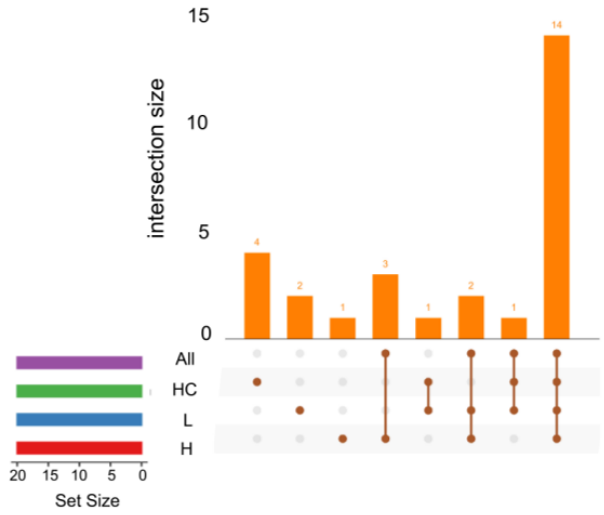

Figure S1. The intersection analysis of the top 20 most abundant bacterial genera among HPV high-risk (H), low-risk (L), HPV-negative (HC) groups, and the overall population (All).

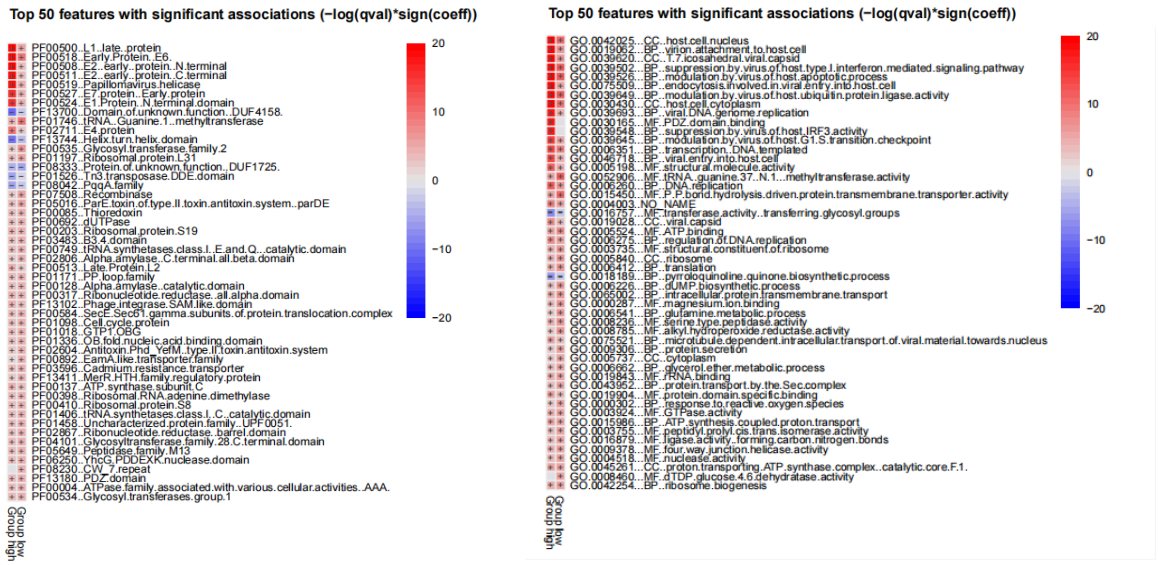

Figure S2. Top-50 enriched PFAM families (left) and Gene Ontology (GO; right) pathways in the H and L groups relative to the HC group. Heatmaps display the β-coefficients obtained from MaAsLin2 differential-abundance analysis, with red and blue indicating increased and decreased abundance in H/L groups versus HC, respectively. The pathways are ordered by the absolute magnitude of the β-coefficient, and only the 50 most significant PFAM families (left panel) and GO terms (right panel) for each comparison are displayed.

