Expanding vaginal microbiome pangenomes via a custom MIDAS database reveals *Lactobacillus crispatus* accessory genes associated with cervical dysplasia

Claire A. Dubin[1,2], Chunyu Zhao[3,4], Katherine S. Pollard[3,4,5], Tomiko Oskotsky[1,2,6], Jonathan L. Golob[7]*, Marina Sirota[1,2,8]*[#]

1. March of Dimes Prematurity Research Center at the University of California San Francisco, San Francisco, CA, USA
2. Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA, USA
3. Institute of Data Science & Biotechnology, Gladstone Institutes, San Francisco, California, USA
4. Data Science, Chan Zuckerberg Biohub Network, San Francisco, California, USA
5. Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, California, USA
6. Division of Clinical Informatics and Digital Transformation, Department of Medicine, University of California, San Francisco, San Francisco, CA, USA
7. Division of Allergy and Infectious Diseases, University of Washington School of Medicine, Seattle, WA, USA
8. Department of Pediatrics, University of California San Francisco, San Francisco, CA, USA

*These authors contributed equally to this work.

[#]Corresponding author: marina.sirota@ucsf.edu

ABSTRACT

The vaginal microbiome plays a central role in reproductive health. Vaginal microbiome dysbiosis is associated with many adverse reproductive health outcomes, but most studies have focused on associations at the species level. The potential contribution of intraspecies microbial variation, especially gene content differences across bacterial strains, remains underexplored in reproductive health contexts. The Metagenomic Intra-Species Diversity Analysis (MIDAS) framework enables such analyses but depends on comprehensive reference databases. We constructed a MIDAS-compatible pangenome database from over 18,000 genomes in the Vaginal Microbiome Genome Collection (VMGC). Compared to the Genome Taxonomy Database (GTDB)-derived reference, the VMGC database expanded pangenomes of prevalent vaginal species, better capturing vaginal-specific intraspecies diversity. Applying this database to vaginal samples from a cervical dysplasia cohort, we identified thirteen *Lactobacillus crispatus* accessory genes significantly associated with cervical dysplasia, including a HicAB toxin-antitoxin system, three transcriptional regulators, and three phage-derived genes. These findings highlight the utility of body site-specific reference resources and shotgun metagenomic sequencing for uncovering intraspecies microbial variation relevant to reproductive health.

IMPORTANCE: The vaginal microbiome plays a critical role in reproductive health, and different bacteria from the same species can carry different genes that influence how the strains interact with the host and other microbes. These strain-level differences are often overlooked when microbiomes are analyzed only at the species level. Existing genomic reference databases are heavily biased toward gut and environmental bacteria, leaving the genetic diversity of vaginal microbes understudied. We built a specialized reference database from over 18,000 vaginal bacterial genomes that better reflects this diversity. We then applied this resource to quantify gene-level variation in vaginal samples from a cervical dysplasia cohort. Focusing on *Lactobacillus crispatus*, a prevalent and often beneficial vaginal species, we identified thirteen genes that were more common in women with cervical dysplasia than in controls. This work demonstrates that body site-specific genomic resources are essential for uncovering strain-level bacterial differences relevant to reproductive health.

INTRODUCTION

The microbiome is increasingly recognized as an essential factor in health and disease throughout the human body.[1] While the gut is the most well-studied system in microbiome research, microbial communities in other body sites also play an important role in host physiology and disease susceptibility.[2,3] Bidirectional interactions between the vaginal microbiome and host play a key role in reproductive health.[4] Disruption to this balance, often called vaginal dysbiosis, is linked to adverse reproductive health outcomes, including preterm birth[5–7] and endometriosis.[8,9] Vaginal dysbiosis involves a shift away from *Lactobacillus*-dominated communities toward more diverse and potentially pathogenic taxa.[4] Pathogenic bacteria in the vaginal microbiome compromise the epithelial barrier and weaken immune responses, heightening susceptibility to infection by other pathogenic bacteria, yeast, and viruses.[4,10] For example, persistent human papillomavirus (HPV) infection is the primary precursor to cervical dysplasia,[11,12] the precancerous growth of abnormal cells on the cervix. Cervical dysplasia is a significant public health concern due to its potential to progress to cervical cancer, the fourth most common cancer in women.[12,13] Further clarifying the role of the vaginal microbiome in preventing or promoting cervical

dysplasia is essential for elucidating its underlying etiology, identifying biomarkers, and developing therapeutics.

While the majority of vaginal microbiome studies have focused on composition at the species level, intraspecies genomic variation also plays a role in the functional potential and ecological dynamics of the vaginal microbiome. *Lactobacillus crispatus*, a widely prevalent species often regarded as a marker of a healthy vaginal microbiome, has a large accessory genome compared to other common vaginal *Lactobacillus* species.[14] Frequent gain and loss of genes across *L. crispatus* strains results in genetic and phenotypic variation among members of the species. *L. crispatus* vaginal isolates vary in their ability to utilize glycogen sources, produce hydrogen peroxide, and survive acidic stress.[15] Genes encoding mucin-binding proteins and bacteriocins, which enhance adhesion to host tissues and inhibit pathogen growth, respectively, are variably present across *L. crispatus* strains.[15–17] Vaginal *L. crispatus* strains differ in their ability to suppress the growth of bacterial and fungal pathogens, and *L. crispatus* and other common vaginal microbiome species exhibit different co-occurrence patterns at the strain level.[18] These observations suggest that *L. crispatus* may not universally inhibit progression to a dysbiotic vaginal microbiome state.[19] Recognizing and characterizing this intraspecies variation is essential for identifying reliable biomarkers of disease and developing probiotics to support reproductive health.

The Metagenomic Intra-Species Diversity Analysis (MIDAS) pipeline is a versatile tool for measuring intraspecific variation at the nucleotide and gene level from shotgun metagenomic sequencing data.[20,21] MIDAS can be used to quantify bacterial gene content in metagenomic samples by mapping metagenomic reads to a preconstructed pangenome reference database. Intraspecies gene presence/absence data generated by MIDAS can be used in metagenome-wide association studies (MWAS) to identify bacterial accessory genes associated with clinical outcomes. For instance, recent work using the MWAS-based model microSLAM identified a fructoselysine utilization operon in *Faecalibacterium prausnitzii* that is negatively associated with inflammatory bowel disease.[22] Given the demonstrated utility of MIDAS and microSLAM in identifying accessory genes associated with disease states and the mounting evidence for intraspecies genotypic and

phenotypic variation in the vaginal microbiome, we sought to optimize this framework to identify associations between accessory genes in common vaginal microbiome species and reproductive health.

