Monteleone *et al. Microbiome* (2026) 14:138
https://doi.org/10.1186/s40168-026-02420-x

Microbiome

RESEARCH

Open Access

Unleashing the potential of mRNA-seq to uncover the microbiome structure and their crosstalk with host cells: the vulvar ecosystem

Emanuele Monteleone[1,2*†], Marco Amato Cianci[1,2†], Alessia Albano[1,2†], Federica Loperfido[1,2], Gloria Griffante[2,3], Luna Brasi[1,2], Fulvio Borella[4], Niccolò Gallio[4], Mario Preti[4], Alessandra Marchi[5], Barbara Gardella[6], Ivan Molineris[1], Giacomo Donati[1,2] and Valentina Proserpio[1,2*]

Abstract

Background To describe both host gene expression and microbiome composition in a single sample, parallel experimental and computational workflows (mRNA-sequencing and either 16S rRNA gene or metagenomics) have been traditionally applied. The vulvar milieu represents an area of emerging research for its role in health and disease. Located at the interface between the vagina and the perineum, the vulvar microbiome displays an intermediate signature, with influx from both ecosystems.

Results Following validation of the reliability of poly(A)-enriched mRNA-sequencing in reconstructing the microbiota composition using both a quantitative microbial standard (mock) and metagenomic analysis, we analyze a full cohort of 30 healthy vulvar samples. Crucially, the analysis of the entire cohort relies solely on mRNA-sequencing without the use of parallel DNA metagenomics. This unified approach allows us to analyze not only the vulvar cell transcriptome, but also the composition and dynamics of microbial communities, including the microbial gene expression signatures. This three-level analysis (host-mRNA, individual bacterial species, bacterial gene pathways) on the very same specimens further enables a gene-level exploration of host-microbe molecular crosstalk. Using this unified framework, we reveal marked heterogeneity and high inter-individual variability in the vulvar microbiota, identifying community state types that mirror those described in the vagina. Importantly, we show that distinct microbial configurations are associated with specific host transcriptional programs: *Lactobacillus crispatus* correlates with epithelial differentiation and barrier integrity, whereas communities enriched in *Gardnerella vaginalis*, or other taxa associated with dysbiosis, exhibit transcriptional signatures linked to inflammation. Interestingly, *Lactobacillus gasseri*, which has been associated with lower protection, shows an intermediate effect on vulvar cells.

Conclusions Beyond providing new biological insights into an understudied anatomical niche, our study introduces a broadly applicable strategy with substantial impact for the field. With tens of thousands of human RNA-seq datasets already available in public repositories, our approach enables retrospective extraction of microbiome information and host-microbe interaction signals from existing transcriptomic data, without the need for additional sequencing

†Emanuele Monteleone, Marco Amato Cianci and Alessia Albano contributed equally to this work.

*Correspondence:
Emanuele Monteleone
emanuele.monteleone@unito.it
Valentina Proserpio
valentina.proserpio@unito.it
Full list of author information is available at the end of the article

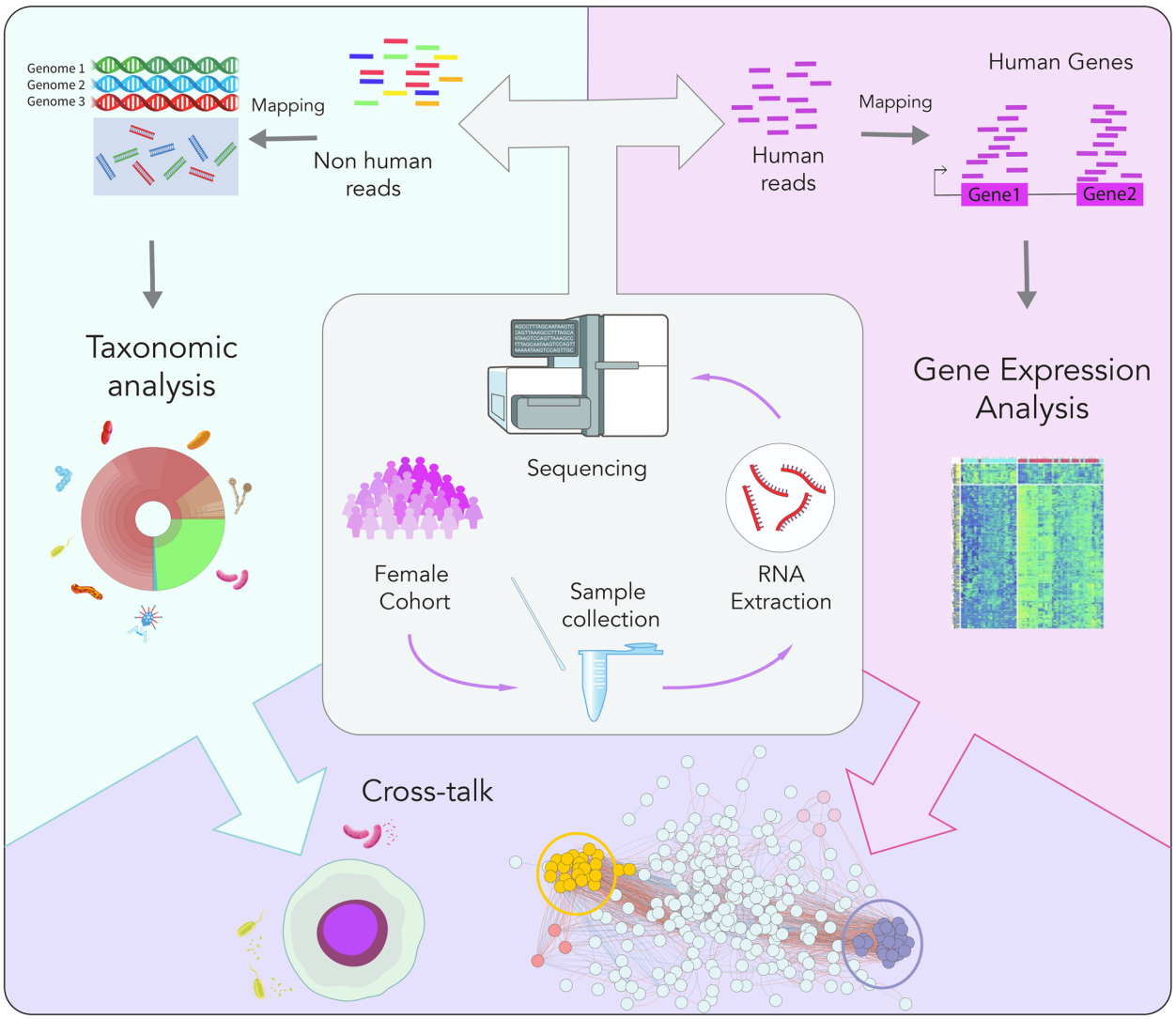
or specialized microbiome protocols. This unlocks a powerful and cost-effective opportunity to revisit archived RNA-seq studies across tissues, diseases, and low-biomass environments, revealing previously inaccessible layers of host-microbiome crosstalk and maximizing the scientific value of published data.

Highlights

- Traditional mRNA-seq can successfully reconstruct the microbiome composition of a given sample.
- Our approach allows us to describe the mutual relationship between changes in the microbiota and the gene expression perturbation in host cells.
- This mutual analysis on vulvar samples highlights the effect of a healthy or dysbiotic vulvar milieu on epithelial cells.

Keywords Vulvar microbiome, Dual RNA-seq, Host-microbiome crosstalk, Community state types (CSTs), Metatranscriptomics, Bias correction, Microbial mRNA-seq, Retrospective studies, Data mining

Graphical Abstract



Background

Microbiome composition analysis has traditionally been performed using two main strategies: amplification of phylogenetic marker genes, commonly referred to as metabarcoding (e.g., 16S rRNA gene for prokaryotes), and whole-genome shotgun metagenomic sequencing (WMS) [1]. Metabarcoding relies on targeted PCR with degenerate oligonucleotides to amplify hypervariable regions of the ribosomal DNA locus from genomic DNA [2], whereas metagenomic sequencing does not focus on a single region, but rather targets all DNA present in a sample [3].

The most widely used approach is 16S rRNA gene profiling, which, despite being fast and cost-effective, has several limitations. These include low taxonomic resolution (typically restricted to the genus level), PCR primer bias, off-target amplification, and the inability to characterize fungi and viruses, as fungi rely on 18S/ITS markers and viruses lack ribosomal genes altogether [4–6]. Moreover, amplification efficiency varies across 16S DNA templates, depending on the variable region of choice (e.g., V1, V2, and others), gene copy number variation, and PCR amplification biases [7]. Minor mismatches between primers and conserved regions introduce amplification bias, whereby some regions are preferentially amplified over others, causing systematic over- or under-representation of certain taxa, or even causing their complete loss [8].

In contrast, WMS captures all DNA present in a sample, including bacterial, viral, and eukaryotic sequences, and provides more comprehensive and less biased information, achieving higher taxonomic resolution down to the species or even strain level, while reducing PCR-related biases [9]. However, WMS is more expensive and cannot always be applied to low-biomass specimens, as it requires relatively large amounts of high-quality microbial DNA, ideally with minimal host-DNA contamination [4, 9, 10].

These two strategies can be combined in optimized experimental designs that leverage their respective strengths to address diverse research questions and accommodate factors such as sample size, longitudinal sampling, and the detection of rare taxa. At the same time, given the limitations of each technique, additional multi-omic data are needed to achieve a comprehensive understanding of microbial functional effects in specific contexts [11]. Indeed, both 16S rRNA gene and metagenomic sequencing describe only the taxonomic composition of a community and therefore fail to measure its functional activity under specific conditions [11].

To overcome this limitation, more recently, metatranscriptomic approaches (which analyze the compendium of the RNA expressed in a given multiorganism niche) have provided deeper insights into the functional activity of microbial communities. As genomics expanded to metagenomics to analyze mixed microbial genomes, transcriptomics similarly expanded to metatranscriptomics to characterize community-wide gene expression patterns across entire microbiomes [12].

Metatranscriptomics leverages the entire RNA pool of a microbial community [13] to offer a direct readout of functional activity and to capture how the community responds to specific conditions or perturbations. In the human microbiome, such gene-expression profiling is essential for understanding how individual taxa contribute to health and disease, since even the same species can adopt different transcriptional states and produce distinct biological effects in host tissues [14].

This omics approach reduces the PCR biases inherent to 16S rRNA gene sequencing, both by minimizing amplification errors (particularly in low-biomass samples) and by offering improved accuracy in species- and strain-level detection. However, it also faces significant challenges, including the high abundance of bacterial rRNAs relative to mRNAs, the short half-life of mRNA, and the absence of poly(A) tails in most prokaryotic transcripts [12, 15]. Furthermore, human samples often exhibit a low ratio of microbial to host cells, requiring careful optimization of sample collection and processing to obtain sufficient microbial data [16, 17].

While all these methods have been widely adopted in scientific research, each relies on a dedicated extraction and library preparation protocol. In particular, in order to successfully extract genomic DNA from different microbial species, harsh mechanical lysis with bead beating is often employed, followed by enzymatic lysis [17, 18], making it incompatible with RNA extraction and subsequent analysis, due to the lower stability of RNA. Overall, several studies showed that different genomic extraction methods can strongly influence the observed community composition, especially for hard-to-lyse organisms such as Gram-positive bacteria and fungi [19, 20].

