nature genetics

Article

https://doi.org/10.1038/s41588-026-02639-2

Genomic landscape of the human vaginal microbiome is linked to host genetics and population of origin

Received: 19 March 2024

Accepted: 14 May 2026

Published online: 11 June 2026

Check for updates

Zhuye Jie[1,2,18], Weiting Liang[1,3,18], Qiuxia Ding[1,2,18], Xiaomin Liu[4,18], Yunhong Zhang[5,18], Na Chen[6,18], Shenghui Li[7,18], Xin Tong[8,9], Hongqin Gao[10], Ruike Lu[10], Xincheng Huang[11], Ruochun Guo[7], Junhong Chen[1], Jie Zhu[1], Zhe Zhang[1], Na Liu[12], Zhangwei Xie[13], Xiaman Wang[12], Le Qi[12], Yumei Li[12], Liang Xiao[11], Shaoqiao Zhang[1], Xin Jin[9], Xun Xu[9], Huanming Yang[11,14], Jian Wang[11,14], Fangqing Zhao[15], Huijie Jia[16,17], Karsten Kristiansen[2,11], Tao Zhang[1,19] ✉, Lilan Hao[1,2,19] ✉, Lan Zhu[6,19] ✉ & Chen Chen[1,19] ✉

The vaginal microbiome is essential for women's health, yet its genomic diversity and interaction with the host remain incompletely characterized. Here we present the Global Vaginal Metagenome-assembled Genomes catalog, an extensive repository of vaginal microbial genomes generated by integrating 10,665 in-house Chinese metagenomes, with 2,967 publicly available metagenomes and 1,433 bacterial isolates. The catalog comprises 65,055 genomes from 890 prokaryotes, 11 eukaryotes and 6,590 viral taxonomic units, many not represented in public reference databases. We investigate virus–bacteria interactions, revealing conserved phages–host associations. We then identify substantial intraspecies genomic and functional variations displaying population-specific patterns. A metagenome-genome-wide association study identifies seven host genetic loci associated with vaginal species at study-wide significance and replicated in at least one independent cohort, notably connecting the gene *OPRK1* with the potential pathogen *Ureaplasma urealyticum*. In summary, our research provides a comprehensive reference for future studies on genotype–phenotype interplay within the human vaginal microbiome.

The vaginal microbiome, comprising bacteria, viruses and eukaryotic microorganisms, is critical to reproductive health and the pathogenesis of urogenital disease[1–3]. Over the past decade, 16S rRNA gene amplicon sequencing has become the predominant method for characterizing vaginal bacterial communities, revealing their correlations with host health status and lifestyle factors[4–7]. However, this approach offers limited taxonomic and functional resolution and excludes nonbacterial microorganisms[8–10]. Shotgun metagenomic sequencing overcomes these limitations by enabling species-level and strain-level discrimination of bacteria, capturing archaea, viruses and eukaryotes, and by allowing functional profiling through gene annotation[9,11]. Because accurate taxonomy assignment relies on reference genomes, continuous database expansion is essential for classification accuracy[12].

Although metagenome-assembled genomes (MAGs) and cultivation[13–15] have substantially expanded the characterization of both cultured and uncultured microorganisms across the gut[16–20], oral cavity[21] and skin[22,23], vaginal microbiome research remains limited by insufficient reference data and small-scale studies[24–27].

A full list of affiliations appears at the end of the paper. ✉e-mail: tao.zhang@genomics.cn; haolilan@genomics.cn; zhu_julie@vip.sina.com; chenchen20192022@163.com

This disparity may stem from the massive proportion of human DNA (>90%) in vaginal samples[28] and low microbial diversity, which hinders genome recovery. Consequently, most reference genomes in existing databases are derived from gut microbiomes (for example, MetaPhlAn)[29], constraining functional and ecological insights into the vaginal ecosystem. Recent efforts have begun to address this gap. Following the Human Microbiome Project[30] metagenomic profiling, vaginal genomic resources grew from an initial 60,699-gene catalog[24] to the 0.95-million-gene nonredundant gene catalog (VIRGO)[26]. More recent work has generated 1,078 MAGs from 705 samples (linking microbial diversity to preterm birth[27]) and 33,804 multikingdom genomes from 4,472 public samples in VMGC[25]. Nevertheless, substantial unexplored diversity remains, particularly in non-Western populations.

While environmental factors predominantly shape microbiome composition, host genetic effect is increasingly recognized, albeit largely with gut microbiome[31–34]. Existing vaginal microbiome genome-wide association studies (GWAS)[35,36], which relied on 16S rRNA sequencing and genotyping arrays, lack the comprehensive resolution of shotgun metagenomics and whole-genome sequencing (WGS). Higher-resolution metagenome-GWAS (M-GWAS) are therefore critical to uncover more nuanced genetic associations. Notably, the high host DNA content (>90%) in vaginal metagenomic data offers an unparalleled opportunity to simultaneously extract host whole-genome variants and high-resolution microbial profiles from a single assay.

Here we present the Peacock cohort, a large non-Western dataset of 10,665 metagenomes from Chinese individuals. By integrating these with global public metagenomes, genomes and in-house isolates, we constructed the Global Vaginal Metagenome-Assembled Genomes (GVMG) catalog. This repository captures 890 prokaryotic species (36,059 genomes), 11 fungal species (43 genomes) and 6,577 species-level viral operational taxonomic units (vOTUs; 28,953 genomes). Leveraging the GVMG, we characterized the taxonomic and functional landscape of the vaginal microbiome at the genomic level, delineated high-resolution population genetics and identified complex host–microbiome interactions using M-GWAS. This catalog provides a valuable foundation poised to accelerate global vaginal microbiome research and to deepen our understanding of host–microbe co-evolution.

Results

Extensive Chinese population dataset integrating microbiome, phenotypic and genomic information

To systematically characterize the vaginal microbiome across diverse Chinese populations, we initiated the Peacock Project and obtained 10,665 qualified cervicovaginal swabs through a standardized multicenter protocol (Fig. 1). The cohort encompassed three urban populations—gynecological clinic attendees from Beijing (BJ-GC; n = 2,878; mean age = 42.4 ± 12.0 years), routine health examinees from Shenzhen (SZ-4D; n = 833; mean age = 31.9 ± 4.9 years) and participants from organized cancer screening programs in Suzhou, spanning two periods, that is, 2018–2019 (SU-CCS2018/2019; n = 3,355; mean age = 48.3 ± 7.5 years) and 2021 (SU-CCS2021; n = 3,527; mean age = 50.8 ± 7.7 years; mean ± s.d.; Fig. 1a, Extended Data Fig. 1a and Supplementary Tables 1 and 2). Additionally, 72 samples were collected from other regions. Sequencing depth varied across centers—SU-CCS2018/2019 (190.1 ± 99.6 million reads), SU-CCS2021 (322.0 ± 51.1 million reads), SZ-4D (161.2 ± 35.2 million reads) and most BJ-GC samples (n = 2,372; 184.9 ± 59.9 million reads); thus, 506 BJ-GC samples were sequenced at higher depth (578.4 ± 105.5 million reads; Fig. 1a). In total, 2.5 T raw paired-end sequencing reads were generated (100 of 150 bp; Supplementary Table 2). Owing to >90% host DNA in vaginal metagenomes, host coverage reached 8.4 ± 5.7×, varying by sequencing depth (Supplementary Table 2). Extensive phenotypic data were collected, covering healthy individuals and those with conditions such as bacterial vaginosis (BV), HPV infection, uterine fibroids and uterine prolapse (Extended Data Fig. 1b).

Construction of microbial genomes from global female vaginal metagenomes

We constructed the GVMG catalog by integrating 10,665 metagenomes from the Peacock cohort and 2,967 publicly available metagenomes from 11 studies across the USA (n = 1,769), France (n = 739) and seven other countries (Extended Data Fig. 1c and Supplementary Tables 1 and 2). After quality control, host read removal and assembly, MAGs were generated using multicoverage (MetaBAT2) and three single-coverage binning strategies (MetaBAT2, MaxBin2 and CONCOCT), refined with dRep to remove within-sample duplicates, yielding 39,816 preliminary prokaryotic genomes (Extended Data Fig. 2).

To enhance representation, we incorporated publicly available vaginal microbial genomes, including 5,600 prokaryotic genomes (972 isolates and 4,628 MAGs derived from 1,477 metagenomes compiled by VMGC from 20 studies, not in our metagenome list) from the VMGC study (up to 2023)[25], 36 eukaryotic genomes (17 isolates and 19 MAGs), 384 prokaryotic isolates cultured from 11 high-risk cervical cancer Peacock participants and 77 additional prokaryotic isolate genomes from 2023–2024 literature (Supplementary Tables 3 and 4). After quality filtering, 36,059 prokaryotic genomes (34,801 MAGs, 1,258 isolates) were retained and classified into 3,810 high-quality, 20,752 near-complete and 11,497 medium-quality genomes (mean completeness = 90.97%, mean contamination = 0.71%; Extended Data Fig. 3a–e and Supplementary Table 5). Notably, Chinese-derived MAGs exhibited superior quality compared to French and American subsamples (Extended Data Fig. 3f–h). In comparison to VMGC[25], our catalog contained 1.85 times more prokaryotic genomes, with 68.1% classified as near-complete genomes or high-quality genomes compared to 48.2% in VMGC (Fig. 2a and Extended Data Fig. 4; all comparisons $P < 0.0001$, a two-sided Wilcoxon rank-sum test).

From metagenomic contigs, we identified 28,953 passed-quality viral sequences, with 2,932 (10.3%) complete, 9,121 (31.5%) high-quality and 16,900 (58.4%) medium-quality genomes (CheckV; median completeness = 84.9 ± 16.9%, median size = 34.2 ± 18.7 kb; Extended Data Fig. 5a,b and Supplementary Table 6). For eukaryotes, we recovered seven fungal MAGs from 39,816 preliminary assemblies >3 Mb, annotated as *Candida albicans* (n = 3) and *Nakaseomyces glabratus* (n = 4). Combined with the aforementioned 36 public eukaryotic genomes, a total of 43 eukaryotic genomes representing 11 fungal species were obtained, with *N. glabratus* and *C. albicans* most prevalent in the Peacock dataset (Supplementary Table 3 and Extended Data Fig. 6).

Overall, the GVMG catalog comprises 65,055 genomes, including 36,059 prokaryotic (55.4%), 43 eukaryotic (0.07%) and 28,953 viral genomes (44.5%; Fig. 1b). This represents a 1.9-fold expansion over VMGC[25] and, to our knowledge, constitutes the most comprehensive genomic repository of the human vaginal microbiome to date.

Characterization and distribution of the vaginal prokaryotic species

From the GVMG catalog, we clustered 36,059 prokaryotic genomes into 890 species-level genome bins (SGBs), representing distinct vaginal microbial species and expanding species-level representation (Supplementary Note). Notably, 78.0% (694 of 890) contained at least one near-complete or high-quality MAG (Fig. 2b,c and Supplementary Table 7). Taxonomic assignment using Genome Taxonomy Database Toolkit (GTDB-Tk) identified 889 bacterial species (spanning 18 phyla, 23 classes, 53 orders, 107 families and 336 genera) and 1 archaeal species (*Methanobrevibacter A smithii*; Fig. 2d). Notably, 116 SGBs (13.0%) were classified as unknown SGBs (uSGBs) due to the absence in public databases, while 773 (87.0%) were known SGBs (kSGBs; Fig. 2b). Most uSGBs were assignable to known high-level taxa (spanning 13 phyla, 16 classes, 23 orders, 36 families and 54 genera),

https://doi.org/10.1038/s41588-026-02639-2

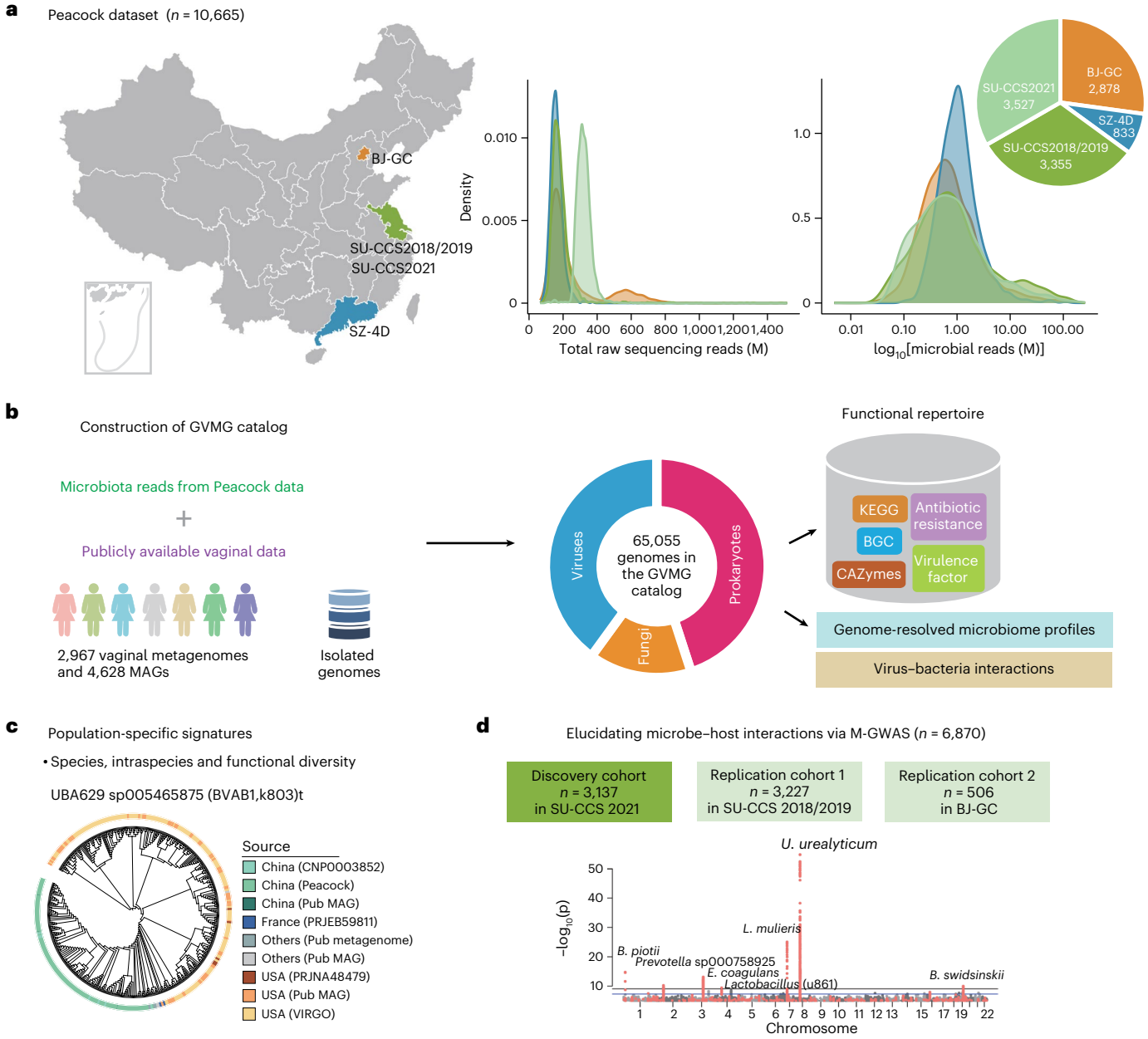


Fig. 1 | Schematic overview of the study design. a, The Peacock dataset featuring metagenomic sequencing of vaginal samples from 10,665 Chinese individuals from three different cities. Samples from women in Suzhou were categorized based on the year of collection, designated as the SU-CCS2018/2019 collection and the SU-CCS2021 collection. Additionally, collections from BJ-GC and SZ-4D were included. The distribution of raw sequencing reads and microbiome reads across the four collections of the Peacock cohort is visualized using density plots, with the area under each curve normalized to 1 and the *y* axis representing probability density, while a pie chart illustrates the distribution of sample numbers across these collections. **b**, The GVMG catalog, integrating 10,665 in-house Chinese metagenomes with 2,967 publicly available metagenomes, 4,628 MAGs and 1,433 bacterial isolates, generating a collection of 65,055 genomes from prokaryotes, eukaryotes and viruses. Leveraging established databases, gene functions within these genomes were annotated, establishing a detailed functional repertoire that enables precise mapping of genomic pathways for biochemical synthesis and their linkage to specific microbial hosts. **c**, Identification of population-specific signatures in species, intraspecies and functional diversity, comparing different populations. **d**, Exploration of the links between host genetic variants and the vaginal microbiome using M-GWAS, incorporating a discovery cohort and two validation cohorts. All statistically significant *P* values for microbial taxa, rather than only the most significant *P* value per taxon, are displayed. Data underlying the plots are provided. M, million; CAZymes, carbohydrate-active enzymes; *E. coagulans*, *Ezakiella coagulans*.

except for 9 uSGBs that could be grouped into 9 new genera and 1 new family (Supplementary Table 7). The uSGBs were widely distributed, comprising 8–13% SGBs in the dominant phyla (Bacillota A, Actinomycetota and Bacteroidota), reaching peak prevalence (48–57%) in Campylobacterota, Patescibacteria and Spirochaetota (Fig. 2e). In Campylobacterota, all uSGBs belonged to *Campylobacter*, associated with campylobacteriosis[37]. In Spirochaetota, most uSGBs belong to *Treponema* genus, including pathogens like *T. pallidum*[38] (Fig. 2e). Furthermore, GVMG contained 344 additional SGBs absent from VMGC (Fig. 2f).