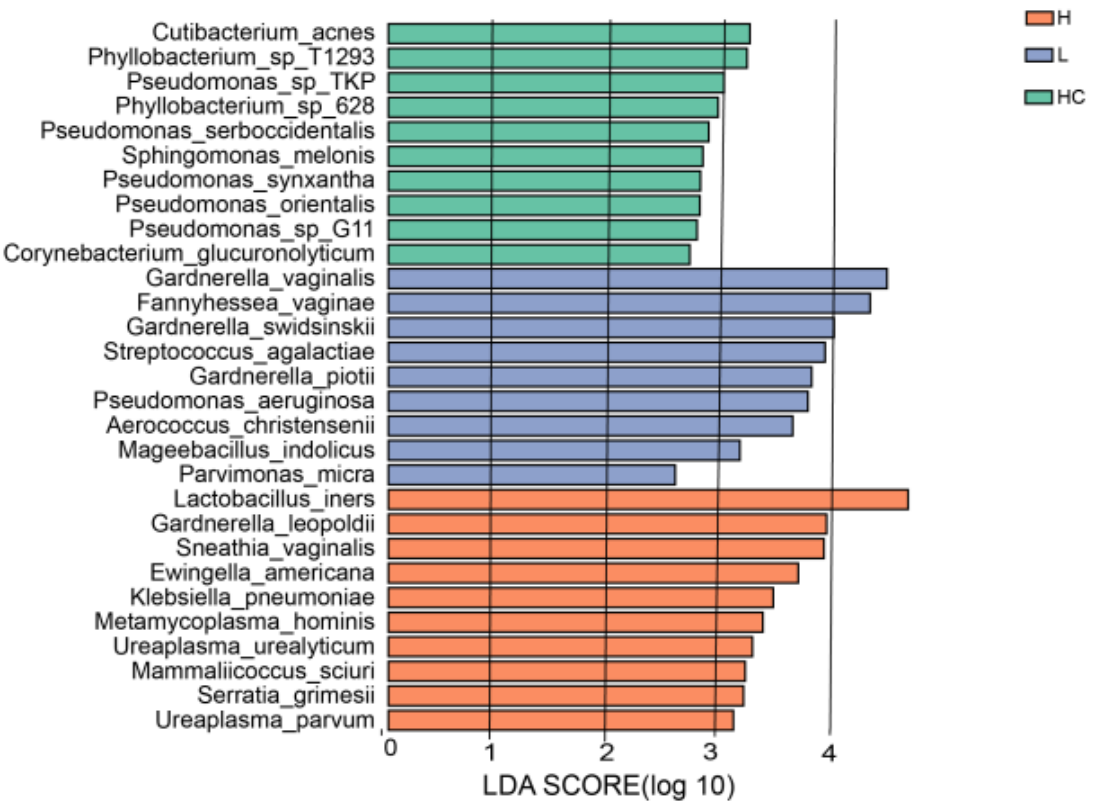

Figure S3. Discriminative taxa identified by LEfSe analysis. Bar plot shows the LDA scores (threshold > 2.5) for microbial features significantly enriched in the H, L, and HC groups. The length of the bar represents a log10 transformed LDA score.

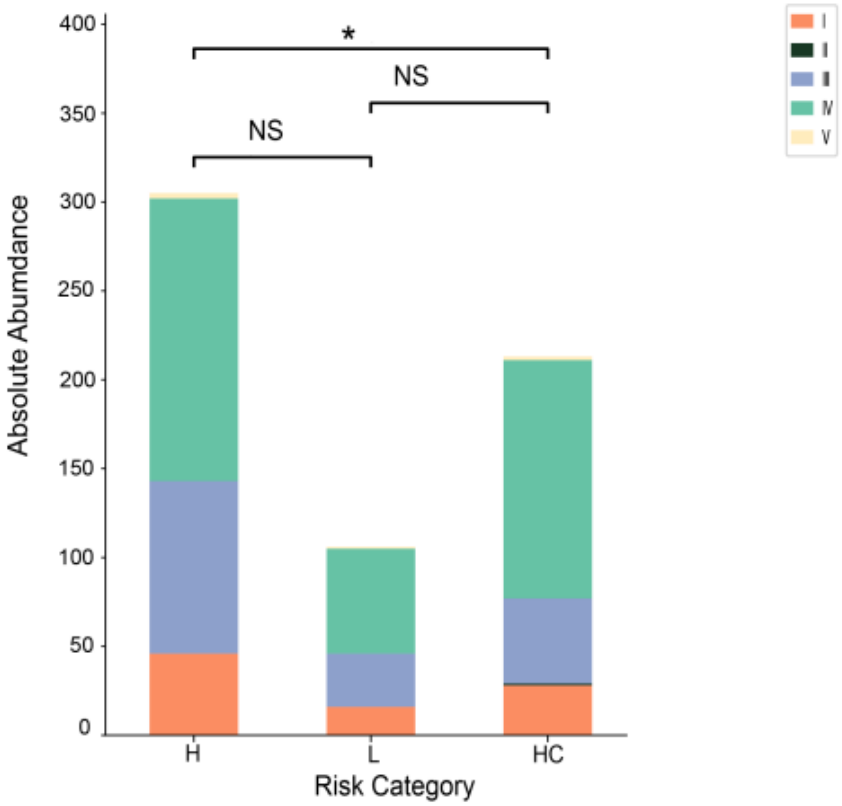

Figure S4. Composition of vaginal microbiota community state types (CSTs) across HPV risk groups. Stacked bar plot shows the proportion of samples in each group (H, L, and HC) classified into CSTs I-V (CST I-III/V: Lactobacillus-dominant; CST IV: non-Lactobacillus predominant microbiota). NS, not significant, *P< 0.05, **P< 0.01, ***P< 0.001.