For metagenomic analysis, libraries are prepared from fragmented DNA, followed by end-repair, adapter ligation, and then sequenced [21]. Instead, for metatranscriptomic workflows, RNA library preparation from microbiome samples is a critical step, as it challenges the preservation of prokaryotic mRNA while depleting the overwhelming abundance of ribosomal RNA [21, 22]. To reduce host contamination and rRNA content in sequencing data, hybridization-capture technologies are usually employed [23].

However, studies addressing multiple research questions often require parallel analyses on separate samples, which can constitute a limitation. Indeed, even when specimens are collected from the same subject at the same time, they may differ in quantity and composition,

introducing biases. Therefore, obtaining multiple layers of information from a single sample is the optimal strategy to ensure consistency and data comparability.

Our analysis shows that a single, unified protocol can be used to assess the microbial composition of a specific sample at the species level, while simultaneously retaining the information on both the host cell and microbial expression profiles. To verify our hypothesis, we employed the vulvar ecosystem as a model system, given the limited knowledge currently available for this organ.

The vulvar case

Our lab focuses on the vulvar ecosystem in health and disease. In the context of microbiome research, female pelvic disorders have frequently been linked to microbial imbalance, generally referred to as dysbiosis [24–26]. Within the female genital tract, the vulvar and vaginal microbiota are distinct yet interconnected ecosystems that play crucial roles in women's health [24–26].

Despite a similar overall burden of diseases affecting both the vagina and vulva, the human vaginal microbiota is far better studied than the vulvar microbiota and is known to be a complex and dynamic ecosystem, primarily dominated by *Lactobacillus* species, resulting in relatively low alpha diversity compared to microbiota from other body sites [27, 28]. These *Lactobacillus* species generally play a protective role by producing lactic acid and bactericidal compounds [29]. A large study of 394 healthy women of reproductive age classified the human vaginal microbiota into five community state types (CSTs), each typically dominated by a specific species. CST-I is dominated by *L. crispatus*, CST-II by *L. gasseri*, CST-III by *L. iners*, CST-V by *L. jensenii*, while CST-IV is dominated by a diverse group of non-*Lactobacillus* taxa with high alpha diversity and high proportions of anaerobes, including *Gardnerella*, *Atopobium*, *Prevotella*, *Dialister*, *Aerococcus*, *Megasphaera*, *Peptoniphilus*, *Sneathia*, *Eggerthella*, *Finegoldia*, and *Mobiluncus*.

A stable acidic vaginal environment dominated by *Lactobacillus* is associated with protection against bacterial vaginosis, sexually transmitted infections, and conditions such as cervical cancer [30–32]. Within the *Lactobacillus* species, *L. gasseri* is health-associated, yet it appears to occupy an intermediate position along the protection spectrum. Epidemiological studies in HPV and preterm birth studies [33, 34] show that *L. gasseri* dominated communities confer a modest degree of protection and display risks that lie between those of *L. crispatus* and *L. iners.* In this framework, *L. gasseri* can be considered a beneficial species that contributes to a favorable vaginal environment, but with an intermediate protective profile. In contrast, dysbiosis, commonly characterized by a decrease in *Lactobacillus* and an expansion of anaerobic bacteria, shifts the ecosystem toward CST-IV [26].

Despite *Lactobacillus* dominance, the vaginal microbiome varies considerably both between individuals and within the same individual over time [35]. This variability contributes to its characteristic low alpha and high beta diversity. Factors influencing this variation include intrinsic elements such as ethnicity, genetic background, age, hormonal status, and individual host factors, as well as extrinsic factors such as sexual behavior, hygiene practices, smoking, diet, stress, and obesity [36]. Similarly, the vulvar microbiome comprises a diverse range of bacterial genera, including *Lactobacillus*, *Corynebacterium*, *Staphylococcus*, *Prevotella*, and *Actinobacteria* spp. [28].

This pattern suggests an influence from adjacent vaginal communities, with additional diversity derived from cutaneous and fecal commensals [28].

Moreover, most investigations of the vulvar microbiota rely on 16S rRNA gene sequencing and rarely employ metagenomic approaches [37–39]. To date, only one published study has applied shotgun metagenomic sequencing to analyze the vulvar microbiome in the context of lichen sclerosus (LS) and high-grade squamous intraepithelial lesions (HSIL) [40]. This study reported that LS lesions are characterized by reduced *Prevotella* and *Bacteroidales*, while HSIL lesions show increased *Fusobacteria* and decreased *Actinobacteria* compared to healthy controls. Moreover, both LS and HSIL vulvar lesions harbor higher levels of Papillomaviridae, particularly Alphapapillomaviruses, than healthy skin.

In light of the strengths and limitations of existing microbiome profiling strategies, we sought to address the current gap in functional characterization of the vulvar ecosystem by implementing an optimized metatranscriptomic approach. Specifically, we developed a unified extraction and library preparation protocol that enables simultaneous profiling of both the vulvar microbiome and host epithelial gene expression from a single specimen. This integrated strategy allows us to reconstruct the microbial composition of the vulvar milieu while capturing host transcriptional responses occurring in the same biological context. Host cells and the microbiota coexist and influence one another through continuous and complex crosstalk [41]. This interaction can significantly affect the gene expression of both host and microbial cells, a phenomenon that remains poorly characterized, especially in vulvar tissue.

By coupling microbial information and host expression data, our approach provides a comprehensive framework to investigate how specific bacterial communities influence vulvar epithelial activity, thereby offering a more efficient, biologically coherent, and cost-effective method

for studying host-microbe interactions in this understudied anatomical site.

Results

Poly(A)-enriched and total RNA-seq approaches successfully reconstruct the microbiome complexity in experimentally controlled samples

To verify the ability of RNA-seq data to reconstruct microbial composition, we first employed a commercially available microbial mock community (ZymoBIOMICS Microbial Community Standard; Zymo Research, Irvine, CA, USA), which contains three easy-to-lyse Gram-negative bacteria, five tough-to-lyse Gram-positive bacteria, and two tough-to-lyse yeast species with abundances determined by cell number and genomic DNA quantity (Table 1).

In parallel, we also extracted total RNA from a commercially available human vulvar cancer epithelial cell line (VLS-VSCC). From all samples, we extracted total RNA using TRIzol and standard protocols for eukaryotic cells, without mechanical or enzymatic digestion of the bacterial cell wall, to mimic traditional mammalian gene expression profiling analysis. We prepared libraries using both poly(A)-enriched and total RNA kits according to the manufacturers' instructions on one microbial RNA sample and three mixed samples of human and microbial RNA with different relative proportions (25% and 60% human RNA, respectively; Fig. 1A and Methods for details).

We sequenced the samples and aligned the raw reads to the human genome to remove host contamination. The remaining unmapped reads were processed through a custom Kraken2/Bracken pipeline to obtain taxonomic abundances per sample (Fig. 1B). As expected, four out of five Gram-positive bacterial strains were underrepresented relative to their theoretical concentrations, likely because we did not perform dedicated cell-wall lysis. Importantly, the detection of relative abundances of the 10 different species was consistent across samples, irrespective of the ratio between human and microbial reads (Fig. 1C, S1A).

Our results closely resemble those obtained using the Human Microbiome Project (HMP) protocol, as reported in the mock community manufacturer's documentation. Despite discrepancies with the real abundances, which are nevertheless consistent across samples (Fig. 1C), our preliminary results suggest that eukaryotic mRNA-sequencing workflows can also be successfully applied to reconstruct microenvironmental niche composition, as demonstrated by the confident detection of the predominant species using both approaches.

Having assessed the reliability of this approach, we next performed RNA-seq on vulvar swabs collected from the vestibule of healthy individuals. Specimens were collected by gently swabbing the inner side of the labium minus, specifically the vestibular area of the vulva, while carefully avoiding the vaginal introitus (Fig. S1B).

We chose the vulva as our study system because, although the vaginal microbiome has been extensively characterized, the complexity of the vulvar microbiome remains poorly understood and represents an emerging area of research [27, 28]. In total, 30 healthy women aged 23 to 44 years, with no history of recurrent pelvic disorders, were included in the study (Table 2). For a subset of individuals (Fig. 2), we also extracted and processed DNA for metagenomic analysis to compare microbiome profiling results. We did not perform 16S rRNA metabarcoding for several reasons, including the low biomass of the samples, PCR biases inherent to this approach, and the need for species-level microbiota resolution to allow comparison with our metatranscriptomic data. Quality control metrics, genomic mapping efficiency, and distribution of human uniquely mapped, microbial kraken assigned and unmapped reads for both DNA and RNA libraries are provided in Fig. S2A-D and Supplementary material. Using this in vivo dataset, we then sought to systematically understand the mechanisms allowing the recovery of bacterial mRNA in poly(A)-selected libraries. We computationally scanned the captured reads for internal poly-A/T stretches (8, 10, and 12 nt) that could facilitate oligo-dT mispriming prior to RNA fragmentation. We compared the observed frequencies

Table 1 Precise composition of the ZymoBIOMICS Microbial Community Standard

Species	Genomic DNA	16S Only	16S & 18S	Genome Copy	Cell Number	Morphology	Lysis
Pseudomonas aeruginosa	12	4.2	3.6	6.1	6.1	Rod_gram_neg	Easy
Escherichia coli	12	10.1	8.9	8.5	8.5	Rod_gram_neg	Easy
Salmonella enterica	12	10.4	9.1	8.7	8.8	Rod_gram_neg	Easy
Lactobacillus fermentum	12	18.4	16.1	21.6	21.9	Rod_gram_pos	Hard
Enterococcus faecalis	12	9.9	8.7	14.6	14.6	Ovoid_gram_pos	Hard
Staphylococcus aureus	12	15.5	13.6	15.2	15.3	Ovoid_gram_pos	Hard
Listeria monocytogenes	12	14.1	12.4	13.9	13.9	Ovoid_gram_pos	Hard
Bacillus subtilis	12	17.4	15.3	10.3	10.3	Rod_gram_pos	Hard
Saccharomyces cerevisiae	2	NA	9.3	0.57	0.29	Fungi	NA
Cryptococcus neoformans	2	NA	3.3	0.37	0.18	Fungi	NA