Phylogenetic analysis of *Bifidobacterium* genus revealed two clusters (Fig. 2g)—one containing typical gut-associated species (*Bifidobacterium breve*, *Bifidobacterium longum* and *Bifidobacterium bifidum*)[39–42] and another reclassified from *Gardnerella*[39,40]. While *Gardnerella*

https://doi.org/10.1038/s41588-026-02639-2

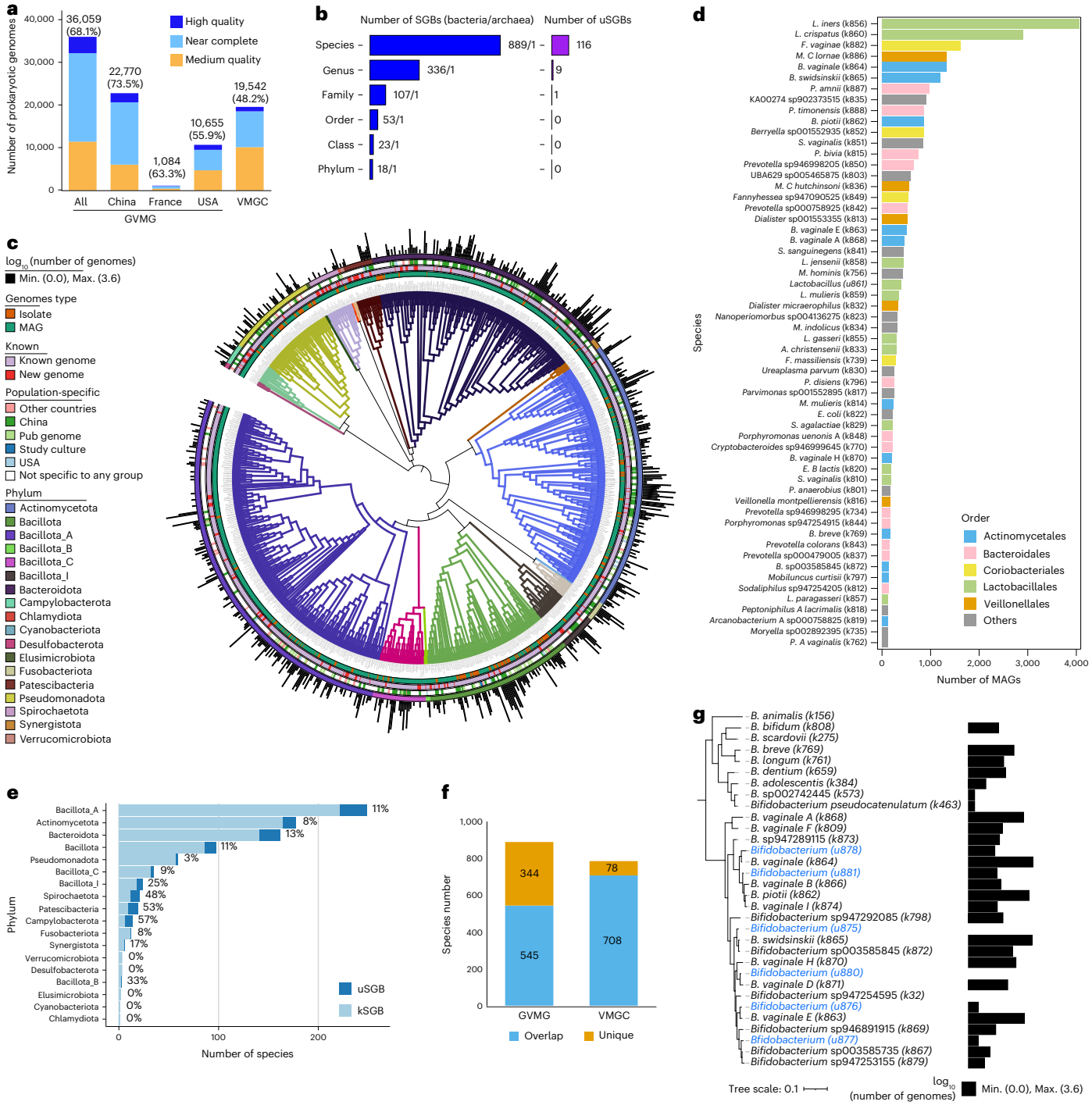


Fig. 2 | Prokaryotic genomes derived from global female vaginal metagenomes form the prokaryotic component of the GVMG catalog. a, Distribution of prokaryotic genomes quality in the GVMG catalog (including China, France and the USA) compared to the VMGC catalog. The number above each bar shows the total number of genomes counts and the percentage in brackets shows the proportion of high-quality and near-complete genomes in total. **b**, Detailed taxonomic composition of 890 total prokaryotic SGBs and uSGBs showing the number of taxa at various levels. Left: the quantity of bacteria before the dash, and the quantity of archaea after the dash. Right: the quantity of uSGBs. **c**, A phylogenetic tree constructed using 889 bacterial representative genomes, each from a distinct SGB. **d**, Number of MAGs within each SGB, highlighting the 58 SGBs with over 100 genomes. **e**, Distribution of kSGBs and uSGBs across phyla, with percentages to the right of each bar representing the proportion of uSGBs within the corresponding taxa. **f**, Comparative analysis of prokaryotic species overlaps between the GVMG catalog and VMGC catalog. Species from the two catalogs with ANI >95% are considered to be the same species. **g**, The *Bifidobacterium* genus is phylogenetically divided into two clusters. The lower cluster, originally classified under the *Gardnerella* genus, was recently reclassified into *Bifidobacterium* and designated as *B. vaginale* genomospecies. This cluster contains 23 SGBs, of which 6 uSGBs are highlighted in blue. Data underlying the plots are provided.

vaginalis was long considered the sole *Gardnerella* species[39,40], our findings revealed 23 SGBs (17 kSGBs and 6 uSGBs) in this second cluster, including 8 subtypes of *Bifidobacterium vaginale*, *Bifidobacterium piotii*, *Bifidobacterium swidsinskii* and *Bifidobacterium* spp., reflecting known genetic heterogeneity among *Gardnerella* members[40–42]. We unified this cluster as *B. vaginale* genomospecies (Fig. 2g and Extended Data Fig. 7) and demonstrated their association with BV (Supplementary Note).

At the species level, 58 SGBs contained >100 genomes (Fig. 2d and Supplementary Table 7). *Lactobacillus iners* and *Lactobacillus crispatus* were the two with the highest number of genomes, constituting 14.2% and 8.1% prokaryotic genomes, respectively. Several reproductive tract diseases-associated species[43] were also prevalent, including seven *B. vaginale* genomospecies (for example, *B. vaginale*, *B. swidsinskii* and *B. piotii*), *Fannyhessea vaginae*, *Megasphaera C lornae*, *Prevotella* spp. (for example, *Prevotella amnii*, *Prevotella timonensis* and *Prevotella bivia*), BVAB1 (*Lachnospiraceae*-UBA629 sp005465875 (ref. 44)) and BVAB3 (*Mageeibacillus indolicus*[45]), and less-characterized species like Fastidiosipilaceae-KA00274 sp902373515 (905 MAGs) and Eggerthellaceae–*Berryella* sp001552935 (857 MAGs). Taxonomic abundance profiles identified 12 vaginal community state types (CSTs[46]; Extended Data Fig. 8), among which 5 were predominantly characterized by distinct *B. vaginale* genomospecies (*B. vaginale*, *B. swidsinskii*, *B. piotii*, *B. vaginale* with co-occurring *B. swidsinskii*, and *B. vaginale* E with co-occurring *B. vaginale*). This refined classification contrasts with prior studies that did not differentiate *B. vaginale* into distinct genomospecies[7,47]. Notably, the BVAB1-dominated CST exhibited the highest Shannon diversity. Its prevalence was markedly lower in the Chinese cohorts (0.46%) compared to the USA cohort (12.38%; $P < 0.001$; Extended Data Fig. 8), suggesting potential geographic influences that warrant further investigation.

Taxonomic landscape of vaginal virus species

In investigating the vaginal virome, we classified 28,953 viral sequences into 6,577 vOTUs (Supplementary Table 8; Methods). Our catalog contained 2.04 times more genomes than VMGC[25]. However, 66.0% vOTUs comprised a single genome, mirroring the VMGC and suggesting that, although multigenome vOTUs were nearing saturation, the overall accumulation curve has not yet plateaued (Extended Data Fig. 5c). Cross-referencing with VMGC[25] and five large-scale nonvaginal virome databases, including NCBI RefSeq, human gut (GVD, GPD, MGV)[48–50] and oral (OVD) database[51], 56.5% GVMG vOTUs were unique, markedly expanding known viral diversity (Fig. 3a). While the composite nonvaginal database contained 29.0 times more vOTUs, only 16.2% overlapped with the GVMG. Furthermore, our GVMG captured 52.0% VMGC vOTUs, while the remainder were excluded due to stringent quality filters and the inclusion of only 66.3% VMGC samples, highlighting considerable interindividual viral heterogeneity[50]. Taxonomic annotation identified 195 vOTUs across 20 prokaryotic viral families and 87 vOTUs across 11 eukaryotic viral families (Supplementary Table 8). Notably, 79.0% vOTUs lacked family-level annotations, highlighting gaps in current vaginal virome taxonomy (Fig. 3b). Among these, 12.6% were annotated to *Aliceevansviridae*, followed by *Microviridae* (2.2%) and *Papillomaviridae* (1.0%; Fig. 3b,c). Of the 87 eukaryotic vOTUs, we focused on 64 clinically relevant *Papillomaviridae* (HPV) vOTUs (Fig. 3d and Supplementary Table 9). Using the major structural L1 open reading frame[52], these comprised 58 known HPV types, 4 unclassified types and 2 potentially new HPV homologs. Among the four most prevalent types, HPV52 exhibited the strongest population stratification in genomic evolution (Fig. 3e).

Viral-prokaryotic co-occurrence networks confirmed that most viruses correlated with their predicted bacterial hosts (Fig. 3f, Extended Data Fig. 5d and Supplementary Note). The three most abundant vOTUs aligned with dominant prokaryotes—vOTU0198 and vOTU0488 correlated positively with *L. iners*, while vOTU0627 correlated with *L. crispatus* (Fig. 3f and Supplementary Fig. 3). Assessing associations between viral profiles and phenotypes (HPV infection, BV and menopause) in the SU-CCS cohort revealed that vOTUs negatively correlated with these phenotypes were predominantly predicted to infect *Lactobacillus* (menopause = 73.8%, HPV = 75.0%, BV = 100%). Specifically, vOTU0627, vOTU0193 and vOTU0512 consistently ranked as the top vOTUs targeting *Lactobacillus* species across all three phenotypes, exhibiting the strongest correlations with *L. crispatus* in healthy reproductive-age women. HPV infection correlated positively with 27 HPV vOTUs, particularly HPV52, HPV58 and HPV16. Notably, HPV52 was also significantly correlated with menopause ($P_{adj} < 0.05$, Fig. 3g and Supplementary Table 10). Here vOTU0023 (phage targeting *Metamycoplasma hominis*) was strongly positively correlated with all three phenotypes ($P_{adj} < 10^{-11}$, coefficient = 0.20–0.34). A consistent pattern emerged—disease-associated phages avoided *Lactobacillus* but selectively targeted pathobionts dominating in specific states. For instance, BV-associated phages preferentially targeted *B. vaginale* genomospecies (33.3%) and *Dialister* species (20.0%), menopause-associated phages predominantly targeted *Prevotella* (50%) and HPV-associated phages targeted *Prevotella* (26.3%) and *B. vaginale* genomospecies (15.8%). These findings suggest that phages may adapt to the altered microbial environment in different pathological conditions.

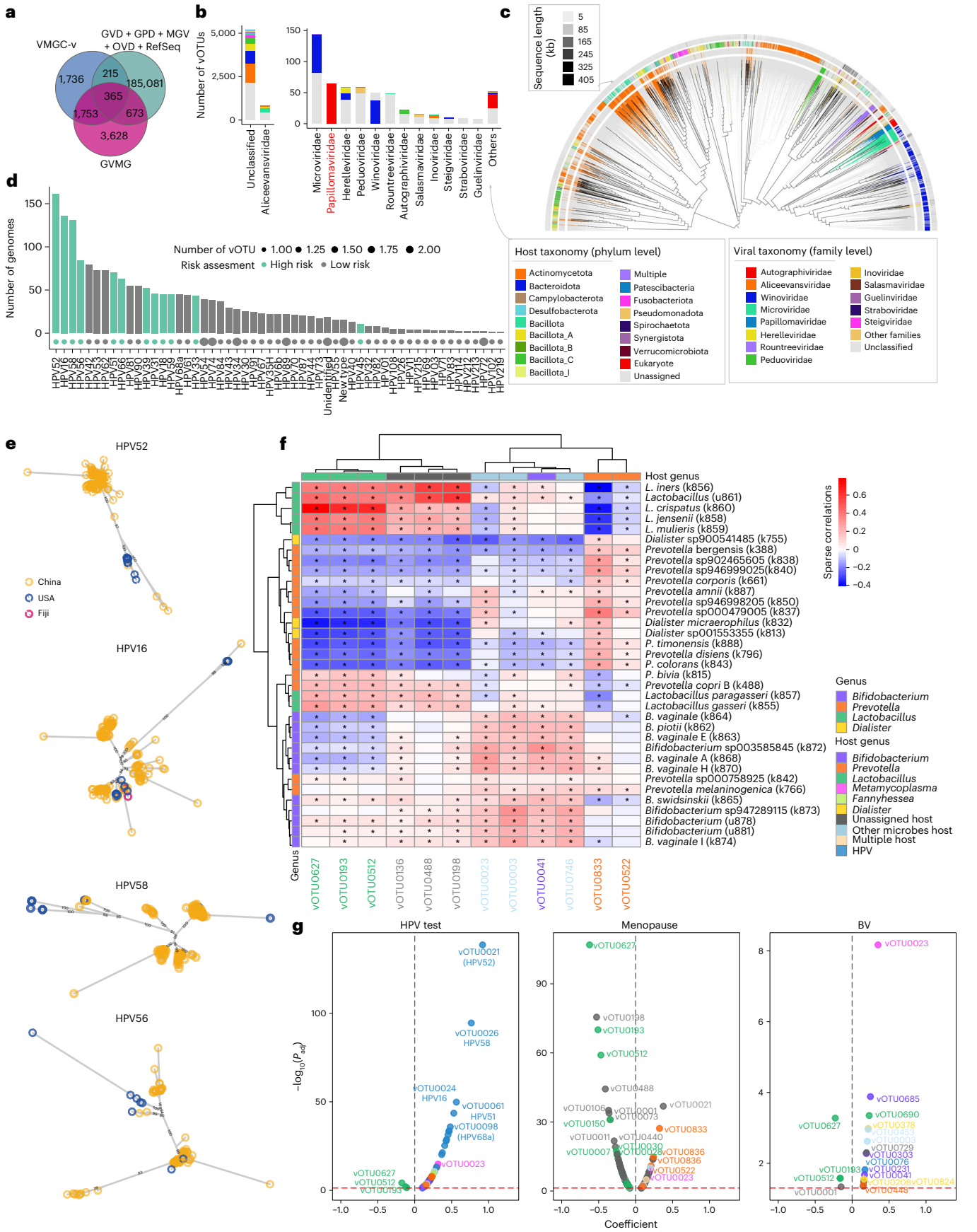
Functional repertoire of the female vaginal microbiome

To elucidate functional potential, we annotated genes in prokaryotic and viral genomes. Among 312,189 viral genes, 41.2% were assigned to Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs, primarily involving genetic information processing (44.1%), environmental information processing (25.0%) and viral auxiliary metabolic genes (14.3%) linked to peptidases and inhibitors, nucleotide and peptidoglycan metabolism (Extended Data Fig. 9a,b). For prokaryotes, 59.4% protein-coding genes mapped to the KEGG database[53], and 2.15% to the carbohydrate-active enzymes database[54]. Smaller subsets encoded virulence factors (0.79%), antibiotic resistance (0.11%) and antimicrobial peptides (0.05%), demonstrating functional diversity.

To delineate the functional landscape of BV, we analyzed key enzymatic activities (for example, sialidases and cytolysin), along with metabolic pathways involved in biogenic amines and short-chain fatty acids (for example, lactate, succinate and butyrate metabolism; Fig. 4a) categorized using VMGC[25] as a reference. *L. crispatus* was characterized by high prevalence of lactate production genes but minimal

Fig. 3 | Features of viral communities in the GVMG. a, Overlap of the viral species among GVMG, VMGC and integrated data from five large-scale nonvaginal virome databases. **b**, Distribution of host phyla for 6,577 vOTUs. The bar plot shows the distribution of host phyla for different viral families, with the eukaryotic viral families colored red. **c**, Tree illustrating the phylogenetic relationships among 6,577 vOTUs. **d**, Number of genomes for different HPV types among 64 vOTUs annotated as Papillomaviridae. **e**, Phylogenetic trees of the top four HPV types in the GVMG, revealing distinct population stratification. **f**, Co-occurrence heatmap constructed to evaluate correlations between prokaryotic SGBs and vOTUs in the SU-CCS2021 (Peacock, $n = 3,527$ samples) cohort with SparCC, with 1,000 iterations of permutation testing. Row labels indicate the genus of each SGB, while column labels and text denote the predicted host for each vOTU. Each cell color indicates the magnitude and direction of the SparCC correlation. Asterisks indicate a significant correlation with two-sided FDR $P_{adj} < 0.05$ (BH-adjusted method). **g**, Volcano plots illustrating the relationships between viral profiles and phenotypes (including clinically diagnosed HPV infection (cases, $n = 861$; controls, $n = 4,579$), menopause (cases, $n = 2,053$; controls, $n = 3,055$) and BV (cases, $n = 360$; controls, $n = 5,031$)) within the SU-CCS cohort, analyzed using the Generalized Linear Model method. The dots above the red dashed lines indicate significant vOTUs with two-sided BH $P_{adj} < 0.05$, the color of the dots and corresponding vOTU IDs matching the predicted host genera in **f**. Data underlying the plots are provided. BH, Benjamini–Hochberg.

https://doi.org/10.1038/s41588-026-02639-2



sialidase and cytolysin genes; these enzymes disrupt the vaginal epithelium, with sialidase degrading mucin glycosylation chains to facilitate cytolysin-mediated host cell lysis and resources mobilization[55]. These findings suggest that *L. crispatus* does not compromise the protective mucus barrier. Conversely, 96.1% *L. iners* genomes harbored cytolysin genes, consistent with its higher expression in diverse communities at low-to-moderate abundances[56]. Notably, *B. vaginale* genomospecies and *Prevotella* species exhibited generally higher prevalence in genes encoding sialidases and cytolysins, with substantial inter-SGB variation (Fig. 4b). Over 90% genomes in *B. vaginale* H and *Prevotella* species (*P. bivia*, *Prevotella* sp946999045 and *Prevotella* sp946998205) harbored genes encoded with both enzymes. In contrast, *B. swidsinskii*, *B. vaginale* A, *B. vaginale* D, *Bifidobacterium* sp947292085 and *Prevotella* sp946998295 predominantly exhibited high prevalence of cytolysin genes (>90%) but minimal sialidase (<1%), whereas species like *Prevotella* sp001553265, *P. amnii*, *B. vaginale* B, *B. breve* and *B. bifidum* were contained high prevalence of sialidase genes. Similar patterns were observed in *Mobiluncus curtisii*, *Gemella asaccharolytica*, *Streptococcus agalactiae*, *Winkia anitrata* and *Porphyromonas gingivalis* (Fig. 4b), indicating complex interspecies interactions in sialidase-mediated and cytolysin-mediated vaginal pathology. We also evaluated BV-associated biogenic amine metabolism and antibiotic resistance (Supplementary Note).

Exploring prokaryotic biosynthetic gene clusters (BGCs), we annotated 2,211 nonredundant gene cluster families (GCFs; Extended Data Fig. 9c and Supplementary Table 11). After *Escherichia coli*, *Prevotella* species (*P. timonensis*, *Prevotella* sp946998205 and *P. amnii*) exhibited the highest GCFs counts and were particularly enriched for genes encoding enzymes with the potential to synthesize arylpolyene and resorcinol (Extended Data Fig. 9d,e). Next, evaluating intraspecies BGC variation between the Chinese and American populations (Fig. 4c), BVAB1 exhibited the most pronounced population divergence. Specifically, BGCs were significantly elevated in the American cohort—LAP (BGC0031/0281/0134), RiPP-like (BGC0122/0034) and ranthipeptide (BGC0381) (Wilcoxon rank-sum test, $P_{adj} < 0.05$). Integrating BGCs with KEGG ortholog annotations identified the following four key pathogenicity-related functional modules in BVAB1: (1) a toxin production—erythrogenic toxin A gene arginine decarboxylase (*speA*; K01585 from BGC003) related to scarlet fever rash[57], membrane-damaging toxin *Tly*C (K03699 from BGC0031) and the *sag*BCD protoxin formation gene cluster (BGC0031/0134/0281)[58,59]; (2) antibacterial defense mechanisms featuring enterocin A immunity and class II bacteriocin production (BGC0034/0122); (3) adhesion enhancement—alcohol dehydrogenase encoded by aldehyde-alcohol dehydrogenase (*adhE*; K04072 from BGC0031) mediating flagella expression[60], the *Scf*AB membrane protein complex (BGC0031)[61]; and (4) stress adaptation—the *DinJ*–*yaf*Q toxin-antitoxin module (K07473 from BGC0031 (ref. 62) and Supplementary Table 11). These findings highlight population-specific variations in BVAB1's pathogenic potential.

Intraspecies phylogenetic analysis highlights population genetic diversification

Using patristic distances inferred from the core-gene phylogenetic trees within each SGB, we analyzed the influence of geography on species phylogeny. Among 75 SGBs containing over 50 genomes, 62 exhibited significant geographic genetic variation (PERMANOVA, $P_{adj} < 0.05$). Geography exerted the strongest explanatory power on non-*Lactobacillus* species, primarily *Arcanobacterium* A sp000758825, *B. vaginale* D, *Fannyhessea massiliensis*, BVAB1, *B. breve* and BVAB3. Among six *Lactobacillus* species, geographic factors most strongly influenced *Lactobacillus jensenii* and *Lactobacillus mulieris*, followed by *L. iners*, while *L. crispatus* was least affected, indicating substantial genomic conservation in *L. crispatus* compared to the other *Lactobacillus* species (Fig. 5a–d and Supplementary Table 12). Notably, genomes from the same region clustered together regardless of their study cohort or sequencing platform, confirming that true geographic divergence supersedes technical batch effects at the intraspecies level (Fig. 5b–d and Supplementary Fig. 5).

