing finding of our Isala citizen-science study is how dietary choices could have a small, but significant impact. For example, intake of vegetable fibers, alcohol consumption and being a pescatarian had a significant beneficial impact on the *L. crispatus*-module, while drinking sugary beverages had a negative impact. These associations should obviously be interpreted with care and not taken as one-on-one directions towards lifestyle improvements. Alcohol consumption, for example, was associated with a higher abundance of the *L. crispatus* module, but has an established detrimental impact on the gut microbiome[77]. By contrast, limiting intake of sugary drinks appears a lifestyle intervention that benefits multiple habitats that make up the human body. Another intriguing finding was the different associations found for probiotic capsules versus yoghurts, possibly because different strains and species are consumed with these products. Consumption of probiotic capsules was associated with a lowering of the *Anaerococcus*-module, probiotics in general in the last 24 hours with an increase of *L. gasseri* levels, while probiotic yoghurt decreased *L. gasseri* levels. Unfortunately, our questionnaires lacked detailed information on the specific species and strains in the probiotic products consumed by the Isala participants. Ultimately, dedicated intervention studies with specific foods or diets, hygienic measures and/or probiotic species and strains should further substantiate the associations found here, and help the design of dedicated pharmaceutical and microbiome interventions.

Conclusion

In this large-scale remote-sampling study, we showed that the vaginal microbiome of women from Belgium is mainly dominated by lactobacilli. We demonstrated that the vaginal microbiome is a continuum, where taxon compositions that are in-between classical community state types are frequently observed. Furthermore, we showed that most vaginal taxa show small to moderate positive or negative abundance correlations with other taxa, and that positively interacting vaginal taxa can be summarized by grouping them into modules of intercorrelated taxa. In addition, we measured 166 participant covariates through questionnaires. Our results showed that some of these factors explain a small but significant part of vaginal microbiome variation, with “having had children” explaining the largest fraction of the variation, after age. Finally, we highlighted that given conscious communication tools and style, women are eager to participate in taboo-breaking conversations as well as scientific studies aimed at improving their health. We therefore endorse citizen science as a powerful approach to facilitate large-scale intimate microbiome research and to empower citizens to impact their individual and community-level health by promoting open science-based communication on taboo subjects.

Acknowledgements

We would like to first and foremost thank all Isala volunteers for their enthusiastic participation and for donating samples. The following colleagues and students helped a lot with the Isala sampling campaign and sample processing: Ines Tuyaerts, Nele Van Vliet, Leen Van Ham, Annelize Groenwals, Samira El Messaoudi, Jana Hiers, Laura Van Dyck, Lize Delanghe, Caroline Dricot, Lore Leysen, Lara Martin Diaz, Marie Legein, Dieter Vandenheuvel, Ilke De Boeck and Eline Cauwenberghs. Strategic communicative support was provided by Liesbeth Talboom and Liesbeth Haesevoets (Studio Maria), Csaba Varszegi (Little Big Things, website), Ruth Broms and Svenja Vergauwen (Sensoa vzw), Elly Den Hond and Carmen Franken (Provinciaal Instituut voor Hygiëne), Nina Van Eekert, Naomi Biegel and Leen De Kort (Sociology Department, UAntwerpen) and Camille Allonsius (Biology Department, UAntwerpen). We much appreciate the guidance provided by Jeroen Raes on logistic lessons learned from the Flemish Gut Flora project and the Antwerp Biobank (University Hospital Antwerp) for their administrative support with biobanking the large number of samples. The Centre of Medical Genetics (University Hospital Antwerp) and Neuromics Support Facility (VIB-UAntwerpen) provided sequencing support. Lastly, we would like to thank Isala ambassador Evi Hanssen, collaborative influencers on social media and supportive science journalists for spreading the word on Isala and helping us build an online community that openly discusses vaginal health with the aim to break the taboo.

Author contributions

SL, SA, EO, SW, GD, VV and CDB designed the study and worked on the conceptualization of the research project. SL, SA, TG, TE, JD, SC, EO, IS, SW, CM and WVB worked on the questionnaire set-up and cleaned the answers. SA, SL, JD, EO, TE and LVD carried out the experimental and logistical work. SW and TG processed the sequencing data and performed the biostatistical analyses. TG, SW, SA, SC and SL worked on the visualizations. SL, SW, TG, SA, VV, GD, SC, JD, IS, PAB and CM contributed to the interpretation of the results. SL, SA, SW and TG wrote the original manuscript. All authors contributed to reviewing and editing of the final manuscript.

Funding

The authors wish to acknowledge the following funding bodies: the European Research Council (ERC; starting grant Lacto-Be 26850 of SL), the Special Research Fund of the Universiteit Antwerpen (UA BOF; DOCPRO 37054 grant of SA), the Inter-university Special Research Fund of Flanders (iBOF; POSSIBL project) and the Research Foundation - Flanders (FWO; aspirant fundamental research grant 11A0620N of SW, senior postdoc research grant 1277222N of IS and Research project FN701000004 of SL).

Competing interests

SL is a voluntary academic board member of ISAPP (the International Scientific Association on Probiotics and Prebiotics, www.isappscience.org) and chairperson of the scientific advisory board of YUN (yun.be). PAB is an independent consultant for several companies in the food and pharmaceutical industry. GD is the chairperson of Femicare vzw (femicare.net) and has worked as a medical consultant for various industries. However, none of these organizations or companies was involved in the design, communication or data analysis of this Isala study, which was fully funded by university, governmental and European funding, with the largest part funded by the ERC StG project Lacto-Be.

Methods

Study cohort and data collection

The study was approved by the Ethical Committee of the Antwerp University Hospital/University of Antwerp (B300201942076) and registered online at clinicaltrials.gov with the unique identifier NCT04319536. The call for participants was launched on March 24th, 2020 with the only inclusion criteria were being not pregnant and at least 18 years old. Within ten days, 6,007 women registered through the Isala website (https://isala.be/en/) by filling five questions on age, postal code, previous pregnancies, residence country in first three years and contraceptive use. After obtaining a digital informed consent, these participants were invited to fill out a large online questionnaire that included 137 relevant and GDPR-compliant questions on the Qualtrics platform (Qualtrics, Provo, UT, USA). The 4,681 participants that filled out the entire questionnaire were invited to fill out their address on the website to receive an Isala self-sampling kit. Eventually, 4,106 self-sampling kits were sent out and 81.5% of the kits were returned to the University of Antwerp between July-October 2020. Two vaginal swabs were self-collected in a standardized way by non-pregnant participants (n = 3,323). And 3,294 participants filled out a short follow-up questionnaire with 39 questions within 24 hours of sampling.

Each kit contained two vaginal swabs. First the eNAT^{TM} (Copan, Brescia, Italy), intended for microbiome profiling, was collected and immediately afterwards the ESwab^{TM} (Copan, Brescia, Italy), intended for culturomics and metabolomics, was collected. In the insert it was stipulated that both swabs had to be turned around 2-3 times to acquire enough biomass. Immediately after sampling swabs were to be transferred to a vial which contained the commercial transport buffer of the eNAT or ESwab and stored at home in the fridge. At last, all samples were transported on room temperature with prepaid services by the national parcel service (Bpost) with an average transport time of 2,9 +- 3,3 days (n = 3,306) from which 92,8% arrived within 7 days from sampling. Upon arrival, the eNAT swabs were stored at -20°C until further processing in the lab[78]. The ESwab was vortexed for 15 seconds and separated in two aliquots of 500µL, the first of which was stored at -80°C in a 96 tube Micronic plate with 500µL 50% glycerol, the other being centrifuged for 3 min at 13,000 g, and its supernatant stored in a 96 tube Micronic plate at -80°C as well.