In this work, we address the limited representation of vaginal-derived genomes in existing MIDAS pangenome databases by developing a pipeline for constructing MIDAS-compatible custom pangenome reference databases and applying it to construct a pangenome database from a collection of over 18,000 vaginal genomes. We demonstrate that this database expands the pangenomes of six prevalent vaginal microbiome species to better reflect the genetic diversity of the vaginal microbiome compared to existing MIDAS pangenome databases. We then apply the newly constructed database to quantify gene content in vaginal microbiome samples and identify thirteen *L. crispatus* accessory genes associated with cervical dysplasia, including a toxin-antitoxin system and phage-derived genes of unknown function.

RESULTS

Construction of a MIDAS-compatible pangenome reference database from vaginal genomes

Strain-level association analyses rely on comprehensive and representative reference genomes. Reference-based gene content analyses are dependent on the quality of the reference database; genes that are absent from the reference cannot be measured, and an excess of irrelevant genes can introduce unnecessary noise. We set out to develop a MIDAS-compatible pangenome database specifically for use with vaginal microbiome specimens to refine and expand the landscape of genes available for reproductive health analyses. Two prebuilt MIDAS-compatible databases were previously constructed from genomes in the Unified Human Gastrointestinal Genome collection (UHGG)[23] and the Genome Taxonomy Database r202 (GTDB).[24] UHGG exclusively contains genomes from the human gut microbiome, and although GTDB contains a broader set of genomes, genomes from gastrointestinal samples make up approximately 35% of human-derived GTDB genomes with specified origin. In contrast, genomes from other human body sites such as the female

reproductive tract, oral cavity, and skin make up only 1%, 2%, and 4%, respectively. Because more than 80% of GTDB r202 genomes are derived from cultured isolates, reference databases built from these genomes may not fully capture the uncultured and strain-diverse taxa prevalent in these underrepresented body sites. Recent collections comprised primarily of metagenome-assembled genomes (MAGs) have increased the number of available genomes for species in these less-studied niches.[25–27] Expanding pangenome databases to include these MAGs may better reflect the microbial genetic diversity within these body sites.

We constructed a custom MIDAS-compatible pangenome reference database using 19,542 prokaryotic genomes in the Vaginal Microbiome Genome Collection (VMGC).[25] The vast majority of genomes in VMGC (95%) are custom-assembled MAGs. Only 3% of the genomes in the VMGC are also present in the existing GTDB-based MIDAS reference, and these shared genomes are all derived from isolate sequencing. The VMGC prokaryotic genome assemblies and their GTDB 214.1 taxonomic assignments were downloaded from Huang et al (2024).[25] There is a median of 19 input genomes per species (IQR 7.0-93.5), and genome quality ranges from medium-quality (52%), to high-quality (43%), and near-complete (5%) as calculated in Huang et al. (2024).[25] We developed a Nextflow pipeline to streamline construction of a MIDAS-compatible pangenome database from a user-specified set of genomes and their species assignments as described in Zhao et al. (2023)[21] and Smith et al. (2025)[20] (Figure 1A). Using this pipeline, we created a pangenome reference database for each prokaryotic species with at least five assigned genomes, totaling 18,970 genomes across 179 species, 81 genera, and 42 families. Predicted genes from all genomes assigned to a species were clustered by average nucleotide identity (ANI) thresholds ranging from 75-99%. Higher ANI thresholds can distinguish subtle differences among highly similar genes, while lower thresholds can group more distant homologs. To balance detection of gene variants and grouping of gene families, we focused our pangenome analysis on 90% ANI gene clusters. The number of 90% ANI gene clusters per species ranges from 713 to 18,533 (IQR 2358.0-4409.5) and is positively correlated with the number of input genomes (Figure 1B, Spearman $\rho = 0.57$, $p = 5e\text{-}17$, Table S1). This trend is consistent across all tested ANI thresholds, with pangenome size decreasing as gene clusters are merged at lower ANI thresholds (Figure S1, Table S2).

Comparison to GTDB-only database

We next compared the ability of the VMGC-derived MIDAS pangenome reference database to reflect the genetic diversity of organisms in the vaginal microbiome as compared to the GTDB-derived reference. Pangenome expansion widens the set of prokaryotic genes that can be tested for differential prevalence across strains and associated with clinical data. We compared pangenome sizes (number of 90% ANI gene clusters) of 129 species with one-to-one mappings across GTDB versions r202 and r214.1, which were used to determine taxonomic classification in the VMGC and GTDB databases, respectively (Table S1). Of these species, 66% (85/129) have a larger pangenome in the database constructed from VMGC genomes, and 34% (44/129) have a larger GTDB pangenome (Figure 2A). Differences in pangenome size are positively correlated with differences in the number of genomes per species in each database (Figure S2, Spearman $\rho = 0.92$, $p = 8\times10^{-55}$). The five species with the greatest difference in pangenome size between GTDB and VMGC, all with larger GTDB pangenomes, are gut-associated species (*Phocaeicola vulgatus, Enterococcus faecalis, Klebsiella pneumoniae*) and/or widely sampled pathogenic species (*Acinetobacter baumannii, Escherichia coli*). Species with a larger VMGC pangenome include all four hallmark *Lactobacillus* species used to classify vaginal community state types (*L. crispatus, L. gasseri, L. iners,* and *L. jensenii*),[28] as well as species associated with vaginal microbiome dysbiosis, including six *Bifidobacterium vaginale* species (collectively referred to as *Gardnerella vaginalis* in NCBI taxonomy) and *Fannyhessea vaginae* (formerly named *Atopobium vaginae*).

We used MIDAS to quantify the relative abundance of 129 species with one-to-one mappings across GTDB versions among 352 vaginal samples from Norenhag et al. (2024)[29] (Figure S3). Of the 86 species with relative abundance greater than 1% in at least one vaginal sample, 76% had larger pangenomes in the database built from VMGC genomes. VMGC has a larger pangenome and more input genomes for all six species with relative abundance greater than 5% in at least 5% of samples (Figure 2A-B, S4). We compared pangenome completeness between databases using rarefaction analysis on these six species. Tail slopes from collector's curves (calculated using the four largest subsampling points per database) were lower in VMGC than GTDB for these six species (Wilcoxon $p = 0.0312$), indicating that VMGC-derived pangenomes are

approaching saturation while GTDB pangenomes continue to accumulate novel gene content as more genomes are added (Figure 2C-D). When comparing pangenome sizes at matched genome counts, GTDB and VMGC were similar across species, except for *L. crispatus*, where the GTDB-based pangenome was larger at lower genome counts. The same trend persisted after stratifying *L. crispatus* genomes from the female reproductive tract by genome type (i.e., MAGs vs. isolates) (Figure S5A-B), indicating the discrepancy may be driven by incompleteness in *L. crispatus* MAGs. In *L. crispatus* and other species, isolate genomes from the female reproductive tract harbored a higher number of predicted mobile elements and viral sequences than female reproductive tract MAGs (Figure S5C-D). Although MAGs are generally less complete than isolate genomes, the substantial quantity of vaginal MAGs in the VMGC database offsets their lower per-genome contribution, resulting in a larger total pangenome across all six species.