Pronounced single-nucleotide polymorphism (SNP) disparities further highlighted this population divergence. In BVAB1, we identified 39,052 population-differential SNPs from 1,066 genes between Chinese and USA populations (Fig. 5e and Supplementary Table 13; chi-squared test). Notably, 694 of these genes (harboring 1–350 SNPs each) carried specific functional annotations. Abundantly mutated genes in prokaryotes were heavily enriched in key metabolic pathways, including DNA replication genes such as DNA polymerase III subunit α (*dnaE*), DNA polymerase III PolC-type (*polC*) and DNA polymerase I (*polA*)[63]; DNA damage response and repair genes including UvrABC system protein A (*uvrA*)[64], DNA mismatch repair protein MutL (*mutL*)[65], DNA repair protein RadA (*radA*)[66] and transcription-repair-coupling factor (*mfd*)[67]; glycogen metabolism genes including glycogen debranching enzyme (*glgX*)[68] and PTS system fructose-specific EIIABC component (*fruA*); tRNA ligase genes valine–tRNA ligase (*valS*) and alanine–tRNA ligase (*alaS_2*); and the protease gene ATP-dependent Clp protease proteolytic subunit (*clpP*)[69] in prokaryotic cells. Significant variations (Bonferroni-corrected $P \leq 0.05$/(number of SNPs tested)) also emerged in antibiotic resistance genes, including penicillin-binding protein 1A (*mrcA*) and penicillin-binding protein 4 (*pbpD*) associated with penicillin resistance[70], mupirocin-resistant isoleucine–tRNA ligase MupB (*mupB*)[71] and DNA-directed RNA polymerase subunit β that confers rifampin resistance (*rpoB*)[72]. Variation was further detected in key regulatory or virulence-associated genes, including *speA* associated with erythrogenic toxin A, *adhE* expressing acetate-stimulated flagella, flagellar hook-associated protein 2 (*fliD*) associated with mucin adhesion[73] and serine/threonine-protein kinase PrkC (*prk*C) regulating spore germination and biofilm formation[74,75]. Similarly, *L. iners* and *L. crispatus* also exhibited significant population-specific differences (Supplementary Note). Together, these findings indicate profound variability in geographic strain, underscoring the GVMG as a robust resource for high-resolution intraspecies exploration.

Host genetics strongly associated with vaginal bacteria

We assessed host genetic influence on the vaginal microbiome through M-GWAS in a discovery cohort (SU-CCS2021, $n = 3{,}137$; average = 32.3 Gb) and two validation cohorts (SU-CCS2018/2019, $n = 3{,}227$, 18.2 Gb; BJ-GC, $n = 506$, 86.7 Gb; Extended Data Fig. 10). The high reliability of metagenome-derived host genotypes was evaluated (Supplementary Note). We tested 5.46 million human genetic variants (minor allele frequency ≥ 5%) against 54 vaginal microbial species meeting prevalence >10% and mean relative abundance $>1 \times 10^{-4}$ (Supplementary Table 14). We identified 18 independent loci (linkage disequilibrium of $r^2 < 0.1$ within 1 Mb) showing genome-wide significance ($P < 5 \times 10^{-8}$; Fig. 6a and Supplementary Table 15). Seven genomic loci reached the study-wide significance ($P < 9.26 \times 10^{-10}$ after adjusting for 54 M-GWAS tests) in the discovery cohort and were robustly replicated in at least one validation cohort with consistent effect direction ($P < 0.0028$ after correcting for 18 lead loci).

The strongest and replicated association was between *OPRK1* and *Ureaplasma urealyticum* (chr8: 53,244,232, $\beta = 1.24$, $P = 1.50 \times 10^{-55}$; Fig. 6b and Supplementary Fig. 10). The lead variant is associated with *OPRK1* expression levels in the brain cerebellum and cerebellar hemisphere (Supplementary Table 16), aligning with prior evidence implicating *OPRK*1 in stress responses modulation, stress-induced craving and relapse susceptibility in cocaine dependence[76]. The variant has also been associated with elevated *Ig*E levels in asthmatic individuals[77] (Supplementary Table 17), increased cervical cancer risk ($\beta = 0.16$, $P = 0.01$) in the BioBank Japan Project cohort and positive association with blood lymphocyte, monocyte and progesterone levels in the Chinese 4D-SZ cohort[31,34]. Summary-data-based Mendelian randomization (SMR) analysis further indicated epigenetic regulation

https://doi.org/10.1038/s41588-026-02639-2

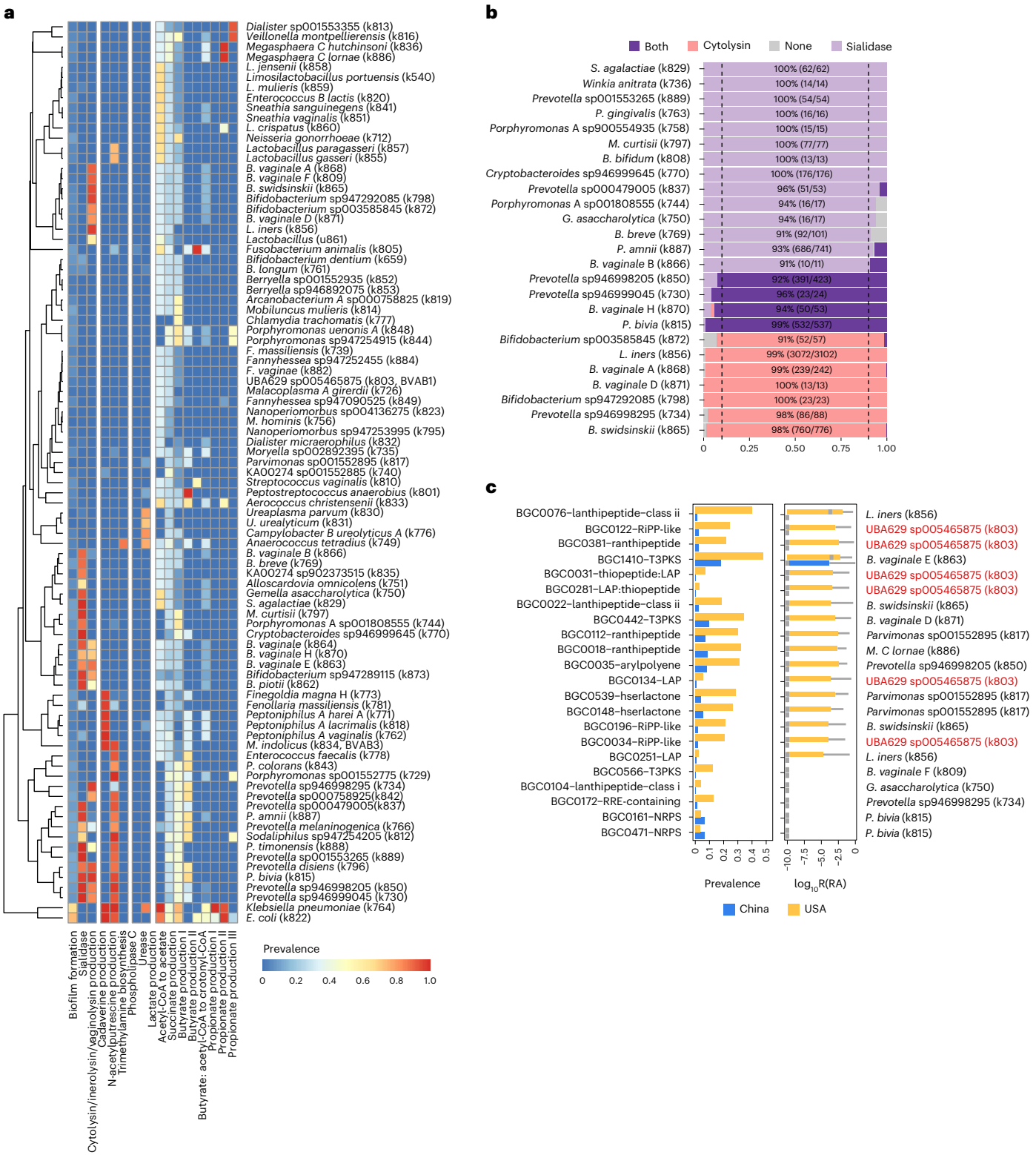


Fig. 4 | Functional characteristics of 889 bacterial SGBs in the GVMG. **a**, Distribution of various BV functional modules and antibiotic resistance modules across SGBs from vaginal samples. The heatmap represents the prevalence, expressed as the percentage of MAGs containing the genes of the corresponding modules, for each SGB with a minimum of 30 genomes. **b**, Gene prevalence of SGBs encoding sialidases and cytolysins. We collected all high-quality and near-complete MAGs associated with a given SGB. The prevalence of these MAGs harboring only genes encoding either sialidase or cytolysin, both or none were calculated, labeled by color. **c**, Association of bacterial GCFs with the host origin. Bar plot illustrating the disparity in GCF prevalence across countries (BH $P_{adj} < 1 \times 10^{-10}$, two-sided Pearson's chi-squared test; USA, $n = 1{,}769$ individuals; China, $n = 10{,}385$ individuals), while the box plot depicts the differences in GCF relative abundance between countries (BH $P_{adj} < 1 \times 10^{-10}$, two-sided Wilcoxon rank-sum test; USA, $n = 1{,}769$; China, $n = 10{,}385$). Relative abundance values are $\log_{10}$-transformed after an increment of 1×10^{-10}. GCFs from UBA629 sp005465875 (BVAB1) with notable toxicological features highlighted in red. The box plots show the median (gray line) and the IQR (yellow box), and whiskers extend to the values no larger than 1.5× the IQR (upper whisker) or smaller than 1.5× the IQR (lower whisker). Outliers are omitted. Data underlying the plots are provided.

https://doi.org/10.1038/s41588-026-02639-2

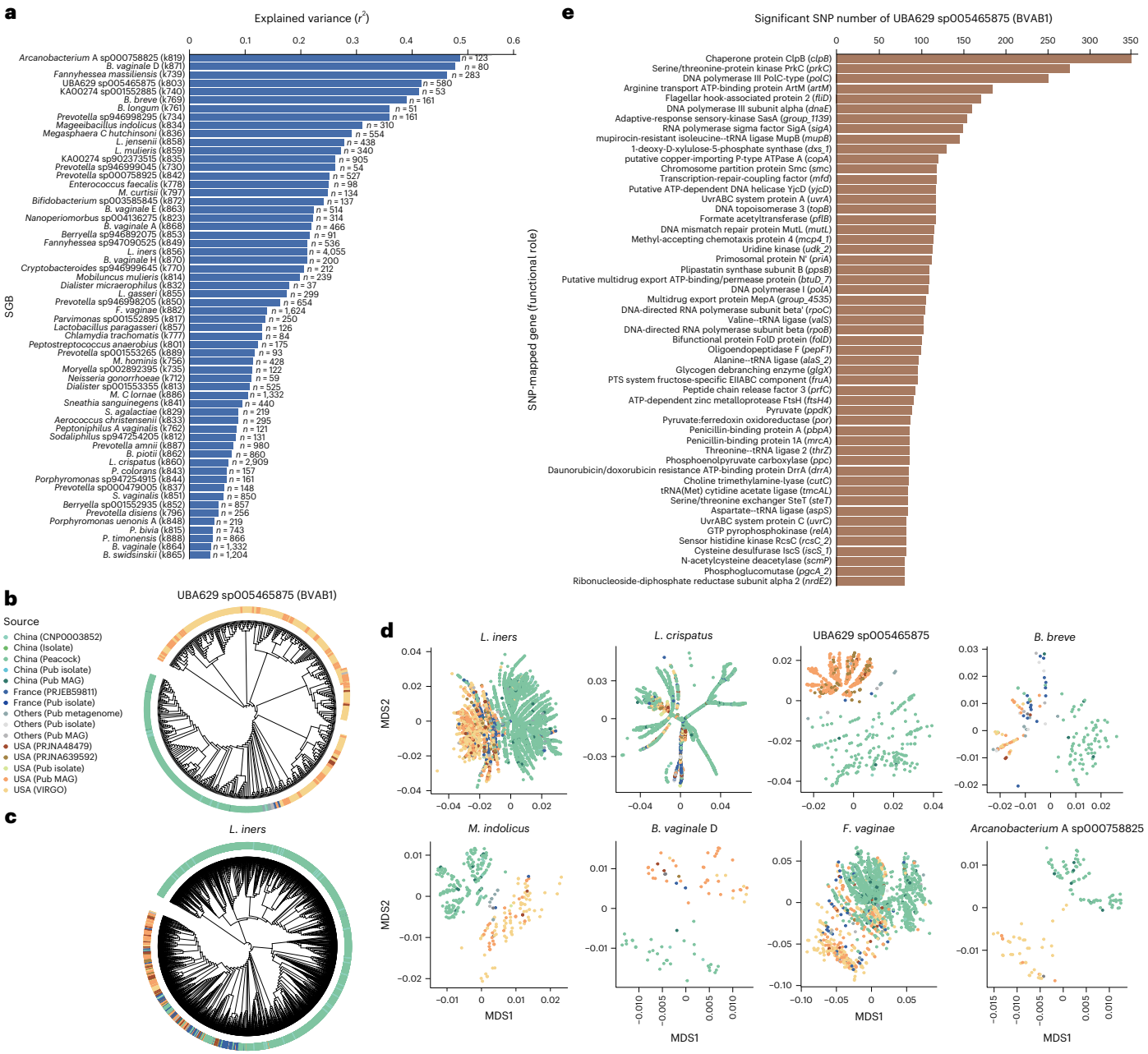


Fig. 5 | Intraspecies genetic diversity across different geographic populations. **a**, Variance explained by geography (r^2) in genetic diversity across 60 SGBs (>50 genomes), exhibiting significant differences (PERMANOVA test with 1,000 iterations of permutation testing, BH $P_{adj} < 0.05$), based on patristic distances derived from each SGB's core-gene phylogenetic tree. The number of genomes within each SGB is indicated in parentheses. **b**,**c**, Approximately maximum-likelihood phylogenetic trees for BVAB1 (**b**) and *L. iners* (**c**) demonstrating clear geographic genetic differentiation among populations, including Chinese, American and French individuals and isolate genomes of undefined origin from public repositories. Intraspecies-level differences caused by study cohorts or sequencing platforms are not included. **d**, NMDS analyses of genetic distances within eight representative SGBs revealing pronounced genetic divergence across populations. **e**, Top 45 genes ranked by the number of population-specific significant SNPs (Bonferroni-corrected $P \leq 0.05/$(number of SNPs tested)) within BVAB1, revealing genes with the greatest differentiation across populations. Data underlying the plots are provided. NMDS, nonmetric multidimensional scaling.

of *OPRK1* (Supplementary Table 18). These findings corroborate the established role of *U. urealyticum* in urogenital infections[78–81] and cervical carcinogenesis[82–84].

The second strongest and replicated association was between *ADAP1* and *L. mulieris* (chr7: 912,256, $\beta = -0.95$, $P = 7.63 \times 10^{-26}$; Fig. 6b and Supplementary Fig. 10). This SNP modulated *ADAP1* expression in multiple tissues and represented a strong expression-based SMR association with *L. mulieris* (Supplementary Tables 16 and 18). Additionally, this significant SNP was associated with neutrophil and lymphocyte counts in the GWAS Catalog (Supplementary Table 17) and vitamin E and mercury levels in 4D-SZ cohort[31,34]. Notably, *ADAP1* has previously been implicated in Alzheimer's disease pathogenesis, cancer progression and human immunodeficiency virus (HIV) reactivation[85].

The third strongest and replicated signal was for the SNP chr1: 12,806,515 near *PRAMEF1*, associated with *B. piotii* ($\beta = 1.05$, $P = 1.05 \times 10^{-16}$; Fig. 6b and Supplementary Fig. 10). The SNP is associated with *PRAMEF11* and *LINC01784* expression levels in testis, correlates with BMI (GWAS Catalog) and with oestrone, serum uric acid and red blood cell counts (4D-SZ, $P < 0.05$). *PRAMEF11* methylation

was also associated with *B. piotii*, suggesting a potential epigenetic regulatory mechanism.

In addition to the three well-replicated associations in both validation cohorts, four additional associations were confirmed in either of them (Fig. 6b). The fourth significant SNP chr3: 103,624,312 near *MIR548AB* exhibited an association with *Ezakiella coagulans* ($\beta = 0.67$, $P = 6.42 \times 10^{-14}$). The fifth significant SNP chr2: 48,843,266, in the intergenic region of *STON1–GTF2A1L*, *LHCGR* and *FSHR*, was associated with *Prevotella* sp000758925 ($\beta = -0.67$, $P = 5.74 \times 10^{-11}$). SMR analysis confirmed the association between *LHCGR* methylation and *Prevotella* sp000758925 (Supplementary Table 18). *LHCGR* is critical for ovulation and corpus luteum maintenance[86]. It is also expressed in various tissues of the reproductive tract such as endometrial cells[87,88]. The sixth significant SNP chr19: 28,732,092 near *MAN1A2P1* and *UQCRFS1* was linked to *B. swidsinskii* ($\beta = -0.34$, $P = 9.14 \times 10^{-11}$). This SNP demonstrated a methylation QTL effect on *UQCRFS1*, a gene implicated in breast and ovarian cancer pathogenesis[89,90]. The seventh significant SNP chr4: 53,916,830 in the intergenic region of *RPL21P44*, *CHIC2* and *PDGFRA* was associated with the expression of these genes. This SNP was associated with *Lactobacillus* (u861). These findings reinforce host genetic contributions to vaginal microbial composition and call for further experimental research to delineate the underlying molecular interactions.

Discussion

As the largest metagenomic shotgun sequencing study of the female vaginal microbiome to date, we constructed a comprehensive catalog comprising 65,055 microbial genomes, including 890 prokaryotic SGBs, 6,577 vOTUs and 11 fungal species. Notably, 13.0% prokaryotic species and 79.0% vOTUs were absent from public repositories. While pioneering resources like VIRGO[26] and VMGC[25] established foundational genomic sources, the largest gene repository and multikingdom genome collection, they were also constrained by demographic and geographical biases. Specifically, these studies predominantly focused on Western populations and pregnant women (54.8% of 3,107 public samples), which critically lack healthy individuals (<25%), menopausal cohorts and disease spectra. To address these gaps, our GVMG catalog integrated public data and more than 10,000 newly sequenced Chinese vaginal samples, totaling 13,632 metagenomic samples. Achieving a threefold increase in sample size and a 1.9-fold increase in genome count compared to the VMGC, our catalog provided broad lifecycle coverage (3,890 menopausal women aged 50–85 years), expanded disease spectrum coverage combining hospital-based and population-based cohorts, and deepened geographical representation for Asian populations. Ultimately, the GVMG fills a critical gap for Asian populations and provides a foundational platform for global vaginal microbiome research and microbe–host interaction studies.

Expanding upon recent systematic profiles of the broader vaginal virome (as VMGC[25]), our GVMG catalog reveals fundamental ecological patterns in vaginal phage–bacteria interactions. We then observed phenotype-specific targeting—in healthy reproductive-age women, *Lactobacillus* species and their corresponding phages were co-enriched, whereas under clinical conditions (HPV infection, BV and menopausal status), phages preferentially target non-*Lactobacillus* bacterial groups. Comparative analysis further suggests that phages infecting *B. vaginale* genomospecies may possess broader host ranges than those targeting *Lactobacillus*, potentially enhancing *B. vaginale*'s ecological adaptability. These findings collectively indicate that vaginal phages have important but phenotypic context-dependent roles in shaping microbial communities, although further research is needed to fully elucidate these dynamics.

