CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

| | |
|---|---|
| Title | Strain-level variation among vaginal Lactobacillus crispatus and Lactobacillus iners as identified by comparative metagenomics |
| Authors | Nori, Sai Ravi Chandra;Walsh, Calum J.;McAuliffe, Fionnuala M.;Moore, Rebecca L.;Van Sinderen, Douwe;Feehily, Conor;Cotter, Paul D. |
| Publication date | 2025-01-01 |
| Original Citation | Nori, S. R. C., Walsh, C. J., McAuliffe, F. M., Moore, R. L., Van Sinderen, Douwe, Feehily, C. and Cotter, Paul D. (2025) 'Strain-level variation among vaginal Lactobacillus crispatus and Lactobacillus iners as identified by comparative metagenomics', npj Biofilms and Microbiomes, 11(1), p.49. https://doi.org/10.1038/s41522-025-00682-1 |
| Type of publication | Article (peer reviewed) |
| Link to publisher's version | 10.1038/s41522-025-00682-1 |
| Rights | © 2025, the Author(s). Open Access. - https://creativecommons.org/licenses/by-nc-nd/4.0/ |
| Download date | 2026-11-15 13:56:32 |
| Item downloaded from | https://hdl.handle.net/10468/17268 |

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

npj | biofilms and microbiomes

Article

Published in partnership with Nanyang Technological University

https://doi.org/10.1038/s41522-025-00682-1

# Strain-level variation among vaginal *Lactobacillus crispatus* and *Lactobacillus iners* as identified by comparative metagenomics

Check for updates

Sai Ravi Chandra Nori [1,2,3,4], Calum J. Walsh[5], Fionnuala M. McAuliffe[6], Rebecca L. Moore[6], Douwe Van Sinderen [2,3], Conor Feehily[7] ✉ & Paul D. Cotter [1,2,3] ✉

The vaginal microbiome, a relatively simple, low diversity ecosystem crucial for female health, is often dominated by *Lactobacillus* spp. Detailed strain-level data, facilitated by shotgun sequencing, can provide a greater understanding of the mechanisms of colonization and host-microbe interactions. We analysed 354 vaginal metagenomes from pregnant women in Ireland to investigate metagenomic community state types and strain-level variation, focusing on cell surface interfaces. Our analysis revealed multiple subspecies, with *Lactobacillus crispatus* and *Lactobacillus iners* being the most dominant. We found genes, including putative mucin-binding genes, distinct to *L. crispatus* subspecies. Using 337 metagenome-assembled genomes, we observed a higher number of strain-specific genes in *L. crispatus* related to cell wall biogenesis, carbohydrate and amino acid metabolism, many under positive selection. A cell surface glycan gene cluster was predominantly found in *L. crispatus* but absent in *L. iners* and *Gardnerella vaginalis*. These findings highlight strain-specific factors associated with colonisation and host-microbe interactions.

Lactobacilli play a major role in modulating the composition of the human vaginal microbiome. Production of lactic acid by these bacteria is an important trait that results in a lowering of the vaginal pH, preventing colonisation by unfavourable bacteria like those associated with bacterial vaginosis (BV)[1–3]. Additionally, production of hydrogen peroxide[4] and antimicrobial peptides can provide a competitive advantage to certain lactobacilli within this niche[5,6]. The dominance of lactobacilli within the vaginal microbiome is distinct to humans, not being observed in other non-human mammals[7]. The evolutionary pressure driving this phenomenon is unclear, however both dietary and hormonal cycling have been suggested[7,8]. Indeed, the high concentration of glycogen within the vaginal lumen, as well as an abundance of mucin, may be important as *Lactobacillus* species have the capacity to use the former glycan as a carbon source, while mucin is believed to provide a distinct interface for interaction between vaginal microbes and the host[9].

The vaginal microbiome in collective datasets has previously been categorised into community state types (CSTs)[10] primarily based on the dominant bacteria present, such as *Lactobacillus crispatus* (CST-I), *Lactobacillus gasseri* (CST-II), *Lactobacillus iners* (CST-III), and *Lactobacillus jensenii* (CST-V), with a further *Lactobacillus*-reduced community of obligate and facultative anaerobes such as *Gardnerella vaginalis* and *Fannyhessea vaginae* classified as CST-IV[10,11]. Due to the diversity of the dominant species, subgroupings or extensions of CSTs are often required, leading to discordance between reported studies. To address this, a nearest centroid based approach (VALENCIA) was developed to help standardise classification[12]. Ultimately, these stratification approaches are based on taxonomic abundance, whereby taxonomy is often assigned based on limited marker genes (as is the case for MetaPhlAN4)[13]. Therefore, the functional potential of the bacterial communities within the vaginal microbiome is not considered, nor are discrete genetic differences within species.

[1]Teagasc Food Research Centre, Fermoy, Co, Cork, Ireland. [2]APC Microbiome Ireland, National University of Ireland, Cork, Ireland. [3]School of Microbiology, University College Cork, Cork, Ireland. [4]SFI Centre for Research Training in Genomics Data Science, School of Mathematics, Statistics & Applied Mathematics, University of Galway, Galway, Ireland. [5]The Centre for Pathogen Genomics, Department of Microbiology & Immunology, Peter Doherty Institute for Infection & Immunity, University of Melbourne, Melbourne, Australia. [6]UCD Perinatal Research Centre, School of Medicine, University College Dublin, National Maternity Hospital, Dublin, Ireland. [7]School of Infection and Immunity, University of Glasgow, Glasgow, G12 8TA, United Kingdom. ✉e-mail: conor.feehily@glasgow.ac.uk; paul.cotter@teagasc.ie

Recently, Holm and colleagues described metagenomic community state types (mgCSTs) based on both taxonomic and functional composition of the vaginal microbiomes from North American women using the non-redundant gene database, VIRGO[14,15]. Assignment of a microbiome to any mgCST is based on the relative abundances of metagenomic subspecies (mgSs). These mgSs are derived from distinct co-occurring genetic variation within a given species and thus provides more granular resolution of a community structure. Based on this approach a total of 27 mgCSTs are currently recognized for the vaginal microbiome. The mgCST-based approach enables a more comprehensive classification of the vaginal microbiome by integrating taxonomic information and functional potential that are crucial to appreciate subtle differences in the communities. For example, the authors define six distinct *L. crispatus* mgCSTs using the mgCST classification scheme, compared to just one CST using the classical CST method.

Multiple factors including microbial interactions, antibiotic treatment, nutrient availability, hormonal changes, and pregnancy can alter the composition of the vaginal microbiome by acting as selective pressures that further affect microdiversity[16–22]. Previous studies have shown that the bacterial communities of the vagina are stable during pregnancy compared to non-pregnant women[11,16]. However, strain-level studies of the vaginal microbiome are needed to understand the subtle genomic variations in the context of pregnancy. Within other niches, analysis of metagenome assembled genomes (MAGs) has been shown to be a powerful approach with respect to describing community function[23,24]. Given the increasing evidence of the impact of subtle strain-level differences on phenotypic characteristics, it is important to investigate genotypes, especially those potentially involved in host-microbe interactions, at a resolution greater than genus and species[9,25,26]. Notably in this regard, differences with respect to mucin binding genes and vaginal cell adhesion have been reported between strains of *L. gasseri*[27], while Argentini et al. have demonstrated that even closely related strains of *L. crispatus* can vary greatly in terms of antimicrobial and competitive abilities[28]. Taken together, there is much to learn about the microbial interactions in this niche by employing sub-species/strain resolved approaches.

The goal of this study was to perform a high-resolution profiling of vaginal metagenomes from pregnant women by analyses of mgCSTs and MAGs in order to understand observed genetic differences at both mgCST- and individual strain-level for genes predicted to be involved in adaptation to the vaginal niche.