VMGC-based pangenome database expands accessory gene repertoire in prevalent vaginal bacteria

Pangenome expansion is most meaningful when it uncovers biologically relevant genes, rather than merely increasing the number of genes associated with a species. To further characterize differences between pangenome reference databases constructed from VMGC and GTDB genomes, we analyzed the overlap and distinctiveness of the gene clusters across databases. We rebuilt pangenome databases with the set of all genomes in VMGC and GTDB for the six prevalent species identified above. The resulting 90% ANI gene clusters were classified as follows: VMGC-only (containing sequences from genomes found exclusively in VMGC), GTDB-only (containing sequences from genomes found exclusively in GTDB), or shared (containing sequences from genomes found in VMGC and GTDB; i.e., at minimum, a single sequence from a genome found in both GTDB and VMGC, or two sequences: one from a genome found only in VMGC and one from a genome found only in GTDB). The number of VMGC-only genes was larger than the number of GTDB-only genes for all six species (Figure 3A).

To see how commonly the database-specific genes are detected in vaginal samples, we used MIDAS and the combined reference database to measure gene presence and absence in vaginal samples from Norenhag et

al. (2024).[29] While *L. crispatus* has a relatively large number of GTDB-only genes, less than 3% (74/2686) of these genes are detected among the 130 vaginal samples with *L. crispatus* present (Figure 3B). Across all six species, shared genes are more frequently detected than VMGC-only and GTDB-only genes, and more VMGC-only genes are detected than GTDB-only genes (Figure 3B-C). Among genes detected in at least one sample, VMGC-only genes are found at a significantly higher frequency than GTDB-only genes in *L. crispatus*, *L. gasseri,* and *B. vaginale* (two-sided Mann Whitney U $p = 7x10^{-6}$, 0.029, and 0.005 respectively). Higher-frequency accessory genes are particularly valuable in association analyses, as they provide greater statistical power to detect associations with clinical traits.

Accessory genes often encode functions that are relevant to fitness in a specific environment. We used a permutation analysis to identify differentially prevalent PFAM domains in VMGC-only and GTDB-only gene clusters. *L. crispatus* GTDB-only centroids were enriched for eleven PFAM domains, including glycosyltransferase domains, cell surface-related domains, and a bacteriophage abortive infection AbiH domain (Table 1). In contrast, *L. crispatus* VMGC-only centroids were enriched for three domains associated with integrated phage sequences: BRO family N-terminal domain, AAA domain, and Terminase_1 (Table 1). *L. crispatus* VMGC-only genomes had a higher number of predicted viral regions compared to GTDB-only genomes (Mann Whitney U $p = 0.001$, Figure 4A). More broadly, *L. crispatus* genomes from human hosts had a higher number of viral regions than those from non-human hosts ($p = 1x10^{-5}$, Figure 4B), and genomes from the human female reproductive tract had more predicted viral regions than genomes from the human gastrointestinal tract ($p = 0.012$) and the combined set of genomes from all other human body sites ($p = 0.014$) (Figure 4C). These trends remained consistent when subsetting to only high-quality genomes or only isolate genomes (Figure S6). These findings align with the well-documented abundance of bacteriophages in the vaginal microbiome.[30,31] Furthermore, they demonstrate that the site-specific database can capture genomic virus-bacteria interactions at the genomic level, even when accounting for the potential limitations of binning-based assembly[32] and the lower quality of VMGC MAGs relative to isolate genomes.

Thirteen *Lactobacillus crispatus* accessory genes are associated with increased cervical dysplasia risk

We applied the VMGC pangenome database to shotgun metagenomic data from a Swedish cervical dysplasia cohort[29] to evaluate its effectiveness in detecting intraspecies associations with clinical phenotypes.
To model these associations, we employed the MWAS-based tool microSLAM.[22] *L. crispatus* was chosen for analysis as the sole species with sufficient genome coverage to meet microSLAM thresholds (>80 samples) for robust gene detection. Of 131 samples with *L. crispatus*, 82 are from healthy controls and 49 are from individuals with cervical dysplasia. We focused our association analysis on 75% ANI gene clusters, representing broader functional gene groups, to increase statistical power by reducing multiple testing. We identified 832 75% ANI gene clusters detected in 15% to 75% of these samples. We used microSLAM to model the associations between the presence/absence vectors of these genes and disease state, also including *L. crispatus* relative abundance, a propensity score calculated from five clinical metadata variables, and population structure as covariates (see Methods). We identified thirteen accessory genes associated with cervical dysplasia (locFDR < 0.2; OR = 2.48-6.19; Figure 5A, Table S3). These genes encode six hypothetical proteins, a protein with a domain of unknown function, three transcriptional regulators, an ion transporter, and both components of the HicAB type II toxin-antitoxin system. All significant genes are more frequently detected in cervical dysplasia samples than in samples from healthy controls (Figure 5B). The ion transporter is a VMGC-only gene. Two groups of genes have similar co-occurrence patterns across samples (Jaccard similarity >= 0.8) (Figure 5C).

Genomic context of cervical dysplasia-associated genes

To further elucidate potential functions of the cervical dysplasia-associated genes, we analyzed their genomic context in high-quality *L. crispatus* isolates (Table S4). We identified all *L. crispatus* genome assemblies available on NCBI constructed from single-isolate (non-metagenomic) sequencing and a ‘complete genome’ assembly level. We used BLASTn to identify predicted coding sequences from these 22 strains with homology to the thirteen significant gene sequences. The significant genes are variably detected across the *L. crispatus* isolate genomes (Figure 6A), consistent with our variable detection of them in genomes. The sets of genes

with similar co-occurrence patterns across samples were co-located in syntenic regions of the isolate genomes (Figure 6B and 6C). We used Operon-mapper[33] to predict co-regulated genes in these regions. A group of five cervical dysplasia-associated genes encoding two hypothetical proteins, an XRE family transcriptional regulator, a MarR family transcriptional regulator, and a protein with a domain of unknown function are genomically adjacent, but only the latter two genes are predicted to be co-regulated (Figure 6B). The group of three co-occurring genes encoding hypothetical proteins is predicted to be a part of an operon within an approximately 40 kb region predicted to be a *Caudoviricetes* provirus (Figure 6C, S7). The HicAB toxin-antitoxin system components are predicted to be part of an operon that also includes two hypothetical proteins (Figure 6D). The LacI transcriptional regulator is part of a potential carbohydrate utilization operon (Figure 6E). The remaining hypothetical protein is predicted to be co-regulated with a GNAT family N-acetyltransferase (Figure 6F), and the *nhaC* ion transporter is not predicted to be co-regulated with any other genes (Figure 6G).