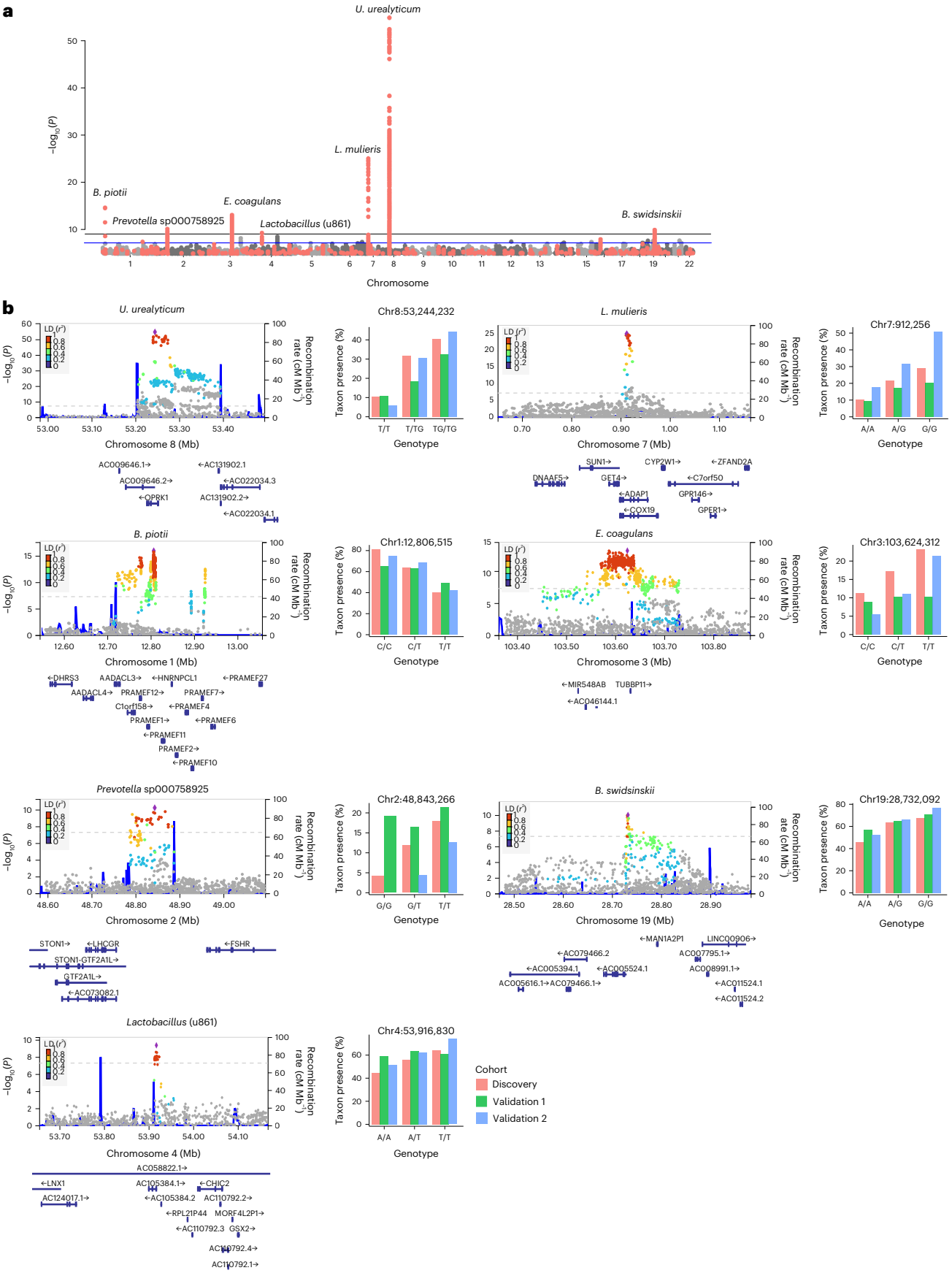
Microbial genomes within the same species can exhibit substantial variation due to the adaptation to diverse environments[27,91]. Although factors such as pregnancy, ethnicity and reproductive tract infections have been extensively studied in relation to the vaginal microbiome[43,55,92], most research has emphasized species-level abundance, with limited exploration of intraspecies-level or genomospecies-level variation[27,93]. The GVMG catalog addresses this gap by providing a comprehensive resource that captures both phylogenetic and functional diversities, enabling finer-scale resolution. For instance, population genetic analysis revealed that most vaginal species exhibit significant but variable degrees of geographic genetic divergence, highlighting ethnogeographic variation at the intraspecies level. *L. iners* showed greater geographic divergence than *L. crispatus*, supporting the hypothesis that *L. iners* diversified around the time of human migration out of Africa, whereas *L. crispatus* diverged after European settlement[93–96]. In contrast, *B. vaginale* likely diversified before human migration, resulting in pronounced divergence[27,93,97,98], reflected in the 23 *B. vaginale* genomospecies identified in GVMG. From a viral perspective, phages infecting different *B. vaginale* genomospecies exhibit greater cross-genomospecies connectivity than those targeting *Lactobacillus*, suggesting enhanced adaptability and evolution in *G. vaginalis*. Notably, geographic divergence varies considerably among genomospecies, with the three most prevalent (*B. vaginale*, *B. swidsinskii* and *B. piotii*) showing the least impact. Functionally, BV-associated bacteria such as *B. vaginale* and *Prevotella* spp. frequently carry genes encoding sialidase and cytolysin, which synergistically disrupt the epithelial barrier by degrading mucin and damaging epithelial cells[55,99]. The co-occurrence of multiple *B. vaginale* genomospecies within the same niche aligns with ecotype theory[93,100–102], suggesting that these closely related lineages may partition resources to coexist.

Another clinically notable bacterium is BVAB1, whose role in vaginal dysbiosis and preterm birth risk is increasingly recognized[3]. Microbial communities dominated by BVAB1 showed the highest Shannon diversity among the 12 CSTs. Interestingly, while this CST accounts for 12.4% samples in the VIRGO cohort, its prevalence was markedly lower (0.46%) in the Peacock cohort, likely explaining the rare reports of BVAB1 in studies of Chinese populations[26,28,92,102,103]. Additionally, BVAB1 from the American population exhibited significantly greater enrichment of virulence-related BGC functions compared to Chinese population, including toxin production, antimicrobial activity, adhesion and stress resistance. Population-specific SNP differences were also observed in genes involved in antibiotic resistance, spore germination and mucin adhesion. These genetic signatures suggest that BVAB1's pathogenic potential may differentially influence reproductive tract infection outcomes across ethnic populations. Together, GVMG provides a high-resolution vaginal microbiome resource that enables researchers to explore strain-level variation across large sample panels and specific clinical contexts, facilitating targeted investigation of biomedically relevant species and their functions in the human vagina.

Leveraging host sequences in vaginal metagenomes, we applied M-GWAS in a large-scale cohort to explore host genetic influences on the microbiome. *U. urealyticum*, *L. mulieris* and *B. piotii* exhibited robust associations with host genetic variations across all three independent

Fig. 6 | Genome-wide associations of host genetics and the vaginal microbiome identified by M-GWAS. a, Manhattan plot showing *P* values from the primary discovery M-GWAS (SU-CCS2021, $n = 3{,}137$) and the three replicated signals in both validation cohorts (U-CCS2018/2019, $n = 3{,}227$; BJ-GC, $n = 506$). All statistically significant *P* values for microbial taxa are displayed. The *y* axis indicates $-\log_{10}$-transformed *P* values, while the *x* axis indicates the genomic positions of the host genetic variants. Study-wide ($P < 9.26 \times 10^{-10}$) and genome-wide ($P < 5 \times 10^{-8}$) significance thresholds are marked by black and blue horizontal lines, respectively. Variants associated with taxa that were replicated in the two validation cohorts are indicated by pink dots. **b**, Regional plots displaying M-GWAS results of the top seven well-replicated signals and bar plots based on the presence–absence status of taxa, along with their corresponding host genetic loci in the discovery cohort and two validation cohorts. Data underlying the plots are provided. LD, linkage disequilibrium.

https://doi.org/10.1038/s41588-026-02639-2



datasets. *U. urealyticum*, detected in 26% samples, exhibited a strong association with the *OPRK1*, an opioid receptor gene implicated in HIV outcomes[104]. *L. mulieris*, the fourth most abundant *Lactobacillus* species, associated with *ADAP1*, a gene involved in HIV-1 reactivation through T cell signaling[85]. *B. piotii*, an important *B. vaginale* genomospecies present in 42.5% samples, also displayed significant genetic associations in our study. These findings underscore the role of host genetics in shaping the vaginal microbiome and are validated across multiple cohorts.

Additional biologically plausible associations included links between IL-5 and both *P. timonensis* and *Prevotella colorans*, consistent with prior reports connecting IL-5 to *Prevotella*[81]. A *FUT8* variant previously associated with *L. iners*[105] was associated with *Prevotella* sp946998205 in our study. Similarly, *MBL2*, previously linked to *L. iners* and *B. vaginale*[81], was associated with *Lawsonella* sp018376445, *F. vaginae*, *E. coli* and *Porphyromonas* sp947254915. Although some species-level discrepancies emerged, likely reflecting population heterogeneity, the underlying biological pathways warrant further investigation.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-026-02639-2.

References

1. Anahtar, M. N. et al. Cervicovaginal bacteria are a major modulator of host inflammatory responses in the female genital tract. *Immunity* **42**, 965–976 (2015).
2. Gosmann, C. et al. *Lactobacillus*-deficient cervicovaginal bacterial communities are associated with increased HIV acquisition in young South African women. *Immunity* **46**, 29–37 (2017).
3. Fettweis, J. M. et al. The vaginal microbiome and preterm birth. *Nat. Med.* **25**, 1012–1021 (2019).
4. Mitra, A. et al. The vaginal microbiota associates with the regression of untreated cervical intraepithelial neoplasia 2 lesions. *Nat. Commun.* **11**, 1999 (2020).
5. Elovitz, M. A. et al. Cervicovaginal microbiota and local immune response modulate the risk of spontaneous preterm delivery. *Nat. Commun.* **10**, 1305 (2019).
6. Klatt, N. R. et al. Vaginal bacteria modify HIV tenofovir microbicide efficacy in African women. *Science* **356**, 938–945 (2017).
7. Lebeer, S. et al. A citizen-science-enabled catalogue of the vaginal microbiome and associated factors. *Nat. Microbiol.* **8**, 2183–2195 (2023).
8. Yarza, P. et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).
9. Hillmann, B. et al. Evaluating the information content of shallow shotgun metagenomics. *mSystems* **3**, e00069-18 (2018).
10. Langille, M. G. et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).
11. Jovel, J. et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front. Microbiol.* **7**, 459 (2016).
12. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
13. Human Microbiome Jumpstart Reference Strains Consortium et al. A catalog of reference genomes from the human microbiome. *Science* **328**, 994–999 (2010).
14. Forster, S. C. et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol.* **37**, 186–192 (2019).
15. Zou, Y. et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* **37**, 179–185 (2019).
16. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662 (2019).
17. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
18. Stewart, R. D. et al. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961 (2019).
19. Zeng, S. et al. A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome. *Nat. Commun.* **13**, 5139 (2022).
20. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
21. Zhu, J. et al. Over 50,000 metagenomically assembled draft genomes for the human oral microbiome reveal new taxa. *Genomics Proteomics Bioinformatics* **20**, 246–259 (2022).
22. 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).
23. Shen, Z. et al. A genome catalog of the early-life human skin microbiome. *Genome Biol.* **24**, 252 (2023).
24. Li, F. et al. The metagenome of the female upper reproductive tract. *Gigascience* **7**, giy107 (2018).
25. Huang, L. et al. A multi-kingdom collection of 33,804 reference genomes for the human vaginal microbiome. *Nat. Microbiol.* **9**, 2185–2200 (2024).
26. 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).
27. Liao, J. et al. Microdiversity of the vaginal microbiome is associated with preterm birth. *Nat. Commun.* **14**, 4997 (2023).
28. Jie, Z. et al. Life history recorded in the vagino-cervical microbiome along with multi-omes. *Genomics Proteomics Bioinformatics* **20**, 304–321 (2022).
29. Blanco-Miguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
30. Huttenhower, C. et al. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
31. Liu, X. et al. Mendelian randomization analyses support causal relationships between blood metabolites and the gut microbiome. *Nat. Genet.* **54**, 52–61 (2022).
32. Turpin, W. et al. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **48**, 1413–1417 (2016).
33. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
34. Liu, X. et al. A genome-wide association study for gut metagenome in Chinese adults illuminates complex diseases. *Cell Discov.* **7**, 9 (2021).
35. Mehta, S. D. et al. Host genetic factors associated with vaginal microbiome composition in Kenyan women. *mSystems* **5**, e00502-20 (2020).
36. Fan, W. et al. Association between human genetic variants and the vaginal bacteriome of pregnant women. *mSystems* **6**, e0015821 (2021).
37. Samanta, I. & Bandyopadhyay, S. in *Antimicrobial Resistance in Agriculture* 241–251 (Academic Press, 2020).

38. Radolf, J. D. et al. *Treponema pallidum*, the syphilis spirochete: making a living as a stealth pathogen. *Nat. Rev. Microbiol.* **14**, 744–759 (2016).
39. Alessandri, G., van Sinderen, D. & Ventura, M. The genus *Bifidobacterium*: from genomics to functionality of an important component of the mammalian gut microbiota. *Comput. Struct. Biotechnol. J.* **19**, 1472–1487 (2021).
40. Tarracchini, C. et al. Assessing the genomic variability of *Gardnerella vaginalis* through comparative genomic analyses: evolutionary and ecological implications. *Appl. Environ. Microbiol.* **87**, e02188-20 (2020).
41. Mendes-Soares, H. et al. Fine-scale analysis of 16S rRNA sequences reveals a high level of taxonomic diversity among vaginal *Atopobium* spp. *Pathog. Dis.* **73**, ftv020 (2015).
42. Potter, R. F., Burnham, C. D. & Dantas, G. In silico analysis of Gardnerella genomospecies detected in the setting of bacterial vaginosis. *Clin. Chem.* **65**, 1375–1387 (2019).
43. Łaniewski, P., Ilhan, Z. & Herbst-Kralovetz, M. The microbiome and gynaecological cancer development, prevention and therapy. *Nat. Rev. Urol.* **17**, 232–250 (2020).
44. Holm, J. B. Comparative metagenome-assembled genome analysis of '*Candidatus* Lachnocurva vaginae', formerly known as bacterial vaginosis-associated bacterium-1 (BVAB1). *Front. Cell. Infect. Microbiol.* **10**, 117 (2020).
45. Austin, M. N. et al. *Mageeibacillus indolicus* gen. nov., sp. nov.: a novel bacterium isolated from the female genital tract. *Anaerobe* **32**, 37–42 (2015).
46. DiGiulio, D. B. et al. Temporal and spatial variation of the human microbiota during pregnancy. *Proc. Natl Acad. Sci. USA* **112**, 11060–11065 (2015).
47. France, M. et al. VALENCIA: a nearest centroid classification method for vaginal microbial communities based on composition. *Microbiome* **8**, 166 (2020).
48. Gregory, A. C. et al. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe* **28**, 724–740 (2020).
49. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098–1109 (2021).
50. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
51. Li, S. et al. A catalog of 48,425 nonredundant viruses from oral metagenomes expands the horizon of the human oral virome. *iScience* **25**, 104418 (2022).
52. Pradini, G. W. et al. Phylogeny and in silico structure analysis of major capsid protein (L1) human papillomavirus 45 from Indonesian isolates. *Asian Pac. J. Cancer Prev.* **21**, 2517–2523 (2020).
53. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
54. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
55. France, M., Alizadeh, M., Brown, S., Ma, B. & Ravel, J. Towards a deeper understanding of the vaginal microbiota. *Nat. Microbiol.* **7**, 367–378 (2022).
56. Kwak, W. et al. Complete genome of *Lactobacillus iners* KY using flongle provides insight into the genetic background of optimal adaption to vaginal econiche. *Front. Microbiol.* **11**, 1048 (2020).
57. Yu, C. E. & Ferretti, J. J. Molecular epidemiologic analysis of the type A streptococcal exotoxin (erythrogenic toxin) gene (*speA*) in clinical *Streptococcus pyogenes* strains. *Infect. Immun.* **57**, 3715–3719 (1989).
58. Mitchell, D. A. et al. Structural and functional dissection of the heterocyclic peptide cytotoxin streptolysin S. *J. Biol. Chem.* **284**, 13004–13012 (2009).
59. Tabata, A. et al. Novel twin streptolysin S-like peptides encoded in the sag operon homologue of β-hemolytic *Streptococcus anginosus*. *J. Bacteriol.* **195**, 1090–1099 (2013).
60. Beckham, K. S. et al. The metabolic enzyme AdhE controls the virulence of *Escherichia coli* O157:H7. *Mol. Microbiol.* **93**, 199–211 (2014).
61. Le Breton, Y. et al. Genome-wide discovery of novel M1T1 group A streptococcal determinants important for fitness and virulence during soft-tissue infection. *PLoS Pathog.* **13**, e1006584 (2017).
62. Harrison, J. J. et al. The chromosomal toxin gene *yafQ* is a determinant of multidrug tolerance for *Escherichia coli* growing in a biofilm. *Antimicrob. Agents Chemother.* **53**, 2253–2258 (2009).
63. Inoue, R. et al. Genetic identification of two distinct DNA polymerases, DnaE and PolC, that are essential for chromosomal DNA replication in *Staphylococcus aureus*. *Mol. Genet. Genomics* **266**, 564–571 (2001).
64. Agostini, H. J., Carroll, J. D. & Minton, K. W. Identification and characterization of uvrA, a DNA repair gene of *Deinococcus radiodurans*. *J. Bacteriol.* **178**, 6759–6765 (1996).
65. Ban, C. & Yang, W. Crystal structure and ATPase activity of MutL: implications for DNA repair and mutagenesis. *Cell* **95**, 541–552 (1998).
66. Marie, L. et al. Bacterial RadA is a DnaB-type helicase interacting with RecA to promote bidirectional D-loop extension. *Nat. Commun.* **8**, 15638 (2017).
67. Guillemet, E. et al. The bacterial DNA repair protein Mfd confers resistance to the host nitrogen immune response. *Sci. Rep.* **6**, 29349 (2016).
68. Dauvillee, D. et al. Role of the *Escherichia coli glgX* gene in glycogen metabolism. *J. Bacteriol.* **187**, 1465–1473 (2005).
69. Lebeer, S., Vanderleyden, J. & De Keersmaecker, S. C. Genes and molecules of lactobacilli supporting probiotic action. *Microbiol. Mol. Biol. Rev.* **72**, 728–764 (2008).
70. Popham, D. L. & Setlow, P. Cloning, nucleotide sequence, mutagenesis, and mapping of the *Bacillus subtilis pbpD* gene, which codes for penicillin-binding protein 4. *J. Bacteriol.* **176**, 7197–7205 (1994).
71. Seah, C. et al. MupB, a new high-level mupirocin resistance mechanism in *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **56**, 1916–1920 (2012).
72. Zaw, M. T., Emran, N. A. & Lin, Z. Mutations inside rifampicin-resistance determining region of rpoB gene associated with rifampicin-resistance in *Mycobacterium tuberculosis*. *J. Infect. Public Health* **11**, 605–610 (2018).
73. Arora, S. K., Ritchings, B. W., Almira, E. C., Lory, S. & Ramphal, R. The *Pseudomonas aeruginosa* flagellar cap protein, FliD, is responsible for mucin adhesion. *Infect. Immun.* **66**, 1000–1007 (1998).
74. Arora, G. et al. Ser/Thr protein kinase PrkC-mediated regulation of GroEL is critical for biofilm formation in *Bacillus anthracis*. *NPJ Biofilms Microbiomes* **3**, 7 (2017).
75. Madec, E., Laszkiewicz, A., Iwanicki, A., Obuchowski, M. & Seror, S. Characterization of a membrane-linked Ser/Thr protein kinase in *Bacillus subtilis*, implicated in developmental processes. *Mol. Microbiol.* **46**, 571–586 (2002).
76. Xu, K. et al. A variant on the κ opioid receptor gene (*OPRK1*) is associated with stress response and related drug craving, limbic brain activation and cocaine relapse risk. *Transl. Psychiatry* **3**, e292 (2013).