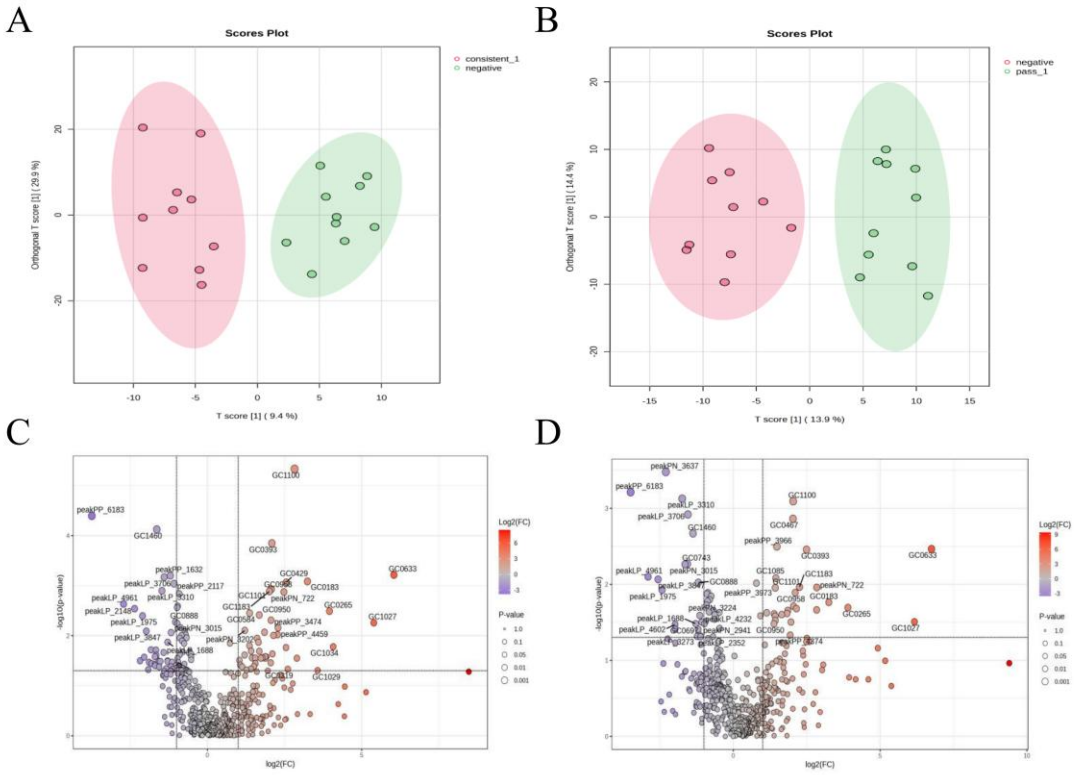

Figure S5. OPLS-DA analysis of metabolite profiles and identification of differential metabolites. (A) OPLS-DA analysis between the persistent infection (consistent_1) and negative groups at baseline. (B) OPLS-DA analysis between the transient infection (pass_1) and negative groups at baseline. (C) Volcano plot of persistent infection vs. negative. (D) Volcano plot of transient infection vs. negative. Each point representing a metabolite (x-axis: $\log_2$ Fold change; y-axis: $-\log_{10}$ p-value). Significantly altered metabolites were identified based on variable importance in projection (VIP) score and screening criteria of fold change (FC) > 2.0 or < 0.5 with P value < 0.01.