16S rRNA amplicon sequencing

Before further processing, all samples were vortexed for 15-30 seconds and extracted with the DNeasy PowerSoil Pro Kit of which some manually and other automated with the QIAcube (Qiagen, Hilden, Germany) according to the instructions of the manufacturer. DNA concentration of all samples was measured using the Qubit 3.0 Fluorometer (Life Technologies, Ledeberg, Belgium) according to the instructions of the manufacturer. No less than 2 µl of each bacterial DNA sample was used to amplify the V4 region of the 16S rRNA gene, using standard barcoded forward (515F) and reverse (806R) primers[78]. These primers were altered for dual index paired-end sequencing, as described in Kozich *et al.* (2013)[79]. The resulting PCR products were checked on a 1.2% agarose gel. The PCR products were then purified using the Agencourt AMPure XP Magnetic BeadCapture Kit (Beckman Coulter, Suarlee, Belgium) and the concentration of all samples was measured using the Qubit 3.0 Fluorometer. Next, a library was prepared by pooling all PCR samples in equimolar concentrations. This library was loaded onto a 0.8% agarose gel and purified using the NucleoSpin Gel and PCR clean-up (Macherey-Nagel). The final concentration of the library was measured with the Qubit 3.0 Fluorometer. Afterwards the library was denatured with 0.2N NaOH (Illumina, San Diego California United States), diluted to 6 pM and spiked with 10-15% PhiX control DNA (Illumina). Finally, dual-index paired-end sequencing was performed on a MiSeq Desktop sequencer (Illumina). All DNA samples as well as negative controls of both PCR (PCR grade water) and the DNA extraction runs were included on the sequencing runs. In total, samples were sequenced across nine different MiSeq runs.

In order to validate our amplicon sequencing pipeline, including *Lactobacillus* subgenus classification, we sequenced samples from the Isala pilot study in Ahannach, Delanghe, *et al.* (2021)[78] with both amplicon and shotgun sequencing. These samples were processed in the same way as the Isala samples, except that the DNA extraction was performed with the HostZERO Microbial DNA Kit (Zymo Research, California, United States). These samples were sequenced across two different MiSeq sequencing runs.

Metagenomic shotgun sequencing (Isala pilot study samples)

For the metagenomic shotgun sequencing of samples from the Isala pilot study, library preparation was performed using the Nextera™ DNA Flex Library Prep or Nextera™ XT DNA Library Preparation kit (Illumina), according to the instructions of the manufacturer. For the Nextera™ DNA Flex Library Prep, 2 – 30 µL DNA sample was used to obtain input DNA with a start amount between 1 and 100 ng. For the Nextera™ XT DNA Library Preparation kit, 1 ng

DNA samples in 5 µL was used as input DNA. For both protocols, when the 1 ng input DNA could not be obtained for a certain DNA sample, the library preparation was continued with the highest available amount of input DNA. Pooling of the libraries was done individually using the Qubit 3.0 Fluorometer. During library preparation, library quality was checked using the 5200 Fragment Analyzer System with Agilent High Sensitivity NGS Fragment Kit (DNF-474). 22µL NGS Diluent Marker solution was mixed with 2µL library and ran on the Fragment Analyzer, according the instruction of the manufacturer. The NGS DNA Ladder was used as standard. Finally, the library was sequenced on a MiSeq desktop sequencer. In total, shotgun samples were sequenced on two MiSeq runs.

Creation of custom taxonomic reference databases

In order to increase taxonomic resolution for the genus *Lactobacillus*, the genus was split into nine subgenera. These subgenera were defined in three steps. First, a maximum-likelihood species phylogeny of the genus was constructed using amino acid sequences of 100 single-copy core genes from representative genomes, using the software IQ-TREE[80]. Second, the subgenera were manually defined as the minimum number of clades in the species phylogeny that would be needed to discriminate the four major vaginal *Lactobacillus* species. Finally, the subgenera were checked for monophyly against the species phylogeny of release 05-RS95 of the Genome Taxonomy Database (GTDB)[31].

To be able to classify amplicon sequences to the *Lactobacillus* subgenera, a custom 16S rRNA reference database was created. This was done by downloading 16S rRNA sequences extracted from sequenced genomes from the GTDB (release 05-RS95) as well as the GTDB taxonomy hierarchy. This dataset was subsetted to sequences of the family *Lactobacillaceae* only, and the genus *Lactobacillus* in the taxonomy hierarchy was replaced by the respective subgenera of the species. Finally, these files were converted into a DADA2-compatible reference database.

To be able to validate our amplicon data processing pipeline, including classification to *Lactobacillus* subgenera, we also created a custom reference database for the classification of metagenomic shotgun sequencing data. This database was created from three pieces of data: (1) representative genomes for all bacterial species, downloaded from release 05-RS95 of the GTDB, (2) the GTDB taxonomy hierarchy updated with the *Lactobacillus* subgenera, and (3) version GRCh38 of the human genome, downloaded from NCBI RefSeq[81]. These files were used to create a database in Kraken2-compatible format.

Processing and quality control of amplicon sequencing data

Quality control and processing of amplicon reads was performed with the R package DADA2, version 1.6.0[82]. First, reads with more than two expected errors were removed (no trimming was performed). Next, paired reads were merged; in this process, read pairs with one or more sequence conflicts were removed. Chimeras were then detected and removed with the removeBimeraDenovo function. The merged and denoised reads (amplicon sequence variants or ASVs) were taxonomically annotated from the phylum to the genus level with the assignTaxonomy function using the EzBioCloud reference 16S rRNA database[83]. Next, three different reclassifications were performed. First, ASVs classified to the family *Leuconostocaceae* were reclassified to the family *Lactobacillaceae* to be in line with the recent taxonomic update[13]. Second, the *Lactobacillaceae* ASVs were reclassified on the genus level to the new genera defined by Zheng et al. And finally, ASVs of the updated genus *Lactobacillus* (previously known as the *Lactobacillus delbrueckii* group) were reclassified to nine different subgenera that we manually defined based on the phylogeny of the genus.

Taxon and sample quality control was performed as follows. Non-bacterial ASVs (e.g., mitochondria and chloroplasts) and ASVs with a length greater than 260 bases were removed. Quality control of the samples was based on normalized read concentrations, which were calculated as follows. First, the total read count per sample was divided by the volume of that sample added to the sequencing library of its MiSeq run (there were nine runs in total). Next, these read concentrations were normalized by dividing them by the median read concentration of their respective run. Samples were then filtered using two criteria: (1) the normalized read concentration should be higher than 0.05 and (2) the read count of a sample should be greater than 2,000.

The Isala pilot study samples were processed in the same way as described above, with the following exceptions: (1) ASV classification was performed with a 16S rRNA reference database constructed from version 05-RS95 of the GTDB, followed by reclassification of the *Lactobacillus* ASVs only to the custom *Lactobacillus* subgenera; (2) sample quality control was based on a minimum read count of 1,000 reads.

Processing and quality control of metagenomic sequencing data (pilot study samples)

Metagenomic shotgun sequenced samples from the Isala pilot study were processed as follows. First, paired reads were filtered with the DADA2 R package, version 1.20.0[82], requiring a minimum length of 50 bases, a maximum of two uncalled bases per read and a maximum of two expected errors per read. Next, read pairs were classified from the phylum to the species level with Kraken2[84], using a custom reference database designed to validate our amplicon sequencing pipeline (including *Lactobacillus* subgenus classification). Based on the read classifications against this custom database, a read count table was constructed where the columns represent taxa and the rows represent samples. Taxa were either species or higher-level taxa for reads that were unclassified at one or more ranks. Non-bacterial taxa were removed from the data, as were samples with fewer than 500 bacterial reads.

All processing of amplicon and shotgun datasets was performed in R version 4.1.1[85], using the tidyverse set of packages, version 1.3.0[86], and the in-house package tidyamplicons, version 0.2.1.

Culture analyses

Based on the questionnaire answers a selection of self-reported "healthy" women was made. This selection took place during the course of the study, so it does not include all "healthy" women and included 592 women with: no known infection at the moment of sampling; no use of vaginal probiotics; no current smokers; good general health; no use of antibiotics/antimycotics in the past three months; no vaginal douching; no overall vaginal conditions. The 592 samples were located in the detailed inventory and retrieved from the Micronic plate at -80°C. The individual tubes were gathered to avoid melting of other samples to preserve optimal viability of the microorganisms. To obtain single colonies, 10 µL of each sample was inoculated on a small Petri dish (10mL) with three types of growth media (MRS, MRS + vancomycin, or Colombia blood, all BD Difco™) and grown for 24-48h at 37°C and 5% CO_2. After 24h the plates were checked for colonies and if present one colony of each plate was selected at random, resulting in a maximum of three isolates per participant. A part of this colony was inoculated in 10 mL MRS broth and grown overnight in 37°C and 5% CO_2. Of the overnight grown culture, 800 µL was mixed with 800 µL 50% glycerol in labelled cryovials (Greiner Bio-one Cryo.S^{TM}) and stored in -80°C. At the same time, another part of the colony was also used for colony polymerase chain reaction (colony PCR) for taxonomic identification with 16S Sanger sequencing, using universal primers 27F and 1492R.

