

Supplementary Materials

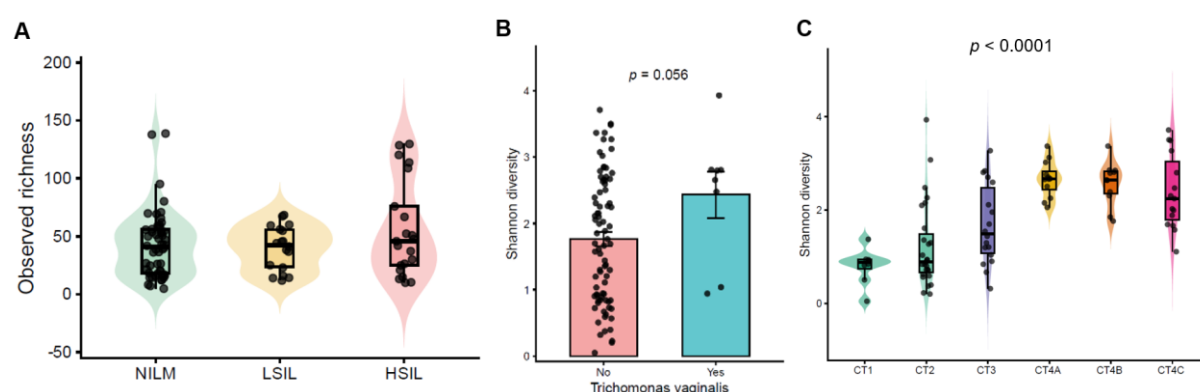
for

Ecological and inferred functional landscape of the cervicovaginal microbiome in relation to HPV infection and disease in Kenyan women

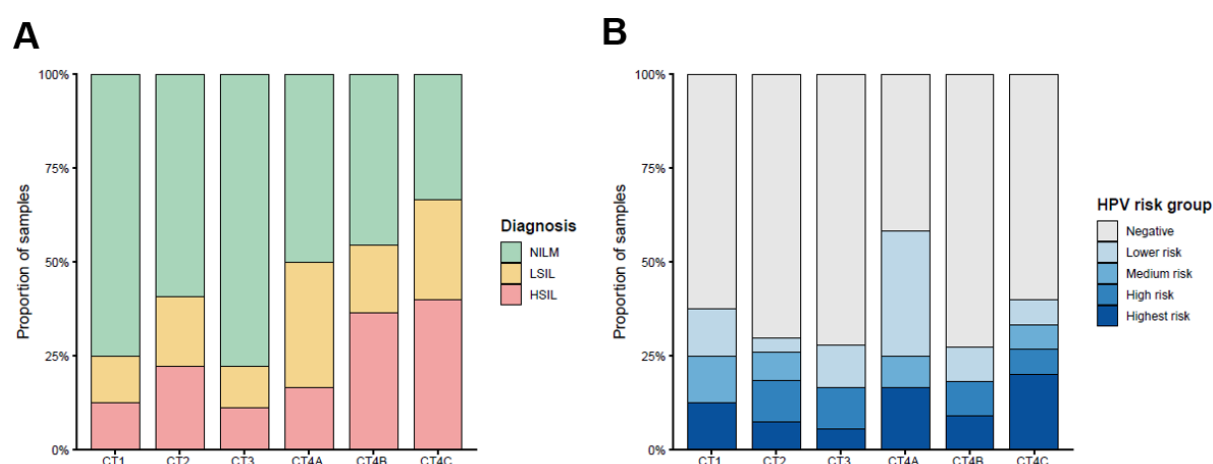
Mariano A. Molina^{1,2,*}, Kosgei Noah³, Ivy Akinyi⁴, Lisse Tavernier⁵, Sadaf Sakina Hassan⁶, Milan Stosic⁷, Davy Vanden Broeck⁵, Hillary Mabeya³, Laila Sara Arroyo Mühr⁶