77. Kim, J. H. et al. A genome-wide association study of total serum and mite-specific IgEs in asthma patients. *PLoS ONE* **8**, e71958 (2013).
78. Noh, E. J. et al. *Ureaplasma urealyticum* infection contributes to the development of pelvic endometriosis through toll-like receptor 2. *Front. Immunol.* **10**, 2373 (2019).
79. Glaser, K. et al. *Ureaplasma* species differentially modulate pro- and anti-inflammatory cytokine responses in newborn and adult human monocytes pushing the state toward pro-inflammation. *Front. Cell. Infect. Microbiol.* **7**, 484 (2017).
80. Paira, D. A. et al. *Ureaplasma urealyticum* and *Mycoplasma hominis* urogenital infections associate with semen inflammation and decreased sperm quality. *J. Leukoc. Biol.* **113**, 18–26 (2023).
81. Si, J., You, H. J., Yu, J., Sung, J. & Ko, G. Prevotella as a hub for vaginal microbiota under the influence of host genetics and their association with obesity. *Cell Host Microbe* **21**, 97–105 (2017).
82. Biernat-Sudolska, M., Szostek, S., Rojek-Zakrzewska, D., Klimek, M. & Kosz-Vnenchak, M. Concomitant infections with human papillomavirus and various *Mycoplasma* and *Ureaplasma* species in women with abnormal cervical cytology. *Adv. Med. Sci.* **56**, 299–303 (2011).
83. Choi, Y. & Roh, J. Cervical cytopathological findings in Korean women with *Chlamydia trachomatis*, *Mycoplasma hominis*, and *Ureaplasma urealyticum* infections. *ScientificWorldJournal* **2014**, 756713 (2014).
84. Lukic, A. et al. Determination of cervicovaginal microorganisms in women with abnormal cervical cytology: the role of *Ureaplasma urealyticum*. *Anticancer Res.* **26**, 4843–4849 (2006).
85. Ramirez, N. P. et al. ADAP1 promotes latent HIV-1 reactivation by selectively tuning KRAS-ERK-AP-1 T cell signaling-transcriptional axis. *Nat. Commun.* **13**, 1109 (2022).
86. Ascoli, M., Fanelli, F. & Segaloff, D. L. The lutropin/choriogonadotropin receptor, a 2002 perspective. *Endocr. Rev.* **23**, 141–174 (2002).
87. Tapia-Pizarro, A. et al. hCG activates Epac-Erk1/2 signaling regulating progesterone receptor expression and function in human endometrial stromal cells. *Mol. Hum. Reprod.* **23**, 393–405 (2017).
88. Sacchi, S., Sena, P., Degli Esposti, C., Lui, J. & La Marca, A. Evidence for expression and functionality of FSH and LH/hCG receptors in human endometrium. *J. Assist. Reprod. Genet.* **35**, 1703–1712 (2018).
89. Ohashi, Y., Kaneko, S. J., Cupples, T. E. & Young, S. R. Ubiquinol cytochrome c reductase (*UQCRFS1*) gene amplification in primary breast cancer core biopsy samples. *Gynecol. Oncol.* **93**, 54–58 (2004).
90. Papke, D. J. Jr., Forgo, E., Charville, G. W. & Hornick, J. L. PDGFRA immunohistochemistry predicts *PDGFRA* mutations in gastrointestinal stromal tumors. *Am. J. Surg. Pathol.* **46**, 3–10 (2022).
91. Liao, J. et al. Nationwide genomic atlas of soil-dwelling Listeria reveals effects of selection and population ecology on pangenome evolution. *Nat. Microbiol.* **6**, 1021–1030 (2021).
92. Serrano, M. G. et al. Racioethnic diversity in the dynamics of the vaginal microbiome during pregnancy. *Nat. Med.* **25**, 1001–1011 (2019).
93. Tortelli, B. A., Lewis, A. L. & Fay, J. C. The structure and diversity of strain-level variation in vaginal bacteria. *Microb. Genom.* **7**, mgen000543 (2021).
94. Nielsen, R. et al. Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
95. Rivera, A. J., Stumpf, R. M., Wilson, B., Leigh, S. & Salyers, A. A. Baboon vaginal microbiota: an overlooked aspect of primate physiology. *Am. J. Primatol.* **72**, 467–474 (2010).
96. Wei, X. et al. Vaginal microbiomes show ethnic evolutionary dynamics and positive selection of *Lactobacillus* adhesins driven by a long-term niche-specific process. *Cell Rep.* **43**, 114078 (2024).
97. Janulaitiene, M. et al. Prevalence and distribution of *Gardnerella vaginalis* subgroups in women with and without bacterial vaginosis. *BMC Infect. Dis.* **17**, 394 (2017).
98. Hill, J. E. & Albert, A. Y. K. Resolution and cooccurrence patterns of *Gardnerella leopoldii*, *G. swidsinskii*, *G. piotii*, and *G. vaginalis* within the vaginal microbiome. *Infect. Immun.* **87**, e00532-19 (2019).
99. Ragaliauskas, T. et al. Inerolysin and vaginolysin, the cytolysins implicated in vaginal dysbiosis, differently impair molecular integrity of phospholipid membranes. *Sci. Rep.* **9**, 10606 (2019).
100. Cornejo, O. E., Hickey, R. J., Suzuki, H. & Forney, L. J. Focusing the diversity of *Gardnerella vaginalis* through the lens of ecotypes. *Evol. Appl.* **11**, 312–324 (2018).
101. Cohan, F. M. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **361**, 1985–1996 (2006).
102. Xu, X. et al. A cross-sectional analysis about bacterial vaginosis, high-risk human papillomavirus infection, and cervical intraepithelial neoplasia in Chinese women. *Sci. Rep.* **12**, 6609 (2022).
103. Naidoo, K., Abbai, N., Tinarwo, P. & Sebitloane, M. BV associated bacteria specifically BVAB 1 and BVAB 3 as biomarkers for HPV risk and progression of cervical neoplasia. *Infect. Dis. Obstet. Gynecol.* **2022**, 9562937 (2022).
104. Proudnikov, D. et al. Polymorphisms of the κ opioid receptor and prodynorphin genes: HIV risk and HIV natural history. *J. Acquir. Immune Defic. Syndr.* **63**, 17–26 (2013).
105. Fan, Q. et al. *Lactobacillus* spp. create a protective micro-ecological environment through regulating the core fucosylation of vaginal epithelial cells against cervical cancer. *Cell Death Dis.* **12**, 1094 (2021).

[1]BGI Research, Wuhan, China. [2]Laboratory of Integrative Biomedicine, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [3]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China. [4]School of Public Health and Emergency Management, School of Medicine, Southern University of Science and Technology, Shenzhen, China. [5]Social Affairs Bureau of Suzhou National New and Hi-tech Industrial Development Zone, Suzhou, China. [6]National Clinical Research Center for Obstetric and Gynecologic Diseases, Department of Obstetrics and Gynecology, Peking Union Medical College Hospital, State Key Laboratory for Complex Severe and Rare Diseases, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China. [7]Puensum Genetech Institute, Wuhan, China. [8]Shenzhen Key Laboratory of Human Commensal Microorganisms and Health Research, BGI Research, Shenzhen, China. [9]State Key Laboratory of Genome and Multi-omics Technologies, BGI Research, Shenzhen, China. [10]Suzhou National New and Hi-tech Industrial Development Zone Center for Maternal and Child Health and Family Planning Service, Suzhou, China. [11]BGI Research, Shenzhen, China. [12]BGI Genomics, Shenzhen, China. [13]Clin Lab, BGI Genomics, Nanjing, China. [14]James D. Watson Institute of Genome Sciences, Hangzhou, China. [15]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing, China. [16]School of Life Sciences, Fudan University, Shanghai, China. [17]Greater Bay Area Institute of Precision Medicine (Guangzhou), Fudan University, Guangzhou, China. [18]These authors contributed equally: Zhuye Jie, Weiting Liang, Qiuxia Ding, Xiaomin Liu, Yunhong Zhang, Na Chen, Shenghui Li. [19]These authors jointly supervised this work: Tao Zhang, Lilan Hao, Lan Zhu, Chen Chen. ✉e-mail: tao.zhang@genomics.cn; haolilan@genomics.cn; zhu_julie@vip.sina.com; chenchen20192022@163.com

Methods

Vaginal metagenomic datasets

The Peacock Project in China comprises 10,665 cervicovaginal samples collected from different regions. These include samples from cancer screening programs in Suzhou (SU-CCS2018/2019, $n = 3,355$; SU-CCS2021, $n = 3,527$; with 414 overlapping individuals), gynecology clinic attendees at Peking Union Medical College Hospital (PUMCH) in Beijing (BJ-GC, 2018–2019, $n = 2,878$), health check-ups in Shenzhen (SZ-4D, 2018, $n = 833$) and other regions ($n = 72$). All the samples were collected from the cervicovaginal site using sterile swabs by gynecologists following protocols approved by the Institutional Review Board of PUMCH in Beijing (ZS-1683), the Institutional Review Board of Suzhou Municipal Hospital (IEC-C-008-A07-V1.0) and the Institutional Review Board of BGI (BGI-IRB 19027-T3), and with written informed consent from all participants. Upon collection, samples were promptly immersed in an N-octylpyridinium bromide-based stabilizer reagent (MGIEasy Collection Kit)[106], transported to the laboratory at ambient conditions and preserved at −80 °C for long-term storage.

Additionally, 11 cervicovaginal samples from patients with cervical intraepithelial neoplasia grade 2/3 (CIN2/3) from BJ-GC were collected by gynecologists preserved in glycerol for cultivation. The samples were temporarily stored at −80 °C and then transported on dry ice to the laboratory for cultivation. Anaerobic cultivation was performed as previously described[15] (Supplementary Note).

DNA extraction and metagenomic sequencing

DNA extraction of samples from the Peacock cohort was performed as previously described[107] (Supplementary Note). Library construction and metagenomic shotgun sequencing were performed using the BGI-DIPSEQ platform with 100 bp or 150 bp paired-end reads[108]. The 2.5 T raw sequencing reads were obtained for Peacock cohort.

For the 384 isolates, DNA extraction process was performed as described above, except that the addition of glass beads and the shaking step were omitted. Sequencing libraries were constructed as above and sequenced on BGI-DIPSEQ platform to obtain 15,665,856,030 raw sequencing reads.

Genome assembly and binning

To construct the GVMG catalog, we included, in addition to the Chinese sample sequenced in this study, 2,967 metagenomes from 11 different studies publicly available, spanning seven countries (Supplementary Table 1). The 13,204 metagenomes, including 10,237 Peacock metagenomes and 2,967 publicly available metagenomes, were processed through a comprehensive pipeline that involved quality control, host removal, assembly and binning for each raw read. Briefly, sequences were subjected to quality control using fastp (v0.20.1; minimum length of 51 bp). Reads originating from humans were removed using Bowtie 2 (v2.4.2) with the '--end-to-end --very-sensitive' options. For Peacock metagenomes, filtering was performed sequentially based on the GRCh38 and CHM13v2.0 reference databases, while for publicly available metagenomes, only the CHM13v2.0 reference database was used. As a result, approximately 84.5 Gb reads of high-quality nonhuman metagenomic data were obtained for further analysis (Supplementary Table 2). The clean reads were de novo assembled into contigs using different assemblers based on the read types—metaSPAdes (v3.14.0) with the parameters '-k 21,33,55,77 --memory 81' for samples with paired-end reads, and MegaHIT (v1.2.9) with parameters '--presets meta-sensitive --min-contig-len 100' for samples with single-end reads, resulting in 40.2 Gb contigs (Supplementary Table 2).

Assembled contigs were binned and dereplicated within each single sample using both multicoverage and single-coverage binning approaches to ensure maximum genome recovery. For multicoverage binning, Mash was used to identify the 19 closest samples for read alignment using Burrow-Wheeler Aligner (BWA, v0.7.19), with results processed through MetaBAT2 (v2.15). Single-coverage binning was performed using three complementary methods—MetaBAT2, MaxBin2 (v2.2.7) and CONCOCT (v1.1.0). Finally, all generated MAGs were dereplicated at the strain level (99% average nucleotide identity (ANI)) using dRep (v3.4.9) to eliminate redundancy within each sample (Supplementary Note). As a result, we obtained a total of 39,816 quality-controlled MAGs.

For isolates, sequences were subjected to quality control using fastp (v0.20.1; minimum length of 51 bp). The clean reads were assembled using SPAdes (v3.15.2) with the parameters '-k 21,33,55,77 --memory 81' to form scaffolds.

Retrieval of publicly available prokaryotic MAGs and genomes

Given that the VMGC[25] collected a comprehensive set of publicly available vaginal microbial genomes from the NCBI database, we aimed to ensure that our genome collection was even more comprehensive. Therefore, we included 4,628 MAGs derived from 1,477 metagenomes unique to the VMGC study, as well as the genomes of 972 isolates from the VMGC study. Because the VMGC database search was conducted up to April 2023, we further reviewed literature from April 2023 to August 2024, acquiring genomes of additional 77 isolates.

Quality assessment of prokaryotic genomes

Through in-house assembly and publicly available sources, we obtained a total of 44,444 MAGs and 1,433 isolated genomes. The quality of prokaryotic genomes was assessed using CheckM2 (v1.0.1) with the database 'uniref100.KO.1.dmnd'. Genomes with a completeness ≥50%, contamination <5% and a quality score (calculated as completeness − 5 × contamination) ≥50 were retained for further analyses. Subsequently, the genome chimerism of prokaryotic genomes was evaluated using GUNC (v1.06) with the 'gunc_db_progenomes 2.1.dmnd' database. Genomes with a clade separation score ≥0.45 were excluded from subsequent evaluation. Next, we decided to categorize the integrity of the obtained 36,059 prokaryotic genomes[20]. The metagenomic community has unfortunately presented two different definitions concerning near-complete and high-quality MAGs (Supplementary Table 5). We followed the generally accepted MIMAG standards[109], and changed the classification of the VMGC accordingly. Thus, medium-quality MAGs were identified by ≥50% completeness and <5% contamination. Near-complete MAGs were distinguished by ≥90% completeness and <5% contamination. High-quality MAGs were specified as near-complete MAGs that additionally included the 5S, 16S and 23S rRNA genes, along with at least 18 types of transfer RNA. The rRNA genes within MAGs were identified using cmsearch in INFERNAL (v1.1.5), whereas tRNA genes were detected using tRNAscan-SE (v2.0.12) with the parameters -B. A total of 2,955 high-quality MAGs, 20,368 near-complete MAGs and 11,478 medium-quality MAGs were obtained, amounting to 36,059 prokaryotic genomes in our comprehensive collection.

SGBs

To delineate SGBs, we clustered 36,059 prokaryotic genomes using the Galah (v0.4.0), with parameters --fragment-length 1500, --min-aligned-fraction 50, --ani 9, resulting in 890 SGBs. Among these, 774 species were annotated, while 116 remained unannotated, as determined by GTDB-Tk (v2.4.0) using the GTDB release 220. The phylogenetic relationships among these genomes were elucidated using the 'infer' module of GTDB-Tk and the resulting phylogenetic tree was visualized with the Interactive Tree of Life (iTOL) platform.

Functional analysis of prokaryotic genomes

The annotation of 36,059 prokaryotic genomes comprised virulence factors, antibiotic resistance genes, genes involved in carbohydrate metabolism (CAZy), KEGG metabolic pathways, antimicrobial peptides and secondary metabolite gene clusters (BGC). Initially, gene prediction was performed using Prodigal (v2.6.3, -p meta) to

obtain protein-coding sequences for each prokaryotic genome, yielding 53,679,109 protein-coding sequences, with an average of 1,488 per MAG. The comprehensive functional annotation and profiling were performed using a multidatabase annotation pipeline (Supplementary Note).

First, the protein-coding sequences were labeled as BGC. Second, BGC data were filtered using a mapping coverage threshold of 0.5 for all core biosynthetic genes. We compared the presence and abundance based on reads per kilobase of transcript, per million mapped reads of GCFs between Chinese and USA samples using chi-squared and two-sided Wilcoxon rank-sum tests, respectively. Given the observed species specificity, BGCs corresponding to multiple SGBs were assigned to the taxon with the highest proportion of genomes harboring the cluster. The top 15 most significantly enriched BGCs per population ($P_{adj} < 1 \times 10^{-10}$) were identified and visualized using bar plots. If fewer than 15 BGCs were available, all were displayed.

Fungal MAGs

We selected 39,816 in-house bins exceeding 3 Mb for the identification of fungal MAGs. Subsequently, we used EukRep (v0.6.7) with the parameter '-min 2000' to filter out noneukaryotic sequences from each bin[25], resulting in the generation of 121 MAGs, each with a genomic size exceeding 3 Mb. To evaluate the genome quality of these MAGs, we used EukCC (v2.1.0) with the eukcc2_db_ver_1.1 database to filter genomes with a completeness ≥50% and contamination <10%. Next, we removed strain-level duplication (99% ANI) from MAGs originating from the same sample using dRep (v3.4.0) with the parameters -pa 0.9, -sa 0.99, -nc 0.3, and -S_algorithm fastANI, ultimately obtaining 13 eukaryotic MAGs. Taxonomic classification was conducted using BLAST (v2.11.0) with NCBI/bast_nt database, filtering for ≥90% identity. Seven MAGs were successfully annotated, while the remaining MAGs had overly complex annotations and were excluded from further analysis. Combining these with the 11 eukaryotic MAGs and 25 isolated genomes from the VMGC[25], a total of 43 eukaryotic genomes were included in the GVMG. Using the fastANI software (v1.34), pairwise ANI values between eukaryotic genomes were calculated. Based on these ANI values, the distances across genomes were determined. A phylogenetic tree was constructed using the upgma function from the R package phangorn, based on the maximum-likelihood method. The tree was then ladderized using the ape package, and a phylogram-type phylogenetic tree was plotted.

Analysis of the genome for viral populations

Virus sequences from vaginal metagenomic data were processed following the previous studies[25] (Supplementary Note). The viral genomes were then clustered into vOTUs based on a 95% nucleotide similarity threshold over at least 85% genome length. This clustering was performed using BLASTn (v2.12.0) with parameters '-evalue 1e-10 -word_size 20 -num_alignments 999999' and used a greedy incremental method similar to the CD-HIT tool[25,110]. Within each vOTU, the largest viral sequence was designated as the representative genome. To reveal the phylogenetic relationships among the viral genomes, a proteomic tree of the viruses using ViPTreeGen (v1.1.2) with default settings[111] was generated. Next, we conducted virus-host predictions based on all prokaryotic genomes within the GVMG (Supplementary Note).

Papillomaviridae members were categorized according to traditional papillomavirus types by analyzing their L1 structural protein sequences. First, the protein-coding sequences labeled as L1 structural proteins were extracted from all Papillomaviridae members available in the NCBI database. Then, we selected putative L1 structural genes in our Papillomaviridae genomes that were ≥1,300 bp in length and share ≥40% nucleotide identity with known L1 structural genes. Finally, based on the degree of similarity in the L1 sequences, we categorized the Papillomaviridae members into known types (with >90% similarity), new type (>70% similarity) and new species (>60% similarity)[112].

GVMG construction and species abundance profiling

For species abundance profiling, we analyzed 13,632 metagenomic sequencing datasets, which included the previously mentioned 13,204 metagenomes and additional 428 Chinese metagenomes from CNP0006125 (MRKH). The datasets underwent rigorous filtering and trimming, retaining reads with an average Phred quality score ≥20 and a length ≥30, as processed by fastp (v0.19.4)[113]. Human-derived reads were excluded using Bowtie 2 (v2.3.5; ref. 114; aligned against the human reference genome GRCh38 and CHM13v2.0) with the parameters '--end-to-end --very-sensitive', excluding the default settings.