To assess the broader phylogenetic distribution of these genes, we used BLASTx[34] to search for homologous sequences across bacterial species (Table S5). The two top-scoring hypothetical proteins (1 and 2) showed no significant hits outside of *L. crispatus*, suggesting they may be species-specific. Hypothetical proteins 3-6, the XRE family transcriptional regulator, and the DUF771 domain-containing protein had homologs within other members of the *Lactobacillaceae* family. The *nhaC* antiporter, LacI family transcriptional regulator, and MarR family transcriptional regulator were found throughout *Lactobacillaceae* and detected in a limited number of species outside this family. The two toxin-antitoxin system genes showed the broadest distribution, with homologs detected across diverse bacterial species. Collectively, the cervical dysplasia-associated genes span diverse functional categories, genomic contexts, and degrees of taxonomic specificity; the prevalence of proteins with unknown functions in and near the set of cervical dysplasia-associated genes highlights the need for functional studies to investigate the mechanistic basis of these associations.

DISCUSSION

In this study, we constructed a vaginal microbiome-specific pangenome database using the Vaginal Microbiome Genome Collection and demonstrated its utility in identifying associations between bacterial accessory genes and clinical phenotypes. Previously published MIDAS pangenome databases use genomes largely derived from gastrointestinal and environmental organisms, limiting their ability to reflect the genetic diversity of bacteria from other body sites such as the human vaginal microbiome. The integration of MAGs is especially valuable for expanding pangenomes without the prerequisite of *in vitro* cultivation. This is a critical consideration for the vaginal microbiome, as many resident species are anaerobic and experimentally challenging to culture and isolate.[35] For hallmark *Lactobacillus* species, as well as dysbiosis-associated organisms like *Fannyhessea vaginae* and *Bifidobacterium vaginale*, use of VMGC genomes increased pangenome size and completeness, with percent increase in size ranging from 21% to 332%. Pangenome expansion uncovered accessory genes that were frequently detected in vaginal samples and enriched for environment-specific functions. For *L. crispatus*, vaginal genomes were enriched for regions of viral origin and PFAM domains associated with integrated phage, consistent with prior observations that viral elements are widely prevalent in the vaginal niche.[31,36] One of these cervical dysplasia-associated genes (*nhaC* ion transporter) is found in the VMGC pangenome database but not in the GTDB database, highlighting the importance of using a body site-specific database in MWAS analyses.

We identified thirteen *L. crispatus* accessory genes associated with increased risk of cervical dysplasia, despite the species' overall protectiveness against cervical dysplasia and other adverse reproductive health outcomes.[37] These include transcriptional regulators, toxin-antitoxin system components, and proteins of unknown function, several of which were embedded within phage-associated regions or putative operons. The HicAB toxin-antitoxin system promotes biofilm formation and inhibits translation to slow bacterial growth in response to stress, promoting persistence during antibiotic treatment and acid stress in *Burkholderia pseudomallei* and *Acetobacter pasteurianus.*[38–40] Other type II antitoxin systems are involved in abortive infection of phage DNA, and a similar role has been proposed for the HicAB system via comparative genomics.[41] Given the well-documented relationship between HPV infection and cervical dysplasia,[11] the

significance of three genes of unknown function in a phage-derived region is particularly notable and warrants further study of virus-bacteria interactions in this context. HPV infection and cervical dysplasia are frequently accompanied by dysbiosis, inflammation, pH changes, and barrier disruption; these factors are known to induce stress responses, alter metabolic demands, and drive prophage induction.[42,43] The significant genes involved in acid stress (*nhaC* ion transporter), sugar metabolism (LacI family transcriptional regulator), and persistence (HicAB toxin-antitoxin system) may promote survival in the hostile, high-inflammation environment associated with cervical dysplasia. Experimental work is needed to understand any mechanistic role of these genes. They may be directly involved in the etiology of cervical dysplasia, act as a response to the environmental changes associated with dysbiosis, or make *L. crispatus* more permissive to the growth of pathogenic species, increasing the likelihood of transitioning to a dysbiotic state that favors viral infection and persistence.

The heterogeneity within *L. crispatus* populations underscores the need to move beyond species-level assessments when studying host-microbiome relationships. In particular, distinguishing beneficial from neutral or potentially pathogenic sublineages may be critical for understanding why *L. crispatus* dominance is not universally protective against dysbiosis and adverse outcomes. Norenhag, et al. 2024[44] identified amino acid biosynthesis and sugar degradation pathways enriched in healthy controls and nucleotide and peptidoglycan biosynthesis pathways enriched in cases of cervical dysplasia.[29] *L. crispatus* strongly contributed to the pathways enriched in healthy controls, but within-species variation of these pathways was not reported. This is the first study to our knowledge to analyze bacterial intraspecies variation in relation to cervical dysplasia and points to the analytic gain from a richer, more tailored, reference database.

Unlike existing scripts for constructing a MIDAS-compatible pangenome database, our Nextflow pipeline is automated, highly parallelized, and compatible with multiple executors. This workflow may be particularly useful for constructing pangenome databases from genome collections recently published for the oral and skin microbiomes,[26,27] as well as keeping reference sets updated and up-to-date with new data. VIRGO2, an expansive catalog of non-redundant vaginal microbiome genes pooled across species, is also an excellent resource for measuring genetic diversity in the vaginal microbiome.[45] Our reference database contains

pangenomes constructed for each species individually, and adjustable ANI clustering thresholds to suit the desired level of granularity. Our approach may be particularly advantageous for studies focused on specific species or for seamless integration with the MIDAS pipeline.