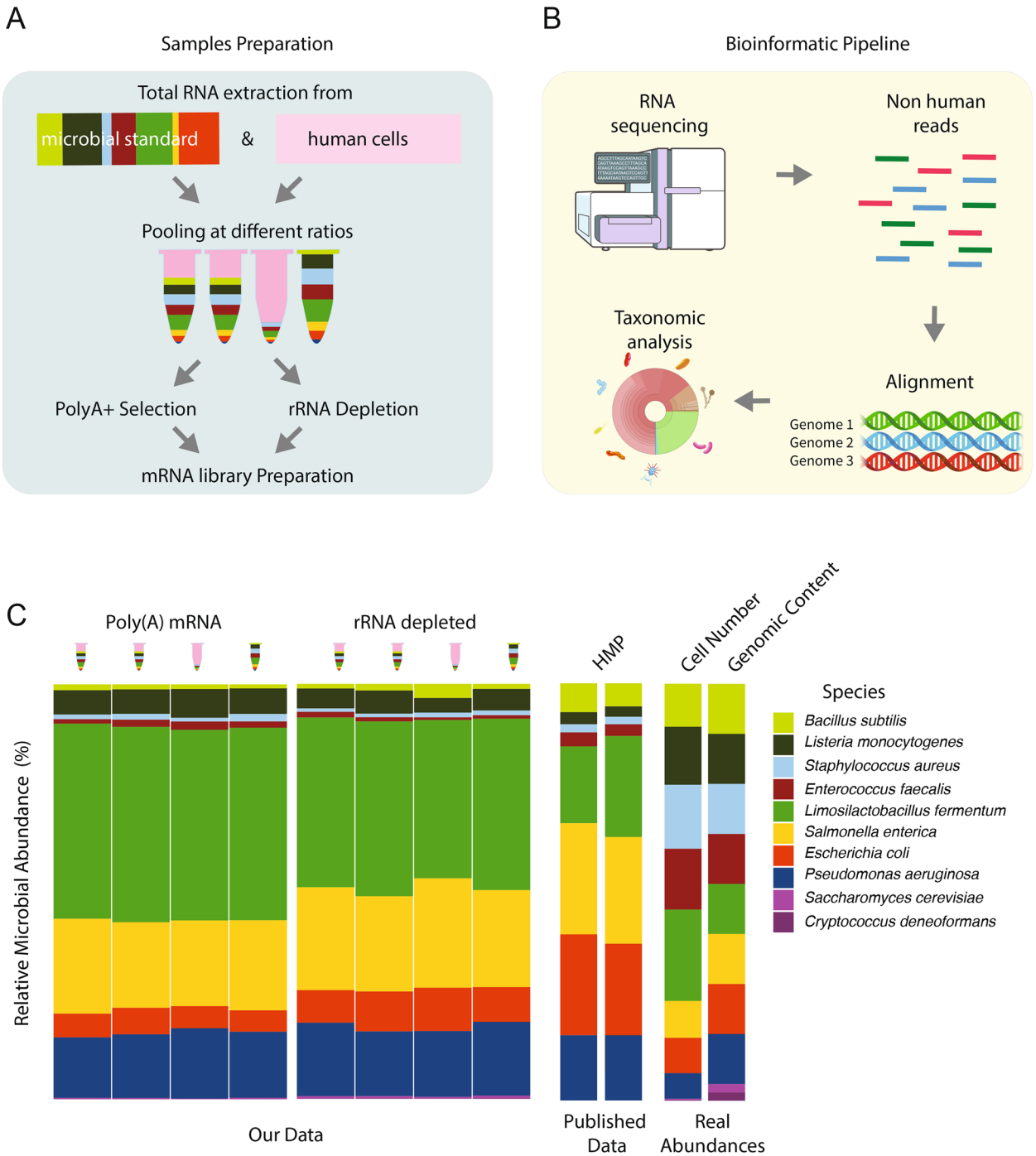


Fig. 1 Poly(A)-enriched and total RNA-seq approaches successfully reconstruct the microbiome complexity in experimentally controlled samples. **A**, **B** Workflow of the experimental approach and bioinformatic pipeline, respectively. **C** Taxonomic profiling of the microbial mock community. Stacked bar charts showing the relative microbial abundances (%) of the ten species present in the ZymoBIOMICS Microbial Community Standard. The panel compares the taxonomic profiles obtained in this study ("Our Data") using two library preparation strategies: poly(A) mRNA selection and rRNA depletion. These experimental results are compared with "Published Data" (HMP) and the theoretical composition of the standard ("Real Abundances") based on cell number and genomic content. Each color represents a distinct microbial species as indicated in the legend on the right

against an empirical background probability and, as expected, human host reads exhibit a high frequency of A/T stretches (Fig. S1C). Conversely, in the bacterial fraction, the observed frequency of 8-nt A/T tracts is significantly lower (~2.5%) than the expected background probability (~5.5–6%), reflecting the evolutionary suppression of long A/T homopolymers in bacterial coding sequences [42, 43]. This indicates that while rare opportunistic mispriming on internal sites does occur, it cannot fully account for the entirety of the recovered microbiome. Rather, the bacterial RNA is likely retained through non-specific physical carryover, a capture that explains the compositional identity of poly(A) and rRNA-depleted libraries observed in our quantitative standards (Fig. 1C).

Comparative metagenome and metatranscriptome analyses

To initially validate the reliability of our approach in ex vivo specimens, six samples were subjected to both DNA and RNA extraction followed by sequencing (Fig. 2). The comparative analysis of metagenomic (DNA) and metatranscriptomic (RNA) datasets demonstrates a high degree of concordance ($R \sim 0.8$) between the taxonomic profiles obtained with the two techniques (Fig. 2B–D, Fig. S2E), despite a major difference in the distribution of sequencing reads across the genome (Fig. 2A). Principal coordinate analysis (PCoA) (Fig. 2B) shows close clustering of matched DNA and RNA samples, with minimal separation observed between the two data types. Hierarchical clustering (Fig. 2D) further supports this concordance, as paired DNA and RNA samples cluster together in all cases, reflecting strong similarity between the metagenomic and metatranscriptomic profiles at the community level. Taxonomic composition comparison (Fig. 2C, D) reveals consistent representation of dominant species such as *Lactobacillus crispatus* and *Lactobacillus iners* in both DNA and RNA datasets.

However, some differences in relative abundances are noted, especially for less abundant taxa (*Prevotella bivia*), but also in the case of *G. vaginalis*, which exhibit increased representation in the RNA profiles with respect to the DNA-based approach (Fig. S2F). We also analyzed the concordance in RNA/DNA abundances for each gene/pathway (Fig. 2E). We observe that genes belonging specifically to two pathways are over-represented (ribosome and metabolic pathways) and two under-represented in the RNA analysis (NER and BER). Although each sample is dominated by different species (*L. iners*, *L. crispatus*, and *G. vaginalis*), this trend is overall maintained across all samples, suggesting that this is a shared feature across different species (Fig. 2E).

Taken together, these data demonstrate substantial concordance between metagenomic and metatranscriptomic approaches in determining the composition of the microbiota.

The vulvar milieu at the crossroads between the anal and the vaginal microbiome, and the vulvar community state types

Having assessed the reliability of our method, we applied it to mRNA-seq data obtained from 30 vulvar swabs collected from 30 premenopausal healthy individuals (Table 2). It is well established that the vulvar microbiota closely resembles the vaginal microbiota, although the vulvar microbiome generally exhibits greater diversity due to the presence of commensal organisms of cutaneous and fecal origin [28]. To further investigate similarities and differences in microbial composition, we performed NMDS analysis on batch-corrected data, comparing vaginal, vulvar, and anal microbiomes from both published metagenomic datasets [40] and our own samples (Fig. 3A).

Our analysis revealed that samples from each body site cluster together, with clearly separated centroids for each group. This distinct separation underscores the unique microbial communities associated with each anatomical environment. Notably, our metatranscriptomic samples clustered closely with published vulvar metagenomic data, further validating the reliability of our approach. Interindividual variation may be influenced by sampling location (e.g., vulvar mucosa, cutaneous folds, anal or perianal areas), which in our dataset was restricted to the vestibular area, whereas published datasets included more diverse sites.

Interestingly, the centroid of the vulvar microbiome occupies an intermediate position between the anal and vaginal clusters, reflecting its central anatomical location. Upon examining species shared across the three environments, we found that among the top 100 expressed species in each compartment, the vulva shares more species with the anus than with the vagina (Fig. 3B). Notably, there are no species shared exclusively between the anus and vagina without also being present in the vulva, further supporting the idea that the vulva serves as a microbial intermediary between these otherwise distinct environments (Fig. 3B). Analysis of relative abundances (Fig. 3C) similarly illustrates that the vulva represents an intermediate environment, sharing microbial features with both the anus and the vagina.

Together, these findings further validate the robustness of our metatranscriptomic approach and reinforce the hypothesis that the vulvar microbiome is a heterogeneous environment shaped by microbial communities of vaginal, cutaneous, and fecal origin. We sought to