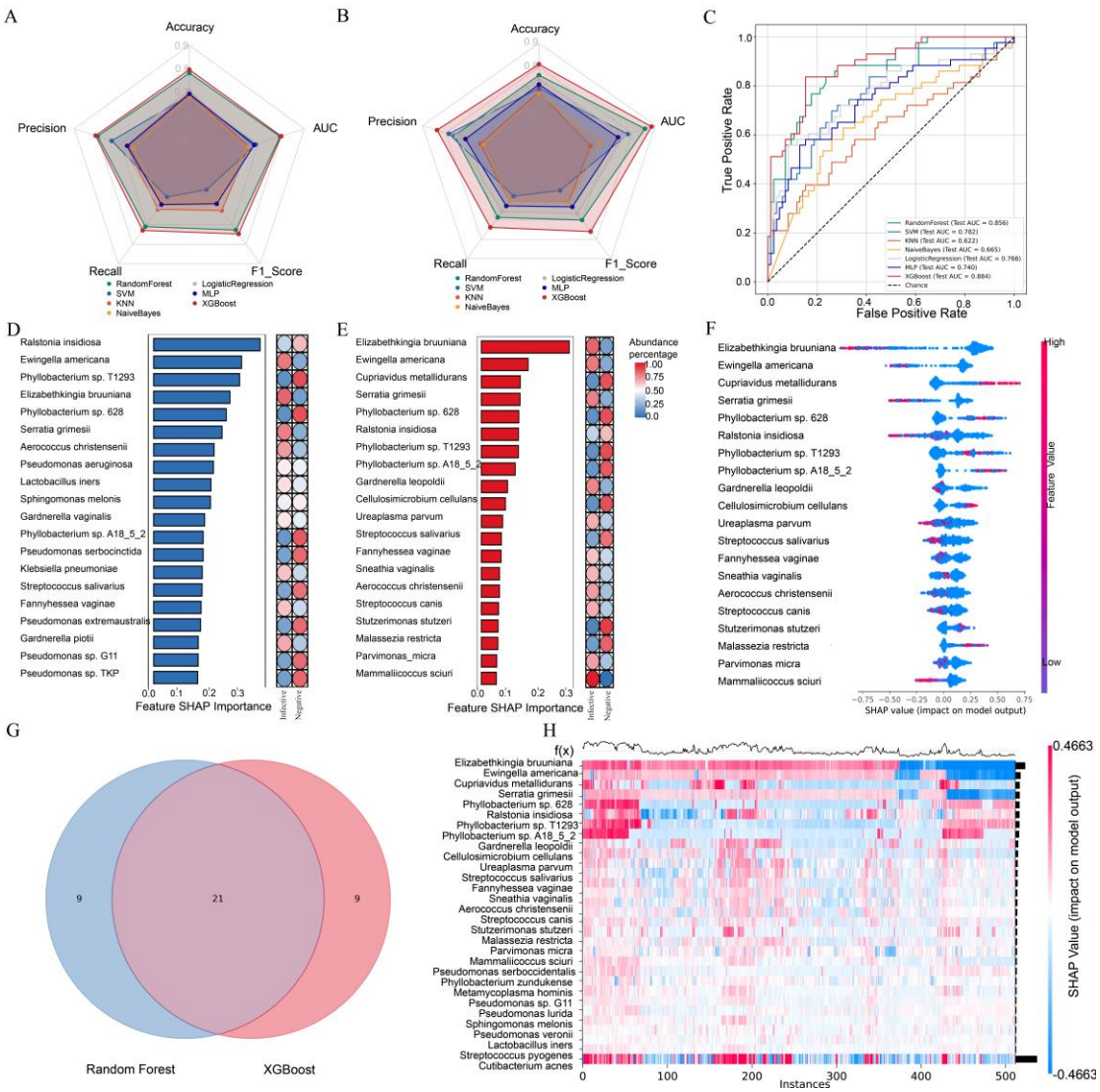

Figure S6. Development and evaluation of machine learning models for HPV infection diagnosis. (A-B) Radar charts comparing seven algorithms across five performance metrics (Accuracy, Precision, Recall, F1-score, and AUC) in the cross-validation set (A) and the independent test set (B). (C) ROC curve of the seven models on the test set. (D-E) SHAP summary plots displaying the top 20 feature importance rankings for the Random Forest (D) and XGBoost (E) models. (F) Beeswarm plot of SHAP values for the top 20 features in the XGBoost model. Each point represents a sample, colored by normalized feature value (red: high, blue: low). Its position on the x-axis indicates the impact of the feature on the model output (SHAP value). (G) Venn diagram comparing the top 30 features selected by the Random Forest (blue) and XGBoost (pink) models. (H) Heatmap of the mean absolute SHAP values for the top 30 features in the XGBoost model across the test set. Intensity of red color corresponds to greater feature importance.

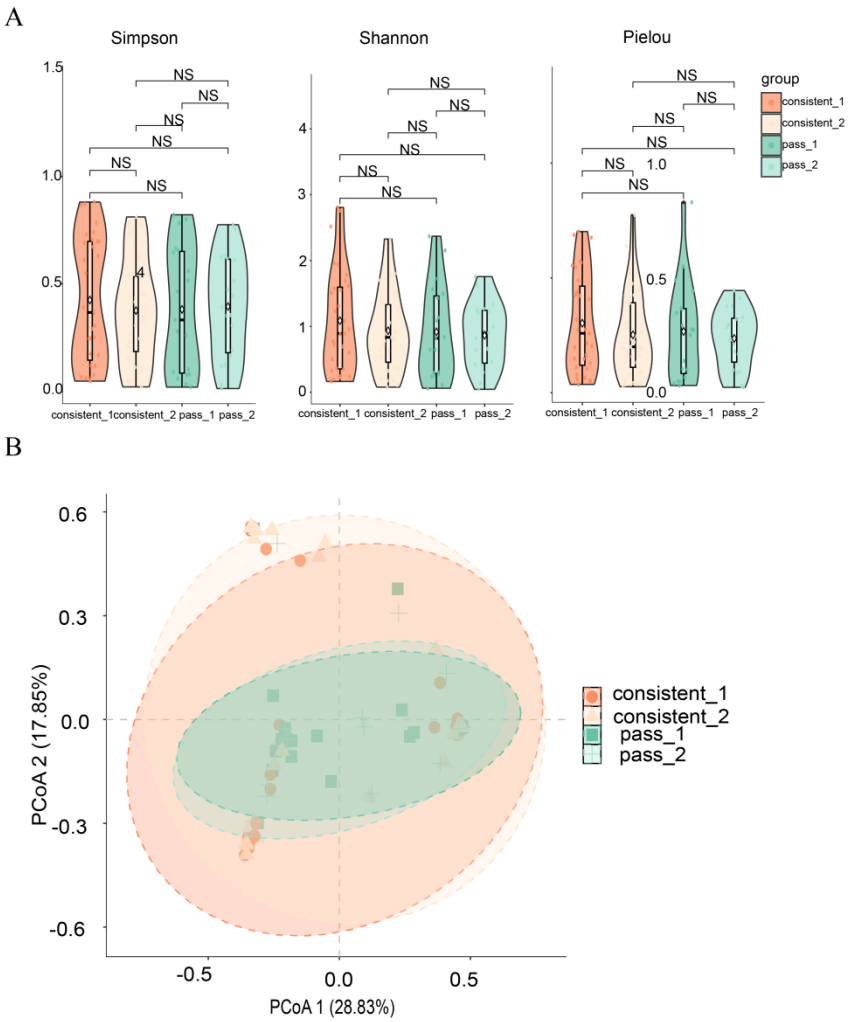

Figure S7. Longitudinal analysis of vaginal microbiota α- and β-diversity in persistent versus transient HPV infection groups. (A) Analysis of α-diversity (Simpson, Shannon, and Pielou's indices) between the two groups at both baseline and 1-year follow-up. (B) Analysis of β-diversity (PCoA) between the two groups at both baseline and 1-year follow-up. Groups are defined as: consistent_1, persistent infection (baseline); consistent_2, persistent infection (1-year follow-up); pass_1, transient infection (baseline); pass_2, transient infection (1-year follow-up).