This study has several limitations. We did not stratify cases by severity of dysplasia, as the sample sizes within each subgroup would be too small for statistical analysis. Access to healthcare influences the availability of cervical dysplasia screening and screening practices vary across countries,[46] so individuals included in this study may not fully represent the broader population of people with and without cervical dysplasia; generalizing these results to other populations will require validation in independent cohorts. Furthermore, users of the VMGC reference should consider several limitations inherent to reference-based approaches. Contributing isolate genomes and MAGs in VMGC are predominantly derived from samples from the United States, China, and Europe,[47] which may limit the representation of genetic diversity from other geographic regions. The use of MAGs may introduce bias from assembly errors, chimeric contigs, or incomplete gene predictions.[32] However, these challenges can be mitigated using improved assembly algorithms, and our reproducible Nextflow pipeline allows pangenomes to be easily reconstructed as MAGs are refined. Future users should evaluate pangenome size, saturation, and quality of contributing genomes when deciding whether the pangenome reference for a particular species is sufficiently comprehensive for their application or whether de novo gene content characterization would be more appropriate. Our methodology for accurate measurement of gene content relies on sufficient sequencing depth across at least 80 samples. This technical constraint limited our scope to *L. crispatus*, preventing exploration of potential gene-disease associations in other bacterial species. Moreover, while our model accounts for the relative abundance of *L. crispatus*, it does not incorporate the broader taxonomic composition of the vaginal microbiome. The top dysplasia-associated gene is a hypothetical protein unique to *L. crispatus* with no known protein domains, underscoring the need for approaches to improve the functional annotation of vaginal pangenomes. Finally, the significant genes identified here are associations rather than causal determinants of cervical dysplasia; experimental validation is necessary to confirm their roles and characterize the genes with unknown function.

Together, our work establishes both a resource and a methodological framework for expanding MIDAS pangenome databases to better represent body site-specific microbiomes. The VMGC-based database improves the resolution of metagenomic analyses in the vaginal microbiome, reveals novel accessory genes of potential clinical importance, and demonstrates that intraspecies diversity within key taxa like *L. crispatus* may shape reproductive health outcomes. Extending similar approaches to other body sites and diseases will help uncover overlooked functional variation within the human microbiome and elucidate its role in health and disease.

TABLES

Table 1: PFAM enrichment

Species	Enrichment	PFAM	BH-adjusted *p*
Lactobacillus crispatus	Enriched in GTDB-only centroids	Bacteriophage abortive infection AbiH	0.0080
		YSIRK type signal peptide	0.0080
		LPXTG cell wall anchor motif	0.0080
		Methylase_S	0.0080
		Glycosyl transferase family 2	0.0080
		Glycosyl transferases group 1	0.0080
		N-6 DNA Methylase	0.0140
		Polysaccharide biosynthesis C-terminal domain	0.0224
		Glycosyltransferase Family 4	0.0224
		Hexapep	0.0420
		SLAP (surface layer-associated protein) domain	0.0440
	Enriched in VMGC-only centroids	BRO family, N-terminal domain	0.0080
		AAA domain	0.0356
		Terminase_1	0.0430
Fannyhessea vaginae	Enriched in VMGC-only centroids	CHAP domain	0.0190
		Methylase_S	0.0190

METHODS

Categorization of GTDB genome sources

We downloaded metadata for the 258,405 GTDB genomes used in the GTDB-based pangenome database via the NCBI datasets command line tool version 18.0.2.[48] We identified 98,680 genomes with the words “human” or “homo sapiens” in the Assembly BioSample Host, Assembly BioSample Source type, Assembly BioSample Isolation source, Assembly BioSample Models, and/or Assembly BioProject Lineage Title. We manually categorized the 187 most common isolation sources into the following categories: skin, gastrointestinal, respiratory, female reproductive tract, oral, blood, urine, other, unknown, and non-human. These manual annotations categorized the source of 91% (89,483/98,680) of human-associated genomes. We used the gemma-2-9b-it[45] large language model to categorize the 3,022 remaining isolate sources corresponding to 9% of genomes. We removed genomes whose sources remained unknown (n = 33,573) or were recategorized as non-human (n = 532) before calculating the proportions of genomes isolated from each body site. For the six species depicted in Figure 2B and S4, we manually verified and re-annotated genomes whose sources were categorized as unknown and non-human through a literature search.

Genome assembly download

Prokaryotic vaginal genomes and their taxonomic assignments to GTDB r214.1 were downloaded from Huang et al. (2024).[25] Accession numbers and GTDB r202 species assignments for 258,405 genomes in the preconstructed pangenome database generated from GTDB r202 genomes were downloaded via MIDAS version 3.0.1.[20] We used gtdb-taxdump version v0.5.0[50] to map taxonomic assignments across versions r202 and r214.1 of GTDB. Species not present in both database versions and species with many-to-one mappings across versions were excluded from the comparison of pangenome databases. We verified that genomes from the remaining species were assigned to the same species in GTDB versions r202 and r214.1, and we downloaded GTDB genome assemblies via the NCBI datasets command line tool version 18.0.2.[48]

Nextflow pipeline for constructing custom pangenome databases

Nextflow is an HPC and workflow orchestration system widely used in biomedical research notable for versatility and scalability. We created a Nextflow pipeline to streamline construction of a pangenome database from a user-specified set of genomes and their species assignments. The steps for constructing a pangenome reference database are described in Zhao et al. (2023)[21] and Smith et al. (2025)[20] and Figure 1A. Briefly, each genome is annotated via Prokka,[51] and predicted protein-coding sequences are clustered at 99% ANI using VSEARCH.[52] The centroids of each cluster are reclustered at 95, 90, 85, 80, and 75% ANI, and refined using CD-HIT.[53] Centroid sequences of 99% ANI clusters are functionally annotated using eggNOG.[54] For each genome, regions associated with antimicrobial resistance, viral elements, and plasmid/mobile element-associated regions are identified using ResFinder,[55] geNomad,[56] and MobileElementFinder,[57] respectively. We verified that our Nextflow pipeline outputs identical results to the original database build scripts published in Smith et al. (2025)[20] for two sample species. The database build from VMGC genomes took 199 CPU days (16 real-time days) on a high-performance computing cluster, distributed across 506 nodes equipped with Intel and AMD processors and 12-256 CPUs and 48-1512 GiB of RAM per node. Most processes used 4 or fewer CPUs and less than 16GB of RAM; the Nextflow pipeline dynamically increased CPU and RAM requests for more computationally intensive processes. While this work was completed on a Sun Grid Engine-managed cluster, Nextflow enables easy transfer to SLURM, AWS, GCP, and other HPC systems.

Cervical dysplasia samples

We downloaded shotgun metagenomic sequencing reads from European Nucleotide Archive projects PRJEB72778 and PRJEB72779 and corresponding metadata from Norenhag et al. (2024).[29] This is an age-matched case-control study conducted in Sweden (N=177 healthy controls and N=177 cervical dysplasia). We used Geneshot[58] to run a quality control workflow on the sequencing data, using Trim Galore version 0.6.6[59] to remove adapter sequences, bwa version 0.7.18[60] to identify sequences mapping to the human genome version hg38, and samtools version 1.20[61] to filter out the non-microbial reads. We removed samples with fewer than 250,000 reads remaining after quality control, leaving 352 samples for analysis (175 samples from healthy controls and 177 samples from individuals with cervical dysplasia).