## Results

### Metagenomic community state typing reveals multiple sub-species

A total of 354 vaginal metagenomic samples from two Irish cohorts, high-risk preterm ($n = 87$) (PRJEB34536) and MicrobeMom ($n = 267$) (PRJEB48251), with a median count of 336,878 (interquartile range of 176,702–796,475 reads) quality trimmed microbial read pairs per sample, were analysed. Using the metagenomic community state type (mgCST) classifier approach, we grouped the metagenomes at subspecies resolution and established that 18 distinct mgCSTs were present across the dataset. A total of five metagenomic subspecies (mgSs) from *Lactobacillus crispatus* (mgCST 1, mgCST 3-6) (Fig. 1a) were identified, while three and two metagenomic subspecies (mgSs) were shown to correspond to *Lactobacillus iners* (mgCST 10-12 and 27) and *Lactobacillus jensenii* (mgCST 15 and 16), respectively. There were three mgSs identified for *Lactobacillus gasseri* (mgCST 7-9) as well as for *Gardnerella vaginalis* (mgCST 23-25), while only one subspecies was identified for *Bifidobacterium breve* (mgCST 26). In contrast, the conventional CST assignment approach, classified the vaginal metagenomes to only 6 CSTs (CST-1,2,3,4,5 and 8) at species level resolution (Supplementary Table 1a). To investigate the dynamics and stability of mgCST throughout pregnancy, we compared assignment at both the 2$^{nd}$ and 3$^{rd}$ trimester in a subset of individuals ($n = 48$). Our analysis revealed that 27% (13/48) of the samples exhibited intra-species mgCST changes (mgCST shifts with the same dominant taxon) compared to the initial sample. We found inter-species mgCST in 17% (8/48) of the samples, indicating shifts to a different dominant taxon. No differences were observed in 56% of the samples (27/48) (Supplementary Table 1b). Furthermore, we investigated the differences in pan-mgCST gene content between the trimesters using paired samples and no major gene content differences were observed (Supplementary Fig. 1).

### Pan-metagenomics reveal mucin-binding genes in *L. crispatus*

Metagenomic community state types dominated by *L. crispatus* (48% of samples (173/354)) and *L. iners* (13% of samples (47/354)) accounted for the majority of all profiled metagenomes (Table 1). We therefore reconstructed pan-metagenomes of *L. crispatus* and *L. iners* using non-redundant genes of each species to understand differences in mgCSTs between and within these species (Fig. 1b, c). The *L. crispatus* pan-metagenome consisted of 5943 vaginal orthologous groups (VOGs), with 4184 VOGs in the *L. iners* pan-metagenome. Both D- and L- lactate dehydrogenase-encoding genes were present in all *L. crispatus* mgCSTs (5/5 mgCSTs, 173/173 samples) while, *L. iners* mgCSTs lacked the gene encoding a D-lactate dehydrogenase (Supplementary Table 1c and 1d). Similarly, whilst the glycogen debranching gene (*pulA*) was found in all mgCSTs in both *L. crispatus* (5/5 mgCSTs, 168/173 samples) and *L. iners* (2/2 mgCSTs, 47/47 samples), mucin binding genes (*mucBP*) were only present in *L. crispatus* mgCSTs (5/5 mgCSTs, 172/173 samples) (Supplementary Table 1c and 1d).

Within the pan-metagenomes, we identified a total of 405 genes unique to individual *L. crispatus* mgCSTs, with mgCSTs 1 and 6 having the largest number of unique genes, 199/405 and 113/405, respectively (Supplementary Table 1e). Most of these genes were annotated as 'hypothetical' for mgCST 1 (167/199) and mgCST 6 (53/113). Nonetheless, of those that could be functionally annotated, most of the unique genes assignable to COG functional categories were predicted to have roles in transcription, carbohydrate transport and metabolism, replication and repair and cell wall-related activities (Supplementary Fig. 2a). A total of 462 unique genes were identified in *L. iners* mgCSTs, corresponding to mgCST 10 (273), mgCST 11 (53) and mgCST 12 (136) (Supplementary Table 1f). Similar to *L. crispatus*, the majority of these genes, i.e., 194 genes in mgCST 10, 45 genes in mgCST 11 and 114 genes in mgCST 12, had no assigned function (Supplementary Table 1f). The top COG categories in *L. iners* were comparable to *L. crispatus* except for the defense mechanism category where *L. iners* has more mgCST-specific genes (Supplementary Fig. 2b).

### The accessory genome shapes the phylogenetic diversity of both *L. crispatus* and *L. iners*

To investigate the vaginal microbiome at strain-level resolution, we reconstructed 337 high-quality metagenome-assembled genomes (MAGs; completeness >90%, contamination <5%) from 354 metagenomes. *L. crispatus* (154), *L. iners* (57), and *L. jensenii* (31) represented the most frequently reconstructed species (Supplementary Table 2). We found that ~100,000 *L. crispatus* reads were required to recover a high-quality MAG with only ~50,000 reads required for *L. iners* MAG recovery (Fig. 2b, Supplementary Table 2). The median recovery was 1 MAG per sample with a maximum recovery of 8 MAGs in a single sample, reflecting the distribution of species-level diversity expected from the vaginal microbiome. There were no instances where multiple strains/MAGs of the same species were recovered from a single sample.

High-quality MAGs from *L. crispatus* and *L. iners* were used for phylogenetic analysis. Core genome (genes present in 95-100% of MAGs of that species) phylogenetic comparison revealed the presence of three major clades for *L. crispatus* whilst there were two major clades for *L. iners* (Fig. 2c and d). There did not appear to be any relationship between mgCST classification and phylogenetic clades of the core genomes for either *L. crispatus or L. iners*. Furthermore, we reconstructed accessory genome (genes present in <15% of MAGs) phylogenies for *L. crispatus* and *L. iners* to investigate the concordance with the core phylogeny. This analysis revealed a topological discordance between core and accessory genome phylogenies in both *L. crispatus* and *L. iners* with a robinsonfould-distance of 0.58 and 0.70,