To establish the microbial profile in GVMG catalog, we incorporated 35,915 prokaryotic genomes (spanning 746 SGBs, excluding medium-quality SGBs represented by a single MAG), 43 fungal genomes (spanning 11 fungal species) and the protozoan parasite genome of *Trichomonas vaginalis* (Supplementary Table 19). Given the substantial presence of host DNA in vaginal samples, the human reference genome (CHM13v2.0) was also integrated into the GVMG catalog to mitigate the impact of host reads. Then GVMG database without virus was performed through the Phanta workflow[115] (Supplementary Note). The sylph[116] was used to estimate the relative abundance of viruses involved in 839 vOTUs having more than five MAGs (Supplementary Note).

The occurrence of specific SGBs or vOTUs was calculated as the relative abundance of at least 0.0001 across all samples in the Chinese and U.S. populations.

Phylogenetic analysis of *B. vaginale* genomospecies

To construct a phylogenetic tree, we used the 'infer' module of GTDB-Tk (v1.5.1) for all genomes from *B. vaginale* genomospecies. The trees were visualized and annotated using the iTOL online version.

BV-specific signatures in bacterial SGBs

Three independent subcohorts were used, derived from the clinically BV samples of a Chinese public project (CNP0003852; $n = 48$), SU-CCS ($n = 279$) and BJ-GC ($n = 83$), with the same sample size of age-matched and menopause-matched healthy controls from the public project ($n = 48$), SU-CCS ($n = 279$) and BJ-GC ($n = 83$). The signatures between samples from individuals with BV and healthy individuals for each subcohort were determined through Analysis of Composition of Microbiome[117] for SGB profiling with a threshold ≥0.7. A co-occurrence matrix was calculated to evaluate correlations across bacterial SGBs in all of the above BV samples with Sparse Correlations for Compositional data (SparCC)[118]. The SGBs were filtered out based on−(1) prevalence (relative abundance at least 0.0001) <10%; (2) mean relative abundance <0.0001; and (3) the presence in <5 samples with relative abundance less than 0.0001. Significant correlations with adjusted *P* value (false discovery rate, FDR) <0.05 were visualized using the R package pheatmap.

Co-occurrence between SGBs and vOTUs in populations

To assess the correlation between SGBs and vOTUs in populations, we constructed a co-occurrence matrix using the SparCC method for two populations, SU-CCS2021 (Peacock) and the USA, obtaining empirical *P* values from 100 bootstraps. The SGBs and vOTUs were filtered out in the data of all samples from each population based on (1) prevalence (relative abundance at least 0.0001) <10%; (2) mean relative abundance <0.0001; and (3) the presence in <5 samples with relative abundance less than 0.0001. Significant correlations with adjusted *P* value (FDR) <0.05 were visualized using the R package pheatmap.

Association between vOTUs in population and phenotypes

The relationships between viral profiles and phenotypes (including clinically diagnosed HPV infection, BV and menopause) were analyzed using the Generalized Linear Model with the SU-CCS cohort. Significant vOTUs with adjusted *P* value (FDR) <0.05 were visualized using the R package ggplot2.

Microbial community and diversity analysis

The vaginal microbial CSTs were clustered through the partitioning around medoids algorithm, which was applied to the Bray–Curtis distance matrix derived from the relative taxonomic abundance profiles of all samples (using the pam function from R package cluster (v2.1.6))[46,96]. The optimal number of clusters was determined by calculating the average silhouette width (using the fviz_nbclust function from R package factoextra (v1.0.7)). A total of 13,632 metagenomic samples were first clustered into three clusters that were *L. iners* dominated, *L. crispatus* dominated and non-*Lactobacillus* dominated. Subsequently, the samples belonging to the non-*Lactobacillus*-dominated cluster were clustered into nine additional clusters. After this process, 12 CSTs were identified. The difference in proportion of CST between the Peacock cohort and the VIRGO cohort was determined using the 'prop.test' functions from the base R package.

The α diversity (richness and the Shannon–Wiener Index) was calculated based on the relative taxonomic abundance by vegan R package (v2.6-8). The difference in diversity between the Peacock cohort and the VIRGO cohort was evaluated using the 'wilcox.test' functions from the base R package.

Genome annotation and pangenome and core-genome analysis

Prokaryotic genome annotation was performed using Prokka (v1.14.6). The core genome and pangenome for each SGB (comprising more than 50 genomes) were computed with Roary (v3.13.0) taking annotated assemblies in the GFF3 format produced by Prokka as input, with options '-i 90 -cd 80 -e -n -p 8 -z -g 10000000' (minimum identity of protein-coding sequences as a positive match at 90%, being present at no less than 80% input conspecific genomes is defined as a core gene).

Population-specific signatures identification

To investigate the population-specific intraspecies phylogenetic characteristics of each SGB (comprising more than 50 genomes) in the GVMG at both strain and SNP levels, we first inferred approximately maximum-likelihood phylogenetic trees using FastTree ('-nt -gtr') based on multiFASTA alignments of core genes produced by Roary (-e). The pairwise genetic distances (patristic distances) across intraspecies genomes were estimated by their branch lengths in the phylogenetic tree using the cophenetic.phylo function in the ape R package. Then we performed PERMANOVA (adonis2) analysis in the 'vegan' R package using patristic distances, with adjustment for confounding factors (MAG quality score, age and clean reads count). The number of permutations in PERMANOVA was set to 999. The *Q* values and r^2 were visualized using the pheatmap function in the R package. The intraspecies phylogenetic trees of investigated SGBs and population information were visualized and annotated by the iTOL online version. The nonmetric multidimensional scaling plots were computed with the nmds function in the ecodist R package.

SNP calling was performed using SNP sites (v2.5.1) to extract SNPs from the multiFASTA alignment of core genes generated by Roary. SNP and population group association analysis was conducted by PLINK using the --assoc option, which performs a single-locus association test for each SNP based on a chi-squared statistic. The threshold for significance was calculated using a Bonferroni correction as $P \leq 0.05/$ (total number of SNPs detected). SNPs with *P* values below this corrected threshold were considered significant. To assign the significant SNPs to their corresponding genes and to identify their functional roles, we referenced the core_alignment_header.embl files generated by Roary. These files provided detailed information about the core genes, including their genomic coordinates and gene labels, enabling us to map the significant SNPs to their corresponding genes and retrieve their annotations from Prokka.

Extracting host WGS data from vaginal metagenomic samples

The raw sequencing reads for the discovery dataset (mean depth 10.3×) and validation cohort 2 (mean depth 27.9×) were aligned to the GRCh38/hg38 reference using BWA and used GATK for variant calling. GVCFs containing SNVs and INDELs from GATK HaplotypeCaller were combined (CombineGVCFs), genotyped (GenotypeGVCFs), recalibrated (VariantRecalibrator) and filtered (ApplyRecalibration). Variant Quality Score Recalibration was trained by the following four standard SNP sets: (1) HapMap3.3 SNPs; (2) dbSNP build 150 SNPs; (3) 1000 Genomes Project SNPs from Omni 2.5 chip and (4) 1000G phase1 high-confidence SNPs. Sensitivity thresholds of 99.5% to SNPs and 95% to INDELs were applied for variant selection after optimizing transition-to-transversion ratios using the GATK ApplyRecalibration command. From 55,606,543 raw variants, we then demanded variants to meet the following criteria: (1) depth >3×; (2) Hardy–Weinberg equilibrium, $P > 10^{-5}$; and (3) genotype calling rate >98%. We also excluded related individuals by calculating pairwise identity by descent (Pi-hat threshold of 0.1875) in PLINK. Finally, the discovery dataset retained 3,137 individuals with 5,456,968 common variants (minor allele frequency ≥ 5%) for M-GWAS analysis, while validation cohort 2 retained 506 individuals with 5,751,704 common variants.

Due to its lower sequencing depth (mean 5.7×), validation cohort 1 required prephasing and imputation. After identical initial calling, 7,809,087 high-quality variants (<2% missing frequency, minor allele count >3 and Hardy–Weinberg equilibrium $P > 10^{-5}$) were then phased and imputed using BEAGLE 5 (ref. 119), against a custom 1,992 high-depth WGS datasets (mean depth 42×) as the reference panel[31]. After filtering for an imputation information score above 0.7 (retaining 11.52 million variants), we excluded population stratification and kinship (pairwise identity by descent Pi-hat > 0.1875) in PLINK. Ultimately, 3,227 individuals with 6,026,457 high-quality common variants were used for M-GWAS analysis.

Microbiome association analysis

For the M-GWAS analysis, we selected 54 species meeting the following criteria: the presence in more than 10% samples and a relative abundance of 0.001 or greater. Additionally, species exhibiting a correlation coefficient greater than 0.99 were excluded from the analysis (Supplementary Table 14).

We examined the association between host genetic variants and the vaginal microbiome using linear or logistic models. The abundance of species with an occurrence rate exceeding 90% in the cohort was log-transformed and then analyzed as linear model as quantitative traits. Otherwise, we categorized bacteria into the presence/absence (1/0) patterns as a dichotomous trait for logistic model (Supplementary Table 14). Subsequently, standard single-variant M-GWAS analysis was conducted by using PLINK (v1.9) with a linear or logistic regression analysis, adjusting age, sequencing read counts and the top ten host principal components as covariates.

Next, a meta-analysis was performed on association results of the three cohorts, using inverse-variance-weighted fixed-effect meta-analysis in METAL software (https://genome.sph.umich.edu/wiki/METAL).

The gene of genetic variants was annotated using ANNOVAR tool[120] and required for eQTL signals by using the GTEx (v8) dataset. The association of significant genetic variants with reported phenotypes was investigated by searching in the GWAS Catalog (v1.0.2; https://www.ebi.ac.uk/gwas/), BioBank Japan dataset and Chinese 4D-SZ dataset. The regional plot was created with our own GWAS results at https://statgen.github.io/localzoom/.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The metagenomic data in the Peacock Dataset are available in the CNGB Nucleotide Sequence Archive (https://db.cngb.org/cnsa) under accessions CNP0005953 (ref. 121) and CNP0006125 (ref. 122). The genome databases of GVMG, including prokaryotic, eukaryotic and viral genome sequences, and the updated Kraken database, have been deposited in the Zenodo repository[123]. The GWAS summary statistics of microbial taxa are publicly available in the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/) from accession GCST90573122 to accession GCST90573174. All these data are freely available for download and analysis without login requirements or usage restrictions. Source data are provided with this paper.

Code availability

Analysis code can be accessed through GitHub (https://github.com/weiting-liang/GVMG/) and Zenodo repository[124].

References

106. Han, M. et al. A novel affordable reagent for room temperature storage and transport of fecal samples for metagenomic analyses. *Microbiome* **6**, 43 (2018).
107. Chen, C. et al. The microbiota continuum along the female reproductive tract and its relation to uterine-related diseases. *Nat. Commun.* **8**, 875 (2017).
108. Fang, C. Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. *Gigascience* **7**, 1–8 (2018).
109. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
110. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
111. Nishimura, Y. et al. ViPTree: the viral proteomic tree server. *Bioinformatics* **33**, 2379–2380 (2017).
112. Bernard, H. U. et al. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* **401**, 70–79 (2010).
113. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
114. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
115. Pinto, Y., Chakraborty, M., Jain, N. & Bhatt, A. S. Phage-inclusive profiling of human gut microbiomes with Phanta. *Nat. Biotechnol.* **42**, 651–662 (2024).
116. Shaw, J. & Yu, Y. W. Rapid species-level metagenome profiling and containment estimation with sylph. *Nat. Biotechnol.* **43**, 1348–1359 (2025).
117. Mandal, S. et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).
118. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
119. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
120. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
121. Hao, L. Genomic landscape of the human vaginal microbiome is linked to host genetics and population of origin. *CNGBdb* https://doi.org/10.26036/CNP0005953 (2026).
122. Hao, L. The vaginal microbiome of 472 women in Peacock cohort. *CNGBdb* https://doi.org/10.26036/CNP0006125 (2024).
123. Liang, W. Global vaginal microbial genomes (GVMG)—genomic landscape in the human vaginal microbiome links to host geographics and genetics. *Zenodo* https://doi.org/10.5281/zenodo.14708991 (2025).
124. Liang, W. weiting-liang/GVMG: v1 (v1.0). *Zenodo* https://doi.org/10.5281/zenodo.18923908 (2026).

Acknowledgements

We thank all participants for agreeing to take part in this study. We are also very grateful to the colleagues at BGI-Shenzhen for sample collection, DNA extraction, library construction, sequencing and discussions.

Author contributions

C.C., L.Z. and L.H. conceived of and directed the Peacock cohort construction. C.C. and Z.J. conceived of and directed this study. Y.Z., H.G. and R.L. provided samples from routine Cervical and Breast Cancer Screening in Suzhou. L.Z. and N.C. provided samples from diagnosis, sample collection and result analyses. X.H. had established a detailed full process for sample management, including sample collection, transportation and storage, ensuring standardization throughout the entire sample reception process. Z.J. led the bioinformatics analysis with contributions from W.L, L.H., Q.D., X.L., X.T., R.G., J.Z. and J.C. C.C. and Z.J. conceived the framework of the article and wrote the paper with contribution from S.L. K.K. thoroughly revised the paper. All these authors, including Z.Z., N.L., Z.X., X.W., L.Q., Y.L., L.X., S.Z., X.J., X.X., H.Y., J.W., F.Z. and H.J., contributed to the revision of the paper.

Funding

N.C. discloses support for participant enrollment and sample collection of this work from CAMS Initiative Fund for Medical Sciences (2025-I2M-C&T-B-013). L.Z. acknowledges support for participant enrollment and sample collection of this work from the National Natural Science Foundation of China Key Project (82530054) and the National High Level Hospital Clinical Research Funding (2025-PUMCH-C-037). Z.J., X.L., X.T. and L.X. disclose support for the research analysis and publication of this work from Shenzhen Science and Technology Program (grant SYSPG20241211173845014). Y.Z. received support for participant enrollment and sample collection of this project from Suzhou Major Disease and Infectious Disease Prevention and Key Technology Research Fund (GWZX201805).

Competing interests

The authors declare no competing interests.

Additional information

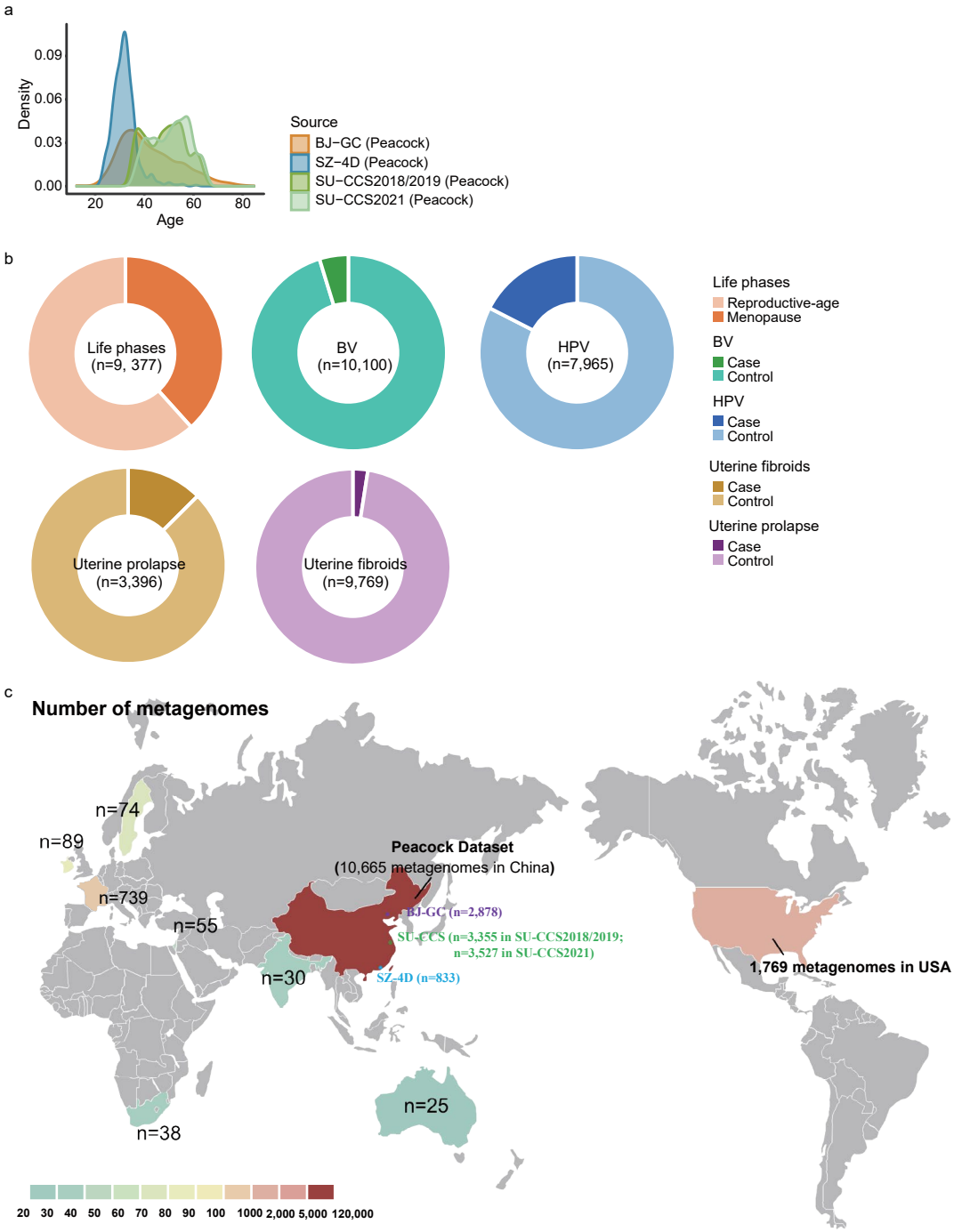
Extended data is available for this paper at https://doi.org/10.1038/s41588-026-02639-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-026-02639-2.

Correspondence and requests for materials should be addressed to Tao Zhang, Lilan Hao, Lan Zhu or Chen Chen.