We quantified species-level composition in each sample using the MIDAS run_species command and merged results across samples with the merge_species command. We considered a species to be present in a sample if the median coverage across single-copy marker genes was at least 0.75 and relative abundance was at least 1%. 138 and 137 species were detected in at least one sample when using the VMGC and GTDB-based pangenome database as a reference, respectively.

Rarefaction analysis

We selected six species with a relative abundance greater than 5% in at least 5% of samples for rarefaction analysis. For each species and database, rarefaction was performed by selecting a subset of genomes, rebuilding the pangenome database with only those genomes as input, and calculating the number of 90% average nucleotide identity gene clusters in the resulting pangenome database. The number of genomes selected began with n-1 genomes, where n is the total number of genomes available for that species, and decreased by 10% until 2 genomes remained. Ten replicates at each number of genomes were performed.

Gene cluster uniqueness

Genomes from the six species selected above were assigned categories: VMGC-unique (only present in the VMGC database), GTDB-unique (only present in the GTDB database), or common (present in both the VMGC and GTDB databases). For each species, we rebuilt the pangenome database with the combined set of VMGC-unique, GTDB-unique, and common genomes as input. 90% ANI gene clusters were categorized as “VMGC-only” if they only contained sequences derived from VMGC-unique genomes, “GTDB-only” if they only contained sequences derived from GTDB-unique genomes, and shared if they contained sequences from a common genome or sequences from both VMGC-unique and GTDB-unique genomes.

PFAM enrichment

We used eggNOG-generated PFAM domain annotations to test for functional enrichment in VMGC- and GTDB-only gene clusters for each species. $N_{clusters}$ represents the total number of VMGC-only or GTDB-only 90% ANI gene clusters for a species, and N_{PFAM} represents the number of database-unique gene clusters with a given PFAM domain. For each PFAM with N_{PFAM} greater than or equal to ten, we took 10,000 samples of

$N_{clusters}$ gene clusters from the set of all gene clusters for the species. We used the proportion of samples with more than $_{NPFAM}$ gene clusters with the given PFAM annotation as an empirical *p* value. *p* values for each species were adjusted for multiple hypothesis testing via the Benjamini Hochberg procedure.[62]

Gene presence/absence

The MIDAS run_genes command was used to measure gene content in each vaginal microbiome sample for all species detected in the sample with median coverage of marker genes greater than 2 and reads mapping to at least 50% of marker genes. Gene clusters with copy number of 0.35 or greater were considered present in a sample. We observed that gene detection varied significantly at low sequencing depths. To ensure reliable gene presence/absence determination, we identified the minimum mean depth required for gene counts to stabilize for each species. For most species (*L. crispatus, L. jensenii, L. gasseri, F. vaginae*, and *B. vaginale*), we set the threshold at 5X, where the correlation between depth and gene count was not significant (Spearman $p > 0.05$). For *L. iners*, we iteratively increased the cutoff and determined that a higher threshold of 8X was necessary to eliminate this significant correlation (at 5X coverage Spearman $p = 0.0005$; at 8X coverage Spearman $p = 0.21$). Results of the gene-level analysis were merged across samples via the merge_genes command, with genes clustered at 90% ANI for pangenome database comparisons and 75% ANI for the cervical dysplasia association analysis. The six-species combined database was used as a reference for the database comparisons (Figure 3-4) and the pangenome database constructed from VMGC genomes was used as a reference in the cervical dysplasia association analysis (Figures 5-6).

Propensity score calculation

For each of 21 metadata variables reported in Norenhag et al. (2024), we constructed a logistic regression model: outcome ~ variable + 1, with outcome = 0 representing healthy controls and outcome = 1 representing cervical dysplasia. Five variables with $p < 0.1$ in the univariate models were used in the propensity score: LowRisk_HPV, PotentialHR_HPV, Gyn_problem_24hrs, Menses_yesno, and Gyn_BloodyDischarge. Each of these variables is binary. LowRisk_HPV and PotentialHR_HPV did not have missing values for any samples. The variable Menses_yesno was missing data for 7% of samples; we imputed missing values for this column based on subject age. Subjects with AgeAtSampling < 52 were imputed with True, and subjects with

AgeAtSampling >= 52 were imputed with False. 52 was selected as the threshold because the average age of menopause is 51-52.[63] Gyn_problem_24hrs and Gyn_BloodyDischarge had low frequency of nonzero values (16% and 2% of samples, respectively), and for these columns we imputed missing values (10% of samples for both variables) with the mode (0). We fit a logistic regression model with the matrix of selected variables as features and outcome as the response variable. The predicted probability of each sample belonging to the disease group based on this model was used as a propensity score.

Cervical dysplasia association analysis

We used the microSLAM R package version 1.0.0.9[22] to model associations between *L. crispatus* gene content and outcome (*y*). We calculated fixed effects of covariates (α) using a baseline generalized linear model (GLM): y ~ propensity score + *L. crispatus* relative abundance. We constructed a gene presence/absence matrix from 75% ANI gene clusters detected in 15-75% of samples with mean depth of at least 5 across detected *L. crispatus* genes (n = 702 genes and 131 samples). Next, we calculated a gene relatedness matrix (GRM), a pairwise similarity matrix based on gene presence/absence data used as a proxy for population structure across samples. To calculate the random effects between population structure and outcome (*b*), we combined the baseline GLM and the GRM in a random-effects generalized linear model. Permutation of the tau variable from this model indicated there was not a significant correlation between population structure and outcome ($p = 0.34$). Lastly, we fit a mixed effects model for each gene presence/absence vector *g:*

$$logit(y_i) = X_i\alpha + \beta g_i + b_i + \varepsilon.$$

where *i* represents a sample, *X* is the covariate matrix, and *β* represents the effect size of the gene. All models and statistical tests except for the baseline GLM were implemented via microSLAM. We used the locfdr R package version 1.1.8[64] to calculate false discovery rates from the *z* scores derived from the mixed effects model. Significant *L. crispatus* gene clusters were annotated via BLASTx[34] of the centroid sequences against the ClusteredNR protein database. To determine if these genes were present in other species, we filtered for homologs meeting UniRef50 criteria (>50% identity, >80% query coverage) while excluding matches to *L. crispatus.*