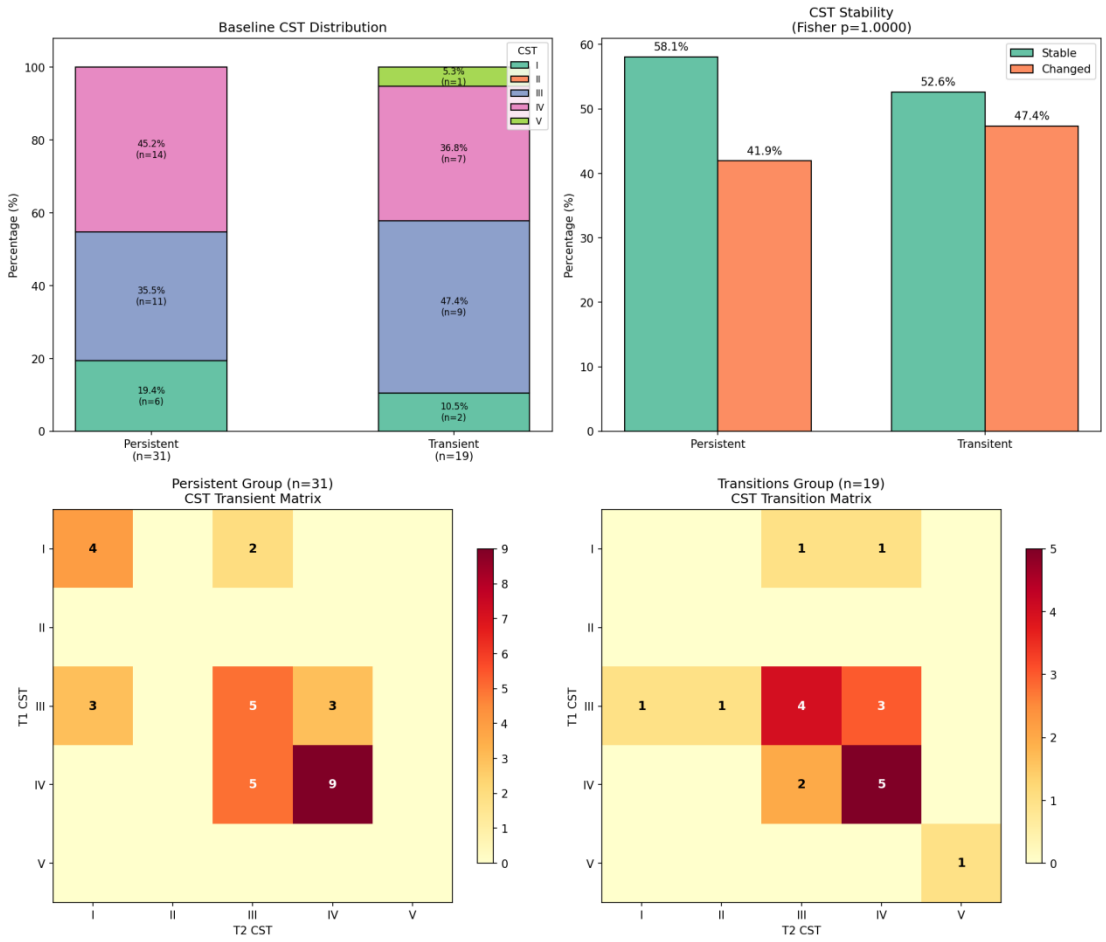

Figure S8. Baseline CST distribution and longitudinal CST transition patterns in hrHPV persistent and transient infection groups. (A) Baseline community state type (CST) distribution in the persistent infection group (n = 31) and transient infection group (n = 19). The overall CST distribution was comparable between the two groups (P = 0.42). (B) Comparison of CST stability between the persistent infection and transient infection groups. Stable indicates that the CST remained unchanged between baseline and follow-up, whereas changed indicates a transition to a different CST at follow-up. The proportion of individuals maintaining a stable CST did not differ significantly between the persistent infection and transient infection groups. (C) CST transition matrix in the persistent infection group. Rows indicate baseline CSTs and columns indicate follow-up CSTs. Numbers in the matrix represent the number of individuals with each transition pattern. (D) CST transition matrix in the transient infection group. Rows indicate baseline CSTs and columns indicate follow-up CSTs. Overall, baseline CST composition, CST stability, and CST transition patterns were not significantly associated with hrHPV persistence or transient infection in this cohort.

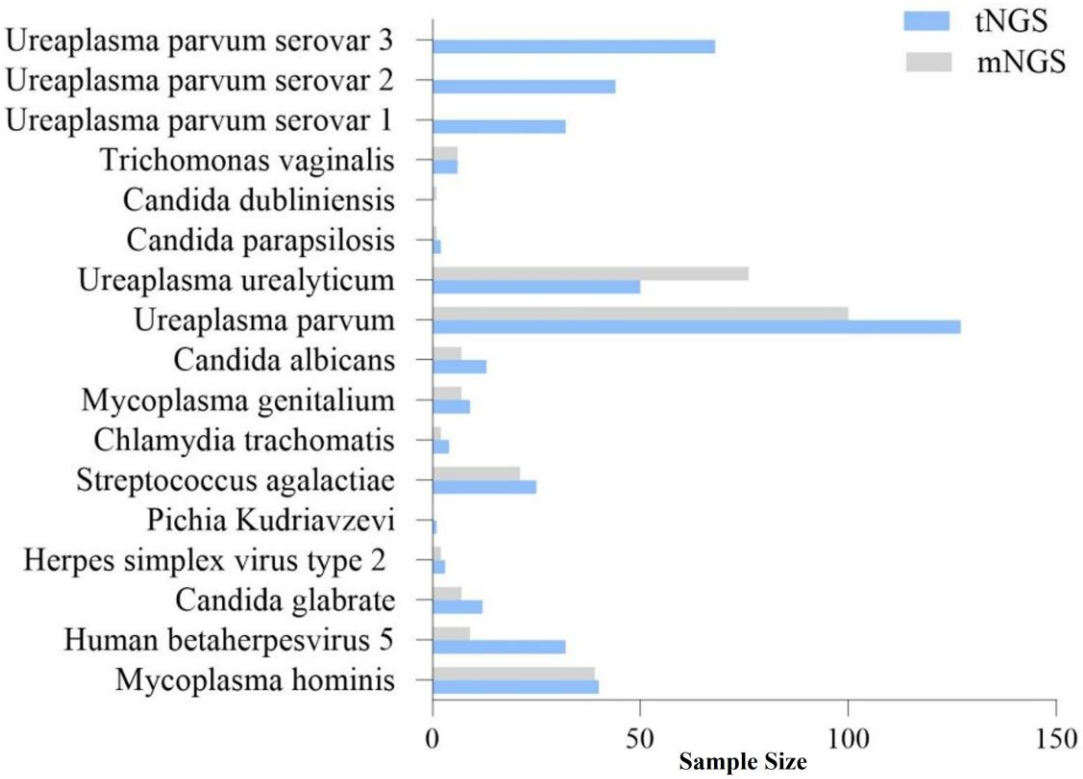

Figure S9. Comparative detection of 17 female reproductive tract pathogens by tNGS and mNGS. A total of 250 samples were tested in parallel using targeted next-generation sequencing and metagenomic next-generation sequencing. After excluding 28 samples that failed tNGS quality control, 222 samples were included in the final analysis. The figure summarizes the detection counts and differences between tNGS and mNGS for 17 common female reproductive tract pathogens.