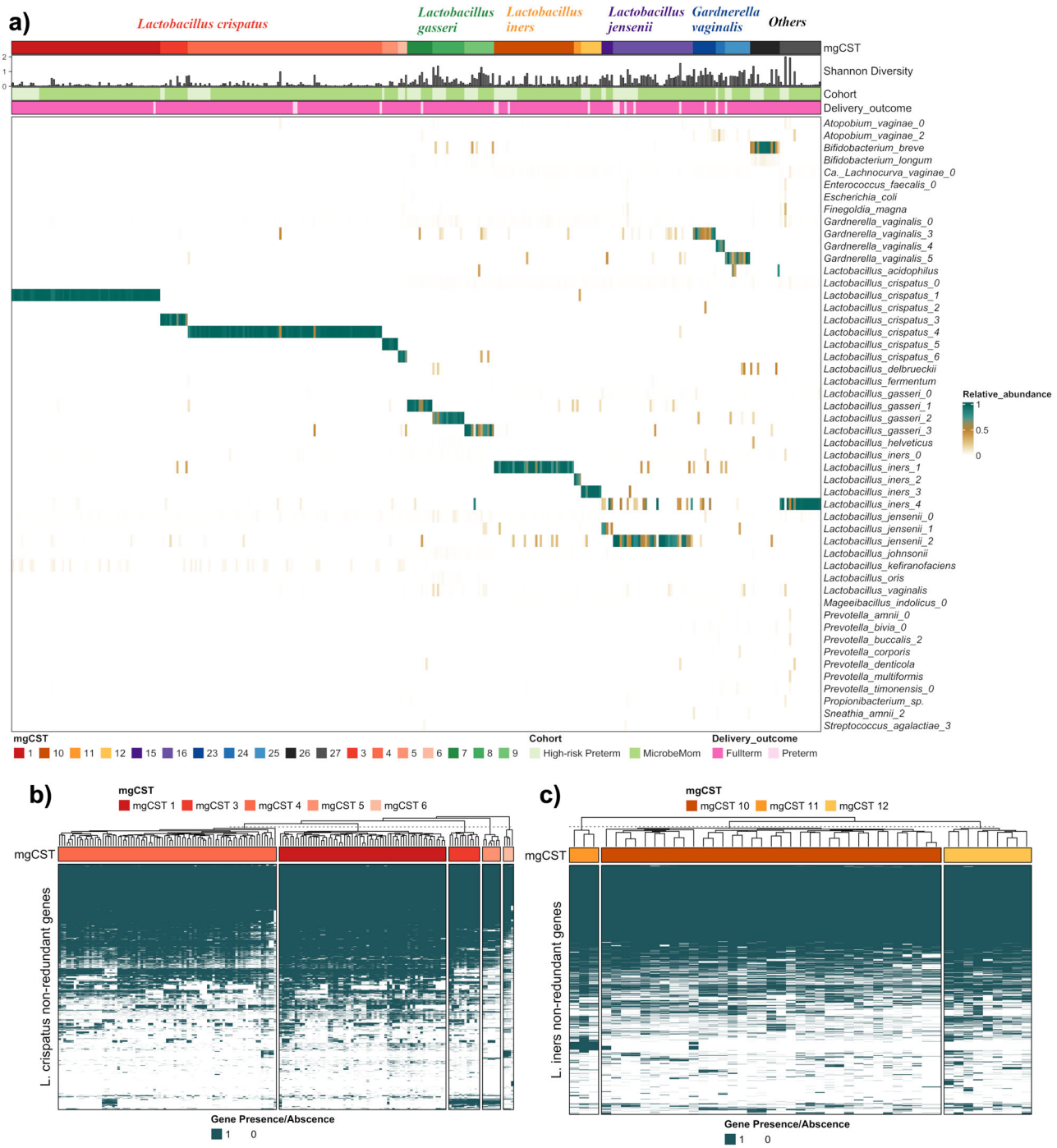



**Fig. 1 | Metagenomic community state types (mgCSTs) identified in 354 vaginal metagenomes. a** Heatmap showing the relative abundance of metagenomic subspecies in each sample. The barplot on the top represents the alpha diversity within the samples and color strip with gradients indicates the mgCSTs in each species. **b**, **c** Pan-metagenome heatmaps of *L. crispatus* and *L. iners* representing the gene content within each mgCSTs. Heatmaps are clustered based on mgCSTs and indicated by color strip on the top. For each non-redundant gene, a dark green cell indicates its presence (1), and a white cell (0) indicates its absence in the corresponding sample.

respectively, indicating accessory genome content might play a role in shaping phylogenetic diversity within these species (Fig. 3a and b).

### The intra-species genetic diversity of vaginal *L. crispatus* is greater than that of *L. iners*

We reconstructed the pangenomes for *L. crispatus* and *L. iners*, as these species are supported by at least 50 high-quality metagenome-assembled genomes (MAGs) each to understand the strain-level differences between the clades (Fig. 4a, b). The pangenomes of *L. crispatus* and *L. iners* contained a total of 8831 and 3221 gene families, respectively. Within *L. crispatus*, 1109 gene families (12.5% of pangenome) were present in more than 95% of the MAGs (core genome) while 5415 gene families (61.3%) were present in less than 15% (cloud genome). For *L. iners*, the core genome comprised of 835 gene families (25.9%) and had a cloud genome of 1805 (56%). Additionally, we randomly sampled three sets of 57 *L. crispatus* MAGs to match *L. iners* sample size and reconstructed the pangenome and found cloud gene content (mean count of 3322 genes - 53% of pangenome) was always larger than the core (mean count of 1116 genes - 18%) in the *L. crispatus* pangenome

**Table 1 | Metagenomic community state types identified in the dataset**

| mgCST | Frequent mgSs (prevalence) | Abundant mgSs | Samples |
|---|---|---|---|
| mgCST 4 | *Lactobacillus crispatus 4* | *Lactobacillus crispatus 4* | 85 |
| mgCST 1 | *Lactobacillus crispatus 1* | *Lactobacillus crispatus 1* | 65 |
| mgCST 10 | *Lactobacillus iners 1* | *Lactobacillus iners 1* | 35 |
| mgCST 16 | *Lactobacillus jensenii 2* | *Lactobacillus jensenii 2* | 35 |
| mgCST 27 | *Bifidobacterium dentium* | *Enterococcus faecalis 3* | 18 |
| mgCST 8 | *Lactobacillus gasseri 2* | *Lactobacillus gasseri 2* | 14 |
| mgCST 26 | *Bifidobacterium breve* | *Bifidobacterium breve* | 13 |
| mgCST 9 | *Lactobacillus gasseri 3* | *Lactobacillus gasseri 3* | 13 |
| mgCST 3 | *Lactobacillus crispatus 3* | *Lactobacillus crispatus 3* | 12 |
| mgCST 25 | *Gardnerella vaginalis 5* | *Gardnerella vaginalis 5* | 11 |
| mgCST 7 | *Lactobacillus gasseri 1* | *Lactobacillus gasseri 1* | 11 |
| mgCST 23 | *Gardnerella vaginalis 3* | *Gardnerella vaginalis 3* | 10 |
| mgCST 12 | *Lactobacillus iners 3* | *Lactobacillus iners 3* | 9 |
| mgCST 5 | *Lactobacillus crispatus 5* | *Lactobacillus crispatus 5* | 7 |
| mgCST 15 | *Lactobacillus jensenii 1* | *Lactobacillus jensenii 1* | 5 |
| mgCST 24 | *Gardnerella vaginalis 4* | *Gardnerella vaginalis 4* | 4 |
| mgCST 6 | *Lactobacillus crispatus 6* | *Lactobacillus crispatus 6* | 4 |
| mgCST 11 | *Lactobacillus iners 2* | *Lactobacillus iners 2* | 3 |

(Supplementary Fig. 3a). Further analysis revealed an open pangenome ($\gamma > 0$) for both *L. crispatus* ($\gamma = 0.25$) and *L. iners* ($\gamma = 0.26$) suggesting high intra-species genetic diversity.

Next, we examined the distribution profiles of core and cloud genes in *L. crispatus* and *L. iners* in various functional categories. Genes putatively involved in cell wall biogenesis (Cloud - 184 vs Core - 35), carbohydrate (160 vs 34) and amino acid transport, and metabolism (151 vs 34) appeared to be more abundant in the cloud of *L. crispatus* compared to the core (Supplementary Fig. 3b). In contrast, we found a similar number of cloud and core genes in the *L. iners* pangenome associated with cell wall biogenesis (Cloud – 30 vs Core - 22), carbohydrate (36 vs 42) and amino acid transport and metabolism (30 vs 22) categories. Inter-clade analysis within the *L. crispatus* species revealed 16 genes that were specific to clade-1. No other *L. crispatus* clades had unique genes. Of the 16 genes from clade-1, all were annotated as encoding for hypothetical proteins. There were no clade-specific genes identified in *L. iners*.

### Metagenomic strain profiling reveals potential genes involved in host-microbe interface in *L. crispatus*

Given the marked differences in identified strain-specific genes between *L. crispatus* and *L. iners*, we aimed to identify the genes driving this by searching for genes under positive selection in the metagenomic data. We found 28 genes under positive selection in *L. crispatus* and four genes in *L. iners* (Fig. 5a). The most prevalent genes under positive selection in *L. crispatus* were predicted to encode transposons (DDE_Tnps), ATP-binding cassette transporters (ABC transporters), S-layer associated protein (SLAP), Gram positive cell wall anchor (Gram_pos_anchor) and mucin-binding protein (MucBP). These genes were found irrespective of the associated mgCSTs assigned to the source metagenome. In *L. iners*, three genes were related to membrane transport (ABC transporters and bPH_2) and the other is annotated as encoding a domain of unknown function (DUF4355).

Five species (*L. crispatus, L. gasseri, L iners, L. jensenii*, and *G.* vaginalis) provided greater than 20 high quality MAGs. Within these we observed that a putative cell surface glycan encoding gene cluster containing *mucBP* was predominantly present in *L. crispatus* (86%, 113/154) and not in *L. iners* (0/57) or *G. vaginalis* (0/26) (Fig. 5c). In *L. jensenii* and *L. gasseri*, we found this gene cluster in 36% (11/30) and 30% (6/20) of the samples, respectively. Moreover, we found that a lower microbiome alpha diversity in a given sample is correlated with a higher probability of detecting this cell surface glycan gene cluster in *L. crispatus* and *L. jensenii* ($p = 0.005$) species (Fig. 5b).

Finally, we extended this investigation to publicly available genomes of *L. crispatus* ($n = 406$) and *L. iners* ($n = 279$) from different countries and sampling sites to survey the presence of this cell surface gene cluster in a more geographically broader context. We found similar results to our dataset, where the gene cluster was present in 63% (256/406) of *L. crispatus* genomes, yet absent in *L. iners* (0/279) (Supplementary Fig. 4).