1. Department of Laboratory Medicine, Karolinska Institutet, 141 52 Huddinge, Sweden.
2. Department of Cellular Therapy and Allogeneic Stem Cell Transplantation (CAST), Karolinska University Hospital, Huddinge, Sweden.
3. Gynocare Women's and Fistula Hospital, Eldoret, Kenya.
4. School of Health Sciences, Jaramogi Oginga Odinga University of Science and Technology, Kisumu, Kenya.
5. Laboratory of Molecular Pathology, AML Sonic Healthcare, Antwerp, Belgium.
6. Center for Cervical Cancer Elimination, F56, Karolinska University Hospital, Karolinska Institutet, 141 86 Huddinge, Sweden.
7. Norwegian HPV Reference Laboratory, Department of Microbiology and Infection Control, Akershus University Hospital, 1478 Lørenskog, Norway.

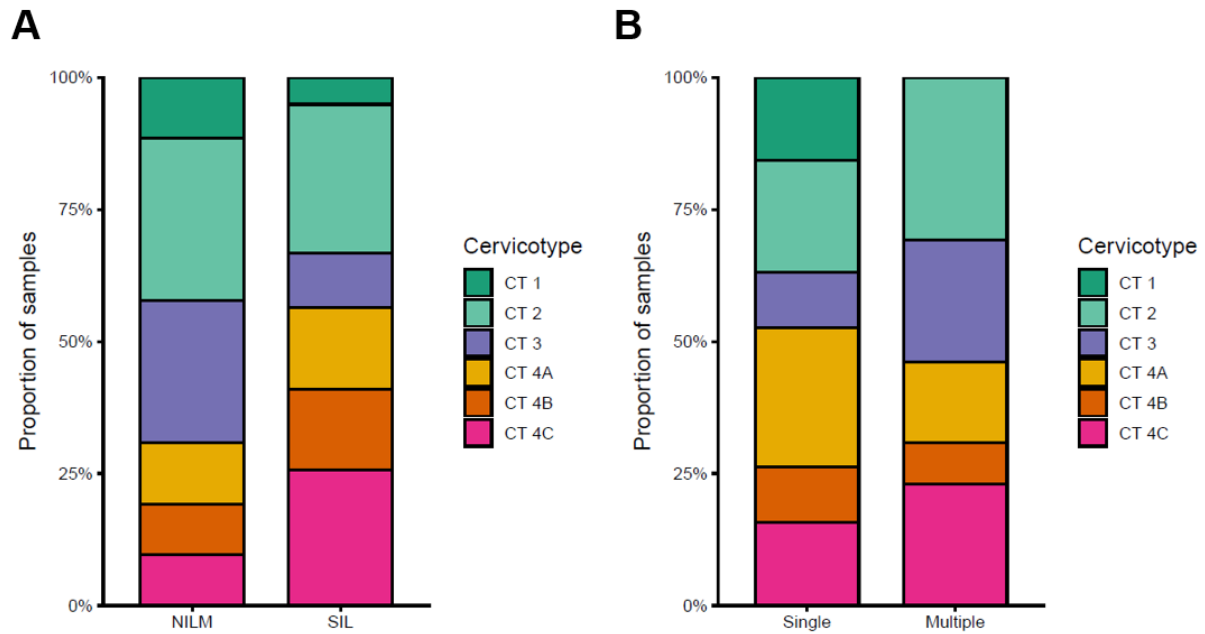
Correspondence: mariano.molina.beitia@ki.se