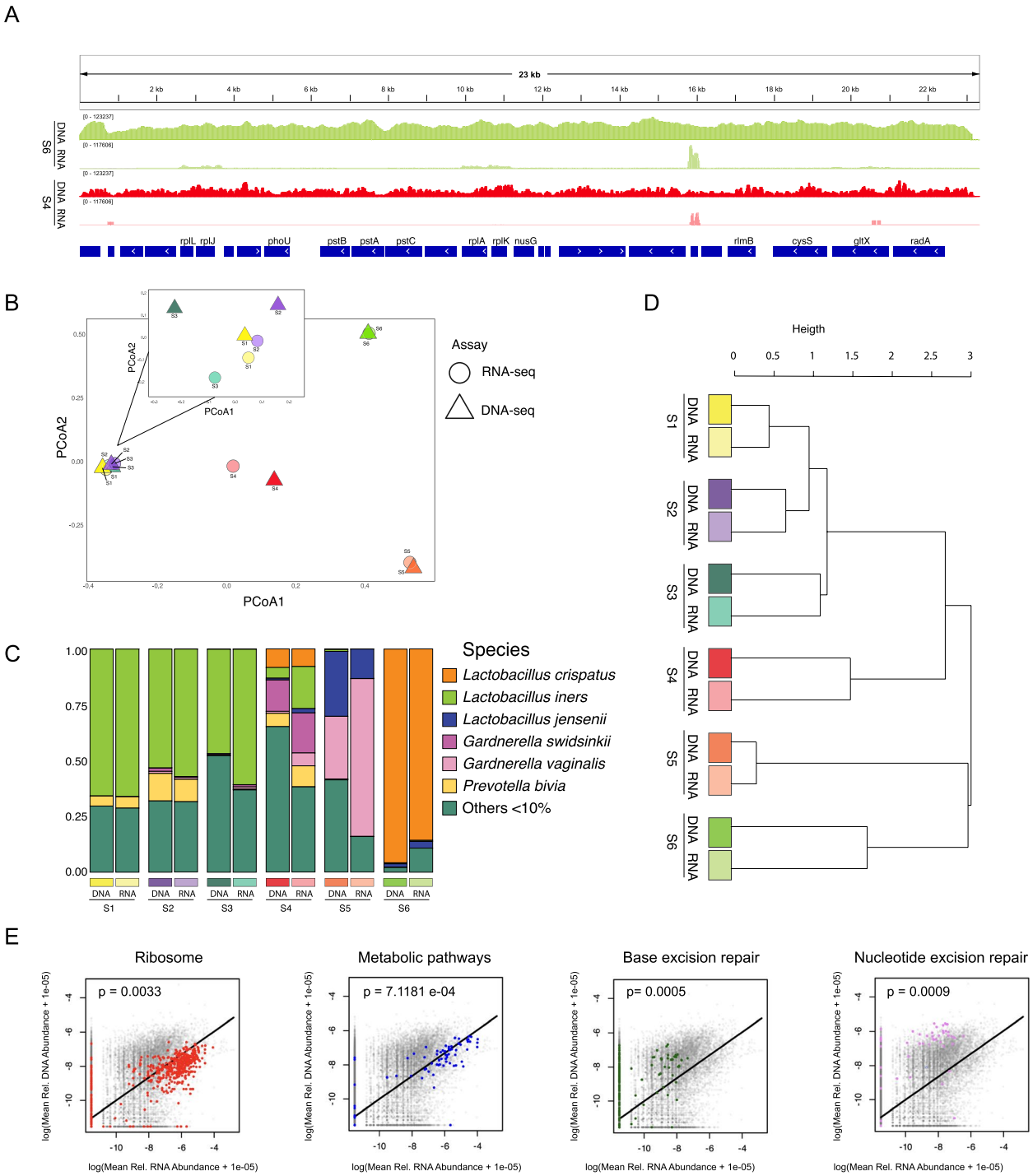


Fig. 2 Comparative metagenome and metatranscriptome analyses. **A** Integrative Genomics Viewer (IGV) display of read mapping to a 23-kb region of the *Lactobacillus crispatus* genome. The specific window chosen allows the visualization of multiple microbial genes at once. Only two representative samples have been chosen, which share *Lactobacillus crispatus* in their ecosystem. For each sample, the upper track displays DNA reads, while the lower track displays RNA reads. **B** Principal coordinate analysis (PCoA) plot based on Bray-Curtis distances, illustrating the similarity between RNA and DNA data. In the inset is the magnification of particularly similar samples. **C** Pair of stacked bar plots showing the relative abundance of the top 10% microorganisms. **D** Hierarchical clustering dendrogram based on Bray-Curtis dissimilarity, showing the relationships between metatranscriptome and metagenome of samples. Samples are labeled with their identifiers, with DNA samples in darker shades and RNA samples in brighter shades. **E** Scatter plots showing linear relationships between relative abundance in RNA (x) and DNA (y) of KEGG pathway ortholog groups (KOs) across all samples. The top two upregulated (ribosome and metabolic) and downregulated (base and nucleotide excision repair) pathways are highlighted, with each dot/KO in a different color

Table 2 Socio-demographic and clinical characteristics of the study cohort ($N = 30$). Data are presented as mean (M) ± standard deviation (SD) for age, and as number (percentage) for categorical variables. Ethnicity was self-reported. "NA" indicates data was not available or the participant declined to answer

Healthy individuals (N=30)		
Age		M = 28.5, SD = 5.7, range 23–44
Ethnicity	Caucasian	29 (97%)
	Asian	1 (3%)
Smoking	No	16 (53%)
	Occasionally	8 (27%)
	Less than 5 cigarettes per day	4 (13%)
	More than 10 cigarettes per day	2 (7%)
Diet	Omnivore	22 (73%)
	Omnivore, No Dairy	1 (3,33%)
	Omnivore, No Gluten	1 (3,33%)
	Vegetarian	5 (17%)
	Vegan	1 (3,33%)
Relationship status	Single	6 (20%)
	Married/In a relationship	21 (70%)
	Occasional partners	2 (7%)
	NA	1 (3%)
Menstruation	Regular	14 (47%)
	Irregular	9 (30%)
	Hormonal contraceptives	6 (20%)
	NA	1 (3%)
Contraception	None	12 (40%)
	Ring	1 (3,33%)
	Male condom	9 (30%)
	Estro-progestin pill	6 (20%)
	Progestin-only pill	1 (3,33%)
	NA	1 (3,33%)
HPV vaccine	No	9 (30%)
	Yes	10 (33%)
	NA	11 (37%)

identify the common features between vulvar and vaginal milieus. We thus hypothesize that the community state types (CST) found in the vaginal microbiome of different individuals can also be reflected by the vulvar milieu.

In order to confirm our hypothesis, we performed a clustering analysis on non-scaled data, analyzing the top 100 most abundant species in the vaginal milieu, thus allowing us to find clusters driven by the most abundant microorganisms (Fig. 3D and Fig. S3A, B). The analysis revealed four out of the five major groups of microbial communities, three of which are dominated by *Lactobacillus* and one group composed of genera other than *Lactobacillus*, and without any individual dominated by *L. jensenii*. In particular, we found 16 samples dominated by *L. crispatus*, 9 samples dominated by *L. iners*, 3 samples dominated by *L. gasseri*, and 3 by other species (Fig. 3E). These four CSTs not only mirror those previously established in vaginal studies [44], but also the distribution of

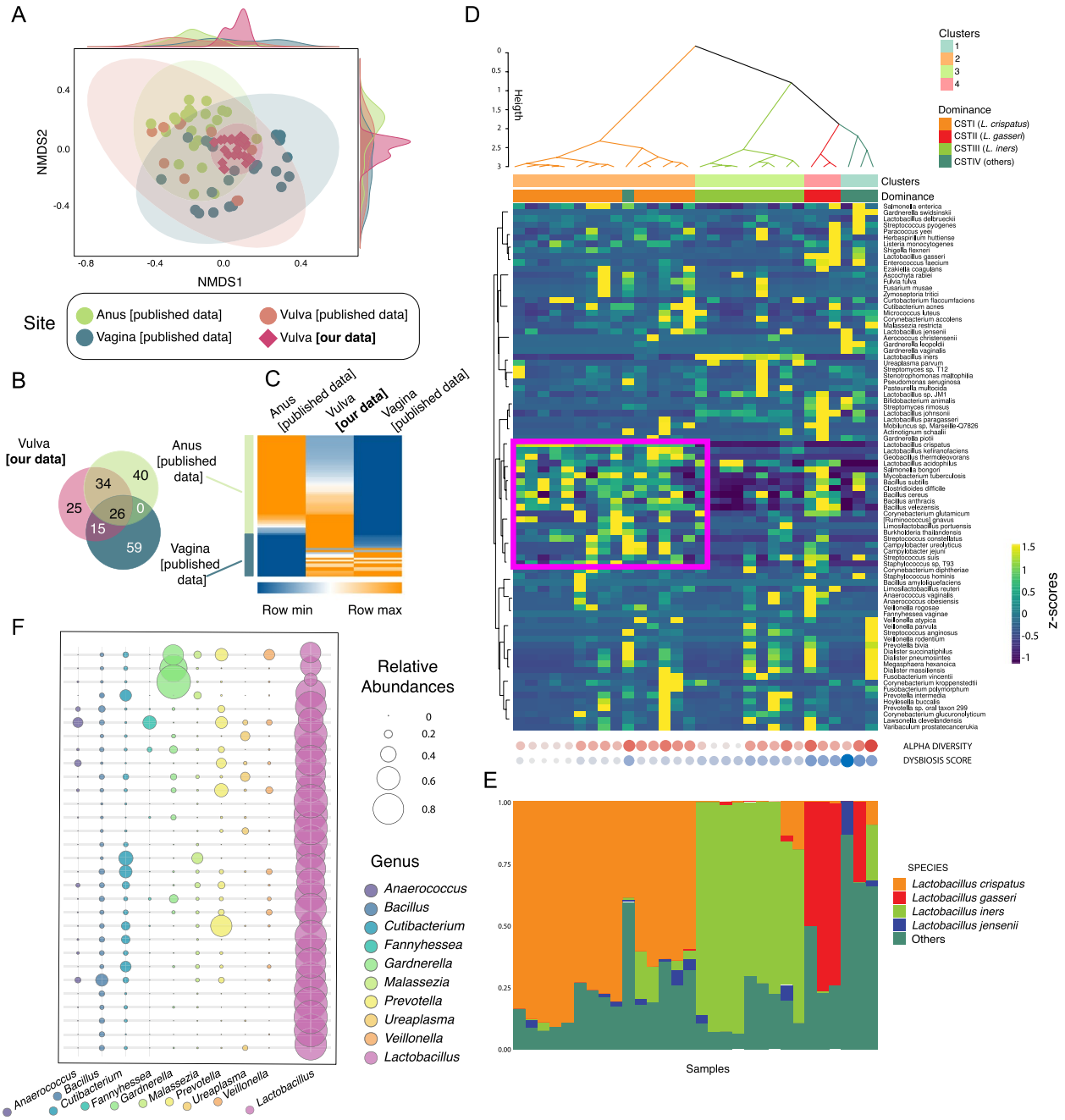


Fig. 3 The CST of the vulvar milieu at the crossroads between the anal and the vaginal one. **A** Non-metric multidimensional scaling (NMDS) by Bray-Curtis dissimilarity plot comparing microbial community compositions across different body sites and studies. Each point represents a sample, with shapes that distinguish samples from our study (diamonds) and published data (circles). Colors indicate the body site of origin. Shaded ellipses represent 95% confidence intervals for each body site. Marginal density plots along the axes illustrate the distribution of samples for each body site along NMDS1 and NMDS2. **B** Venn diagram showing the overlap of the top 100 expressed species for each of the indicated anatomical compartments in our vulvar dataset and in published anal and vaginal datasets. **C** Heatmap of z-score transformed abundances of the 75 species shared in at least two anatomical compartments. The colored sidebar indicates the compartment in which each species is most abundant. **D** Heatmap of scaled relative abundances of most abundant (>0.1% in at least one sample) microbial taxa across samples, with hierarchical clustering and dominance annotations. Each row represents a taxon and each column represents an individual sample, ordered by hierarchical clustering. The dendrogram above is based on Ward's method and Euclidean distances, with branches color-coded into four clusters (cluster 1 to 4). Alpha diversity and dysbiotic score metrics are shown below the heatmap. The dysbiosis score was calculated by weighting taxa abundances against MicroPhenoDB association scores specific to the term "bacteria vaginosis." **E** Stacked bar plot of abundances of species characterizing each sample with the 4 most abundant species. **F** Bubble plot of abundances of genera that are over 5% in at least one sample. The size of each bubble corresponds to the relative abundance of a genus within a sample, while colors indicate different genera

samples within these groups is similar to the vaginal CST types [45, 46], where CST-I and CST-III are the most abundant with respect to CST-II and CST-V within the healthy individuals, again reinforcing the biological connection between these anatomical sites. In the attempt to correct our data from any lysis and extraction biases, we applied the morphology-based computational correction described by Rauer et al. [47] (see Methods section for details).

We therefore estimated the correction biases (Fig. S3C) for different mock communities (poly(A) mRNA samples of Fig. 1). Because the key species driving vulvar ecosystem variance are not heavily affected by the extraction bias (Fig. S3D), and because any morphological-based correction would remove viruses and fungi, we kept original abundances for downstream analysis. From a global perspective, also NMDS confirms that the main species identified through CST analysis are the main drivers of sample separation (Fig. S3E). From the heatmap (Fig. 3D), we do not observe major subclades of concomitant bacteria, with the exception of one cluster of co-occurring species in the *L. crispatus* cluster (in pink). Interestingly, this cluster is mainly populated by *Lactobacilli* and *Bacillus* species (including *B. subtilis*, *B. cereus*, and *B. velezensis*), which are known to have a protective role in the vaginal milieu [48, 49]. In addition, looking at the relative enrichment of different genera across CSTs, we observe no significant differences, with only the exception of the aforementioned *Bacillus* genus which is enriched in CST-I and CST-II with respect to CST-III and CST-IV (Fig. S3F).

We also tried to identify whether any other associated variables (diet, smoking, age etc.) could also explain some of the variability observed in our cohort, but we did not observe any significance (data not shown).

For each individual, we evaluate the biodiversity (measured as alpha diversity per sample) and we compare it among the 4 CST clusters. As expected, CST-IV, which by definition is enriched in "other" species, displays a general trend of higher biodiversity when compared to *L. iners*- and *L. crispatus*-dominated CST (Fig. S3G).