Contraceptives, menstrual cycle and hormonal levels

Upon sampling, participants indicated when their menstrual cycle began, and also the average length of their cycle. Depending upon the contraceptive, we used this data to determine the day in which they are in, and predicted the levels of endo and exogenous levels of estrogen and progestin. Peri and post-menopausal women were excluded from this analysis.

Statistical analyses

t-SNE-embeddings were performed on the relative abundances per sample, using the Bray-Curtis distance metric[87] to calculate distances within the t-SNE[33]. Samples were classified into a "primary type" based on the most dominant taxa, except if that taxon occurred less than 200 times as the most dominant taxon, in which case it was classified into a type "other". To determine correlations between the abundances of taxa across our samples, we used the fastspar implementation of SparCC with 100,000 permutations. We calculated correlations only between taxa which were present at some non-zero abundance in at least 100 samples. We used the same correlation threshold of 0.3 as in the original SparCC manuscript[34]. Clusters were identified with hierarchical clustering with single linkage. Eigentaxa, a summary score for a given set of taxa, (determined by the modules identified in the taxa-taxa correlation networks) were calculated by first CLR-transforming the relative abundance data, and taking the first principle component of the taxa in each cluster. Eigentaxa were multiplied by the sign of the correlation coefficient between the eigentaxa and a representative taxon for each cluster: *Gardnerella, Prevotella* and *Limosilactobacillus* for the BV, AV and *Lactobacillus* modules, respectively.

Associations between microbial community composition and the questionnaire were performed with an Adonis test, as implemented in the vegan package in R. For each effect of interest, we tested three models. 1) ~ e_i, 2) ~ e_t + e_i, and 3) ~ e_t + age + e_i , where e_t are technical effects, e_i is the effect of interest. Technical effects used were identical across all experiments, and consisted of sequencing run, normalized read concentration and library size, which were found to be strongly associated with the principal component s of the relative abundance. In order to optimize computational performance, initially 1,000 permutations were performed for each effect of interest. A total of 10,000 permutations were performed only for those effects which had p-values equal to 0.001.

Associations between Shannon diversity and variable collected via the questionnaire were performed with a multiple linear regression, with three different models, as in the Adonis test, 1) Diversity ~ e_i, 2) Diversity ~ e_t + e_i, and 3) Diversity ~ e_t + age + e_i.

Associations between the relative abundance of specific taxa and the questionnaire were done with a multiple linear regression, with a model CLR(RA_I) ~ e_t + e_i, where RA_i refers to the relative abundance of a taxa of interest, and CLR refers to the centered log ratio[88].

Associations between assigned community types and the questionnaire were performed with a logistic regression, where, for each pair of community types T_A and T_B, we tested the following three models: 1) I_T ~ e_i, 2) I_T ~ e_t + e_i, and 3) T_I ~ e_t + age + e_i, where I_T is an indicator function whereby: I_T = 0 if sample is in T_A else 1 if sample is in T_B. Results in figure 5 show the results for model 3, except for age, in which the results for model 2 are shown.

For the Adonis model analysis of total explained variance, we included all significant factors in a factorial Adonis test (Factors included are shown in figure 5). In order to perform this, missing values in the questions were encoded as separate categories.

All data handling and visualization was performed in python and R version 4.1.0[85] using the tidyverse set of packages and the in-house developed package tidyamplicons (github.com/Swittouck/tidyamplicons).

Data availability

Sequencing data are available at the European Nucleotide Archive (ENA) under bioproject PRJEB50407.

References

1. Weinstein, L., Bogin, M., Howard, J. H. & Finkelstone, B. B. A survey of the vaginal flora at various ages, with special reference to the Döderlein bacillus. *Am. J. Obstet. Gynecol.* **32**, 211–218 (1936).

2. Lash, A. F. & Kaplan, B. A Study of Döderlein ' s Vaginal Bacillus. *Oxford Univ. Press* **38**, 333–340 (2021).

3. Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4680–4687 (2011).

4. Lopes dos Santos Santiago, G. *et al.* Longitudinal qPCR Study of the Dynamics of L. crispatus, L. iners, A. vaginae, (Sialidase Positive) G. vaginalis, and P. bivia in the Vagina. *PLoS One* **7**, e45281 (2012).

5. El Aila, N. A. *et al.* Identification and genotyping of bacteria from paired vaginal and rectal samples from pregnant women indicates similarity between vaginal and rectal microflora. *BMC Infect. Dis.* **9**, 167 (2009).

6. Oerlemans, E. F. M. *et al.* The Dwindling Microbiota of Aerobic Vaginitis, an Inflammatory State Enriched in Pathobionts with Limited TLR Stimulation. *Diagnostics* **10**, 879 (2020).

7. Donders, G. G. G. *et al.* Definition of a type of abnormal vaginal flora that is distinct from bacterial vaginosis: Aerobic vaginitis. *BJOG An Int. J. Obstet. Gynaecol.* **109**, 34–43 (2002).

8. Gosmann, C. *et al.* Lactobacillus-Deficient Cervicovaginal Bacterial Communities Are Associated with Increased HIV Acquisition in Young South African Women. *Immunity* **46**, 29–37 (2017).

9. McClelland, R. S. *et al.* Evaluation of the association between the concentrations of key vaginal bacteria and the increased risk of HIV acquisition in African women from five cohorts: a nested case-control study. *Lancet Infect. Dis.* **18**, 554–564 (2018).

10. Lewis, F. M. T., Bernstein, K. T. & Aral, S. O. Vaginal microbiome and its relationship to

behavior, sexual health, and sexually transmitted diseases. *Obstet. Gynecol.* **129**, 643–654 (2017).

11. Campisciano, G. *et al.* Subclinical alteration of the cervical–vaginal microbiome in women with idiopathic infertility. *J. Cell. Physiol.* **232**, 1681–1688 (2017).

12. Kroon, S. J., Ravel, J. & Huston, W. M. Cervicovaginal microbiota, women's health, and reproductive outcomes. *Fertil. Steril.* **110**, 327–336 (2018).

13. Zheng, J. *et al.* A taxonomic note on the genus Lactobacillus: Description of 23 novel genera, emended description of the genus Lactobacillus Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. *Int. J. Syst. Evol. Microbiol.* **70**, 2782–2858 (2020).

14. Gajer, P. *et al.* Temporal Dynamics of the Human Vaginal Microbiota. *Sci. Transl. Med.* **4**, 1–21 (2012).

15. France, M. *et al.* VALENCIA: A Nearest Centroid Classification Method for Vaginal Microbial Communities Based on Composition&nbsp; 1–15 (2020) doi:10.21203/rs.2.24139/v1.

16. Drell, T. *et al.* Characterization of the Vaginal Micro- and Mycobiome in Asymptomatic Reproductive-Age Estonian Women. *PLoS One* **8**, (2013).

17. Freitas, A. C. *et al.* The vaginal microbiome of pregnant women is less rich and diverse, with lower prevalence of Mollicutes, compared to non-pregnant women. *Sci. Rep.* **7**, 1–16 (2017).

18. Lennard, K. *et al.* Microbial Composition Predicts Genital Tract Inflammation and Persistent Bacterial Vaginosis in South African Adolescent Females. *Infect. Immun.* **86**, (2017).

19. Rhoades, N. S. *et al.* Longitudinal Profiling of the Macaque Vaginal Microbiome Reveals Similarities to Diverse Human Vaginal Communities. *mSystems* **6**, (2021).

20. Miller, E. A., Beasley, D. A. E., Dunn, R. R. & Archie, E. A. Lactobacilli dominance and vaginal pH: Why is the human vaginal microbiome unique? *Front. Microbiol.* **7**, 1–13 (2016).

21. Yildirim, S. *et al.* Primate vaginal microbiomes exhibit species specificity without universal Lactobacillus dominance. *ISME J.* **8**, 2431–2444 (2014).

22. Mirmonsef, P. *et al.* Free glycogen in vaginal fluids is associated with Lactobacillus colonization and low vaginal pH. *PLoS One* **9**, 26–29 (2014).

23. Song, S. D. *et al.* Daily Vaginal Microbiota Fluctuations Associated with Natural Hormonal Cycle, Contraceptives, Diet, and Exercise. *mSphere* **5**, 1–14 (2020).