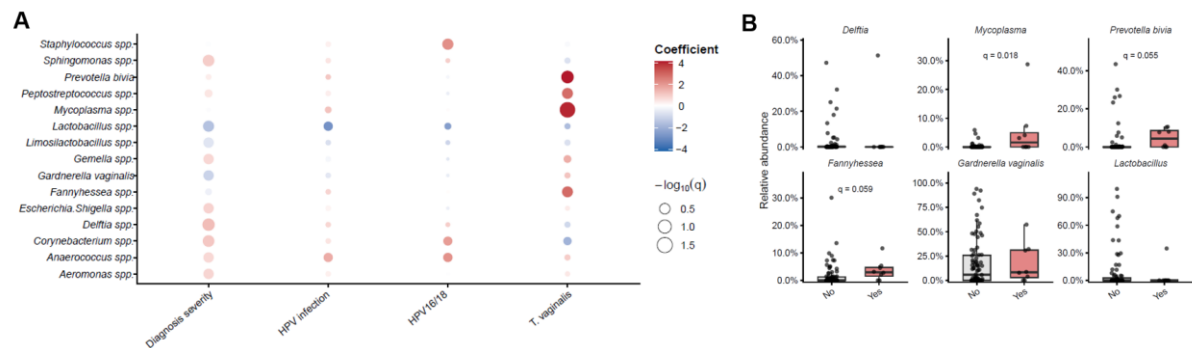
Supplementary Figure S1. Alpha diversity across cytological diagnoses, *T. vaginalis* status, and cervicotypes. (A) Observed richness according to cytological diagnosis (NILM, LSIL, and HSIL). (B) Shannon diversity according to *T. vaginalis* infection status; the displayed *p*-value was calculated using the Wilcoxon rank-sum test. (C) Shannon diversity across cervicotypes (CT1, CT2, CT3, CT4A, CT4B, and CT4C); the displayed *p*-value was calculated using the Kruskal–Wallis test. Violin plots show the distribution of values, boxplots indicate the median and interquartile range, whiskers extend to 1.5× the interquartile range, and individual points represent samples. NILM, negative for intraepithelial lesion or malignancy; LSIL, low-grade squamous intraepithelial lesion; HSIL, high-grade squamous intraepithelial lesion.



Supplementary Figure S2. Distribution of cytological diagnoses and HPV risk groups among cervicotypes. (A) Proportion of samples classified as negative for intraepithelial lesion or malignancy (NILM), low-grade squamous intraepithelial lesion (LSIL), and high-grade squamous intraepithelial lesion (HSIL) within each cervicotype. (B) Proportion of samples within HPV risk groups (see Supplementary Table S1) across cervicotypes.



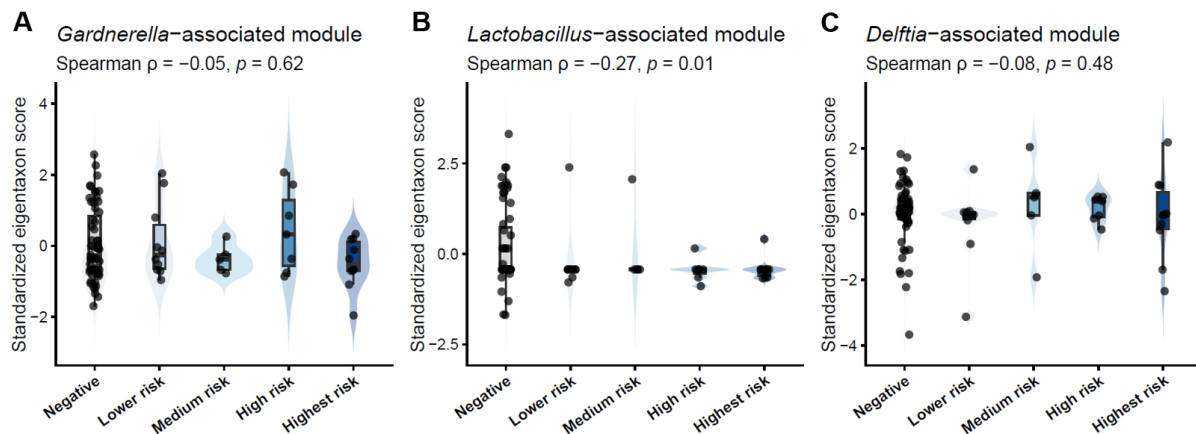
Supplementary Figure S3. Distribution of cervicotypes according to cervical lesion and HPV coinfections. (A) Proportion of cervicotypes among samples classified as negative for intraepithelial lesion or malignancy (NILM) or squamous intraepithelial lesion (SIL; combined LSIL and HSIL). (B) Proportion of cervicotypes among samples with single HPV infection or multiple HPV coinfections.