Analysis of *L. crispatus* isolate genome assemblies

We identified 22 *L. crispatus* genome assemblies available on NCBI with sequencing data from a single isolate, assembly level of "complete assembly," and available coding sequence annotations. Accession numbers for these isolates are provided in Table S4. We used BLASTn version 2.16.0[34] with minimum query and subject coverage of 75% and minimum identity of 75% to identify predicted coding sequences in each strain with homology to the centroids of 75% ANI gene clusters for *L. crispatus*. The gene with the highest bit score was selected for each coding sequence. If a gene did not have any matches from a strain, it was considered absent in that strain. We used the Operon-mapper web tool[33] to predict the location of operons and geNomad[56] to predict viral regions in each strain. PyGenomeViz version 1.6.1[65] was used for visualizations in Figures 6B-G and S8. The PyGenomeViz integration of MMseqs2 reciprocal best hit search identified sequences with identity of at least 50% across at least 80% of the query and target sequences. The phylogenetic tree for these isolates was created using StrainPhlAn version 3.1.68,[66,67] with parameters --diversity low, --trim greedy, and –accurate. The closely related species *Lactobacillus helveticus* was selected as the outgroup[68], and the *L. helveticus* reference strain TCI357 genome assembly was downloaded from NCBI (Table S4). The phylogenetic tree was visualized with ITOL.[69] The full phylogenetic tree scaled by branch length is available in Figure S7.

Data Availability: The VMGC pangenome database is publicly available on Zenodo (DOI: 10.5281/zenodo.18355814), with genes clustered at 75%, 80%, 85%, 90%, 95%, and 99% average nucleotide identity for each of 179 species. The Nextflow pipeline for constructing custom MIDAS-compatible pangenome databases is available at https://github.com/clairedubin/MIDASv3_database_build_nf and code for data analysis presented here is available at https://github.com/clairedubin/pangenome_cervical_dysplasia_analysis. The GTDB-based pangenome database was published in Zhao, et al. 2023[21] and can be downloaded through the MIDAS package.

Funding: This work was funded by the March of Dimes Prematurity Research Center at the University of California, San Francisco.
Conflicts of Interest: The authors declare no conflicts of interest.
STORMS Checklist: The STORMS checklist for this project is available at: https://github.com/clairedubin/pangenome_cervical_dysplasia_analysis/blob/main/STORMS_Excel_1.03.xlsx

REFERENCES

1. Rackaityte E, Lynch SV. The human microbiome in the 21st century. *Nat Commun*. 2020;11(1):5256. doi:10.1038/s41467-020-18983-8

2. Reynoso-García J, Miranda-Santiago AE, Meléndez-Vázquez NM, et al. A complete guide to human microbiomes: Body niches, transmission, development, dysbiosis, and restoration. *Front Syst Biol*. 2022;2. doi:10.3389/fsysb.2022.951403

3. Asangba AE, Mugisha L, Rukundo J, et al. Large Comparative Analyses of Primate Body Site Microbiomes Indicate that the Oral Microbiome Is Unique among All Body Sites and Conserved among Nonhuman Primates. *Microbiol Spectr*. 2022;10(3):e01643-21. doi:10.1128/spectrum.01643-21

4. Zhu B, Tao Z, Edupuganti L, Serrano MG, Buck GA. Roles of the Microbiota of the Female Reproductive Tract in Gynecological and Reproductive Health. *Microbiol Mol Biol Rev*. 2022;86(4):e00181-21. doi:10.1128/mmbr.00181-21

5. Liao J, Shenhav L, Urban JA, et al. Microdiversity of the vaginal microbiome is associated with preterm birth. *Nat Commun*. 2023;14(1):4997. doi:10.1038/s41467-023-40719-7

6. Fettweis JM, Serrano MG, Brooks JP, et al. The vaginal microbiome and preterm birth. *Nat Med*. 2019;25(6):1012-1021. doi:10.1038/s41591-019-0450-2

7. Golob JL, Oskotsky TT, Tang AS, et al. Microbiome preterm birth DREAM challenge: Crowdsourcing machine learning approaches to advance preterm birth research. *Cell Rep Med*. 2023;5(1):101350. doi:10.1016/j.xcrm.2023.101350

8. Qing X, Xie M, Liu P, et al. Correlation between dysbiosis of vaginal microecology and endometriosis: A systematic review and meta-analysis. *PLOS ONE*. 2024;19(7):e0306780. doi:10.1371/journal.pone.0306780

9. MacSharry J, Kovács Z, Xie Y, et al. Endometriosis specific vaginal microbiota links to urine and serum N-glycome. *Sci Rep*. 2024;14(1):25372. doi:10.1038/s41598-024-76125-2

10. Dabee S, Passmore JAS, Heffron R, Jaspan HB. The Complex Link between the Female Genital Microbiota, Genital Infections, and Inflammation. *Infect Immun*. 2021;89(5). doi:10.1128/iai.00487-20

11. Sharifian K, Shoja Z, Jalilvand S. The interplay between human papillomavirus and vaginal microbiota in cervical cancer development. *Virol J*. 2023;20(1):73. doi:10.1186/s12985-023-02037-8

12. Perkins RB, Wentzensen N, Guido RS, Schiffman M. Cervical Cancer Screening: A Review. *JAMA*. 2023;330(6):547-558. doi:10.1001/jama.2023.13174

13. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi:10.3322/caac.21660

14. Bhattacharya A, Das S, Bhattacharjee MJ, Mukherjee AK, Khan MR. Comparative pangenomic analysis of predominant human vaginal lactobacilli strains towards population-specific adaptation: understanding the role in sustaining a balanced and healthy vaginal microenvironment. *BMC Genomics*. 2023;24(1):565. doi:10.1186/s12864-023-09665-y

15. Mancabelli L, Mancino W, Lugli GA, et al. Comparative Genome Analyses of Lactobacillus crispatus Isolates from Different Ecological Niches Reveal an Adaptation of This Species to the Human Vaginal Environment. *Appl Environ Microbiol*. 2021;87(8):e02899-20. doi:10.1128/AEM.02899-20

16. Santarelli G, Sanguinetti M, Delogu G, De Maio F. Comparative genomic analysis of bacteriocin genes in Lactobacillus crispatus strains. *Sci Rep*. 2025;15(1):20798. doi:10.1038/s41598-025-08321-7

17. Nori SRC, Walsh CJ, McAuliffe FM, et al. Strain-level variation among vaginal Lactobacillus crispatus and Lactobacillus iners as identified by comparative metagenomics. *Npj Biofilms Microbiomes*. 2025;11(1):49. doi:10.1038/s41522-025-00682-1