Among the predominant genera (defined as relative abundance >5%), we observed notable diversity (Fig. 3F). Anaerobic and skin-associated genera, such as *Anaerococcus*, *Veillonella*, and *Cutibacterium*, were prevalent. Additionally, genera commonly associated with the vaginal or gastrointestinal microbiomes, including *Fannyhessea*, *Gardnerella*, and *Prevotella*, were detected. Interestingly, *Gardnerella* exhibited a tendency toward mutual exclusivity with the predominant genera, while *Prevotella* and *Ureaplasma* displayed more uniform abundance distributions across samples.

Within the fungal community, only *Malassezia* is present above the threshold of >5% relative abundance and it is followed by *Saccharomyces*, which was present at lower levels (<1%). These findings highlight the richness and complexity of the vulvar microbial ecosystem, comprising contributions from multiple microbial niches and the ability of our approach to investigate both the bacteria and the fungal components of the vulvar milieu.

Multi-omic integration of the vulvar microbiome and host transcriptome

To elucidate the functional crosstalk between the vulvar microbiota and host cellular processes, we constructed an integrated network by calculating correlations between individual human genes and microbes (Fig. 4 and Methods section). This approach is particularly advantageous because both host gene expression and microbiome data are derived from the same RNA-seq samples. Sampling the two parameters simultaneously not only reduces the impact of technical batch effects but also enhances the reliability of the correlations.

The network of microbial interactions reveals that while *Lactobacillus* species show the highest mean relative abundance (node size), they exhibit a lower degree of centrality compared to less abundant taxa, identifying the latter as the primary structural hubs of the network. These taxa are positively connected to *G. vaginalis*, forming a well-integrated hub of pathobionts and dysbiotic species. While connections across the network are predominantly positive, rare negative correlations specifically link *L. crispatus* to *L. iners* and *Gardnerella* spp., representing a mutually exclusive antagonistic relationship (Fig. 4A).

The microbial community structure, characterized by genus-specific correlation clusters, was then mapped onto the host transcriptome, revealing a complex interactome where specific microorganisms act as key modulators of host gene modules (Fig. 4B). The relationship between these transcriptomic modules and CST was further validated through module-trait analysis (Fig. 4C), highlighting significant associations between CST composition and biological processes such as keratinization (green) and immune activation (turquoise). A deeper investigation into the functional landscape (Fig. 4D) revealed that microbial influence is highly specialized and polarized. By quantifying the percentage of interactions across selected Gene Ontology (GO) terms, we observed that potential pathobionts (i.e., *L. gasseri* or *G. vaginalis*) exhibit a negative regulation with epithelial integrity (keratinization and epithelial cell differentiation) and a concomitant activation of immunity and inflammatory pathways. This

specialized regulation is exemplified by their synergistic interaction profiles (Fig. 4E), targeting shared pathways through overlapping genes. In particular, *G. vaginalis* is highly connected to general immunity-related genes, while both bacteria are positively interacting with innate immune genes.

Finally, we identified a subset of "bridge genes," defined as host nodes interacting with multiple microbial taxa that serve as central molecular switches in the vulvar environment (Fig. 4F). The co-regulation heatmap of these bridges demonstrates a binary control mechanism: while most host processes exhibit resilience to microbial variability, key checkpoints in the inflammatory response and epithelial differentiation (e.g., NF-κB signaling [50], IL-1β–mediated pathways [51]) are subject to opposing regulatory pressures from symbionts and pathobionts. This integrated analysis allows us to assign profound biological value to individual genes that might otherwise be overlooked in broader datasets. A prime example is ZDHHC5, a palmitoyl acyltransferase that acts as a requisite for activating gasdermin D-mediated pyroptosis, a cell autonomous defense mechanism against pathogens and inflammatory diseases [52, 53].

Notably, our analysis highlighted ZDHHC5 as a bridge between *G. vaginalis*, *L. gasseri*, and *L. crispatus* (Fig. 4E–F). The positive correlation of ZDHHC5 with *G. vaginalis* and *L. gasseri* suggests that these microbes may prime the vulvar epithelium for pyroptotic signaling. Conversely, its negative correlation with *L. crispatus* indicates a protective suppression of the ZDHHC5-GSDMD axis, potentially preserving barrier integrity and limiting cytokine release. Collectively, these data suggest that vulvar health is maintained through a delicate balance of microbial inputs that converge on specific host bridge genes to regulate tissue homeostasis and immune tolerance.

Discussion

With the parallel evolution of experimental strategies and computational approaches, microbiome analysis has become increasingly detailed, sensitive, and specific, enabling high-resolution characterization of complex microbial communities. Nevertheless, most available protocols have inherent strengths and limitations, making the careful selection of experimental and analytical procedures essential for addressing specific biological questions.

Our laboratory primarily focuses on RNA perturbation in human vulvar diseases, and we recently recognized that our sequencing datasets also contained previously unexplored information regarding microbial community composition.

By validating the concordance between DNA- and RNA-based profiling in both in vitro and ex vivo models, we demonstrate that conventional poly(A)-enriched or ribosomal RNA-depleted mRNA-seq can be repurposed as a unified, metatranscriptomic-like strategy to identify microbial species and estimate their relative abundances.

Minor discordance between microbial abundances inferred from RNA and DNA analyses may also reflect the transcriptional activation within specific taxa. In living cells, one of the most rapid mechanisms to modulate protein synthesis is through variation in ribosomal RNA content, allowing adaptation to changing environmental conditions. The relatively high RNA-level abundance of ribosomal genes, together with genes involved in metabolic pathways, suggests that they are crucial for cell survival in the dynamic vulvar milieu.

The efficiency of our soft-lysis RNA-sequencing–based protocol is skewed toward easy-to-lyse Gram-negative bacteria, whereas Gram-positive species are generally underrepresented. Importantly, this bias is consistent across conditions and samples. Despite this limitation, a single protocol enabled the simultaneous profiling of the vulvar microbiome and host epithelial transcriptome

(See figure on next page.)

Fig. 4 The interdependencies of different microbial species and the gene expression level of vulvar cells. **A** CytoScape visualization of correlation network of bacterial species in the vulvar microbiome, with colors identifying shared genera. **B** Integrated network showing correlation between microorganism species and host genes. Genes (round shape) are color-coded and grouped in WGCNA modules while microbes are depicted as black diamonds. **C** Module trait relationship between modules in B and CST clusters. The color-coded text corresponding to the GO term with the highest significance is indicated next to each module (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). **D** Diverging bar plot illustrating the functional landscape of host-microbe interactions, showing the relative contribution of each microbial taxon to specific biological processes. The data represents the percentage of interactions (Spearman's correlations) associated with each Gene Ontology (GO) term, normalized by the total number of functional occurrences for each microorganism. Positive correlations (Spearman's $\rho > 0$) are shown to the right of the zero-line, while negative correlations (Spearman's $\rho < 0$) are shown to the left. Microorganisms (Y-axis) are listed and ordered to highlight differences in their functional profiles. Colors represent distinct biological processes as indicated in the legend. **E** Network representation of the host-microbiome crosstalk between *G. vaginalis*, *L. gasseri*, and human genes (ellipses). Genes are clustered based on their biological processes: immune system, innate immune response, and keratinization are highlighted (other connected genes in Supp. Material). **F** Heatmap illustrating Spearman's correlations (ρ) between selected microbial taxa (columns) and host genes (rows) identified as "bridges" within the host-microbiome network. Bridge genes are grouped based on their biological process annotations already defined in **D**. These genes are defined as host nodes exhibiting statistically significant interactions with two or more distinct microbial taxa

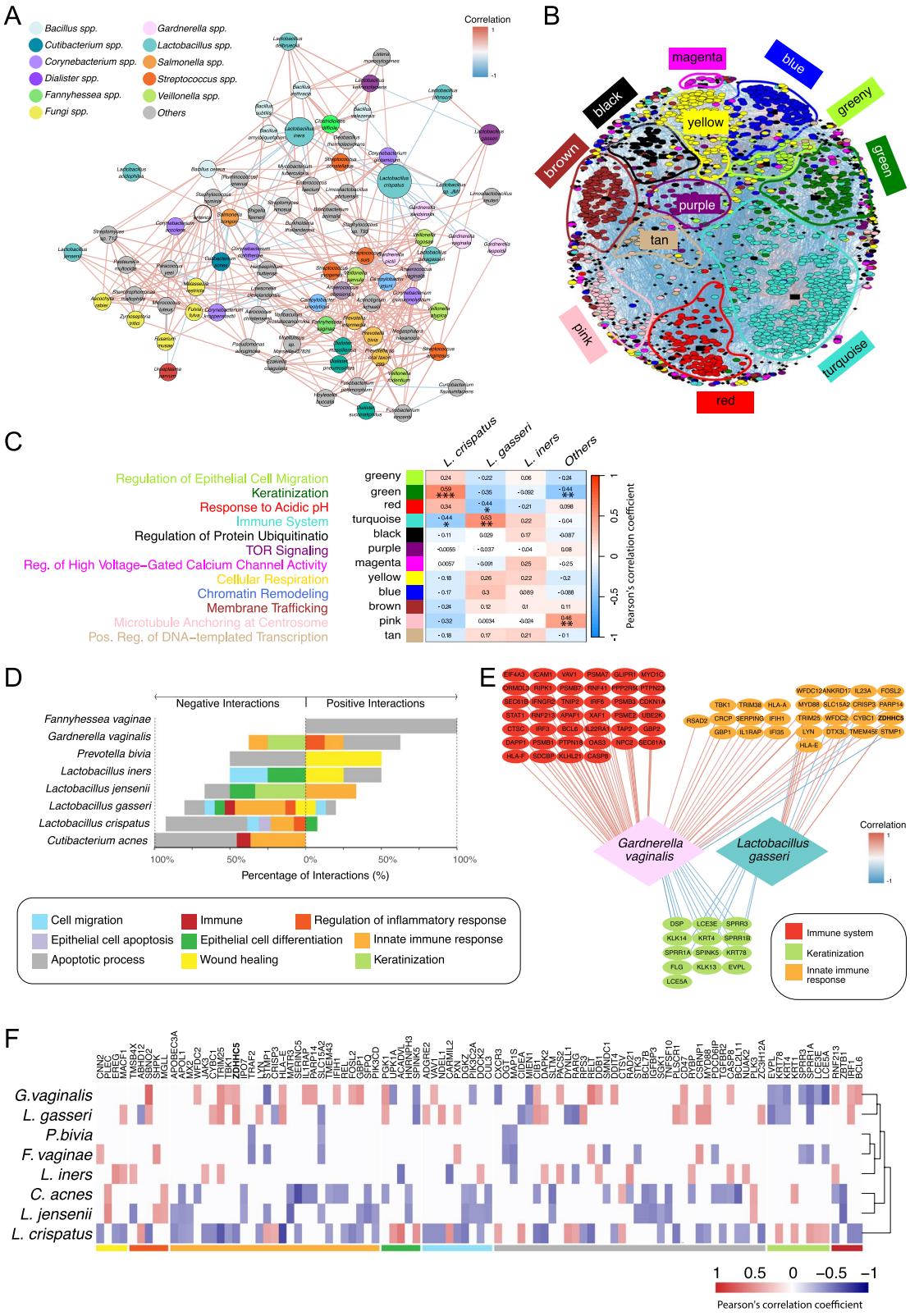


Fig. 4 (See legend on previous page.)

from the same specimen. To generalize soft-lysis protocols to other body sites characterized by more complex ecosystems, the integration of morphology-based computational corrections [47] may be essential to ensure accurate absolute abundance estimations, particularly in contexts with high proportions of hard-to-lyse bacteria.