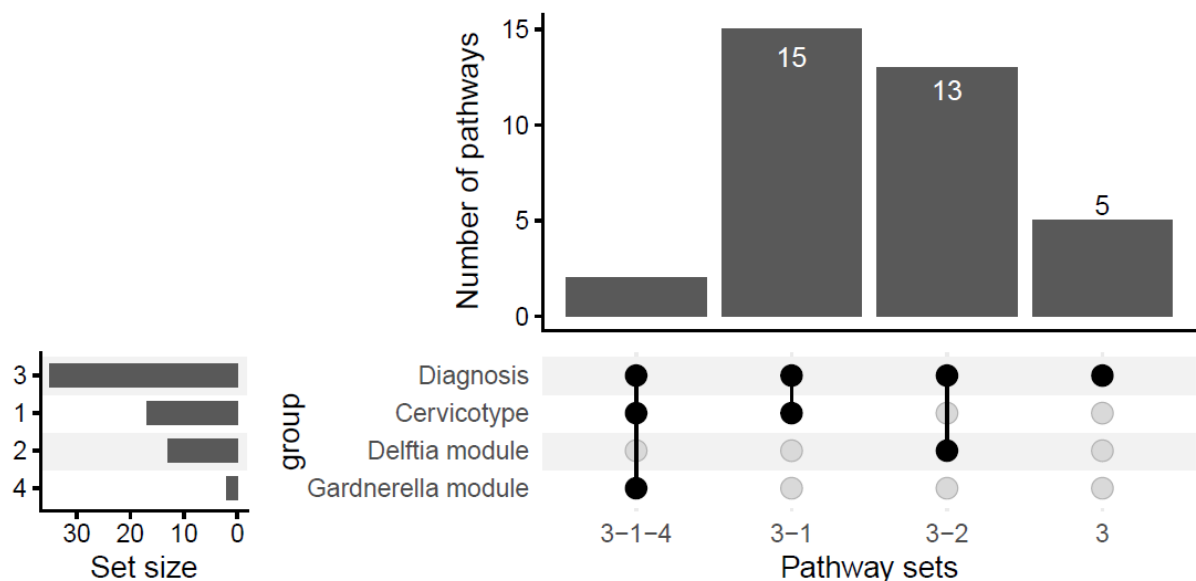
Supplementary Figure S4. Differential bacterial abundance analysis among clinical characteristics. (A) Bubble plot summarizing associations between bacterial species and clinical variables identified using MaAsLin2. Associations are shown for cytological diagnosis severity (NILM–LSIL–HSIL), HPV infection, HPV16/18 infection, and *T. vaginalis* infection. Bubble color indicates the direction and magnitude of the regression coefficient (red, positive association; blue, negative association), while bubble size corresponds to statistical significance. (B) Relative abundances of selected bacterial taxa associated with *T. vaginalis* infection. Points represent individual samples, boxplots indicate the median and interquartile range, and whiskers extend to 1.5× the interquartile range.