## Discussion

Understanding the genetic diversity among strains in a microbial community is often very important[29]. In particular, for putatively beneficial bacteria like lactobacilli, there is a requirement to differentiate specific strains that have greater therapeutic potential from those that do not[30–34]. Within the vaginal tract, it has been widely reported that *L. crispatus* is distinctly associated with better health outcomes[35–38]. As such, the general consensus appears to be to promote a microbiome dominated by this species. This however does not consider diversity within the species, or any other species within the vagina. Approaches to study the vaginal microbiome have more often relied on a metataxonomic approach, which does not provide the resolution required to understand functional differences at strain level[39–41]. Within our metagenomic data we revealed the presence of multiple subspecies of bacteria with a diverse gene repertoire in the vaginal microbiome underscoring the importance of studying functional differences even in compositionally similar data. The presence of mgCST-specific genes in *L. crispatus* and *L. iners* related to cell-wall and defence mechanisms suggests different functions of the strains in the vaginal environment. A previous report has shown that *Limosilactobacillus reuteri* strains with mucus adhesins can exert immunoregulatory effects in the gut[42]. Furthermore, mucus-binding proteins from *Lactiplantibacillus plantarum* have shown inhibition towards enterotoxigenic *E. coli* cells[43]. The presence of mucin-binding genes in metagenomes dominated by *L. crispatus*, but not in *L. iners*, may have similar functional implications in the vaginal microbiome in terms of bacterial adhesion and pathogen resistance.

The genomic diversity among *L. crispatus* strains has recently been reported and reveals that this results in phenotypic variation, particularly with respect to glycogen metabolism, a major carbon source in the vaginal environment that can determine colonization levels and associated persistence, as well as microbe-host communication ability of the strains[9,26]. Our results show that a gene (*pulA*), encoding a putative glycogen degrading activity, pullulanase, is present in all mgCSTs of both *L. crispatus* and *L. iners*, indicating the importance of this enzyme in glycogen metabolism within the vaginal ecosystem. Separately, previous studies have shown that D-lactate is more protective against unfavourable vaginal communities than L-lactate and its levels are highest when *L. crispatus* is dominant[44,45]. The presence of D- and L-lactate dehydrogenase-encoding genes in all *L. crispatus* mgCSTs, but the absence of D-lactate dehydrogenase gene in *L. iners*, is in line with previous literature[46]. However, Holm and colleagues have reported that the D-lactate dehydrogenase gene is missing in mgCST 2, one of the subspecies of *L. crispatus*, and that samples dominated by mgCST 2 were compositionally more diverse and had fewer *L. crispatus* strains compared to other mgCSTs[15]. Interestingly, we did not find any samples with mgCST 2 in our dataset, which contains data from samples collected predominantly from healthy women. Altogether, our analysis of taxonomic composition of the vaginal microbiome, coupled with its functional potential, offers insights that are vital for understanding the mechanisms underlying the maintenance of a healthy vaginal environment.

A recent pangenomic survey of major vaginal lactobacilli showed that genes related to adherence functions in *L. crispatus* and *L. iners* are strain specific[47]. In line with this, we report that a high number of cloud genes compared to core genes were found in cell wall biogenesis, carbohydrate and amino acid metabolism functional categories in *L. crispatus* compared to *L. iners*. Furthermore, our pangenome analysis revealed open-pangenomes indicating high intra-species genetic diversity among vaginal *L. crispatus*

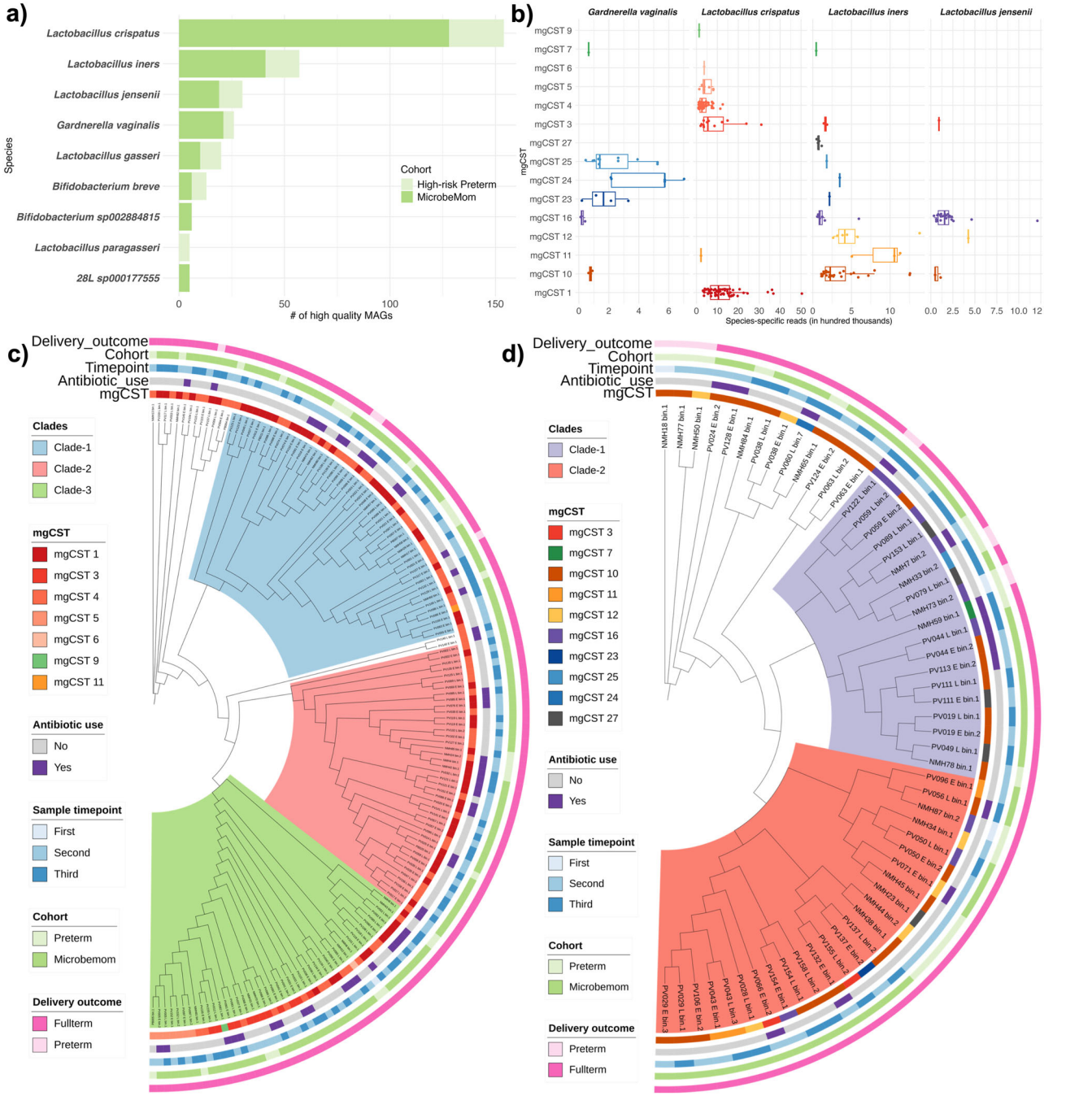



**Fig. 2 | Metagenome-assembled genomes (MAGs) from vaginal metagenomes.** **a** Stacked barplot showing the number of MAGs recovered from each species from the two cohorts. **b** Boxplots are faceted by species and colored by mgCSTs representing the number of species-specific reads needed to recover high-quality MAGs from metagenomic samples. **c**, **d** Maximum likelihood cladograms of *L. crispatus* (**c**) and *L. iners* (**d**) reconstructed based on core genomes. Branches are colored based on the phylogenetic clades and color strips from inner to outermost represent mgCSTs, antibiotic usage, sample collection timepoint, cohort and delivery outcome.

and *L. iners* strains. This strain-specific genetic diversity, especially in critical functional categories, may underpin differences in colonization efficiency, resilience to environmental stresses, and interactions with the host. Interestingly, we observed a higher positive selection signature among genes related to host-microbe interface in *L. crispatus* suggesting a continuous evolutionary pressure for adaptation. Mainly, we noted cell wall anchor genes with YSIRK signal motifs and mucin-binding genes along with transposons and S-layer associated proteins as major genes under positive selection. A recent study has also shown a positive selection signature in *Lactobacillus* adhesin genes in non-pregnancy cohorts suggesting a ubiquitous selection pressure on these genes[48]. Mucin-binding and cell surface genes are of critical importance as they directly interact with the host epithelial cells and immunity systems[49–51].