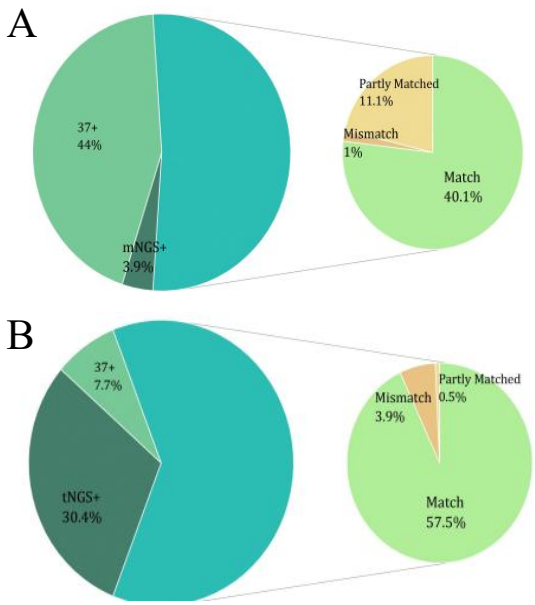

Figure S10. Concordance analysis of HPV genotyping results between mNGS/tNGS and the 37-type HPV genotyping test. (A) Concordance between mNGS and the 37-type HPV genotyping test across all samples. (B) Concordance between tNGS and the 37-type HPV genotyping test across all samples. Samples positive by both methods were classified as complete match, partial match, or mismatch. Complete match indicates identical HPV genotypes detected by both methods; partial match indicates that at least one HPV genotype overlapped; and mismatch indicates that both methods detected HPV but shared no identical genotype. Samples positive only by the 37-type HPV genotyping test were considered missed detections by mNGS or tNGS. The complete concordance rates were 40.1% for mNGS and 57.5% for tNGS, and the missed-detection rates were 44% and 7.7%, respectively.

Supplementary Table 1 Pairwise PERMANOVA analysis of vaginal microbiota community structure among HPV infection status groups.

pairs	R2	p.value	p.adjusted
H vs L	0.003526445	0.13	0.13
H vs HC	0.009375346	0.001	0.0015
L vs HC	0.010485308	0.001	0.0015

Supplementary Table 2 The top 20 genus-level taxa in the H, L, HC, and all-sample groups.

the H group	the L group	the HC group	all-sample group
Lactobacillus	Lactobacillus	Lactobacillus	Lactobacillus
Gardnerella	Gardnerella	Pseudomonas	Gardnerella
Pseudomonas	Streptococcus	Gardnerella	Pseudomonas
Escherichia	Pseudomonas	Escherichia	Streptococcus
Streptococcus	Escherichia	Streptococcus	Prevotella
Prevotella	Fannyhessea	Enterococcus	Fannyhessea
Sneathia	Prevotella	Prevotella	Sneathia
Fannyhessea	Sneathia	Fannyhessea	Enterococcus
Sphingomonas	Sphingomonas	Sphingomonas	Escherichia
Klebsiella	Hoylesella	Staphylococcus	Sphingomonas
Enterococcus	Klebsiella	Sneathia	Klebsiella
Trichomonas	Staphylococcus	Bifidobacterium	Hoylesella
Hoylesella	Aerococcus	Hoylesella	Staphylococcus
Ureaplasma	Bifidobacterium	Citrobacter	Ureaplasma
Aerococcus	Proteus	Campylobacter	Trichomonas
Alphapapillomavirus	Mycoplasmoides	Proteus	Aerococcus
Staphylococcus	Enterococcus	Nakaseomyces	Bifidobacterium
Metamycoplasma	Ureaplasma	Klebsiella	Metamycoplasma
Ewingella	Fusobacterium	Ureaplasma	Ewingella
Acinetobacter	Metamycoplasma	Schaalia	Alphapapillomavirus

Supplementary Table 3 Plasmid-borne genes identified in Lactobacillus iners

Species	gene	marker	annotation_accession	annotation_description
Lactobacillus iners	k45_52231_4	GENOMAD.182904.PC	Predicted small periplasmic lipoprotein YifL	PF02402;COG5567
Lactobacillus iners	k45_45115_1	GENOMAD.138275.PV	plasmid partitioning protein RepB	PF07506;PF02195;TIGR03454;COG1475;K03497
Lactobacillus iners	k45_19617_5	GENOMAD.182904.PC	Predicted small periplasmic lipoprotein YifL	PF02402;COG5567
Lactobacillus iners	k45_79370_1	GENOMAD.138275.PV	plasmid partitioning protein RepB	PF07506;PF02195;TIGR03454;COG1475;K03497

Supplementary Table 4 Socio-demographic characteristics of subjects

Characteristics	Transient infection (n-=19)	Persistent infection （n=31）	p-Valu e
Age (mean ± SD, range)	36.9 ± 4.7 (20-52)	34.5 + 4.5 (21-55)	0.075
Time since last menstrual period (mean ± SD)	15.7±5.7	16.7±7.1	0.589
Menarche age	14.5±1.5	14.4±1.2	0.781
Smoker(s)	0	2/31	0.481
Passive smokers	8/19	16/31	0.876
Number of pregnancies			0.064
<3	12/19	17/31	
≥3	7/19	14/31	

Supplementary Table 5 Comparison of baseline community state type distribution between persistent and transient hrHPV infection groups.