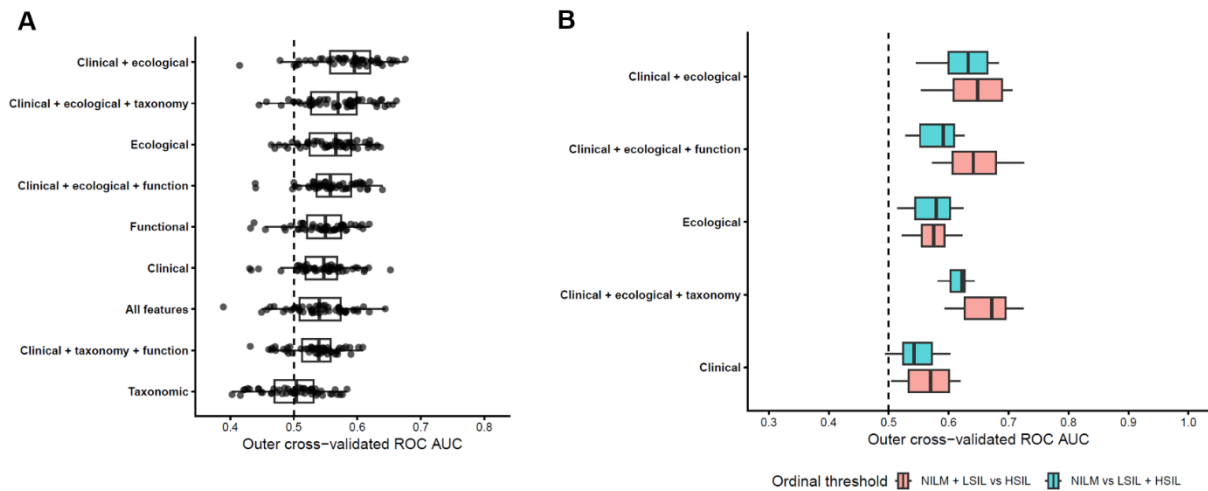
Supplementary Figure S5. Thresholded SparCC association matrix ordered by MCL-defined microbial modules. Pairwise SparCC correlation matrix showing associations among bacterial taxa retained for microbial network analysis. Rows and columns represent bacterial taxa, ordered according to the Markov Cluster Algorithm (MCL)-defined microbial modules. Cell colors indicate SparCC correlation coefficients, with green representing positive associations, purple representing negative associations, and white indicating weak or absent associations below the applied correlation threshold. The three major microbial modules correspond to the *Gardnerella*-, *Delftia*-, and *Lactobacillus*-associated communities identified in the network analysis (Figure 4).



Supplementary Figure S6. Associations between microbial ecological modules and HPV risk groups. Standardized eigentaxon scores for the (A) *Gardnerella*-associated, (B) *Lactobacillus*-associated, and (C) *Delftia*-associated modules across HPV risk groups. Violin plots show the distribution of eigentaxon scores, boxplots indicate the median and interquartile range, whiskers extend to 1.5× the interquartile range, and points represent individual samples. Associations with increasing HPV risk groups were assessed using Spearman rank correlation; corresponding ρ and p values are shown above each panel.



Supplementary Figure S7. Overlap of inferred metabolic pathways associated with microbial ecological modules and clinical variables. UpSet plot summarizing the overlap of PICRUSt2-inferred MetaCyc pathways significantly associated with cytological diagnosis, cervicotype, the *Gardnerella* module, and the *Delftia* module after FDR correction ($q < 0.05$). Vertical bars indicate the number of inferred pathways shared among combinations of variables, while horizontal bars show the total number of significant inferred pathways identified for each variable. Filled circles connected by vertical lines indicate the variables included in each intersection, whereas unfilled circles indicate variables not included.



Supplementary Figure S8. Predictive performance of clinical and microbiome-derived feature sets for binary and ordinal cytological classification. (A) Performance of elastic-net penalized logistic regression models for distinguishing NILM from SIL (LSIL + HSIL), evaluated using repeated nested five-fold cross-validation. Each point represents the area under the receiver operating characteristic curve (AUC) from one cross-validation repeat, and boxplots summarize the distribution across repeats. The dashed vertical line indicates random discrimination (AUC = 0.50). (B) Performance of elastic-net penalized ordinal logistic regression models for predicting increasing cytological diagnosis severity (NILM–LSIL–HSIL), shown for the two cumulative ordinal thresholds: NILM versus LSIL + HSIL and NILM + LSIL versus HSIL. Boxplots summarize AUC values across repeated nested cross-validation. Feature sets included Clinical (*T. vaginalis*, HPV coinfections, and HPV risk group), Ecological (Shannon diversity, cervicotype, and microbial module scores), Taxonomic (relative abundances of bacterial taxa), and Functional (PICRUSt2-inferred pathway abundances). Combined models incorporated the corresponding feature sets indicated in their labels.

Supplementary Table S1. Classification of HPV genotypes according to CIN3+ risk groups.