Our previous genomics work on *Lactobacillus jensenii* highlighted a cell surface gene cluster which harbours identical cell wall anchor genes and mucin-binding genes along with glycosyltransferases (GTs) and gene machinery required for export of proteins to the cell surface[52]. Zeng and colleagues also showed that MucBP-like domains and a similar cell surface gene cluster in a vaginal *Lactobacillus gasseri* strain plays an important role in adhesion to vaginal epithelium through gene knockout experiments[27]. Interestingly, we observed this gene cluster to be predominantly present in *L. crispatus*, both in our dataset and public genomes, but not in *L. iners* and

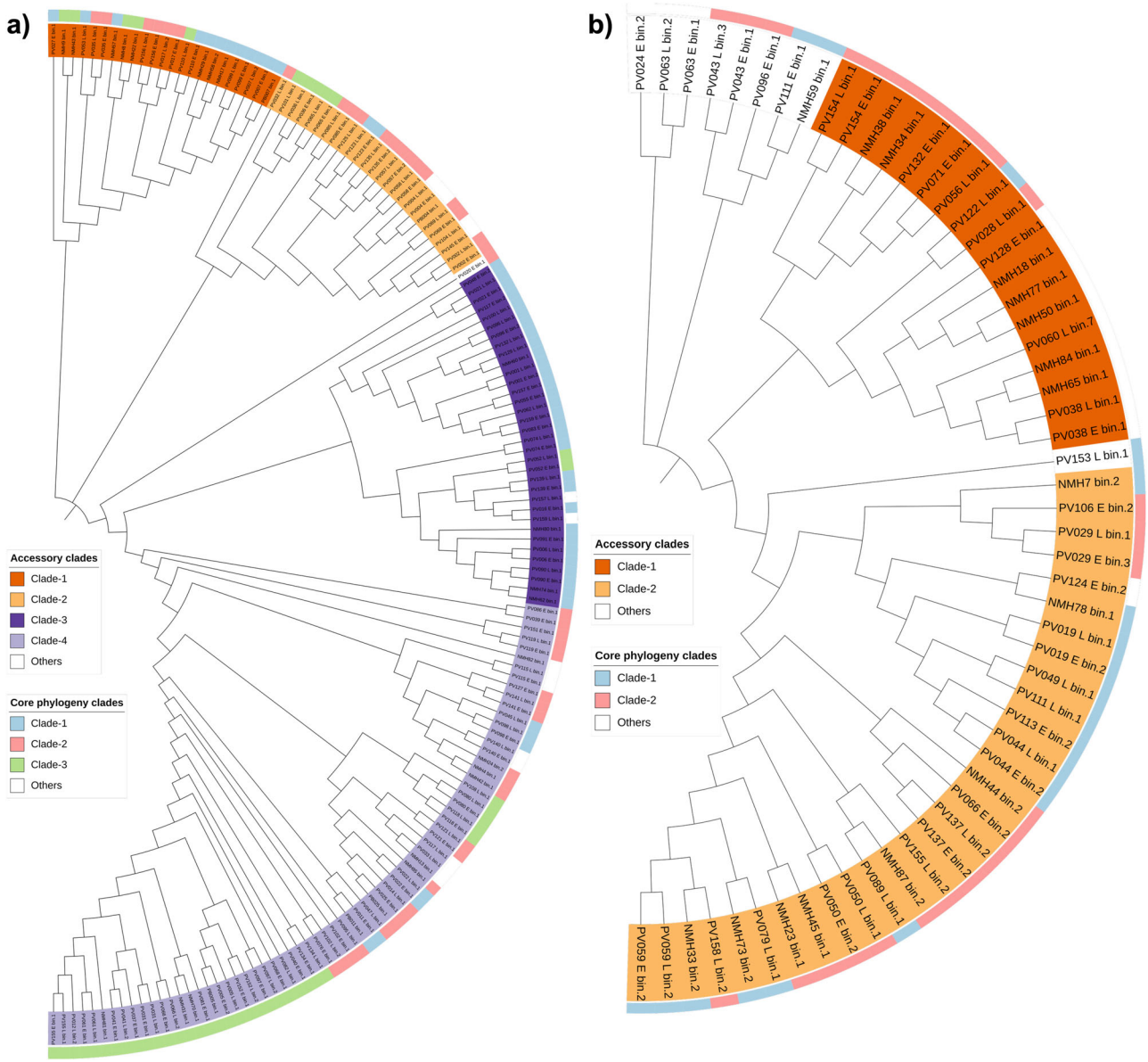



**Fig. 3 | Phylogenetic trees of *L. crispatus* and *L. iners*.** **a**, **b** Maximum likelihood phylogenetic trees of *L. crispatus* (**a**) and *L. iners* (**b**) reconstructed based on accessory genome content. Branches are colored based on the phylogenetic clades and the color strip represents the core genome phylogenetic clades.

*Gardnerella vaginalis*, indicating potential species-specific cell surface adaptations. Furthermore, most of the variations within the gene cluster in *L. crispatus* were found in GTases, suggesting a strain-specific microbial cell surface glycome that could be crucial for interactions with vaginal epithelial cells. Moreover, the gene cluster is mainly detected in samples with low alpha diversity indicating the importance of cell surface adaptations in establishing a dominant and stable colonization. Overall, these observations underscore the significance of the cell surface gene cluster and its associated genes in potential host-microbe interactions.

Taken together, our findings highlight the critical role of *Lactobacillus* strain diversity in vaginal microbial ecology, with a specific emphasis on genes predicted to be involved in glycogen metabolism, mucin-binding, and cell surface adaptations. The observed positive selection signature among host-microbe interface genes and the presence of species-specific cell surface gene clusters in *L. crispatus* underline evolutionary adaptations for the vaginal niche. Whilst all these genomic comparisons reveal distinctions with respect to gene content, it still needs to be determined whether a phenotypic difference is also present within strains. Furthermore, higher sequencing depths are required to detect the subtle genomic variations that might have been missed due to the sequencing depth used in our study. Future RNA-seq experiments both in vitro and direct-from-swabs would provide a valuable insight into the degree of functional variation within the members of the vaginal microbiota. Nonetheless, our study underscores the importance of strain-level genomic analysis but also opens avenues for novel probiotic developments to support the vaginal microbiome. Further studies with a more diverse dataset are needed to fully understand the implications of strain-level variations in the vaginal microbiome and how this may influence measured health outcomes.

## Methods

### Metagenomic community state typing and pan metagenome reconstruction

Raw 354 vaginal metagenomic samples from two Irish pregnancy cohorts, high-risk preterm ($n = 87$) (PRJEB34536) and MicrobeMom ($n = 267$) (PRJEB48251) were used for the analysis[53–55]. Inclusion criteria for both cohorts are previously described (52-54). In brief, samples from the high-risk preterm group were pregnant, >18 years of age, either had previous