Another limitation is represented by low-biomass samples, which inherently amplify the challenge of environmental and reagent contamination. Perfectly matched physical negative controls represent the gold standard for direct background subtraction, and we successfully applied this strategy to our DNA libraries.

However, this approach is often logistically difficult in clinical studies involving complex, multi-batch, and multi-site sample collection. In such scenarios, where continuous physical controls are fragmented, as occurred for our RNA swabs, the use of robust statistical methods like the *decontam* package [54] becomes essential for accurately identifying and removing contaminants. Tailoring the decontamination strategy to the logistical constraints of each dataset is a key step to ensure maximal accuracy.

We applied our approach to the vulvar milieu, which represents an emerging area of research, with only a few recent studies beginning to explore its complexity. Despite the limited number of investigations, current evidence suggests that the vulvar microbiome shares several characteristics with the vaginal environment [26, 29, 36, 55]. However, it generally exhibits a higher degree of colonization by commensals of both cutaneous and fecal origin, endowing the vulva with a distinctive microbial signature. Here, we show that vulvar community state types (CSTs) mirror canonical vaginal CSTs while occupying an intermediate ecological niche between anal and vaginal communities, supporting the concept of the vulva as a distinct yet interconnected microbial ecosystem. Notably, the CST-I cluster is enriched not only in *Lactobacillus crispatus*, but also in *Bacillus* species, which have been linked to vaginal protection, suggesting that the protective profile typically attributed to CST-I may not be driven by *Lactobacillus* alone but is likely reinforced by the concomitant presence of other protective species. Interestingly, vulvar dysbiosis is associated with a higher biodiversity (Fig. S3G, H), which is in contrast with the behavior of the gut where a higher microbiome heterogeneity is overall associated with a healthy status [56, 57], suggesting the existence of a stricter control of the vulvar microenvironment.

Our finding suggests that distinct microbial communities are associated with specific host transcriptional programs. *Lactobacillus crispatus*-dominated communities are positively correlated with genes involved in epithelial differentiation and barrier integrity, whereas communities enriched in *Lactobacillus gasseri* are positively associated with transcriptional signatures of inflammation and negatively with skin keratinization. These trends support the protective role of *L. crispatus* and an intermediate role of *L. gasseri*, in line with data from literature, while providing a more direct molecular link between the host and the microbiome.

The intra-sample framework of our methodology, integrating microbial abundances with host gene expression, uncovers specific host-microbe interaction modules, linking *Lactobacillus*-dominated communities to epithelial differentiation and barrier maintenance and dysbiosis-associated taxa to inflammatory transcriptional programs.

Together, these findings provide both a conceptual and technical framework to mine existing RNA-seq datasets for microbiome and host-microbe crosstalk signals. Although the recovery of bacterial transcripts in poly(A)-selected libraries may seem counterintuitive, our data suggest this is likely driven by a combination of rare internal mispriming and non-specific physical carryover (Fig. S1C), confirmed by the concordant microbial profiles between poly(A)-enriched and rRNA-depleted libraries (Fig. 1C).

Thus, despite these technical peculiarities, this method ensures that the core biological signals are accurately captured and offers a scalable, cost-effective avenue to revisit archived transcriptomic cohorts and illuminate the functional impact of microbiota in health and disease.

Even though comprehensive analysis of the vulvar microbiome functional activity was not the primary aim of our study, the approach we employed could also enable the investigation of the dynamic gene-expression profiles of vulvar microorganisms, thus providing a more accurate snapshot of microbial activity in the vulvar environment.

Conclusions

In conclusion, our work demonstrates that conventional mRNA-seq datasets can be reliably repurposed to simultaneously analyze the microbiota composition and host epithelial transcriptome, providing an integrated view of host-microbe interactions within the same specimen. This integrated strategy: 1. eliminates batch effects associated with the preparation of separate RNA and DNA sequencing libraries; 2. maximizes the informational output of a single experiment; 3. reduces experimental costs; and 4. allows the investigation of mutual relationships between host gene expression and microbial species

and their functional pathways. Applying this framework to the vulvar milieu, we show that vulvar community state types reflect vaginal-like structures while defining a distinct ecological niche, and that specific microbial configurations are tightly linked to host transcriptional programs governing barrier integrity and inflammation. Beyond characterizing the vulvar ecosystem, this study establishes a scalable and cost-effective conceptual and methodological strategy to mine existing RNA-seq cohorts for microbiome composition, activity, and host-microbe crosstalk in both health and disease.

Methods

Mock microbial community RNA extraction and library preparation

For sample preparation, we used both co-extraction and post-extraction mixing of the VLS-VSCC human vulvar squamous cell carcinoma cell line and the ZymoBIOMICS Microbial Community Standard (Zymo Research; Irvine, CA, USA). Four input samples were generated for RNA extraction: (1) a co-extracted sample containing 60% human cells and 40% of the ZymoBIOMICS standard; (2) a Zymo-only positive control; (3) a post-extraction mix containing 75% ZymoBIOMICS and 25% of human RNA; (4) a technical replicate of condition 3. Total RNA was extracted from all four input samples using a TRIzol-based protocol and eluted in nuclease-free water. The two post-extraction mix replicates were generated by manually mixing the human-only and Zymo-only eluates.

Assessment of poly(A) capture mechanisms

To evaluate the capture mechanism of bacterial mRNA in poly(A)-selected libraries, mapped host and microbial reads were computationally scanned for internal poly-A/T stretches (8, 10, and 12 nt). To establish a baseline of expected random occurrence, an empirical background probability was computed via sequence permutation, shuffling the bacterial reads while strictly preserving read length and GC content. The observed frequency of A/T stretches in the actual bacterial reads was then compared against this permuted baseline to estimate the relative contributions of opportunistic internal mispriming versus non-specific physical carryover.

Patients' recruitment

Thirty healthy volunteers were recruited from Centro Salute Pelvi from February 2023 to July 2023. The inclusion criteria were premenopausal women (between 18 and 50 years of age) without any vulvar pain or vulvar symptoms. Participants have been excluded from the study if they were post-menopausal, pregnant, breast-feeding, or gave birth within 4 months. In addition, subjects have been excluded if they received systemic or vaginal antimicrobial or probiotic therapy within 1 week before the visit. The study was revised and approved by the University of Turin Ethical Committee (Prot. N. 0214340, 06/04/2023). The participants provided their written informed consent to participate in this study.

RNA collection and extraction from swabs

We collected vulvar specimens by a gentle sterile dry swab of the vestibule area, which was then placed in a 1.5-ml tube containing RNA/DNA shield reagent to preserve nucleic acids. RNA extraction was performed by adding 1 ml of QIazol/TRIzol Lysis Reagent (Qiagen). After incubation at RT for 3 min, 200 µl of chloroform were added and samples were vortexed for 15 s and incubated at room temperature for 3 min. The samples were then centrifuged at max speed for 15 min at 4 °C, and the obtained aqueous phase was transferred into a new tube and precipitated with 500 µl of isopropanol and 1 µl of glycogen at room temperature. After 10 min, samples were centrifuged at maximum speed for 15 min at 4 °C. The pellet was washed with 70% EtOH, resuspended in DNAse/RNAse-free water and quantified at Nanodrop.

DNA collection and extraction from swabs for microbiome analysis

Vulvar swabs were collected by sterile stick swabs and immediately stored at −80 °C. To resuspend the samples, we soaked the swab in 1 ml of NaCl 0.9% for 5 min, while vortexing. After centrifugation at 13,000 rpm for 10 min at room temperature, the supernatant was discarded and the pellet was resuspended in 80 µl of TrisHCl. Subsequently, 20 µl of lysozyme (10 mg/ml) was added and samples were kept at 37 °C for 30 min to allow cell wall degradation. After incubation, we performed bead beating using TissueLyser II at 30 Hz for 2 min with 1 mm metallic beads. Next, to remove proteins, we added 1 µl of proteinase K (20 µg/µl) and we incubated samples at 56 °C for 30 min. To extract DNA, we added 500 µl of Phenol-Chloro-Isoamyl-Alcohol and 500 µl of NaCl 0.9%, incubated them at room temperature for 10 min and then centrifuge at 11,000 rpm for 5 min. After phase separation, we added to the aqueous phase 1 µl of glycogen, 30 µl of SodiumAcetate 3M, and 800 µl of EtOH 100% and we precipitated at −20 °C overnight. The next day, to purify the DNA, samples were centrifuged at 13,000 rpm at 4 °C for 30 min, the supernatant was removed, and the samples were washed with 500 µl of EtOH 70%. After centrifugation at 13,000 rpm at 4 °C for 5 min, the supernatant was discarded and the pellet was resuspended in

30 µl of RNAse-free water. The samples were then quantified at the Qubit instrument.

Library preparation and sequencing

For RNA-seq library preparation, the quantity and quality of the starting RNA were checked by Qubit and TapeStation (Agilent). For the in vitro experiments presented in Fig. 1, libraries were prepared from 500 ng of total RNA using the Watchmaker mRNA Library Prep Kit and Watchmaker RNA Library Prep Kits with Polaris® Depletion, following the manufacturer's instructions.

Libraries for ex vivo samples shown in Figs. 2, 3, and 4 were prepared exclusively using the Illumina Stranded mRNA prep kit.

For DNA-seq libraries, 100 ng of DNA was sonicated at the Pico Bioruptor (Biosense srl) with 3 cycles of sonication (20 s ON and 30 s OFF). Libraries were prepared using the Watchmaker DNA Library Prep Kit, following the manufacturer's instructions.

Libraries were assessed for the quality at the TapeStation (Agilent) and sequenced on the Illumina NextSeq1000 System (paired-end 60 bp reads).