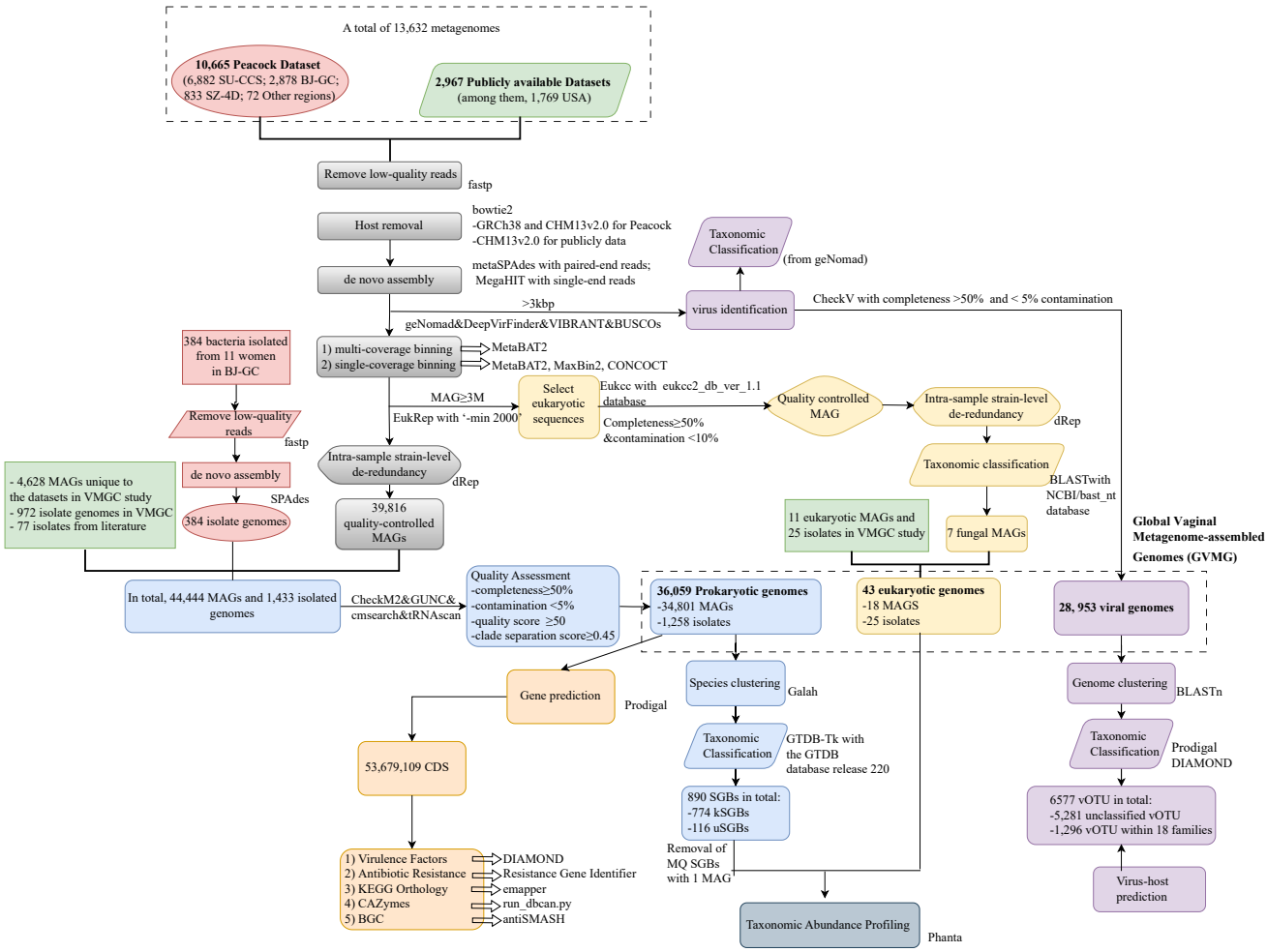
Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-026-02639-2



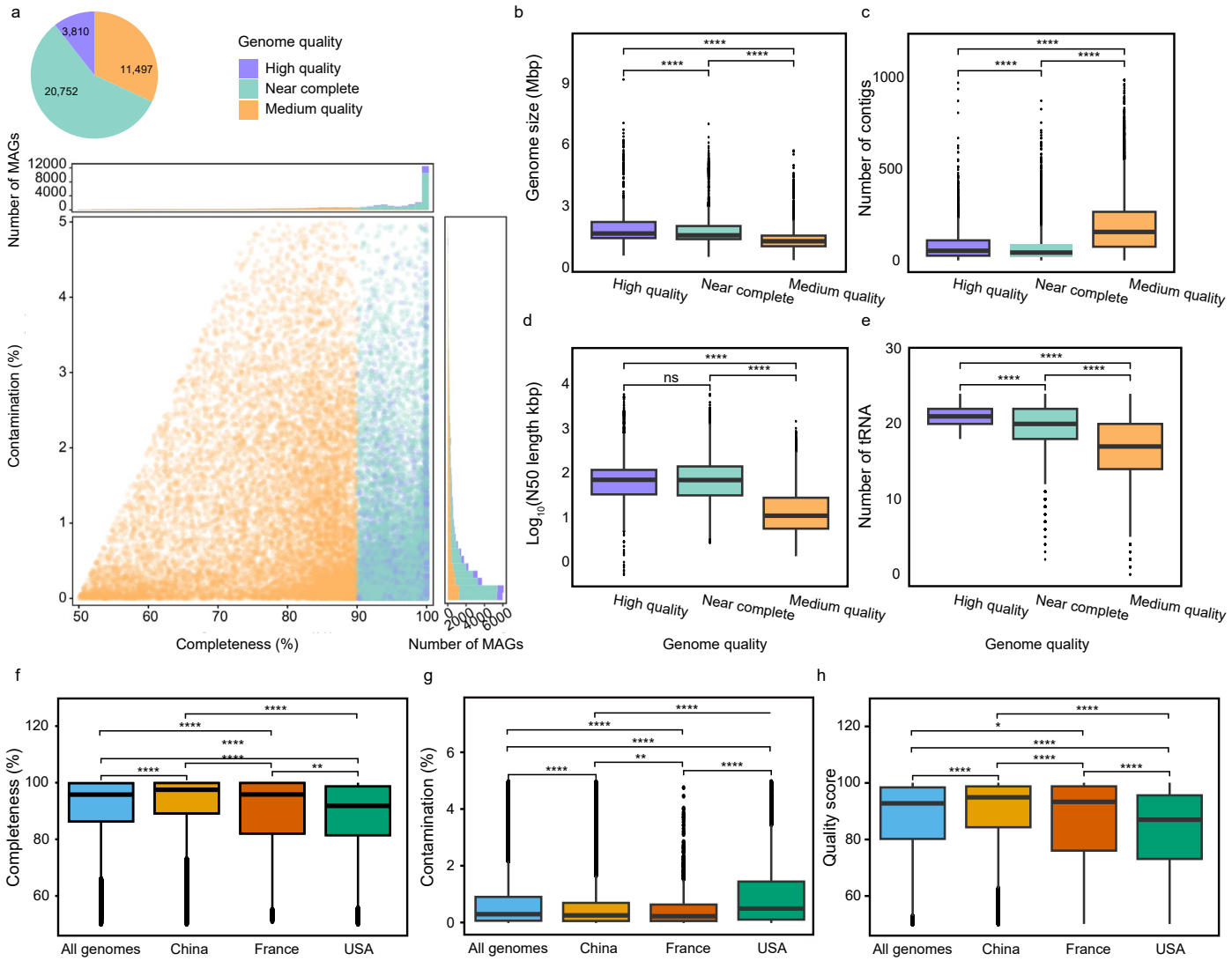
Extended Data Fig. 1 | Phenotypic characteristics of individuals corresponding to vaginal metagenomes in the GVMG. a, Distribution of age across the Peacock cohort is visualized using density plots, with the area under each curve normalized to 1 and the *y*-axis representing probability density. **b**, The number and proportion of vaginal metagenomes categorized by factors such as menopause, bacterial vaginosis (BV), human papillomavirus (HPV), uterine fibroids, and uterine prolapse. The '*n*' within each circle represents the number of samples with corresponding phenotypic information, excluding those with missing values. **c**, The geographic distribution of metagenomes, detailing counts per population, combining in-house Peacock Chinese cohort data with public data from eight additional countries. Data underlying plots are provided.

https://doi.org/10.1038/s41588-026-02639-2



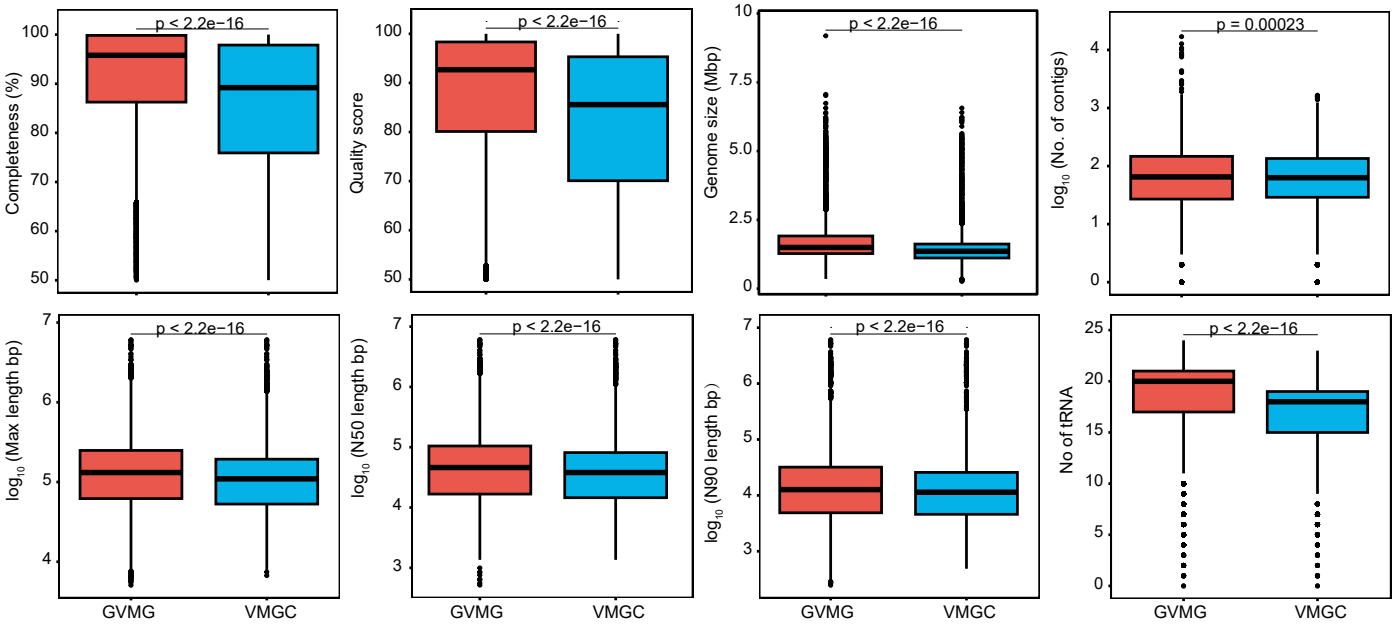
Extended Data Fig. 2 | Multiple data sources and a sequential processing workflow are used to construct GVMG. The schematic plot illustrates the step-by-step procedures and specific software tools used for handling prokaryotes, fungi, and vOTUs within the GVMG construction pipeline.

https://doi.org/10.1038/s41588-026-02639-2



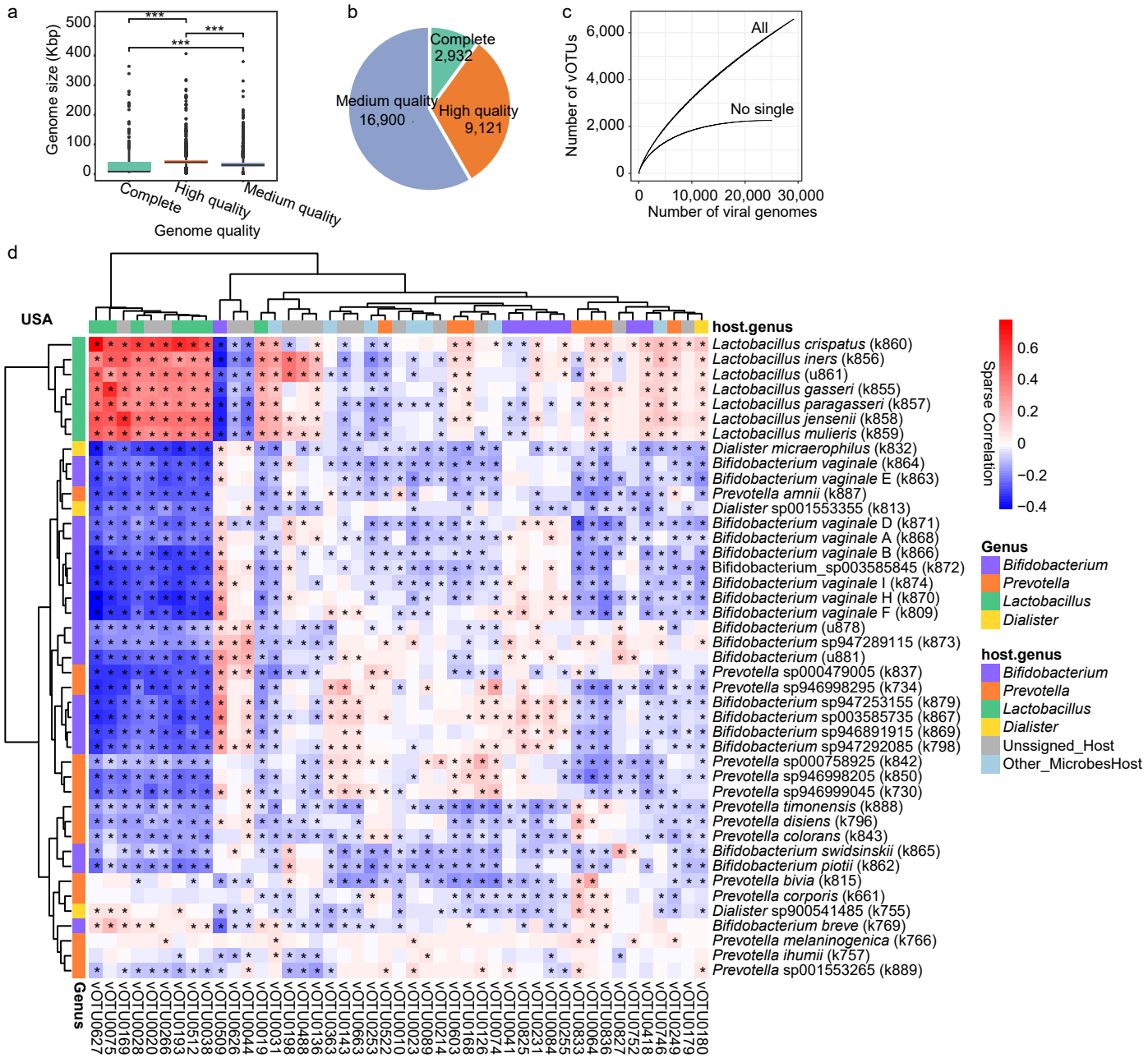
Extended Data Fig. 3 | Quality evaluation of prokaryotic genomes in GVMG. **a**, Completeness and contamination scores for each of 36,059 prokaryotic genomes. **b**–**e**, The distribution of the genome size (**b**), number of contigs (**c**), N50 length (**d**) and number of tRNA (**e**) of 36,059 prokaryotic genomes, including 3,810 high-quality 20,752 near-complete and 11,497 medium-quality genomes. **f**–**h**, The completeness (**f**), contamination (**g**), and quality scores (**h**) are provided for all genomes, as well as for genomes specifically derived from samples of Chinese, French, and American individuals, respectively. In the box plots (**b**–**h**), the centerline indicates the median, the box limits mark the upper and lower quartiles, the whiskers extend to 1.5× the IQR, and points beyond the whiskers are outliers. Two-sided Wilcoxon rank-sum tests were performed in **b**–**h** (*, $P < 0.05$; **, $P < 0.01$; ****, $P < 0.0001$). Data underlying plots are provided.

https://doi.org/10.1038/s41588-026-02639-2



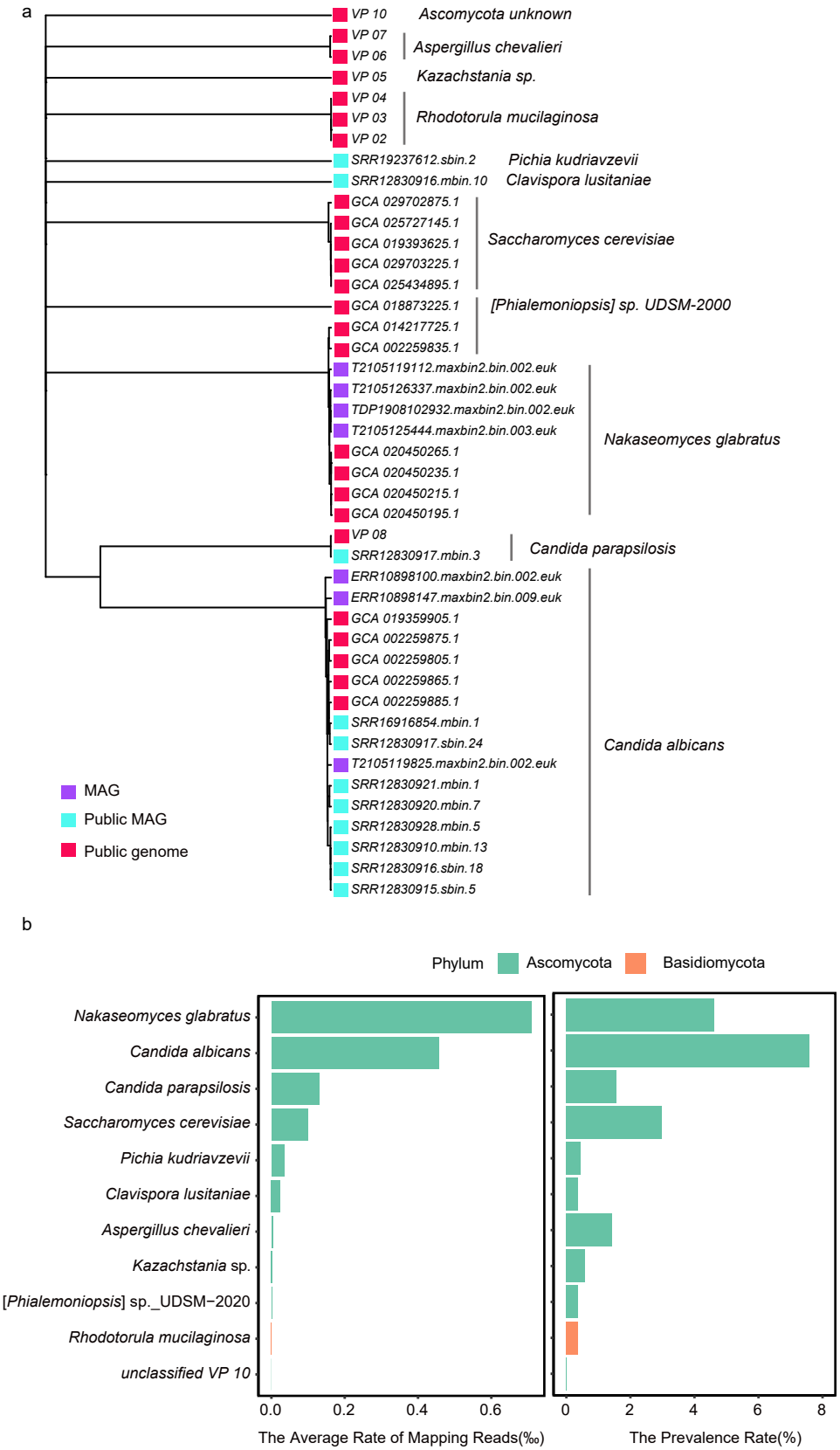
Extended Data Fig. 4 | Comparison of the quality metrics of prokaryotic genomes between GVMG and VMGC catalogs. In the box plots, the centerline indicates the median, the box limits mark the upper and lower quartiles, the whiskers extend to 1.5× the IQR, and points beyond the whiskers are outliers. A two-sided Wilcoxon rank-sum test was performed. Data underlying plots are provided.