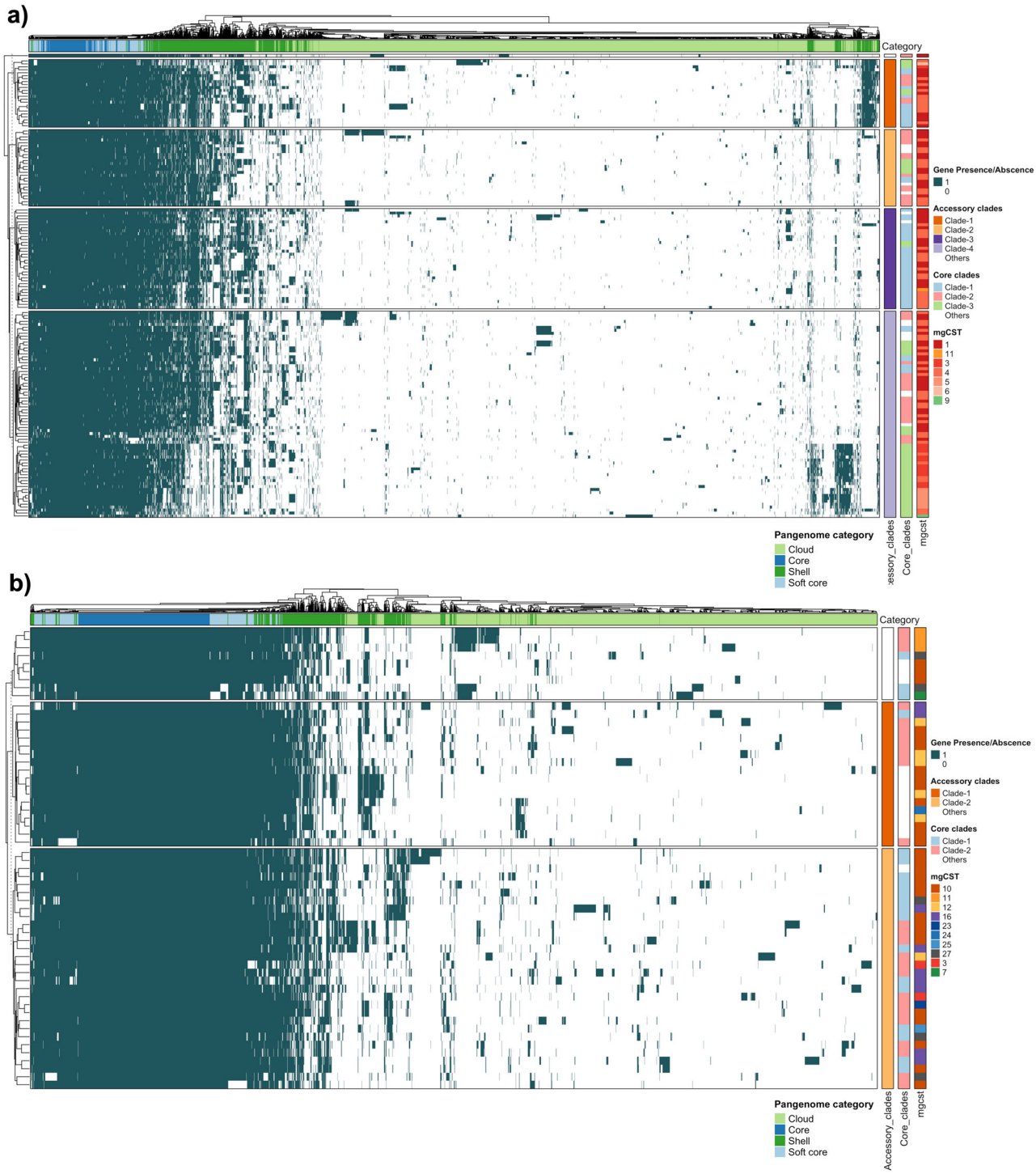



**Fig. 4 | Pangenomes of *L. crispatus* and *L. iners* reconstructed from high-quality MAGs. a**, **b** Pangenome heatmaps of *L. crispatus* (**a**) and *L. iners* (**b**) representing the gene content within each MAG. The heatmaps were clustered based on accessory clades indicated with color strip on the side. For each gene, a dark green cell indicates its presence, and a white cell indicates its absence in the corresponding MAG. The color strip on the top indicates the pangenome category of the corresponding gene and side color strips represents the core genome phylogenetic clades and mgCSTs.

preterm birth or undergone LLETZ surgery, or no known risk factor (as control group), whilst those from the MicrobeMom group were pregnant, >18 years of age, a BMI between 18.5 and 35 kg/m$^2$, no gestational diabetes, and singleton pregnancy. The samples were collected from 240 individual participants at different timepoints throughout pregnancy as detailed in Supplementary Table 1a. 94% (225/240) of the participants delivered full-term while 6% (15/240) were preterm. Quality filtering and host contamination removal was performed using Trim Galore (v 0.6.0) and Hostile (v 0.1.0) respectively[56,57]. Alpha diversity of the samples was calculated using VIRGO output with VEGAN r package[58]. The quality filtered reads were annotated against the VIRGO database using built-in scripts[14]. The annotated files were used as input for metagenomic subspecies identification using mgCST classifier[15]. The non-redundant gene profiles of *Lactobacillus crispatus* and *Lactobacillus iners* dominant samples were used for pan-metagenome reconstruction. The Clusters of Orthologous Genes annotation from VIRGO database were used for identifying functional categories.

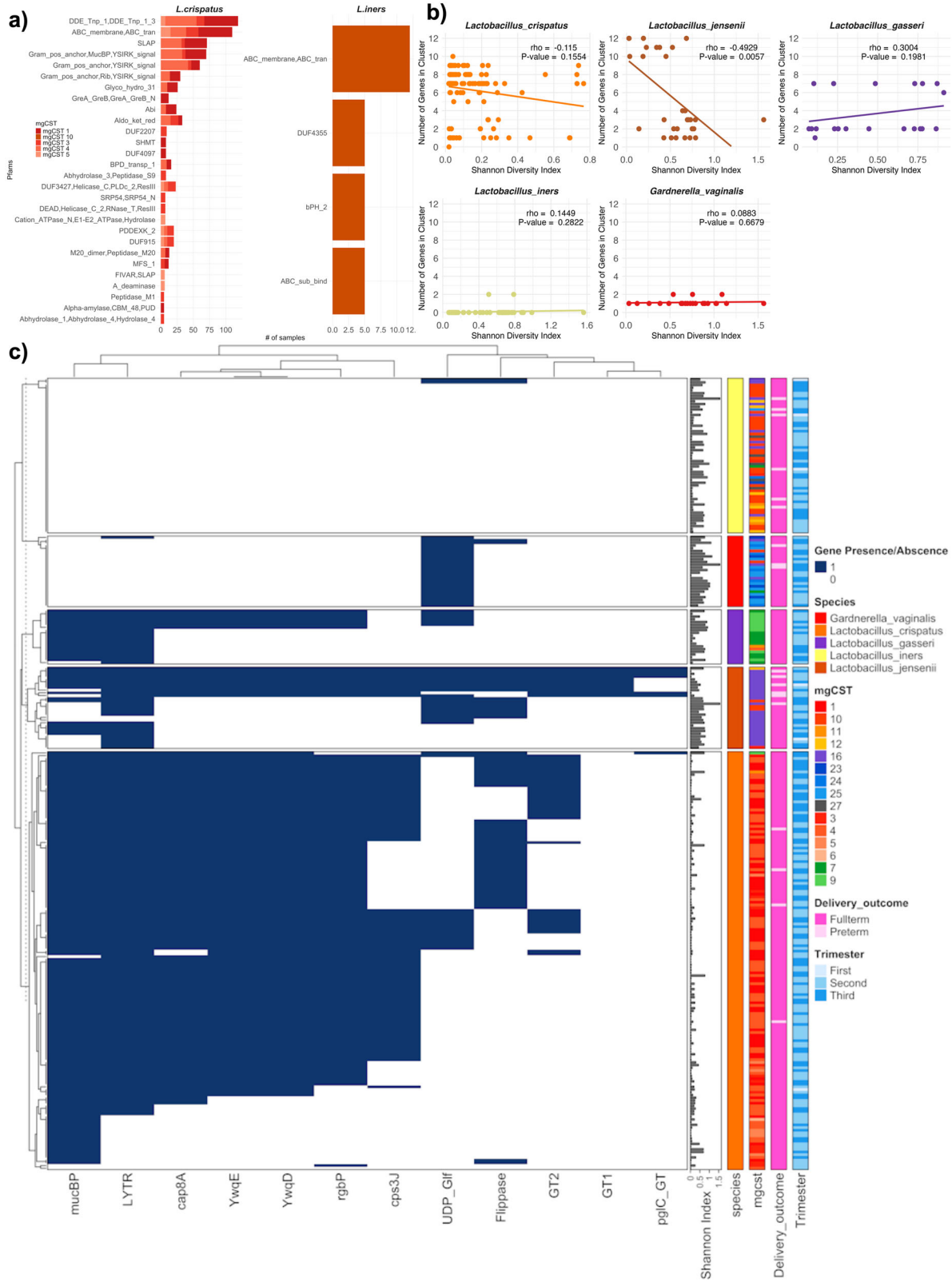



**Fig. 5 | Strain level profiling of major vaginal bacteria. a** Stacked barplots faceted by species and colored by mgCSTs representing the number of genes under positive selection in *L. crispatus* and *L. iners* with Pfams annotation. **b** Scatterplots faceted by species shows the correlation between Shannon diversity index and the number of cell surface glycan cluster genes detected in the MAGs from all the samples, with Spearman's rho and corresponding p-values indicating the strength and significance of each correlation. **c** Heatmap showing the presence and absence of genes from cell surface glycan gene cluster across major vaginal bacteria. The x-axis represents the genes from the gene cluster. For each gene, a blue cell indicates its presence, and a white cell indicates its absence in the corresponding genome. The heatmap was clustered by the similarity of gene content and separated by species indicated by the color strip on the side. The barplot on the side represents the alpha diversity in the metagenomic sample and color strips represent mgCST, delivery outcome and trimester.