Bioinformatic analysis

Metatranscriptome and metagenome pre-processing

The core bioinformatic analysis was performed using a custom Snakemake pipeline. Briefly, the pipeline executes a set of rules in order to:

i) Preprocess and perform quality control for FastQ data using the snakemake wrapper of fastp tool (https://github.com/OpenGene/fastp).
ii) To account for highly repetitive regions, centromeric gaps, and structural variants often unmapped in older reference genomes, we utilized a two-step alignment strategy. RNA-seq paired-end reads were aligned to two human genome assemblies: hg38 (GRCh38) and the T2T-CHM13 (telomere-to-telomere) assembly using the STAR aligner. This dual-mapping approach ensures that even low-complexity or non-canonical human reads are identified. Only reads that did not map to either human assembly were retained for microbial analysis.
iii) Remaining reads were classified using Kraken2 with the pre-configured reference database Kraken2 PlusPF (Standard plus RefSeq protozoa, fungi, and human). The inclusion of the human genome in this database acts as a secondary "safety net"; any human reads bypassing the initial alignment filters (due to sequence divergence or library artifacts) are cross-referenced against the internal human k-mer library and flagged as "Human" or "Unclassified." The Kraken2 classification was executed with additional options (--minimum-hit-groups 3) in order to achieve a more stringent classification. By leveraging Kraken2's lowest common ancestor (LCA) algorithm, our pipeline ensures resilience to chimeric artifacts by either assigning reads with conflicting k-mer signatures to high-level taxonomic nodes or discarding them as unclassified, thereby preventing the erroneous inflation of specific microbial abundances.
iv) Kraken2 report files were processed with Bracken using a pre-configured database (Kraken2 PlusPF). The analysis was performed with the following parameters: -r 50 -l S -t 5. Bracken re-estimated taxonomic abundances by analyzing the k-mer distribution across specified taxonomic levels. The output included a detailed abundance report and a secondary file containing abundance data for downstream analyses.
v) To monitor the performance and capture efficiency of our dual-transcriptomics approach, read distribution statistics were calculated at each step of this pipeline. Specifically, we tracked the total number of raw reads, the reads retained after fastp quality control (QC reads), the proportion of host-derived sequences successfully mapped to the human reference genomes via STAR, and the remaining unmapped fraction. Finally, we quantified the microbial content by tallying the reads successfully classified by Kraken2. The complete read statistics at each processing stage were tabulated for the entire cohort and are visually represented in Supplementary Fig. S2A–D.

Decontamination in RNA data

To address potential contamination in the dataset, a decontamination process was performed using the *decontam* package in R [54]. Contaminants were identified using a frequency-based method, which evaluates the relationship between taxon abundance and DNA concentration. For samples with non-numeric or "too low" values in the DNA concentration measurements (Qubit quantification), these were adjusted by replacing "too low" with a minimal value of 0.01 ng/µl to allow for proper analysis. RNA extraction batches were specified as a batch variable to account for potential differences in processing. Species identified as contaminants were excluded from the dataset. Finally, the dataset was normalized to ensure proportional representation of taxa, and only non-contaminant species were retained for downstream analyses.

Decontamination in DNA data

To address potential contamination from negative controls in DNA samples, a normalization procedure aimed to remove background signals introduced by contaminants in negative control samples, ensuring that only biologically relevant signals were retained for downstream analyses. The dataset was divided into two subsets based on two different sequencing libraries. For both subsets, the values in each sample column were adjusted by subtracting the corresponding negative control values. Specifically, for each sample column, the value of the negative control column was subtracted, and any resulting negative values were set to zero. This ensured that no artificially negative values were introduced during the normalization process. Rows with zero total abundance across all samples were excluded, ensuring that only species with meaningful signals remained.

Filtering

Datasets are filtered in order to maintain only species above the 0.1% in at least 1 sample. In the NMDS showing comparison between two techniques, no filtering on abundance was performed.

Clustering

Clustering was performed on unscaled data, using relative abundances as inputs, to intentionally prioritize the contribution of dominant bacteria and prevent the "double-zero" effect (shared absences among rare taxa) from confounding the analysis. Dissimilarity was calculated using the Euclidean distance metric. To ensure that the identified clusters were not an artifact of this metric, the sample grouping was independently validated using Bray-Curtis dissimilarities (Fig. S3B, E). Hierarchical clustering was then carried out using Ward's minimum variance method (ward.D2), minimizing the total within-cluster variance at each step of the clustering process. A dendrogram visualization was subsequently generated; to enhance interpretability, the branches of the dendrogram were color-coded to represent the four clusters.

Calculation of bacterial vaginosis (BV) dysbiosis score

To assess the pathogenic potential of the microbiome, we established a reference dysbiotic environment based specifically on bacterial vaginosis. We queried the MicroPhenoDB database to assign an association score to each identified taxon, reflecting its known link to BV. A cumulative dysbiosis score was then computed for each sample by summing these association scores, weighted by the relative abundance of the corresponding taxa. A higher score indicates a microbial profile that more closely resembles a BV-associated community.

NMDS

Non-metric multidimensional scaling (NMDS) was used to explore compositional differences among samples in different studies. Batch effects (our study vs other studies) were corrected using the ComBat function from the sva package. Bray-Curtis dissimilarities were calculated to quantify pairwise differences between samples. NMDS ordination was performed with a maximum of 100 iterations to ensure convergence.

Samples were plotted based on their NMDS1 and NMDS2 coordinates, with points color-coded by sampling site and shaped by batch. Ellipses representing 95% confidence intervals for each site were overlaid on the plots to highlight clustering patterns among groups. To provide additional context, marginal density plots along each axis were included showing sample distributions along NMDS1 and NMDS2.

PCoA

Principal coordinate analysis (PCoA) was performed to evaluate compositional dissimilarities between samples coming from either RNA or DNA sequencing. To quantify the dissimilarities between samples, the Bray distance metric was used. To address potential negative eigenvalues, the Cailliez correction was applied during the ordination process.

Shannon index alpha diversity calculation

First, the minimum and maximum sequencing depths were determined across the filtered dataset. To normalize the library sizes, samples were rarefied to the minimum sequencing depth using the "rrarefy" function from the vegan R package. Shannon diversity indices were then computed on the sub-sampled community matrix using the diversity function.

Correlation analysis and network constructions

To identify dependencies within vulvar microbial communities, we constructed correlation networks using Centered Log-Ratio (CLR) transformation to mitigate the spurious correlations inherent in compositional data. Spearman's rank correlation was then applied to these transformed values to capture monotonic dependencies, ensuring the analysis remained robust to the non-normal distributions and outliers common in microbiome datasets. This integrated approach provides a mathematically rigorous framework for detecting biological associations that are independent

of varying sequencing depths and non-linear growth dynamics.

Prior to correlation calculation, we retained microbes displaying at least 0.1% of relative abundance in at least one sample, achieving 84 microbes. The corresponding correlation network was then loaded into R and filtered, allowing only correlation of at least an absolute value of 0.4 and displaying a FDR less than 5%. Networks were then visualized using Cytoscape (https://cytoscape.org/) and annotated using the "microbetag" Cytoscape plugin (https://hariszaf.github.io/microbetag/). Network clustering was performed using the MCODE plugin (https://mcode.readthedocs.io/en/latest/) with parameters (K-core = 2; Node Score Cutoff = 0.2; max depth = 100; Degree CutOff = 2). The host gene-microbiome correlation networks were constructed by first normalizing raw counts, which were normalized and transformed using the normalization method and the *varianceStabilizingTransformation* function of the DESeq2 package. Normalized data were then subject to correlation analysis using the *corAndPvalue* function of the WGCNA R package. Correlations of at least an absolute value of 0.6 and displaying a FDR less than 5% were retained for further analysis. Gene modules and module-trait relationships were obtained using the WGCNA R package. Functional enrichment of gene modules was performed using enrichR (https://maayanlab.cloud/Enrichr/). The resulting network was then exported and visualized using Cytoscape. Interactions between gene modules were inspected by selecting all genes belonging to the specific module and by including the first microbe neighbor of every gene. To evaluate the functional impact of the microbiome on the host, host genes significantly correlated (Spearman's $|R| > 0.5$) with the most abundant microbial taxa were annotated using Gene Ontology Biological Process (GO BP) terms. The relative proportions of positive and negative correlations between individual microbes and specific host epithelial functions were calculated and visualized as a diverging bar plot. Furthermore, a "bridges" analysis was performed to identify host genes exhibiting significant co-regulation with multiple distinct microbial taxa ($n \geq 2$). The correlation profiles of these multi-taxon interacting genes were visualized using a hierarchically clustered heatmap (Ward.D2 method), grouped by their assigned biological functions.

Abbreviations

CST	Community state type
GO	Gene Ontology
HMP	Human Microbiome Project
HSIL	High-grade squamous intraepithelial lesions
LS	Lichen sclerosus
QC	Quality-controlled
VLS-VSCC	Human vulvar lichen sclerosus and vulvar squamous cell carcinoma cell line
WMS	Whole-genome shotgun metagenomic sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-026-02420-x.

Supplementary Material 1.

Acknowledgements
We extend our sincere thanks to all the participants for their contributions to this study. We are grateful to the "*Comitato Vulvodinia e Neuropatia del Pudendo*" association for their support in female pelvic disease research. We would like to thank all the members of the Proserpio Lab, the Donati Lab and the Oliviero Lab for their input and also for the support we received in the last years. We are also grateful to Martino Adamo for his support.

Materials availability
This study did not generate new, unique reagents.

Lead contact
Further information should be directed to Valentina Proserpio (valentina.proserpio@unito.it) and Emanuele Monteleone (emanuele.monteleone@unito.it).

Declaration of generative AI and AI-assisted technologies in the writing process
The authors used ChatGPT or Gemini to improve the readability and language for the original draft. After using these tools, the authors reviewed and edited the content as needed, and they take full responsibility for the content of the publication.

Authors' contributions
V.P., E.M. and I.M. conceptualised the manuscript. E.M. and M.A.C. performed data analysis. A.A., M.A.C., F.L. and L.B. performed the experiments. A.M. collected the samples. V.P., E.M., M.A.C. and F.L. wrote the original draft. V.P. and E.M. supervised the work. V.P. acquired the funding to support the work. All authors reviewed and edited the manuscript.

Funding
Proserpio laboratory is supported by AIRC (MFAG 2023—ID. 29203), PRIN 2022 (COD. 2022CLTAYH), Fondazione Cassa di Risparmio di Torino (CRT 2023 RF = 106089/2023.1841), and Fondazione Compagnia di San Paolo (COD. 2025.0983). Furthermore, financial support for E.M.'s research was provided by Fondazione Umberto Veronesi (FUV). The Proserpio Lab sincerely acknowledges the "***Comitato Vulvodinia e Neuropatia del Pudendo***" association for enabling the donation by Damiano David in support of research on female pelvic diseases. The Donati laboratory is supported by AIRC, Associazione Italiana per la Ricerca sul Cancro (IG 2023—Id. 28831).

Data availability
The high-throughput data generated in this study will be publicly available at the time of publication. The sequencing data discussed in this publication has been deposited in NCBI's Gene Expression Omnibus and is accessible through GEO Series accession number GSE314734 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?&acc=GSE314734) with reviewer access token "ifyteuaojpyrtol" and the European Nucleotide Archive (ENA) repository, PRJEB61325 (https://www.ebi.ac.uk/ena/browser/view/PRJEB61325) [40].
All code used for bioinformatic analysis and visualization is available as follows:
- Project name: vulvarmicrob
- Project home page: https://github.com/marcoamatocianci/vulvarmicrob
- Operating system(s): Platform independent (e.g., Windows, macOS, Linux)
- Programming language: [e.g., R/Python]
- Any restrictions to use by non-academics: None

Declarations

Ethics approval and consent to participate
This study was approved by the Ethical Committee of the University of Turin and from Città della salute di Torino.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Life Sciences and Systems Biology, University of Turin, Turin, Italy. [2]Molecular Biotechnology Centre (MBC) Guido Tarone, University of Turin, Turin, Italy. [3]Department of Sciences of Public Health and Pediatrics, University of Turin, Turin, Italy. [4]Obstetric and Gynecology Unit, Sant'Anna Hospital, Department of Surgical Sciences, University of Turin, Turin, Italy. [5]Centro Salute Pelvi, Turin, Italy. [6]Obstetric and Gynecology Unit, San Matteo Hospital, Department of Pediatric, Surgical, Clinical and Diagnostic Sciences, University of Pavia, Pavia, Italy.