https://doi.org/10.1038/s41588-026-02639-2



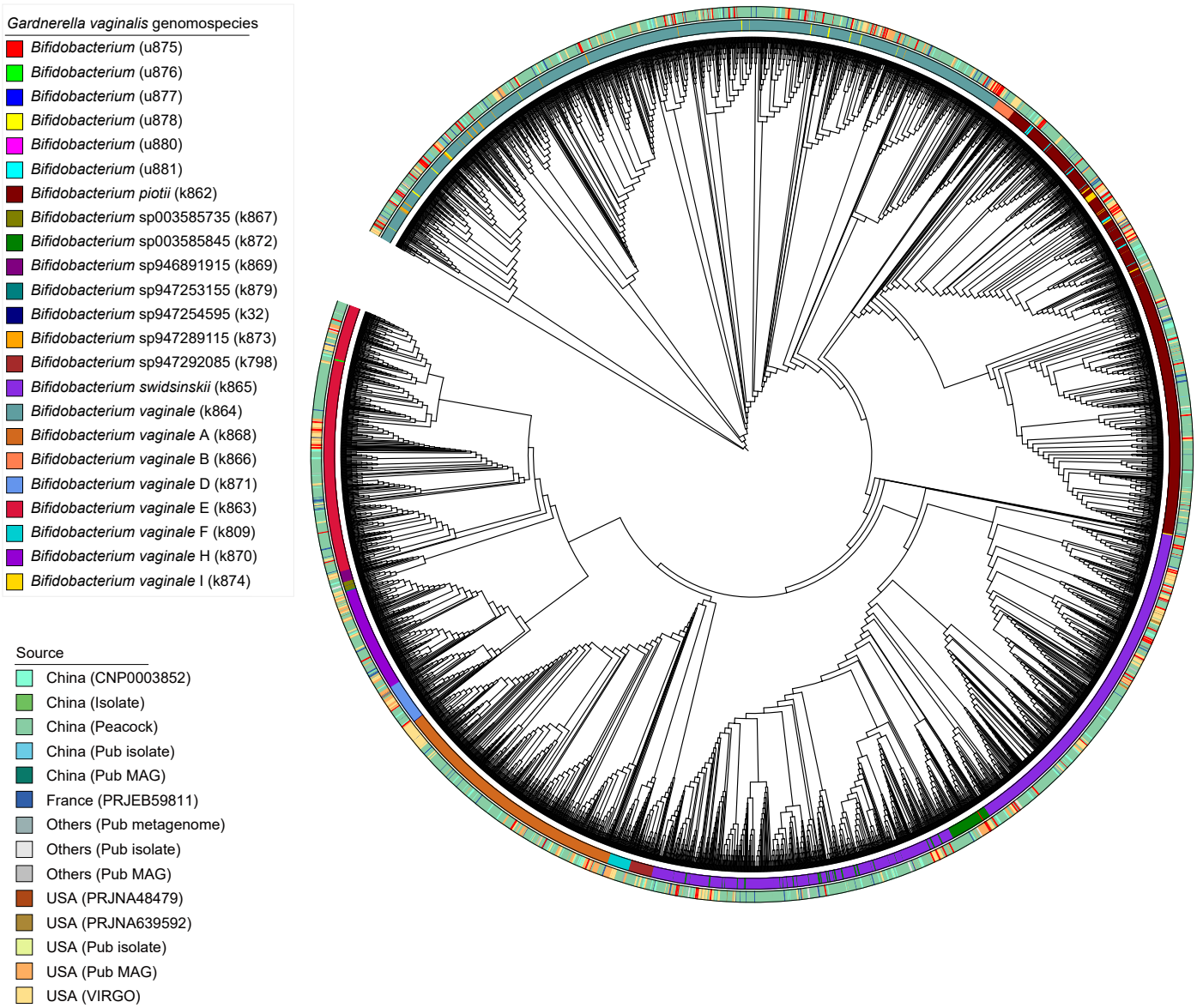
Extended Data Fig. 5 | Genome size and features of viral communities in GVMG. **a**, Genome sizes of viral genomes of different quality. **b**, Genome number of complete, high-quality and medium-quality. **c**, Rarefaction curves estimating the expected number of vOTUs and nonsingle vOTUs for a given number of viral genomes. **d**, Co-occurrence heatmap constructed to evaluate correlations between prokaryotic SGBs and vOTUs in the USA populations ($n = 1,769$) with SparCC. Row labels indicate the genus of each SGB, while column labels denote the predicted host for each vOTU. The color of each cell indicates the magnitude and direction of the SparCC correlation. Asterisks indicate a significant correlation with BH-adjusted P-value < 0.05. Data underlying plots are provided.

https://doi.org/10.1038/s41588-026-02639-2



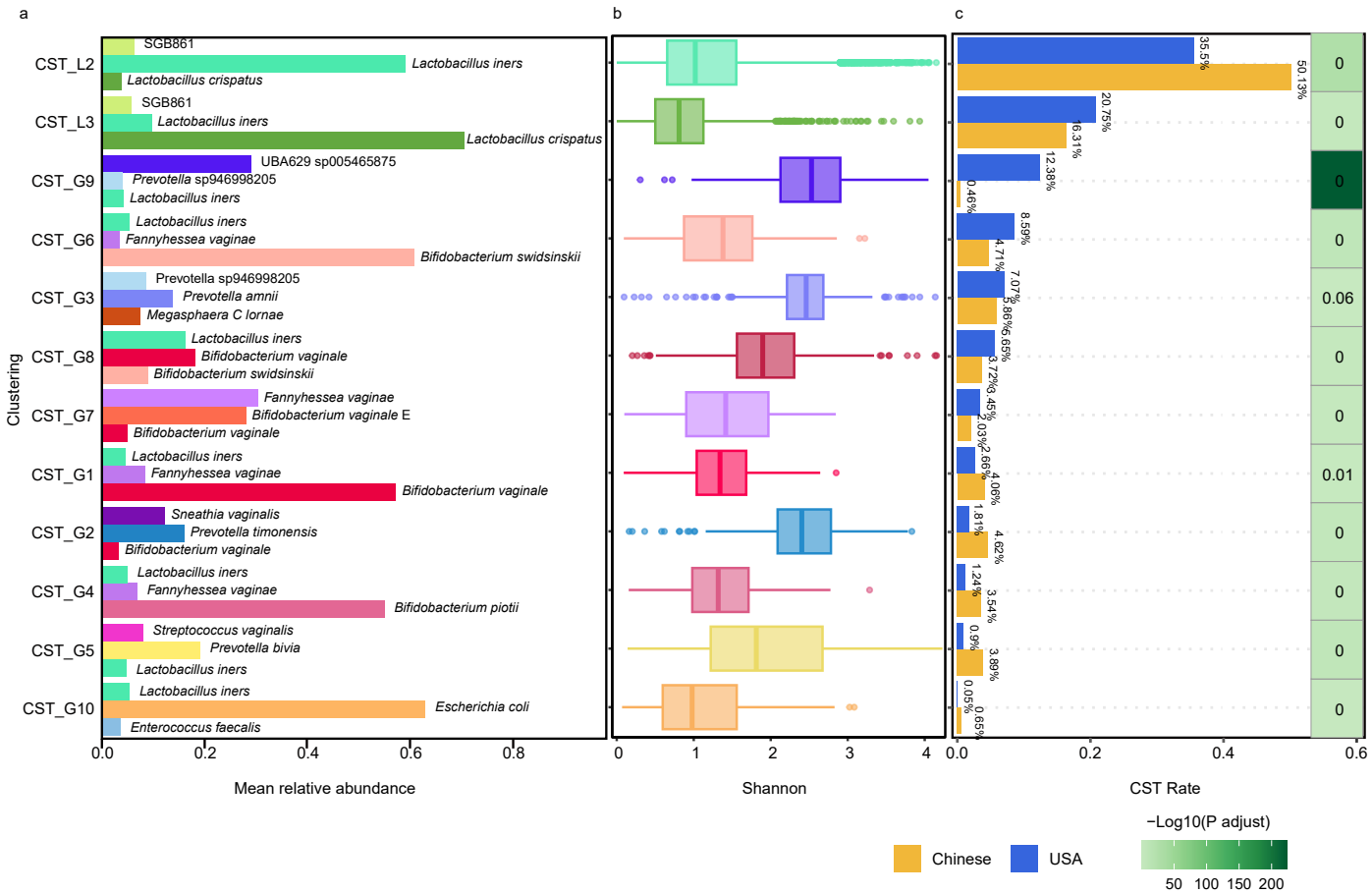
Extended Data Fig. 6 | Characteristics of fungal populations in GVMG. **a**, An approximately maximum-likelihood phylogenetic tree for the 43 fungal genomes. The purple square represents genomes derived from the metagenomic binning algorithm, the blue square represents MAGs from public databases, and the red square represents isolate genomes from public databases. **b**, The average rate of reads mapping and prevalence rates for 11 fungal species in the vaginal mycobiome. Data underlying plots are provided.

https://doi.org/10.1038/s41588-026-02639-2



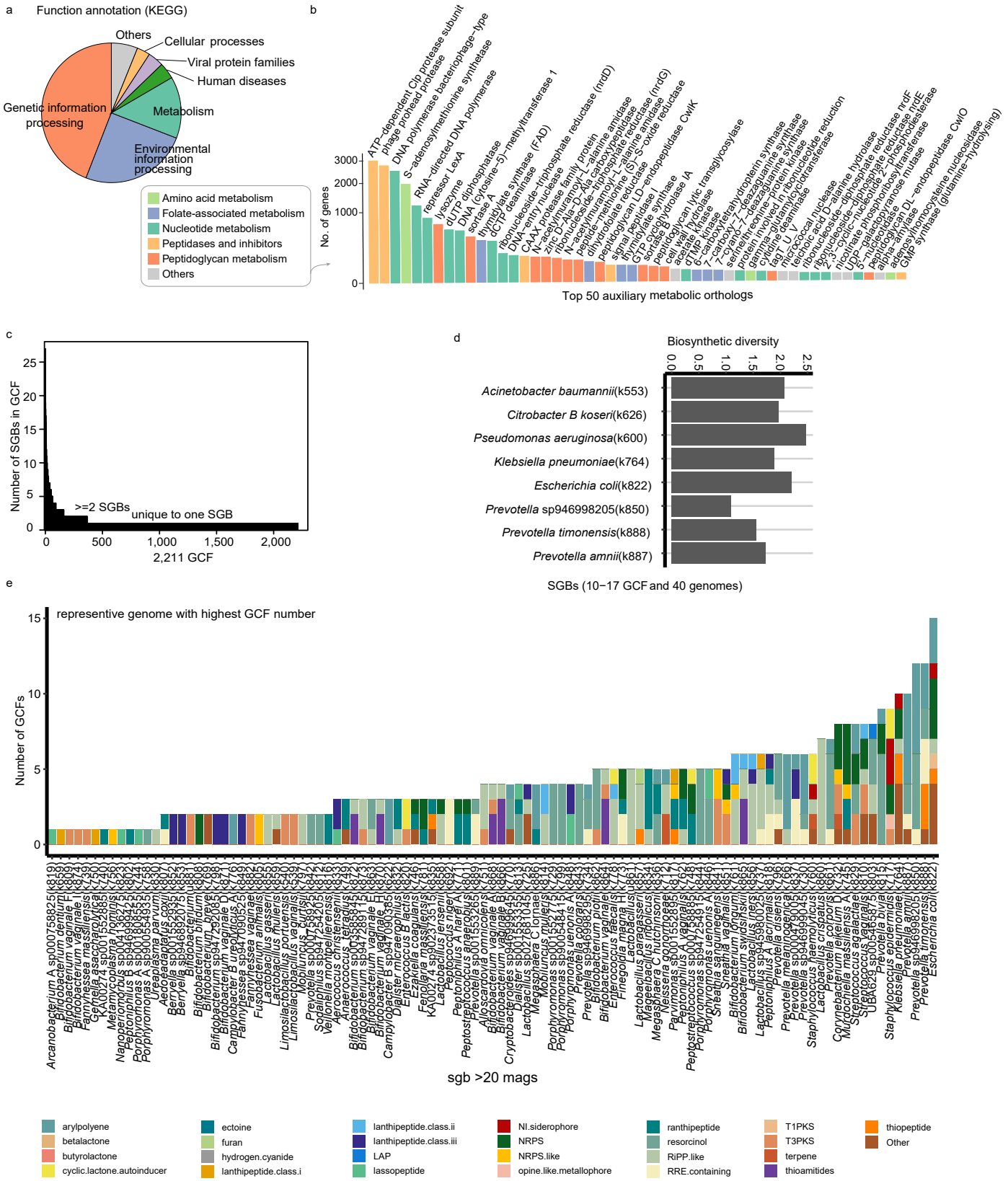
Extended Data Fig. 7 | Phylogenetic diversity for all genomes from *B. vaginale* genomospecies. The phylogenetic tree comprises two concentric rings. The inner ring represents 23 distinct *B. vaginale* genomospecies (i.e., *G. vaginalis* genomospecies), each highlighted with a unique color. The outer ring indicates the sources of the genomes, with different colors distinguishing different origin categories. Data underlying plots are provided.

https://doi.org/10.1038/s41588-026-02639-2



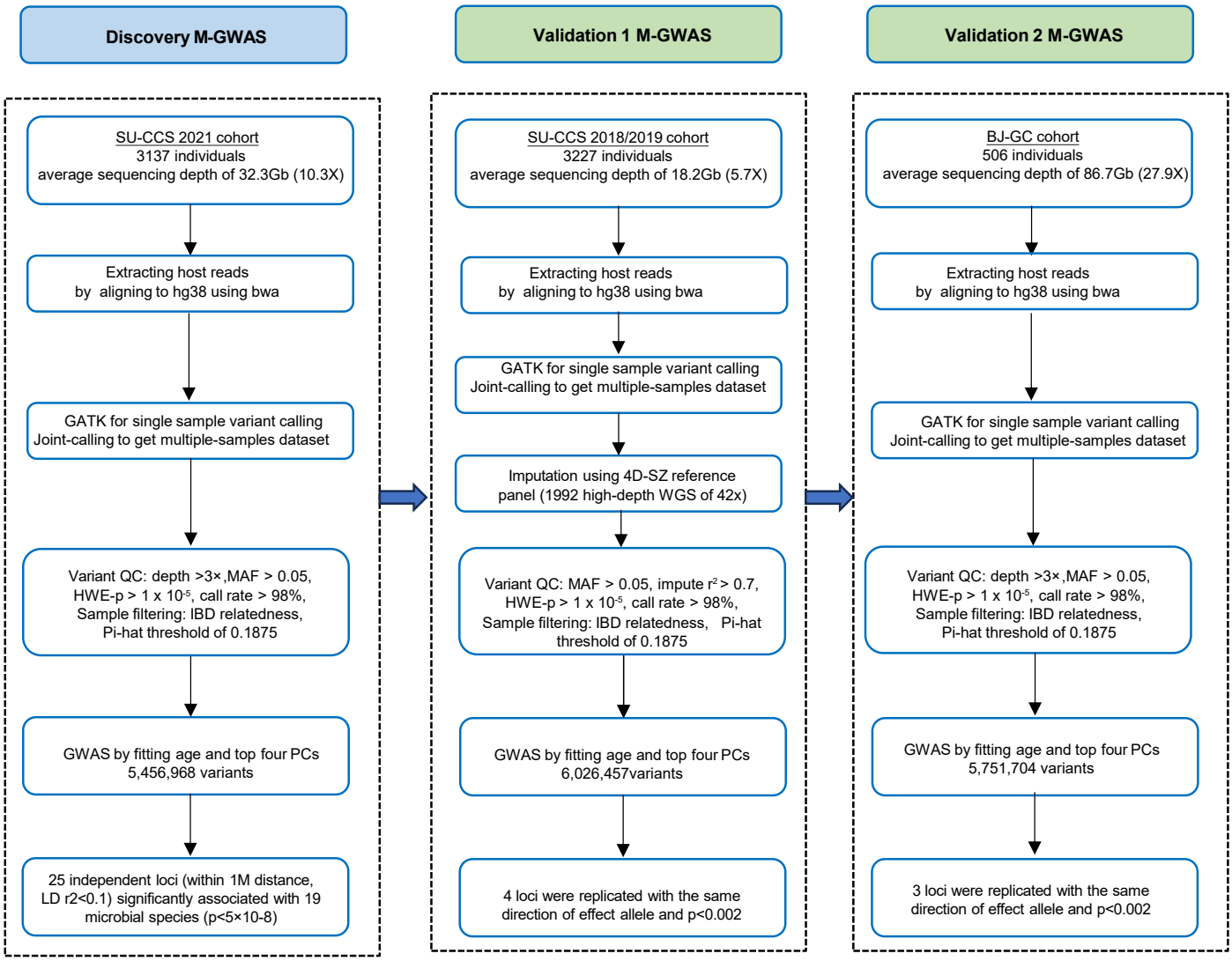
Extended Data Fig. 8 | Community state types (CST) of the nonviral vaginal microbiome. a,**b**, Mean relative abundance of top 3 most abundant species (**a**) and Shannon diversity (**b**) in the 12 CSTs. **c**, Comparison of the CST distributions between Chinese (n = 10,665) and USA (n = 1,769) populations (a two-sided proportion test with FDR-adjusted P-value < 0.05). Data underlying plots are provided.

https://doi.org/10.1038/s41588-026-02639-2



Extended Data Fig. 9 | Functional compendium in GVMG. a, The functional distribution of KEGG-annotated genes for 6,577 vOTUs. **b**, The number of genes from the top 50 auxiliary metabolic orthologs (AMOs) for 6,577 vOTUs. AMOs are defined as KEGG functional ortholog linked the KEGG metabolism pathway. **c**, Comprehensive clustering of BGCs resulting in 2,211 unique, nonredundant gene cluster families (GCFs). Most of the GCFs are unique to one SGB. **d**, The biosynthetic diversity, measured by the Shannon index based on types of natural products, presented for the top 8 BGC-rich SGBs, each containing 10–17 GCFs and more than 40 genomes. **e**, Enumeration of unique GCFs highlights the maximum predicted number of GCFs per genome within each SGB (more than 20 genomes), and GCFs are categorized across different BGC types. Each bar represents the genome with the largest count of BGCs in its SGBs. Data underlying plots are provided.

https://doi.org/10.1038/s41588-026-02639-2



Extended Data Fig. 10 | Flowchart of M-GWAS. Flowchart of the M-GWAS analysis steps for the discovery and two validation datasets.

nature portfolio

Corresponding author(s): Tao Zhang, Lilan Hao, Lan Zhu, Chen Chen

Last updated by author(s): Mar 30, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	All softwares are listed on Methods, Supplementary Note, and analysis code accessed through GitHub (https://github.com/weiting-liang/GVMG/).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Analysis code can be accessed through GitHub (https://github.com/weiting-liang/GVMG). The metagenomic data in Peacock Dataset is available in the CNGB Nucleotide Sequence Archive (CNSA: https://db.cngb.org/cnsa) under accession number CNP0005953 and CNP0006125. The genome database of GVMG, including prokaryotic, eukaryotic and viral genome sequences and the updated Kraken database, have been deposited in the Zenodo repository (https://doi.org/10.5281/

zenodo.14708991). The GWAS summary statistics of microbial taxa are publicly available in the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/) from accession GCST90573122 to GCST90573174. All these data are freely available for download and analysis without login requirements or usage restrictions.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Due to the vagina organ, this study only involve female participants. The profile of the study was explained in detail to all participants and informed consents were obtained from all participants.
Reporting on race, ethnicity, or other socially relevant groupings	We identified the population specific signatures species, intraspecies, and functional diversity, comparing vaginal microbiome from China, USA and France.
Population characteristics	We established the Peacock project in China and recruited 10,665 Chinese participants, from whom we collected cervicovaginal samples (Fig.1a, Supplementary Table 1). The cohort encompassed three urban populations: gynecological clinic attendees from Beijing (BJ-GC; n= 2,878; mean age 42.4 ± 12.0 years), routine health examinees from Shenzhen (SZ-4D; n= 833; mean age 31.9 ± 4.9 years), and participants from organized cancer screening programs in Suzhou, spanning two periods: 2018-2019 (SU-CCS2018/2019; n=3,355; mean age 48.3 ± 7.5) and 2021 (SU-CCS2021; n=3,527; mean age 50.8 ± 7.7).
Recruitment	All the samples in Peacock project were collected following the same protocols approved and with written informed consent from all participants.
Ethics oversight	The study was approved by the Institutional Review Board of Peking Union Medical College Hospital in Beijing (PUMCH) (ZS-1683), the Institutional Review Board of Suzhou Municipal Hospital (IEC-C-008-A07-V1.0) and the Institutional Review Board of BGI (BGI-IRB 19027-T3).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We established the Peacock project in China, collecting 10,665 cervicovaginal samples from various regions, including Suzhou in Eastern China (n=6,882), Beijing in Northern China (n=2,878), Shenzhen in Southern coastal China (n=833), and other regions (n=72).
Data exclusions	Date is not applicable.
Replication	Replication is not applicable.
Randomization	This is not an experimental study. Randomization is not applicable.
Blinding	This is not an experimental study. Blinding is not applicable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a