24. Foxman, B., Muraglia, R., Dietz, J. P., Sobel, J. D. & Wagner, J. Prevalence of recurrent vulvovaginal candidiasis in 5 European countries and the United States: Results from an internet panel survey. *J. Low. Genit. Tract Dis.* **17**, 340–345 (2013).

25. Medina, M. & Castillo-Pino, E. An introduction to the epidemiology and burden of urinary tract infections. *Ther. Adv. Urol.* **11**, 3–7 (2019).

26. Serrano, M. G. *et al.* Racioethnic diversity in the dynamics of the vaginal microbiome during pregnancy. *Nat. Med.* **25**, 1001–1011 (2019).

27. Noppe, J. *et al.* Vlaamse Migratie- en integratiemonitor 2018. *Brussel Agentschap Binnenl. Best.* 311 (2018).

28. Vlaanderen, S. Bevolking onder de armoededrempel - Statistiek Vlaanderen. https://www.statistiekvlaanderen.be/nl/bevolking-onder-de-armoededrempel.

29. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 1–11 (2019).

30. Putonti, C., Shapiro, J. W., Ene, A., Tsibere, O. & Wolfe, A. J. Comparative Genomic Study of Lactobacillus jensenii and the. *Am. Soc. Microbiol.* **5**, 1–5 (2020).

31. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **202**, 1–10 (2021).

32. Rocha, J. *et al.* Lactobacillus mulieris sp. nov., a new species of lactobacillus delbrueckii group. *Int. J. Syst. Evol. Microbiol.* **70**, 1522–1527 (2020).

33. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

34. Watts, S. C., Ritchie, S. C., Inouye, M. & Holt, K. E. FastSpar: Rapid and scalable correlation estimation for compositional data. *Bioinformatics* **35**, 1064–1066 (2019).

35. Rizzo, A., Losacco, A. & Carratelli, C. R. Lactobacillus crispatus modulates epithelial cell defense against Candida albicans through Toll-like receptors 2 and 4, interleukin 8 and human β-defensins 2 and 3. *Immunol. Lett.* **156**, 102–109 (2013).

36. Ojala, T. *et al.* Comparative genomics of Lactobacillus crispatus suggests novel mechanisms for the competitive exclusion of Gardnerella vaginalis. *BMC Genomics* **15**, 1–21 (2014).

37. van der Veer, C. *et al.* Comparative genomics of human Lactobacillus crispatus isolates reveals genes for glycosylation and glycogen degradation: Implications for in vivo dominance of the vaginal microbiota. *Microbiome* **7**, 1–14 (2019).

38. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).

39. Peltola, T. & Arpin, I. Science for everybody?: Bridging the socio-economic gap in urban biodiversity monitoring. *Citiz. Sci. Innov. Open Sci. Soc. Policy* 367–380 (2018).

40. Law, E. *et al.* *The Science of Citizen Science*. (2017). doi:10.1145/3022198.3022652.

41. Petrova, M. I., Reid, G., Vaneechoutte, M. & Lebeer, S. Lactobacillus iners : Friend or Foe? *Trends Microbiol.* **25**, 182–191 (2017).

42. France, M. T. *et al.* Complete Genome Sequences of Six Lactobacillus iners Strains Isolated from the Human Vagina. *Microbiol. Resour. Announc.* **9**, 17–19 (2020).

43. Castro, J., Machado, D. & Cerca, N. Unveiling the role of Gardnerella vaginalis in

polymicrobial Bacterial Vaginosis biofilms: the impact of other vaginal pathogens living as neighbors. *ISME J.* **13**, 1306–1317 (2019).

44. Harwich, M. D. *et al.* Drawing the line between commensal and pathogenic Gardnerella vaginalis through genome analysis and virulence studies. *BMC Genomics* **11**, (2010).

45. Łaniewski, P. & Herbst-Kralovetz, M. M. Bacterial vaginosis and health-associated bacteria modulate the immunometabolic landscape in 3D model of human cervix. *npj Biofilms Microbiomes* **7**, 1–17 (2021).

46. Castro, J., Jefferson, K. K. & Cerca, N. Genetic Heterogeneity and Taxonomic Diversity among Gardnerella Species. *Trends Microbiol.* **28**, 202–211 (2020).

47. Charbonneau, M. R. *et al.* A microbial perspective of human developmental biology. doi:10.1038/nature18845.

48. Koren, O. *et al.* A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. *PLOS Comput. Biol.* **9**, e1002863 (2013).

49. Canon, F., Nidelet, T., Guédon, E., Thierry, A. & Gagnaire, V. Understanding the Mechanisms of Positive Microbial Interactions That Benefit Lactic Acid Bacteria Co-cultures. *Front. Microbiol.* **11**, 1–16 (2020).

50. Blasche, S. *et al.* Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community. *Nat. Microbiol. |* **6**,.

51. Agarwal, K. *et al.* Glycan cross-feeding supports mutualism between Fusobacterium and the vaginal microbiota. *PLOS Biol.* **18**, e3000788 (2020).

52. Mokoena, M. P. Lactic Acid Bacteria and Their Bacteriocins: Classification, Biosynthesis and Applications against Uropathogens: A Mini-Review. *Mol. A J. Synth. Chem. Nat. Prod. Chem.* **22**, (2017).

53. Petrova, M. I., Lievens, E., Malik, S., Imholz, N. & Lebeer, S. Lactobacillus species as biomarkers and agents that can promote various aspects of vaginal health. *Front. Physiol.* **6**, (2015).

54. Lin, X. B. *et al.* The evolution of ecological facilitation within mixed-species biofilms in the mouse gastrointestinal tract. *ISME J.* **12**, 2770–2784 (2018).

55. Hummelen, R. *et al.* Lactobacillus rhamnosus GR-1 and L. reuteri RC-14 to prevent or cure bacterial vaginosis among women with HIV. *Int. J. Gynecol. Obstet.* **111**, 245–248 (2010).

56. Liu, J. J., Reid, G., Jiang, Y., Turner, M. S. & Tsai, C. C. Activity of HIV entry and fusion inhibitors expressed by the human vaginal colonizing probiotic Lactobacillus reuteri RC-14. *Cell. Microbiol.* **9**, 120–130 (2007).

57. Martinez, R. C. R. *et al.* Improved cure of bacterial vaginosis with single dose of tinidazole (2 g), Lactobacillus rhamnosus GR-1, and Lactobacillus reuteri RC-14: A randomized, double-blind, placebo-controlled trial. *Can. J. Microbiol.* **55**, 133–138

(2009).

58. Verhelst, R. *et al. Cloning of 16S rRNA genes amplified from normal and disturbed vaginal microflora suggests a strong association between Atopobium vaginae, Gardnerella vaginalis and bacterial vaginosis*. http://www.biomedcentral.com/1471-2180/4/16 (2004).

59. Hardy, L. *et al.* A fruitful alliance: the synergy between Atopobium vaginae and Gardnerella vaginalis in bacterial vaginosis-associated biofilm. *Sex. Transm. Infect.* **92**, 487–491 (2016).

60. Randis, T. M. & Ratner, A. J. Gardnerella and Prevotella: Co-conspirators in the Pathogenesis of Bacterial Vaginosis. *J. Infect. Dis.* **220**, 1085–1088 (2019).

61. Donders, G. G. G., Bellen, G., Grinceviciene, S., Ruban, K. & Vieira-Baptista, P. Aerobic vaginitis: no longer a stranger. *Res. Microbiol.* **168**, 845–858 (2017).

62. Perrotta, A. R. *et al.* The Vaginal Microbiome as a Tool to Predict rASRM Stage of Disease in Endometriosis: a Pilot Study. doi:10.1007/s43032-019-00113-5.

63. Haggerty, C. L. *et al.* Presence and concentrations of select bacterial vaginosis-associated bacteria are associated with increased risk of pelvic inflammatory disease. *Sex. Transm. Dis.* **47**, 344 (2020).

64. De Seta, F., Campisciano, G., Zanotta, N., Ricci, G. & Comar, M. The vaginal community state types microbiome-immune network as key factor for bacterial vaginosis and aerobic vaginitis. *Front. Microbiol.* **10**, 2451 (2019).

65. Si, J., You, H. J., Yu, J., Sung, J. & Ko, G. P. Prevotella as a Hub for Vaginal Microbiota under the Influence of Host Genetics and Their Association with Obesity. *Cell Host Microbe* **21**, 97–105 (2017).