Received: 4 January 2026 Accepted: 1 April 2026
Published online: 07 May 2026

References

1. Elettrico L, Piacenti G, Levra Levron C, Ansai O, Croce A, Duval C, et al. Omics-based decoding of molecular and metabolic crosstalk in the skin barrier ecosystem. Cell Death Differ. 2026; Available from: https://doi.org/10.1038/s41418-025-01648-8
2. Lopez MLD, Lin Y-Y, Sato M, Hsieh C-H, Shiah F-K, Machida RJ. Using metatranscriptomics to estimate the diversity and composition of zooplankton communities. Mol Ecol Resour. 2022;22:638–52.
3. Jia X, Hu L, Wu M, Ling Y, Wang W, Lu H, et al. A streamlined clinical metagenomic sequencing protocol for rapid pathogen identification. Sci Rep. 2021;11:4405.
4. Liu Y-X, Qin Y, Chen T, Lu M, Qian X, Guo X, et al. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell. 2021;12:315–30.
5. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, et al. Evaluating the information content of shallow shotgun metagenomics. mSystems. 2018. https://doi.org/10.1128/mSystems.00069-18.
6. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol. 2014;12:635–45.
7. Rathod D, Silverman JD. PCR bias impacts microbiome ecological analyses. bioRxiv. 2025. https://doi.org/10.1101/2025.07.31.667904.
8. He J, Zhou T, Shen X, Zhang N, Sun C, Lu S, et al. Primer selection impacts the evaluation of microecological patterns in environmental microbiomes. Imeta. 2023;2:e135.
9. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16:410–22.
10. Virtanen S, Saqib S, Kanerva T, Ventin-Holmberg R, Nieminen P, Holster T, et al. Metagenome-validated combined amplicon sequencing and text mining-based annotations for simultaneous profiling of bacteria and fungi: vaginal microbiota and mycobiota in healthy women. Microbiome. 2024;12:273.
11. Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, et al. Sequencing and beyond: integrating molecular "omics" for microbial community profiling. Nat Rev Microbiol. 2015;13:360–72.
12. Zhang Y, Thompson KN, Branck T, Yan Yan, Nguyen LH, Franzosa EA, et al. Metatranscriptomics for the human microbiome and microbial community functional profiling. Annu Rev Biomed Data Sci. 2021;4:279–311.
13. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, et al. Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci U S A. 2014;111:E2329-38.
14. Duran-Pinedo AE, Chen T, Teles R, Starr JR, Wang X, Krishnan K, et al. Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis. ISME J. 2014;8:1659–72.
15. Westermann AJ, Vogel J. Cross-species RNA-seq for deciphering host-microbe interactions. Nat Rev Genet. 2021;22:361–78.
16. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in low microbial biomass microbiome studies: issues and recommendations. Trends Microbiol. 2019;27:105–17.
17. Kennedy KM, de Goffau MC, Perez-Muñoz ME, Arrieta M-C, Bäckhed F, Bork P, et al. Questioning the fetal microbiome illustrates pitfalls of low-biomass microbial studies. Nature. 2023;613:639–49.
18. Mattei V, Murugesan S, Al Hashmi M, Mathew R, James N, Singh P, et al. Evaluation of methods for the extraction of microbial DNA from vaginal swabs used for microbiome studies. Front Cell Infect Microbiol. 2019;9:197.
19. Zhang L, Chen T, Wang Y, Zhang S, Lv Q, Kong D, et al. Comparison analysis of different DNA extraction methods on suitability for long-read metagenomic nanopore sequencing. Front Cell Infect Microbiol. 2022;12:919903.
20. Akpulu C, Lankapalli AK, Toufiq R, Cook K, Portal EAR, Khalid RY, et al. Comparative evaluation of DNA extraction protocols for neonatal gut microbiota profiling in a resource-limited setting. The Microbe. 2025;7:100398.
21. Peng Z, Zhu X, Wang Z, Yan X, Wang G, Tang M, et al. Comparative analysis of sample extraction and library construction for shotgun metagenomics. Bioinform Biol Insights. 2020;14:1177932220915459.
22. Butowski CF, Dixit Y, Reis MM, Mu C. Metatranscriptomics for understanding the microbiome in food and nutrition science. Metabolites. 2025. https://doi.org/10.3390/metabo15030185.
23. He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, et al. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nat Methods. 2010;7:807–12.
24. Han Y, Liu Z, Chen T. Role of vaginal microbiota dysbiosis in gynecological diseases and the potential interventions. Front Microbiol. 2021;12:643422.
25. Zheng Q, Sun T, Li X, Zhu L. Reproductive tract microbiome dysbiosis associated with gynecological diseases. Front Cell Infect Microbiol. 2025;15:1519690.
26. Chen X, Lu Y, Chen T, Li R. The female vaginal microbiome in health and bacterial vaginosis. Front Cell Infect Microbiol. 2021;11:631972.
27. Mieremet A, van der Wurff M, Pagan L, Ferrer-González E, Seo J, Schuren FHJ. Taxonomic and functional profiling of the vulvar microbiome indicates variations related to ecological signatures, aging, and health status. Front Microbiol. 2025;16:1633147.
28. Pagan L, Ederveen RAM, Huisman BW, Schoones JW, Zwittink RD, Schuren FHJ, et al. The human vulvar microbiome: a systematic review. Microorganisms. 2021. https://doi.org/10.3390/microorganisms9122568.
29. Kwon MS, Lee HK. Host and microbiome interplay shapes the vaginal microenvironment. Front Immunol. 2022;13:919728.
30. Avitabile E, Menotti L, Croatti V, Giordani B, Parolin C, Vitali B. Protective mechanisms of vaginal lactobacilli against sexually transmitted viral infections. Int J Mol Sci. 2024. https://doi.org/10.3390/ijms25179168.
31. Tachedjian G, Aldunate M, Bradshaw CS, Cone RA. The role of lactic acid production by probiotic *Lactobacillus* species in vaginal health. Res Microbiol. 2017;168:782–92.
32. Glick VJ, Webber CA, Simmons LE, Martin MC, Ahmad M, Kim CH, et al. Vaginal lactobacilli produce anti-inflammatory β-carboline compounds. Cell Host Microbe. 2024;32:1897-909.e7.
33. Norenhag J, Du J, Olovsson M, Verstraelen H, Engstrand L, Brusselaers N. The vaginal microbiota, human papillomavirus and cervical dysplasia: a systematic review and network meta-analysis. BJOG. 2020;127:171–80.
34. Gudnadottir U, Debelius JW, Du J, Hugerth LW, Danielsson H, Schuppe-Koistinen I, et al. The vaginal microbiome and the risk of preterm birth: a systematic review and network meta-analysis. Sci Rep. 2022;12:7926.
35. Ravel J, Brotman RM, Gajer P, Ma B, Nandy M, Fadrosh DW, et al. Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. Microbiome. 2013;1:29.
36. Auriemma RS, Scairati R, Del Vecchio G, Liccardi A, Verde N, Pirchio R, et al. The vaginal microbiome: a long urogenital colonization throughout woman life. Front Cell Infect Microbiol. 2021;11:686167.
37. Vargas-Robles D, Morales N, Rodríguez I, Nieves T, Godoy-Vitorino F, Alcaraz LD, et al. Changes in the vaginal microbiota across a gradient of urbanization. Sci Rep. 2020;10:12487.
38. Cobo T, Vergara A, Collado MC, Casals-Pascual C, Herreros E, Bosch J, et al. Characterization of vaginal microbiota in women with preterm labor with intra-amniotic inflammation. Sci Rep. 2019;9:18963.

39. Pramanick R, Mayadeo N, Warke H, Begum S, Aich P, Aranha C. Vaginal microbiota of asymptomatic bacterial vaginosis and vulvovaginal candidiasis: are they different from normal microbiota? Microb Pathog. 2019;134:103599.
40. Pagan L, Huisman BW, van der Wurff M, Naafs RGC, Schuren FHJ, Sanders IMJG, et al. The vulvar microbiome in lichen sclerosus and high-grade intraepithelial lesions. Front Microbiol. 2023;14:1264768.
41. VanEvery H, Franzosa EA, Nguyen LH, Huttenhower C. Microbiome epidemiology and association studies in human health. Nat Rev Genet. 2023;24:109–24.
42. Gu T, Tan S, Gou X, Araki H, Tian D. Avoidance of long mononucleotide repeats in codon pair usage. Genetics. 2010;186:1077–84.
43. Nudler E, Gusarov I. Analysis of the intrinsic transcription termination mechanism and its control. Methods Enzymol. 2003;371:369–82.
44. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, et al. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A. 2011;108(Suppl 1):4680–7.
45. France MT, Ma B, Gajer P, Brown S, Humphrys MS, Holm JB, et al. VALENCIA: a nearest centroid classification method for vaginal microbial communities based on composition. Microbiome. 2020;8:166.
46. Krog MC, Madsen ME, Bliddal S, Bashir Z, Vexø LE, Hartwell D, et al. The microbiome in reproductive health: protocol for a systems biology approach using a prospective, observational study design. Hum Reprod Open. 2022;2022:hoac015.
47. Rauer L, De Tomassi A, Müller CL, Hülpüsch C, Traidl-Hoffmann C, Reiger M, et al. De-biasing microbiome sequencing data: bacterial morphology-based correction of extraction bias and correlates of chimera formation. Microbiome. 2025;13:38.
48. Liu P, Lu Y, Li R, Chen X. Use of probiotic lactobacilli in the treatment of vaginal infections: and investigations. Front Cell Infect Microbiol. 2023;13:1153894.
49. Sahal G, Donmez HG, Beksac MS. Cervicovaginal *Bacillus velezensis* isolate: a potential probiotic and an antagonist against *Candida* and *Staphylococcus*. Curr Microbiol. 2023;80:332.
50. Rahman MM, McFadden G. Modulation of NF-κB signalling by microbial pathogens. Nat Rev Microbiol. 2011;9:291–306.
51. Wang G, Sweren E, Liu H, Wier E, Alphonse MP, Chen R, et al. Bacteria induce skin regeneration via IL-1β signaling. Cell Host Microbe. 2021;29:777-91.e6.
52. Dai Z, Liu W-C, Chen X-Y, Wang X, Li J-L, Zhang X. Gasdermin D-mediated pyroptosis: mechanisms, diseases, and inhibitors. Front Immunol. 2023;14:1178662.
53. Balasubramanian A, Hsu AY, Ghimire L, Tahir M, Devant P, Fontana P, et al. The palmitoylation of gasdermin D directs its membrane translocation and pore formation during pyroptosis. Sci Immunol. 2024;9:eadn1452.
54. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. 2018;6:226.
55. Landolt EF, da Conceição Mendonça J, Behler AE, Lumsdaine SW, Jafar T, Burcham LR. Exploring the vaginal ecosystem: insights into host-microbe interactions and microbial community dynamics. Infect Immun. 2025;93:e0049924.
56. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The long-term stability of the human gut microbiota. Science. 2013;341:1237439.
57. Zeevi D, Korem T, Godneva A, Bar N, Kurilshikov A, Lotan-Pompan M, et al. Structural variation in the gut microbiome associates with host health. Nature. 2019;568:43–8.

Publisher's Note