CST	Persistent (n=31)	Transient (n=19)	p
I	19.4%	10.5%	0.69
III	35.5%	47.4%	0.55
IV	45.2%	36.8%	0.77
V	0%	5.3%	0.38

Note:CST, community state type; CST I, *Lactobacillus crispatus*-dominated; CST III, *Lactobacillus iners*-dominated; CST IV, anaerobe-dominated; CST V, *Lactobacillus jensenii*-dominated. P values indicate comparisons of individual CST proportions between the persistent and transient infection groups. Overall baseline CST distribution was comparable between the two groups (P = 0.42).

Supplementary Table 6 The CST stability and transitions at the two longitudinal time points in the persistent infection and transient infection groups

	Persistent	Transient
CST Stable	58.1% (18/31)	52.6% (10/19)
CST Changed	41.9% (13/31)	47.4% (9/19)

Note: CST, community state type. CST stable indicates that the CST remained unchanged between baseline and follow-up, whereas CST changed indicates a transition to a different CST at follow-up. The proportion of individuals with stable CSTs was comparable between the persistent and transient hrHPV infection groups (58.1% vs. 52.6%, P = 0.77).

Supplementary Table 7 Follow-up CST transition patterns among individuals with baseline CST III in persistent and transient hrHPV infection groups

CST III Change	Persistent (n=11)	Transient (n=9)
CST III	5 (45.5%)	4 (44.4%)
→ CST IV	3 (27.3%)	3 (33.3%)
→ CST I	3 (27.3%)	1 (11.1%)
→ CST II	0	1 (11.1%)

Note: This analysis included only individuals classified as CST III at baseline. Percentages were calculated using the number of baseline CST III individuals in each group as the denominator. CST III indicates a *Lactobacillus iners*-dominated community state type.Because this analysis was restricted to individuals with CST III at baseline, the denominators differ from those used in the overall CST stability analysis.

Supplementary Table 8 Baseline prevalence and relative abundance of *Streptococcus agalactiae* in persistent and transient hrHPV infection groups.

	Persistent_1 (n=31)	Transient_1 (n=19)
Prevalence		
Detected, n (%)	11 (35.5%)	13 (68.4%)
Abundance%		
Mean ± SD	0.0431 ± 0.0666	8.8243 ± 26.0429
Median (IQR)	0.0000 (0.0000–0.0904)	0.0727 (0.0000–0.2689)
Range	0.0000–0.1812	0.0000–84.3915

Supplementary Table 9 Effect-size analysis of baseline *Streptococcus agalactiae* prevalence and abundance between persistent and transient hrHPV infection groups.

Measure		Value	95% CI	P-value
Prevalence				
	Odds ratio (OR)	0.25	[0.08, 0.86]	0.0404
	Cohen's h	-0.67	—	—
Abundance				
	Cliff's δ	-0.37	—	—
	Mann-Whitney U	186	—	0.0199

Note：The odds ratio was calculated for detection in the persistent infection group relative to the transient infection group; therefore, an OR < 1 indicates higher detection in the transient infection group.

Supplementary Table 10 MaAsLin2-based sensitivity analysis of LEfSe-derived species-level associations across hrHPV longitudinal outcome comparisons.

feature	value	coef	pval	qval	LEfSe	pair
Streptococcus_agalactiae	pass_1	2.514	0.011	0.788	√	transient_1_vs_persistent_1
Klebsiella_pneumoniae	pass_1	1.329	0.027	0.788		transient_1_vs_persistent_1
Sneathia_vaginalis	pass_1	-2.527	0.047	0.788		transient_1_vs_persistent_1
Streptococcus_thermophilus	consistent_2	1.008	0.006	0.505		persistent_2_vs_persistent_1
Staphylococcus_agnetis	consistent_2	1.156	0.007	0.505		persistent_2_vs_persistent_1
Pediococcus_acidilactici	consistent_2	0.532	0.010	0.505		persistent_2_vs_persistent_1
Paenibacillus_mucilaginosus	consistent_2	0.549	0.011	0.505		persistent_2_vs_persistent_1
Bacillus_sp._CMF12	consistent_2	0.328	0.012	0.505		persistent_2_vs_persistent_1
Lactobacillus_sp._ESL0684	consistent_2	0.419	0.012	0.505		persistent_2_vs_persistent_1
Lactobacillus_sp._CBA3605	consistent_2	0.273	0.012	0.505		persistent_2_vs_persistent_1
Floricoccus_penangensis	consistent_2	0.764	0.013	0.505		persistent_2_vs_persistent_1
Lentilactobacillus_laojiaonis	consistent_2	0.443	0.014	0.505		persistent_2_vs_persistent_1
Agrobacterium_tumefaciens	consistent_2	1.728	0.017	0.505	√	persistent_2_vs_persistent_1
Photobacterium_profundum	consistent_2	0.526	0.018	0.505		persistent_2_vs_persistent_1
Loigolactobacillus_coryniformis	consistent_2	0.575	0.019	0.505		persistent_2_vs_persistent_1
Sulfidibacter_corallicola	consistent_2	0.381	0.021	0.505		persistent_2_vs_persistent_1
Priestia_megaterium	consistent_2	0.906	0.021	0.505		persistent_2_vs_persistent_1
Ligilactobacillus_salivarius	consistent_2	0.942	0.035	0.670		persistent_2_vs_persistent_1
Companilactobacillus_futsaii	consistent_2	0.530	0.035	0.670		persistent_2_vs_persistent_1
Lactobacillus_paragasseri	consistent_2	0.461	0.039	0.670		persistent_2_vs_persistent_1
Lactobacillus_sp._IBH004	consistent_2	0.411	0.044	0.670		persistent_2_vs_persistent_1
Anaerococcus_vaginalis	consistent_2	-0.354	0.045	0.670		persistent_2_vs_persistent_1
Companilactobacillus_zhachilii	consistent_2	0.534	0.047	0.670		persistent_2_vs_persistent_1
Lactobacillus_sp._ESL0677	consistent_2	0.322	0.048	0.670		persistent_2_vs_persistent_1
Pseudomonas_fluorescens	pass_2	-2.743	0.019	0.676	√	transient_2_vs_transient_1
Ewingella_americana	pass_2	-2.091	0.031	0.676	√	transient_2_vs_transient_1
Pseudomonas_azotoformans	pass_2	-0.906	0.033	0.676		transient_2_vs_transient_1
Bifidobacterium_pseudolongum	pass_2	0.658	0.034	0.676		transient_2_vs_transient_1
Bifidobacterium_longum	pass_2	2.362	0.048	0.676		transient_2_vs_transient_1