66. Vodstrcil, L. A. *et al.* Combined oral contraceptive pill-exposure alone does not reduce the risk of bacterial vaginosis recurrence in a pilot randomised controlled trial. *Sci. Rep.* **9**, 1–13 (2019).

67. Nelson, T. M. *et al.* Cigarette smoking is associated with an altered vaginal tract metabolomic profile. *Sci. Rep.* **8**, 852 (2018).

68. Lewis, C. A. *et al.* Effects of Hormonal Contraceptives on Mood: A Focus on Emotion Recognition and Reactivity, Reward Processing, and Stress Response. *Curr. Psychiatry Rep.* **21**, 1–15 (2019).

69. Lundin, C., Wikman, A., Bixo, M., Gemzell-Danielsson, K. & Sundström Poromaa, I. Towards individualised contraceptive counselling: clinical and reproductive factors associated with self-reported hormonal contraceptive-induced adverse mood symptoms. *BMJ Sex. Reprod. Heal.* **47**, e1–e8 (2021).

70. Burrows, L. J., Basha, M. & Goldstein, A. T. The Effects of Hormonal Contraceptives on Female Sexuality: A Review. *J. Sex. Med.* **9**, 2213–2223 (2012).

71. Khialani, D., Rosendaal, F. & Vlieg, A. V. H. Hormonal Contraceptives and the Risk of Venous Thrombosis. *Semin. Thromb. Hemost.* **46**, 865–871 (2020).

72. Morimont, L., Haguet, H., Dogné, J. M., Gaspard, U. & Douxfils, J. Combined Oral Contraceptives and Venous Thromboembolism: Review and Perspective to Mitigate the Risk. *Front. Endocrinol. (Lausanne).* **12**, 1 (2021).

73. Donders, G. G. G. *et al.* Screening for abnormal vaginal microflora by self-assessed vaginal pH does not enable detection of sexually transmitted infections in Ugandan women. *Diagn. Microbiol. Infect. Dis.* **85**, 227–230 (2016).

74. Achilles, S. L., Meyn, L. A., Austin, M. N., Avolia, H. A. & Hillier, S. L. A longitudinal evaluation of the impact of contraceptive initiation on vaginal microbiota in us women. *Am. J. Obstet. Gynecol.* **219**, 643–644 (2018).

75. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* **24**, 133-145.e5 (2018).

76. DiGiulio, D. B. *et al.* Temporal and spatial variation of the human microbiota during pregnancy. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11060–11065 (2015).

77. Lee, E. & Lee, J. E. Impact of drinking alcohol on gut microbiota: recent perspectives on ethanol and alcoholic beverage. *Curr. Opin. Food Sci.* **37**, 91–97 (2021).

78. Ahannach, S. *et al.* Microbial enrichment and storage for metagenomics of vaginal, skin, and saliva samples. *iScience* **24**, 103306 (2021).

79. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–20 (2013).

80. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

81. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

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

83. Yoon, S. H. *et al.* Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.* **67**, 1613–1617 (2017).

84. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

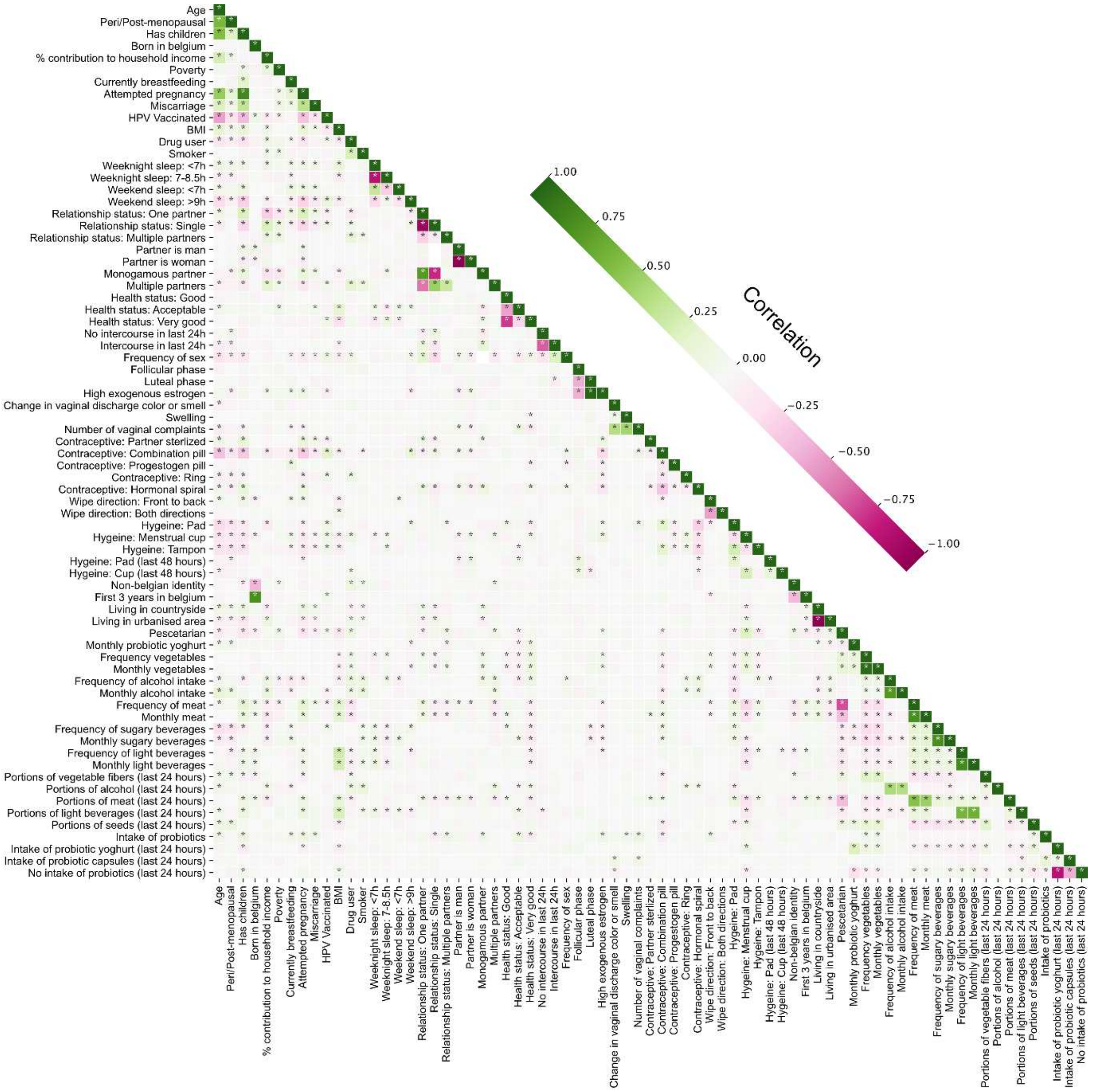
85. R Core Team. R: A Language and Environment for Statistical Computing. (2020).

86. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

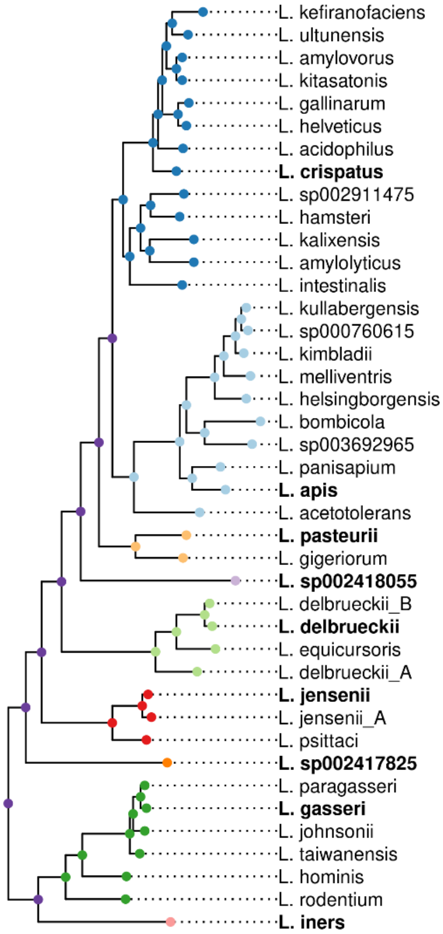
87. van der Maaten, L. Barnes-Hut-SNE. *1st Int. Conf. Learn. Represent. ICLR 2013 - Conf. Track Proc.* 1–11 (2013).