HPV risk group	Genotypes
Negative	No HPV detected
Lower risk	HPV35, HPV39, HPV51, HPV56, HPV59, HPV66, HPV68
Medium risk	HPV45, HPV52, HPV58
High risk	HPV18, HPV31, HPV33
Highest risk	HPV16

Supplementary Table S2. Adjusted *p*-values (*q*-values) from pairwise Wilcoxon rank-sum tests comparing Shannon diversity between cervicotypes.

	CT1	CT2	CT3	CT4A	CT4B
CT2	0.60958				
CT3	0.0191	0.04638			
CT4A	0.00079	0.00079	0.01191		
CT4B	0.00082	0.00079	0.01404	0.82945	
CT4C	0.00079	0.00079	0.04638	0.26675	0.40408

Supplementary Table S3. Sensitivity analyses of *Delftia* association with cytological diagnosis severity after adjustment for HPV risk group and *Trichomonas vaginalis*.

Taxon	Taxonomic level	Model	β	<i>p</i> -value	<i>q</i> -value
<i>Delftia</i>	Genus	Original	1.314	0.0027	0.098
		Adjusted for HPV risk	1.231	0.0055	0.381
		Adjusted for HPV risk + <i>T. vaginalis</i>	1.208	0.0073	0.357

Supplementary Table S4. Overlap of inferred diagnosis-associated metabolic pathways among cervicotypes and microbial ecological modules.

Inferred metabolic pathway	Cervicotype	<i>Gardnerella</i>	<i>Delftia</i>
UMP biosynthesis II	TRUE	TRUE	FALSE
UMP biosynthesis III	TRUE	TRUE	FALSE
2-methylbutanoate biosynthesis	TRUE	FALSE	FALSE
2-methylcitrate cycle I	TRUE	FALSE	FALSE
2-methylcitrate cycle II	TRUE	FALSE	FALSE
4-methylcatechol degradation (ortho cleavage)	FALSE	FALSE	TRUE
D-erythronate degradation II	TRUE	FALSE	FALSE
D-galactarate degradation I	TRUE	FALSE	FALSE
D-glucarate degradation I	TRUE	FALSE	FALSE
L-ornithine degradation II (Stickland reaction)	TRUE	FALSE	FALSE
L-threonate degradation	TRUE	FALSE	FALSE
L-tryptophan degradation X (mammalian, via tryptamine)	TRUE	FALSE	FALSE
adenosylcobalamin biosynthesis II (late cobalt incorporation)	TRUE	FALSE	FALSE
aromatic compounds degradation via β -ketoadipate	FALSE	FALSE	TRUE
catechol degradation III (ortho-cleavage pathway)	FALSE	FALSE	TRUE
catechol degradation to β -ketoadipate	FALSE	FALSE	TRUE
ectoine biosynthesis	TRUE	FALSE	FALSE
fatty acid β -oxidation II (peroxisome)	TRUE	FALSE	FALSE
gallate degradation I	FALSE	FALSE	TRUE
gallate degradation II	FALSE	FALSE	TRUE
glycerophospholipid metabolism	TRUE	FALSE	FALSE
itaconate degradation	FALSE	FALSE	TRUE
methylgallate degradation	FALSE	FALSE	TRUE
nicotinate degradation I	FALSE	FALSE	TRUE
picolinate degradation	FALSE	FALSE	TRUE
protocatechuate degradation I (meta-cleavage pathway)	FALSE	FALSE	TRUE
superpathway of D-glucarate and D-galactarate degradation	TRUE	FALSE	FALSE
superpathway of methylglyoxal degradation	TRUE	FALSE	FALSE
superpathway of salicylate degradation	FALSE	FALSE	TRUE
vanillin and vanillate degradation II	FALSE	FALSE	TRUE
L-histidine degradation II	FALSE	FALSE	FALSE
fatty acid salvage	FALSE	FALSE	FALSE
glucosylglycerol biosynthesis	FALSE	FALSE	FALSE
pyruvate fermentation to acetone	FALSE	FALSE	FALSE