### Reconstruction of metagenome assembled genomes

Metagenome-assembled genomes were reconstructed using quality filtered reads with MetaWRAP (v 1.2.1)[59]. The contigs were assembled using assembly module with metaspades option in the MetaWRAP pipeline. The assembled contigs were binned using binning module with MetaBAT2, Maxbin2 and CONCOCT options. The resulting bins were refined using bin_refinement and reassemble_bins module in the MetaWRAP pipeline. The bins were quality checked with CheckM (v 1.0.18) and only high-quality bins with completeness ≥90% and contamination <5% were used for downstream analyses[60]. The taxonomy of the bins was assigned using GTDB-Tk (v 1.5.0)[61]. The number of reads per species needed for MAG recovery was determined based on the number of reads assigned to that species in Kraken2 (v 2.1.1) and Bracken (v 2.2) outputs[62,63].

### Pangenome reconstruction and phylogenetic analysis

The pangenomes for *L. crispatus* and *L. iners* were reconstructed using Roary (v3.12) with '-e –n --mafft' options with PROKKA (v 1.14) annotation as input[64,65]. The pangenome categories assigned by Roary were used to define the core genes (99% <= strains <= 100%), Soft core genes (95% <= strains <99%), Shell genes (15% <= strains <95%) and Cloud genes (0% <= strains <15%). Heap's law was used to determine whether the pangenome is open ($\gamma > 0$) or closed ($\gamma < 0$) with Heap_law_for_roary script (https://github.com/SethCommichaux/Heap_Law_for_Roary). The concatenated core genome sequences from Roary output were aligned with MAFFT (v 7.475)[66]. The accessory binary tree was used for tree reconstruction. Maximum likelihood phylogeny was reconstructed using multiple sequence alignment with IQ-TREE2 (v 2.1.3) with 500 bootstraps[67]. Output trees were visualised using iTOL[68]. The distance between core genome phylogeny and accessory phylogeny was computed using Robinson-Foulds (RF) distance with Phangorn R package[69,70]. The COG functional categories for gene families in the pangenomes of *L. crispatus* and *L. iners* were assigned using eggNOG-mapper (v 2.1.7)[71].

### Strain profiling of MAGs and publicly available genomes

InStrain (v 1.8.0) was used to identify the genes under positive selection[72]. The high-quality MAGs were dereplicated at 98% identity using dRep (v 3.2.0) and representative MAGs database was built using bowtie2 (v 2.4.4)[73,74]. The quality filtered metagenomic reads were mapped against the MAGs database using bowtie2. The resulting alignment files were used for inStrain gene profiling and further identification of genes under positive selection (dn/ds >1). Pfam annotations from the eggNOG-mapper output was used to assign the function of the genes.

The publicly available high-quality vaginal *L. crispatus* ($n = 406$) and *L. iners* ($n = 279$) genomes were downloaded from PATRIC database on 22nd February 2024[75]. Previously identified cell surface gene cluster genes were used to construct a BLASTp (v 2.8.1) database[52,76]. Protein coding sequences were predicted using Prodigal (v 2.6.3) from high-quality MAGs generated from our dataset and publicly available genomes[77]. The resulting sequences were mapped against the database to identify genes from cell surface gene cluster using BLASTp.

## Data availability

The datasets used in this study can be accessed from ENA using accessions PRJEB34536 (high-risk preterm) and PRJEB48251 (MicrobeMom).