88. Aitchison, J. A concise guide to compositional data analysis. in *2nd Compositional Data Analysis Workshop* (2003).

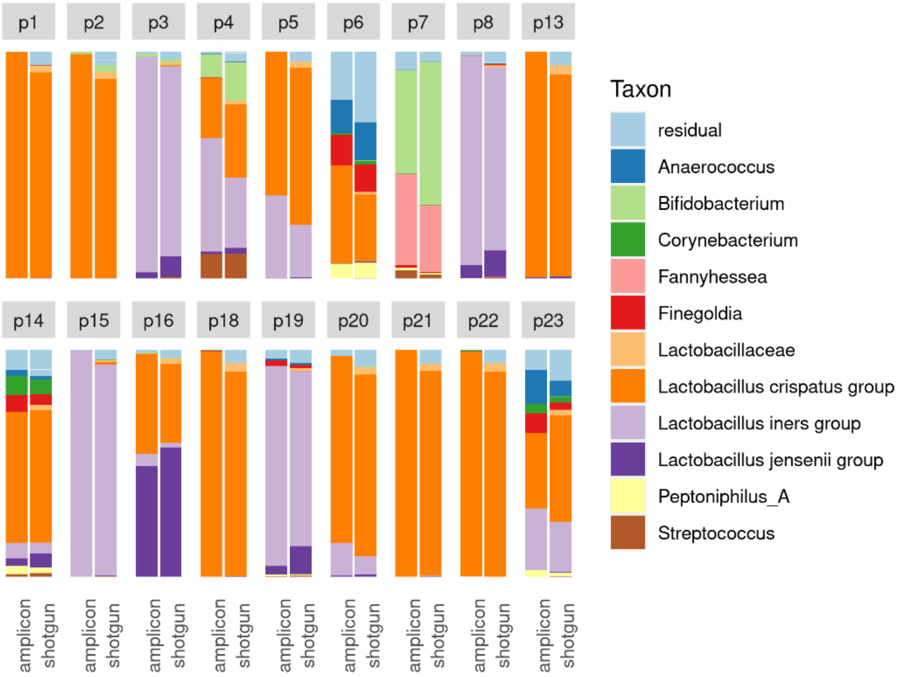
Supplementary figures



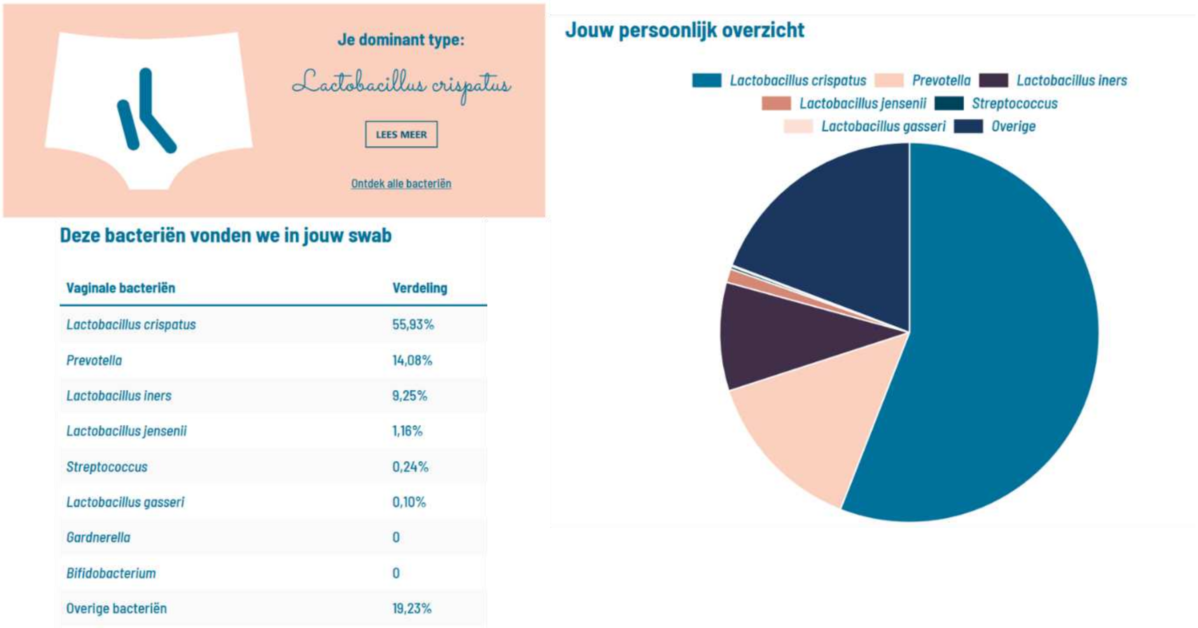
Supplementary Figure 1 - **Correlations between a subset of the questionnaire variables.** Heatmap with correlations between questionnaire variables shown in Figure 5. Positive correlations are indicated with green, negative correlations in red. Significant correlations are marked with an asterisk.



Supplementary Figure 2 - **Species tree of *Lactobacillus* from the Genome Taxonomy Database.** Maximum-likelihood species phylogeny of the genus *Lactobacillus*, obtained by taking a subtree of the species phylogeny of the domain Bacteria inferred by the Genome Taxonomy Database (GTDB), release 05-RS95[31]. Colors indicate the nine custom-defined subgenera used in this study. Bold tip labels indicate representative species of the subgenera.



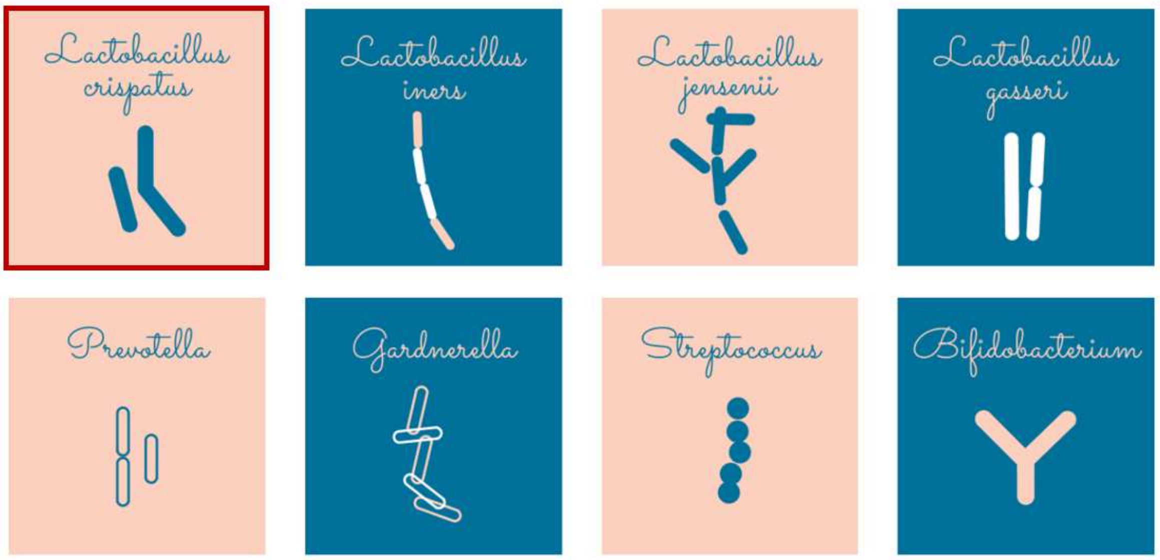
Supplementary Figure 3 - **Comparison between amplicon and shotgun sequencing results for 18 samples.** Relative abundances for the eleven most abundant taxa overall. Each facet shows a vaginal sample from a single participant, sequenced with 16S rRNA amplicon sequencing (left) or metagenome shotgun sequencing (right).