Note: Species-level associations were re-evaluated using MaAsLin2 as an additional sensitivity analysis. The coefficient indicates the direction and magnitude of association with the corresponding group or time-point comparison. Positive coefficients indicate enrichment in the group listed under “value,” whereas negative coefficients indicate depletion. P values are nominal P values, and q values are FDR-adjusted P values. The “LEfSe” column indicates species that were also identified by LEfSe with a consistent direction of association.

Supplementary Table 11 Permutation test validation of machine-learning models for distinguishing persistent hrHPV infection from transient infection.

Model	Observed AUC	Perm. Mean	Perm. 95th Pctile	p-value	Significance
Logistic Regression	0.890	0.339	0.631	0.001	**
Linear SVM	0.970	0.356	0.705	0.002	**
KNN	0.895	0.437	0.700	0.003	**
Naive Bayes	0.880	0.395	0.720	0.003	**
Random Forest	0.820	0.406	0.720	0.010	*
XGBoost	0.860	0.409	0.711	0.005	**
MLP	0.970	0.403	0.700	0.001	**

Significance: *P< 0.05, **P< 0.01, ***P< 0.001

Supplementary Table 12 .632 bootstrap-corrected AUCs of machine-learning models for distinguishing persistent hrHPV infection from transient infection

Model	.632 Corrected AUC	95% CI
Logistic Regression	0.973	[0.831, 1.000]
Linear SVM	0.974	[0.831, 1.000]
KNN	0.938	[0.757, 1.000]
Naive Bayes	0.923	[0.684, 1.000]
Random Forest	0.935	[0.684, 1.000]
XGBoost	0.913	[0.500, 1.000]
MLP	0.981	[0.874, 1.000]

Supplementary Table 13 Comparison of Detection Performance between mNGS and tNGS for 17 Common Pathogens

Pathogen	TP	FN	FP	TN	Sensitivity	Specificity
Mycoplasma hominis	34	5	6	177	87%	97%
Human betaherpesvirus 5	9	0	23	190	100%	89%
Candida glabrata	7	0	5	210	100%	98%
Herpes simplex virus type 2	2	0	1	219	100%	99%
Kodamaea ohmeri	0	0	1	221	/	100%
Streptococcus agalactiae	18	3	7	194	86%	97%
Chlamydia trachomatis	2	0	2	218	100%	99%
Mycoplasma genitalium	6	1	3	212	86%	99%
Candida albicans	7	0	6	209	100%	97%
Ureaplasma parvum	94	6	33	89	94%	73%
Ureaplasma urealyticum	35	41	15	131	46%	90%
Candida parapsilosis	1	0	1	220	100%	100%
Candida dubliniensis	0	1	0	221	0%	100%
Trichomonas vaginalis	6	0	0	216	100%	100%
Ureaplasma parvum serovar 1*	0	0	32	190	/	86%
Ureaplasma parvum serovar 3*	0	0	48	178	/	89%
Ureaplasma parvum serovar 6*	0	0	88	154	/	69%

Note: * indicates that subtyping was not performed by mNGS.

Legends to Supplementary Data

Supplementary Data 1 Complete species-level abundance matrix for the 640 samples

Supplementary Data 2 The genus-level taxa in the H, L, HC, and all-sample groups

Supplementary Data 3 Complete LEfSe marker information for the high-risk (H), low-risk (L), and HPV-negative (HC) groups

Supplementary Data 4 Differences in RPKM of different MAGs among the H group, L group, and HC group.

Supplementary Data 5 Vaginal-secretion metabolomics data of 30 samples measured on three analytical platforms

Supplementary Data 6 Complete list of differential metabolites identified by untargeted metabolomics

Supplementary Data 7 Complete Species-Level Abundance Matrix for Pre-and Post-follow up Patient Samples

Supplementary Data 8 Details of HPV Genotyping Results and Follow-up Detection for Patients

Supplementary Data 9 Normalized counts for functional modules (disease factors and short-chain fatty acids)

Supplementary Data 10 tNGS Reproductive 50 Testing Scope