Received: 7 May 2024; Accepted: 9 March 2025;
Published online: 23 March 2025

## References

1. Tachedjian, G., Aldunate, M., Bradshaw, C. S. & Cone, R. A. The role of lactic acid production by probiotic *Lactobacillus* species in vaginal health. *Res. Microbiol.* **168**, 782–792 (2017).
2. Gong, Z., Luna, Y., Yu, P. & Fan, H. Lactobacilli Inactivate Chlamydia trachomatis through Lactic Acid but Not H2O2. *PLoS ONE* **9**, e107758 (2014).
3. O'Hanlon, D. E., Moench, T. R. & Cone, R. A. Vaginal pH and microbicidal lactic acid when lactobacilli dominate the microbiota. *PLoS ONE* **8**, e80074 (2013).
4. Miko, E. & Barakonyi, A. The Role of Hydrogen-Peroxide (H2O2) produced by vaginal microbiota in female reproductive health. *Antioxid. (Basel)* **12**, 1055 (2023).
5. Fuochi, V., Cardile, V., Petronio Petronio, G. & Furneri, P. M. Biological properties and production of bacteriocins-like-inhibitory substances by Lactobacillus sp. strains from human vagina. *J. Appl. Microbiol.* **126**, 1541–1550 (2019).
6. Aroutcheva, A. et al. Defense factors of vaginal lactobacilli. *Am. J. Obstet. Gynecol.* **185**, 375–379 (2001).
7. Miller, E. A., Beasley, D. E., Dunn, R. R. & Archie, E. A. Lactobacilli dominance and vaginal pH: why is the human vaginal microbiome unique? *Front. Microbiol.* **7**, 1936 (2016).
8. Fuochi, V., Li Volti, G. & Furneri, P. M. Commentary: Lactobacilli Dominance and Vaginal pH: Why Is the Human Vaginal Microbiome Unique? *Front. Microbiol*. **8**, 1815 (2017).
9. Jenkins, D. J. et al. Bacterial amylases enable glycogen degradation by the vaginal microbiome. *Nat. Microbiol* **8**, 1641–1652 (2023).
10. Ravel, J. et al. Vaginal microbiome of reproductive-age women. *Proc. Natl Acad. Sci.* **108**, 4680–4687 (2011).
11. Srinivasan, S. & Fredricks, D. N. The human vaginal bacterial biota and bacterial vaginosis. *Interdiscip. Perspect. Infect. Dis.* **2008**, 1–22 (2008).
12. France, M. T. et al. VALENCIA: a nearest centroid classification method for vaginal microbial communities based on composition. *Microbiome* **8**, 166 (2020).
13. Blanco-Míguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
14. Ma, B. et al. A comprehensive non-redundant gene catalog reveals extensive within-community intraspecies diversity in the human vagina. *Nat. Commun.* **11**, 940 (2020).
15. Holm, J. B. et al. Integrating compositional and functional content to describe vaginal microbiomes in health and disease. *Microbiome* **11**, 259 (2023).
16. Fettweis, J. M. et al. The vaginal microbiome and preterm birth. *Nat. Med.* **25**, 1012–1021 (2019).
17. Serrano, M. G. et al. Racioethnic diversity in the dynamics of the vaginal microbiome during pregnancy. *Nat. Med.* **25**, 1001–1011 (2019).
18. Liao, J. et al. Microdiversity of the vaginal microbiome is associated with preterm birth. *Nat. Commun.* **14**, 4997 (2023).
19. Li, D. et al. Vaginal microbiome analysis of healthy women during different periods of gestation. *Biosci. Rep.* **40**, BSR20201766 (2020).
20. Krog, M. C. et al. The healthy female microbiome across body sites: effect of hormonal contraceptives and the menstrual cycle. *Hum. Reprod.* **37**, 1525–1543 (2022).
21. Ahrens, P. et al. Changes in the vaginal microbiota following antibiotic treatment for Mycoplasma genitalium, Chlamydia trachomatis and bacterial vaginosis. *PLoS One* **15**, e0236036 (2020).
22. O'Neill, I. J. et al. Maternal and infant factors that shape neonatal gut colonization by bacteria. *Expert Rev. Gastroenterol. Hepatol.* **14**, 651–664 (2020).
23. Saheb Kashaf, S. et al. Integrating cultivation and metagenomics for a multi-kingdom view of skin microbiome diversity and functions. *Nat. Microbiol* **7**, 169–179 (2022).
24. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
25. Hertzberger, R. et al. Genetic elements orchestrating lactobacillus crispatus glycogen metabolism in the vagina. *Int. J. Mol. Sci.* **23**, 5590 (2022).
26. Veer, C. V. D. 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).
27. Zeng, Z., Zuo, F. & Marcotte, H. Putative Adhesion Factors in Vaginal Lactobacillus gasseri DSM 14869: Functional Characterization. *Appl. Environ. Microbiol.* **85**, e00800-19 (2019).
28. Argentini, C. et al. Evaluation of modulatory activities of lactobacillus crispatus strains in the context of the vaginal microbiota. *Microbiol. Spectr.* **10**, e02733-21 (2022).
29. Yan, Y., Nguyen, L. H., Franzosa, E. A. & Huttenhower, C. Strain-level epidemiology of microbial communities and the human microbiome. *Genome Med.* **12**, 71 (2020).
30. Abramov, V. et al. Probiotic properties of lactobacillus crispatus 2,029: homeostatic interaction with cervicovaginal epithelial cells and antagonistic activity to genitourinary pathogens. *Probiotics Antimicro. Prot.* **6**, 165–176 (2014).
31. Ansari, J. M., Colasacco, C., Emmanouil, E., Kohlhepp, S. & Harriott, O. Strain-level diversity of commercial probiotic isolates of Bacillus, Lactobacillus, and Saccharomyces species illustrated by molecular identification and phenotypic profiling. *PLOS ONE* **14**, e0213841 (2019).
32. Azad, M. A. K., Sarker, M., Li, T. & Yin, J. Probiotic Species in the Modulation of Gut Microbiota: An Overview. *BioMed Res. Int.* **2018**, 2314–6141 (2018).
33. Huang, Z. et al. Comparative genomics and specific functional characteristics analysis of lactobacillus acidophilus. *Microorganisms* **9**, 1992 (2021).
34. Mu, Q., Tavella, V. J. & Luo, X. M. Role of Lactobacillus reuteri in Human Health and Diseases. *Front Microbiol* **9**, 757 (2018).
35. Gudnadottir, U. et al. The vaginal microbiome and the risk of preterm birth: a systematic review and network meta-analysis. *Sci. Rep.* **12**, 7926 (2022).
36. Mändar et al. Impact of Lactobacillus crispatus-containing oral and vaginal probiotics on vaginal health: a randomised double-blind placebo controlled clinical trial. *Benef. Microbes* **14**, 143–152 (2023).
37. Armstrong, E. et al. Sustained effect of LACTIN-V (Lactobacillus crispatus CTV-05) on genital immunology following standard bacterial vaginosis treatment: results from a randomised, placebo-controlled trial. *Lancet Microbe* **3**, e435–e442 (2022).
38. Cohen, C. R. et al. Randomized Trial of Lactin-V to prevent recurrence of bacterial vaginosis. *Obstetrical Gynecol. Surv.* **75**, 601 (2020).
39. Lebeer, S. et al. A citizen-science-enabled catalogue of the vaginal microbiome and associated factors. *Nat. Microbiol* **8**, 2183–2195 (2023).
40. Wu, M. et al. Leveraging 16S rRNA data to uncover vaginal microbial signatures in women with cervical cancer. *Front. Cell. Infect. Microbiol.* **13**, 1024723 (2023).
41. Virtanen, S., Kalliala, I., Nieminen, P. & Salonen, A. Comparative analysis of vaginal microbiota sampling using 16S rRNA gene analysis. *PLOS ONE* **12**, e0181477 (2017).
42. Bene, K. P. et al. Lactobacillus reuteri surface mucus adhesins upregulate inflammatory responses through interactions with innate C-type lectin receptors. *Front. Microbiol.* **8**, 321 (2017).
43. Singh, K. S., Kumar, S., Mohanty, A. K., Grover, S. & Kaushik, J. K. Mechanistic insights into the host-microbe interaction and pathogen exclusion mediated by the Mucus-binding protein of Lactobacillus plantarum. *Sci. Rep.* **8**, 14198 (2018).
44. Nasioudis, D. et al. α-amylase in vaginal fluid: association with conditions favorable to dominance of lactobacillus. *Reprod. Sci.* **22**, 1393–1398 (2015).
45. Witkin, S. S. et al. Influence of Vaginal Bacteria and d- and l-Lactic Acid Isomers on Vaginal Extracellular Matrix Metalloproteinase Inducer: Implications for Protection against Upper Genital Tract Infections. *mBio* **4**, https://doi.org/10.1128/mbio.00460-13 (2013).
46. France, M. T., Mendes-Soares, H. & Forney, L. J. Genomic Comparisons of Lactobacillus crispatus and Lactobacillus iners Reveal Potential Ecological Drivers of Community Composition in the Vagina. *Appl. Environ. Microbiol.* **82**, 7063–7073 (2016).
47. Bhattacharya, A., Das, S., Bhattacharjee, M. J., Mukherjee, A. K. & Khan, M. R. Comparative pangenomic analysis of predominant human vaginal lactobacilli strains towards population-specific adaptation: understanding the role in sustaining a balanced and healthy vaginal microenvironment. *BMC Genomics* **24**, 1–15 (2023).
48. Wei, X. et al. Vaginal microbiomes show ethnic evolutionary dynamics and positive selection of Lactobacillusadhesins driven by a long-term niche-specific process. *Cell Rep.* **43**, 114078 (2024).
49. Devi, S. M. & Halami, P. M. Diversity and evolutionary aspects of mucin binding (MucBP) domain repeats among Lactobacillus plantarum group strains through comparative genetic analysis. *Syst. Appl Microbiol* **40**, 237–244 (2017).
50. Dharmani, P., Srivastava, V., Kissoon-Singh, V. & Chadee, K. Role of Intestinal Mucins in Innate Host Defense Mechanisms against Pathogens. *J. Innate Immun.* **1**, 123–135 (2008).
51. Van Tassell, M. L. & Miller, M. J. Lactobacillus Adhesion to Mucus. *Nutrients* **3**, 613–636 (2011).
52. Nori, S. R. C. et al. Profiling of vaginal Lactobacillus jensenii isolated from preterm and full-term pregnancies reveals strain-specific factors relating to host interaction. *Microb. Genomics* **9**, 001137 (2023).
53. Feehily, C. et al. Detailed mapping of Bifidobacterium strain transmission from mother to infant via a dual culture-based and metagenomic approach. *Nat. Commun.* **14**, 3015 (2023).
54. Feehily, C. et al. Shotgun sequencing of the vaginal microbiome reveals both a species and functional potential signature of preterm birth. *npj Biofilms Microbiomes* **6**, 1–9 (2020).
55. Moore, R. L. et al. Ability of Bifidobacterium breve 702258 to transfer from mother to infant: the MicrobeMom randomized controlled trial. *Am. J. Obstet. Gynecol. MFM* **5**, 100994 (2023).
56. Krueger, F. et al. FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. https://doi.org/10.5281/ZENODO.7598955 (2023).
57. Constantinides, B., Hunt, M. & Crook, D. W. Hostile: accurate decontamination of microbial host sequences. *Bioinformatics* **39**, btad728 (2023).
58. Oksanen, J. et al. vegan: Community Ecology Package. (2024).
59. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
60. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. https://doi.org/10.1101/gr.186072.114 (2015).
61. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
62. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
63. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
64. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
65. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
66. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
67. Minh, B. Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evolution* **37**, 1530–1534 (2020).
68. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

69. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
70. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
71. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *bioRxiv* 2021.06.03.446934 https://doi.org/10.1101/2021.06.03.446934 (2021).
72. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).
73. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
74. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
75. Gillespie, J. J. et al. PATRIC: the Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species. *Infect. Immun.* **79**, 4286–4298 (2011).
76. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
77. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).

## Acknowledgements

S.R.C. Nori has received funding by Science Foundation Ireland through the SFI Centre for Research Training in Genomics Data Science under Grant number 18/CRT/6214 and supported in part by the EU's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant H2020-MSCA-COFUND-2019-945385. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Numbers SFI/12/RC/2273 and 16/SP/3827.

## Author contributions

Conceptualization: S.R.C.N, C.W., C.F., and P.C Experimental investigation: S.R.C.N. and C.F. Bioinformatics and statistics: S.R.C.N. Writing—original draft: S.R.C.N. Writing—review and editing: S.R.C.N, C.W., F.M., D.V.S., C.F., and P.C. Data curation: S.R.C.N. Figures preparation: S.R.C.N. Funding acquisition: S.R.C.N., D.V.S, F.M., and P.C. C.F. and P.C. are joint senior authors. All authors discussed the results and commented on the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41522-025-00682-1.

**Correspondence** and requests for materials should be addressed to Conor Feehily or Paul D. Cotter.

**Reprints and permissions information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025