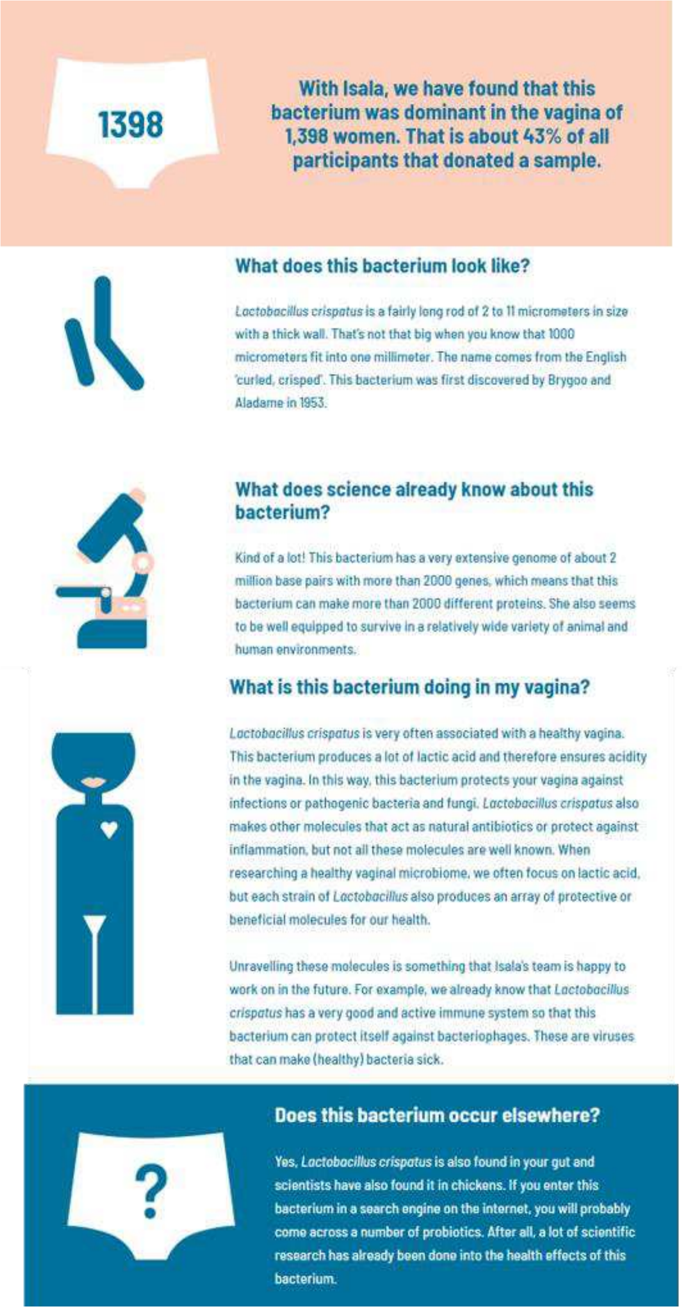


Deze bacteriën vonden we in jouw swab

Vaginale bacteriën	Verdeling
Lactobacillus crispatus	55,93%
Prevotella	14,08%
Lactobacillus iners	9,25%
Lactobacillus jensenii	1,16%
Streptococcus	0,24%
Lactobacillus gasseri	0,10%
Gardnerella	0
Bifidobacterium	0
Overige bacteriën	19,23%

Supplementary Figure 4 - **Example of a personal vaginal microbiome profile result.** Top left figure indicates the dominant type. Bottom left show the percentage ("verdeling") of the top eight taxa identified. Right figure (pie chart) displays the top six taxa plus the remaining ("overage") ones.





1398

With Isala, we have found that this bacterium was dominant in the vagina of 1,398 women. That is about 43% of all participants that donated a sample.

What does this bacterium look like?

Lactobacillus crispatus is a fairly long rod of 2 to 11 micrometers in size with a thick wall. That's not that big when you know that 1000 micrometers fit into one millimeter. The name comes from the English 'curled, crisped'. This bacterium was first discovered by Brygoo and Aladame in 1953.

What does science already know about this bacterium?

Kind of a lot! This bacterium has a very extensive genome of about 2 million base pairs with more than 2000 genes, which means that this bacterium can make more than 2000 different proteins. She also seems to be well equipped to survive in a relatively wide variety of animal and human environments.

What is this bacterium doing in my vagina?

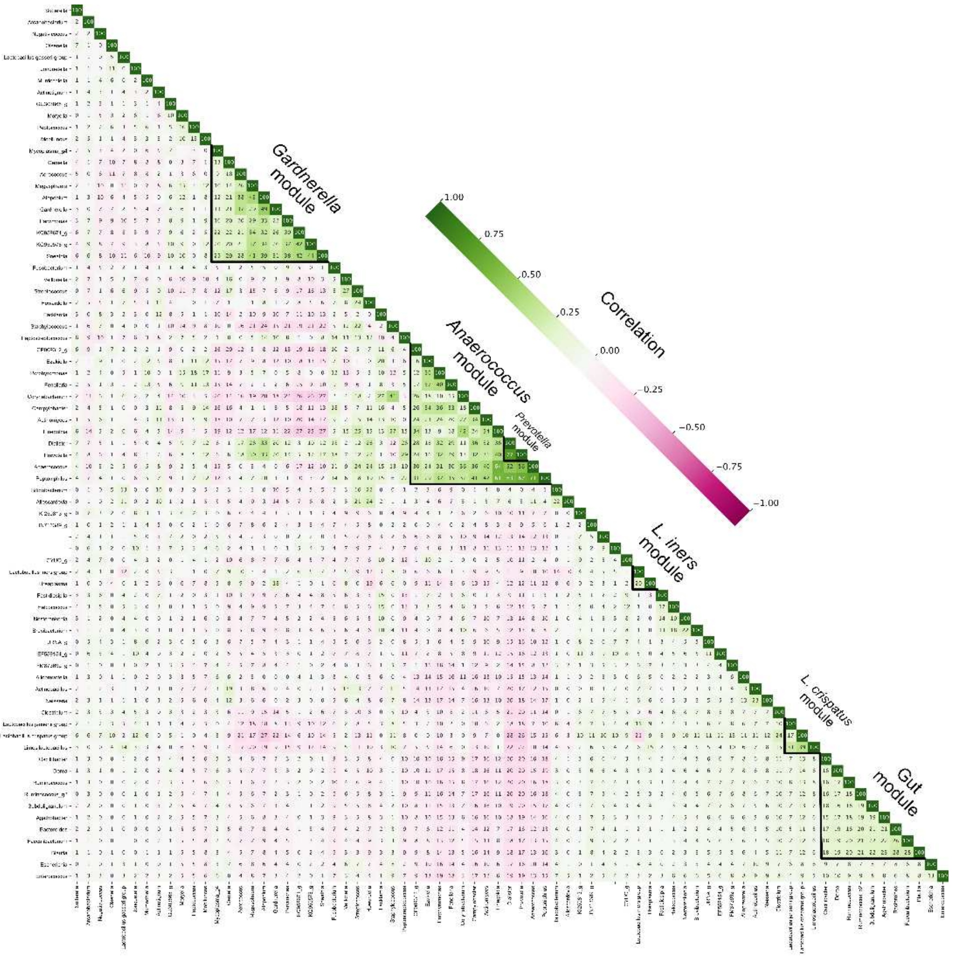
Lactobacillus crispatus is very often associated with a healthy vagina. This bacterium produces a lot of lactic acid and therefore ensures acidity in the vagina. In this way, this bacterium protects your vagina against infections or pathogenic bacteria and fungi. *Lactobacillus crispatus* also makes other molecules that act as natural antibiotics or protect against inflammation, but not all these molecules are well known. When researching a healthy vaginal microbiome, we often focus on lactic acid, but each strain of *Lactobacillus* also produces an array of protective or beneficial molecules for our health.

Unravelling these molecules is something that Isala's team is happy to work on in the future. For example, we already know that *Lactobacillus crispatus* has a very good and active immune system so that this bacterium can protect itself against bacteriophages. These are viruses that can make (healthy) bacteria sick.

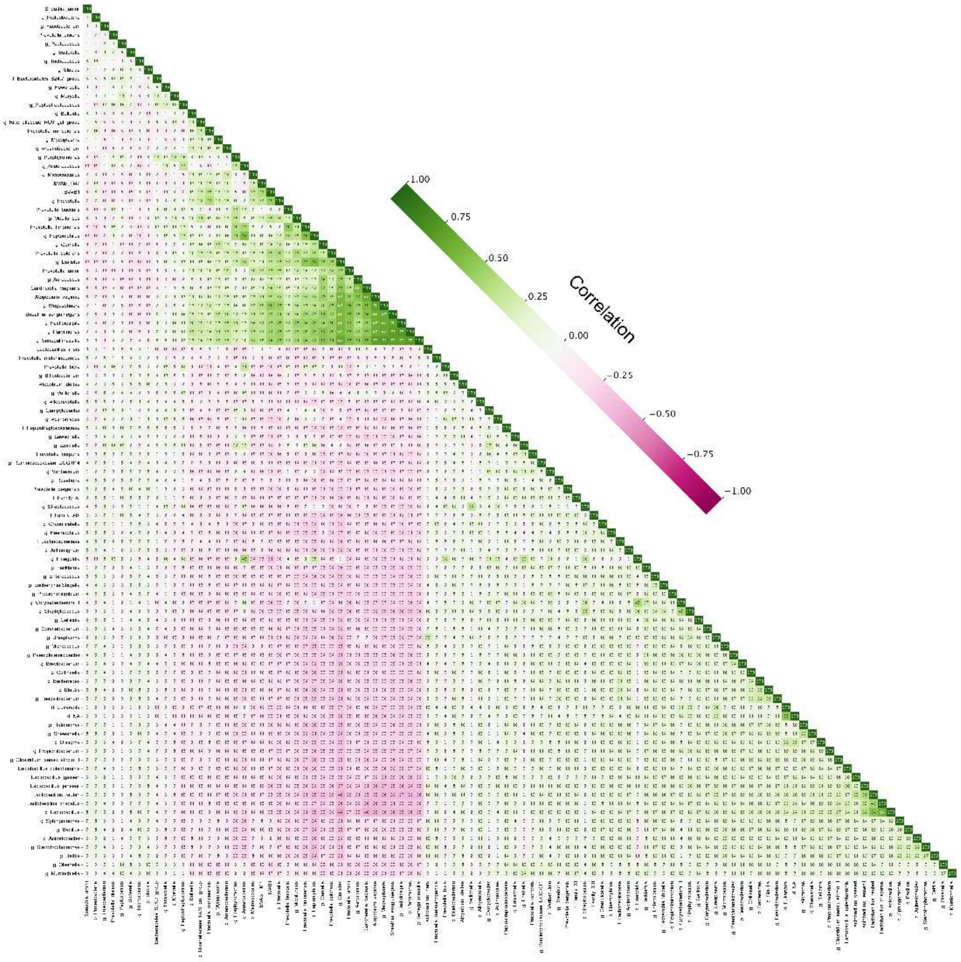
Does this bacterium occur elsewhere?

Yes, *Lactobacillus crispatus* is also found in your gut and scientists have also found it in chickens. If you enter this bacterium in a search engine on the internet, you will probably come across a number of probiotics. After all, a lot of scientific research has already been done into the health effects of this bacterium.

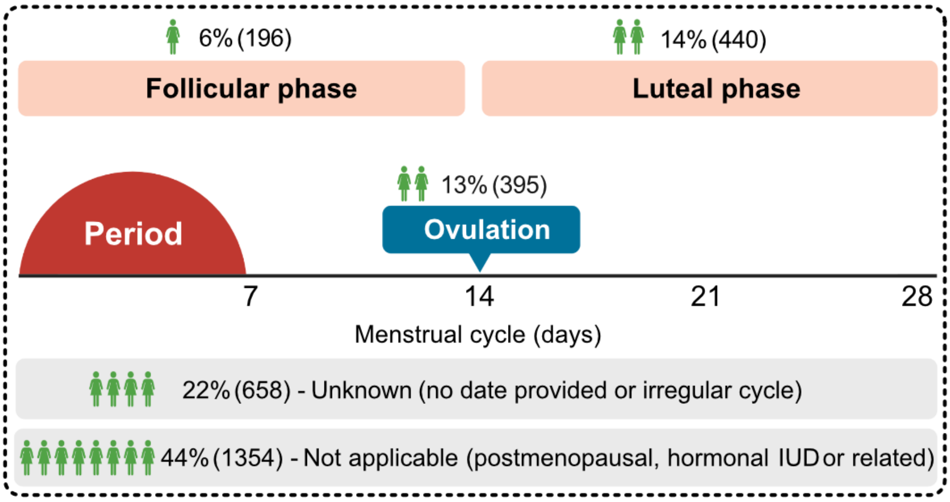
Supplementary Figure 5 - **Received information for non-microbiology experts.** To each of the top eight taxa a webpage was dedicated. Here, an example of the page on *Lactobacillus crispatus* is added. Other taxa can be accessed via https://isala.be/en/category/vaginal-bacteria/



Supplementary Figure 6 – **Full taxon correlation matrix.** SparCC correlation Network between taxa determined in the Isala data. Positive correlations are indicated in green, negative correlations in red. In each cell is given the correlation (*100) for each pair of taxa. The modules identified and shown in figure 4 are indicated with triangles in the figure.



Supplementary Figure 7 – **Full taxon correlation matrix of the Valencia study.** SparCC correlation Network between taxa determined in the Isala data. Positive correlations are indicated in green, negative correlations in red. In each cell is given the correlation (*100) for each pair of taxa.


6% (196)
Follicular phase
14% (440)
Luteal phase
Period
13% (395)
Ovulation
7
14
21
28
Menstrual cycle (days)
22% (658) - Unknown (no date provided or irregular cycle)
44% (1354) - Not applicable (postmenopausal, hormonal IUD or related)

Supplementary Figure 8 – **The menstrual cycle.** Using information about each participant's cycle length and last menstruation, we estimated the stage of the cycle in which the swab was sampled. Participants whose cycles had irregular lengths, or who did not report their last menstruation were classified as "Unknown", and participants using hormonal contraceptives or were peri/post-menopausal were classified as "Not applicable".

Supplementary tables

Supplementary Table 1 – **ASV occurrence and abundance of top 10 lactobacilli and percentage of top 10 isolated lactobacilli from Isala's samples.** The occurrence of the top 10 ASVs of lactobacilli on (sub)genus level over all Isala's samples and their mean relative abundance and the percentage of isolates belonging to the top 10 most isolated lactobacilli (determined by 16S amplicon sequencing) in relation to the total lactobacilli isolates (n = 230) and the total number of isolates per species.

(sub)genus			16S isolates		
(Sub)genus	**Occurrence**	**Mean relative abundance**	**Species**	**Percentage of total lactobacilli isolates on De Man, Rogosa en Sharpe or Columbia Blood media (glucose as main sugar)**	**Number of isolates studied (n = 230)**
Lactobacillus crispatus group	0,897699005	0,399114797	*Limosilactobacillus fermentum*	24,49%	60
Lactobacillus iners group	0,719527363	0,240823923	*Lactobacillus crispatus*	13,88%	34
Limosilactobacillus	0,478544776	0,004111909	*Lactobacillus jensenii*	12,24%	30
Lactobacillus jensenii group	0,467661692	0,04856063	*Lactobacillus paragasseri*	9,80%	24
Lactobacillus gasseri group	0,268345771	0,029760051	*Lacticaseibacillus rhamnosus*	8,98%	22
Lactobacillaceae	0,027052239	0,00199087	*Lacticaseibacillus paracasei*	7,35%	18
Lacticaseibacillus	0,023942786	0,000494929	*Limosilactobacillus reuteri*	6,12%	15
Lactiplantibacillus	0,00528607	1,65417E-05	*Lactiplantibacillus plantarum*	4,90%	12
Ligilactobacillus	0,004975124	2,75966E-05	*Lactobacillus gasseri*	3,67%	9
Apilactobacillus	0,003109453	0,000293372	*Leuconostoc mesenteroides*	2,45%	6

Supplementary Table 2 – **Descriptive statistics of taxa.** Various descriptive statistics for subgenera of the genus *Lactobacillus* and genera detected in this study: number of ASVs within the (sub)genus (n_asvs), occurrence, average relative abundance (mean_rel_abundance), frequency of being the most abundant taxon and greater than 0% abundant (top_and_gt0p), same as previous but greater than 30% abundant (top_and_gt30p), same as previous but greater than 50% abundant (top_and_gt50p), the previous three measures but in terms of relative frequencies (top_and_gtXp_rel).

Supplementary Table 3 – **Association tests between participant characteristics and their vaginal microbiome.** Results of statistical tests for each tested questionnaire responses. Results are provided for the beta-diversity (Adonis), alpha-diversity, taxa relative abundances and eigentaxa level tests. In addition to effect sizes, confidence intervals and p-values the number of participants in each condition are provided.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS2genusstats.csv
- TableS3associationtests20211220.xlsx