18. Pace RM, Chu DM, Prince AL, Ma J, Seferovic MD, Aagaard KM. Complex species and strain ecology of the vaginal microbiome from pregnancy to postpartum and association with preterm birth. *Med N Y N*. 2021;2(9):1027-1049. doi:10.1016/j.medj.2021.06.001

19. Argentini C, Fontana F, Alessandri G, et al. Evaluation of Modulatory Activities of Lactobacillus crispatus Strains in the Context of the Vaginal Microbiota. *Microbiol Spectr*. 2022;10(2):e02733-21. doi:10.1128/spectrum.02733-21

20. Smith BJ, Zhao C, Dubinkina V, et al. Accurate estimation of intraspecific microbial gene content variation in metagenomic data with MIDAS v3 and StrainPGC. *Genome Res*. 2025;35(5):1247-1260. doi:10.1101/gr.279543.124

21. Zhao C, Dimitrov B, Goldman M, Nayfach S, Pollard KS. MIDAS2: Metagenomic Intra-species Diversity Analysis System. *Bioinformatics*. 2023;39(1):btac713. doi:10.1093/bioinformatics/btac713

22. Goldman M, Zhao C, Pollard KS. Improved detection of microbiome-disease associations via population structure-aware generalized linear mixed effects models (microSLAM). *PLOS Comput Biol*. 2025;21(5):e1012277. doi:10.1371/journal.pcbi.1012277

23. Almeida A, Nayfach S, Boland M, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol*. 2021;39(1):105-114. doi:10.1038/s41587-020-0603-3

24. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res*. 2022;50(D1):D785-D794. doi:10.1093/nar/gkab776

25. Huang L, Guo R, Li S, et al. A multi-kingdom collection of 33,804 reference genomes for the human vaginal microbiome. *Nat Microbiol*. 2024;9(8):2185-2200. doi:10.1038/s41564-024-01751-5

26. Cha JH, Kim N, Ma J, et al. A high-quality genomic catalog of the human oral microbiome broadens its phylogeny and clinical insights. *bioRxiv*. Preprint posted online April 21, 2025:2025.03.10.642329. doi:10.1101/2025.03.10.642329

27. Liu Y, Li Z, Zhang C, et al. A human skin microbiome reference catalog and the skin microbial landscape of plateau adults. *iMetaOmics*. 2025;2(2):e70000. doi:10.1002/imo2.70000

28. Ravel J, Gajer P, Abdo Z, et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A*. 2011;108(Suppl 1):4680-4687. doi:10.1073/pnas.1002611107

29. Norenhag J, Edfeldt G, Stålberg K, et al. Compositional and functional differences of the vaginal microbiota of women with and without cervical dysplasia. *Sci Rep*. 2024;14(1):11183. doi:10.1038/s41598-024-61942-2

30. Jakobsen RR, Haahr T, Humaidan P, et al. Characterization of the Vaginal DNA Virome in Health and Dysbiosis. *Viruses*. 2020;12(10):1143. doi:10.3390/v12101143

31. Gupta P, Singh MP, Goyal K. Diversity of Vaginal Microbiome in Pregnancy: Deciphering the Obscurity. *Front Public Health*. 2020;8. doi:10.3389/fpubh.2020.00326

32. Meziti A, Rodriguez-R LM, Hatt JK, Peña-Gonzalez A, Levy K, Konstantinidis KT. The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Appl Environ Microbiol*. 2021;87(6):e02593-20. doi:10.1128/AEM.02593-20

33. Taboada B, Estrada K, Ciria R, Merino E. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. *Bioinformatics*. 2018;34(23):4118-4120. doi:10.1093/bioinformatics/bty496

34. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421. doi:10.1186/1471-2105-10-421

35. Berman H, McLaren M, Callahan B. Understanding and interpreting community sequencing measurements of the vaginal microbiome. *BJOG Int J Obstet Gynaecol*. 2020;127(2):139-146. doi:10.1111/1471-0528.15978

36. Jakobsen RR, Haahr T, Humaidan P, et al. Characterization of the Vaginal DNA Virome in Health and Dysbiosis. *Viruses*. 2020;12(10):1143. doi:10.3390/v12101143

37. Norenhag J, Du J, Olovsson M, Verstraelen H, Engstrand L, Brusselaers N. The vaginal microbiota, human papillomavirus and cervical dysplasia: a systematic review and network meta-analysis. *BJOG Int J Obstet Gynaecol*. 2020;127(2):171-180. doi:10.1111/1471-0528.15854

38. Xia K, Han C, Xu J, Liang X. Toxin-antitoxin HicAB regulates the formation of persister cells responsible for the acid stress resistance in Acetobacter pasteurianus. *Appl Microbiol Biotechnol*. 2021;105(2):725-739. doi:10.1007/s00253-020-11078-w

39. Butt A, Higman VA, Williams C, et al. The HicA toxin from Burkholderia pseudomallei has a role in persister cell formation. *Biochem J*. 2014;459(2):333-344. doi:10.1042/BJ20140073

40. Jørgensen MG, Pandey DP, Jaskolska M, Gerdes K. HicA of Escherichia coli Defines a Novel Family of Translation-Independent mRNA Interferases in Bacteria and Archaea. *J Bacteriol*. 2009;191(4):1191-1199. doi:10.1128/JB.01013-08

41. Gerdes K. Diverse genetic contexts of HicA toxin domains propose a role in anti-phage defense. *mBio*. 2024;15(2):e0329323. doi:10.1128/mbio.03293-23

42. Cui M, Wu Y, Liu Z, Liu Y, Fan L. Advances in the interrelated nature of vaginal microecology, HPV infection, and cervical lesions. *Front Cell Infect Microbiol*. 2025;15. doi:10.3389/fcimb.2025.1608195

43. De Paepe M, Leclerc M, Tinsley CR, Petit MA. Bacteriophages: an underestimated role in human and animal health? *Front Cell Infect Microbiol*. 2014;4. doi:10.3389/fcimb.2014.00039

44. Norenhag J, Edfeldt G, Stålberg K, et al. Compositional and functional differences of the vaginal microbiota of women with and without cervical dysplasia. *Sci Rep*. 2024;14(1):11183. doi:10.1038/s41598-024-61942-2

45. France MT, Chaudry I, Rutt L, et al. VIRGO2: Unveiling the Functional and Ecological Complexity of the Vaginal Microbiome with an Enhanced Non-Redundant Gene Catalog. *bioRxiv*. Published online March 4, 2025:2025.03.04.641479. doi:10.1101/2025.03.04.641479

46. Bruni L, Serrano B, Roura E, et al. Cervical cancer screening programmes and age-specific coverage estimates for 202 countries and territories worldwide: a review and synthetic analysis. *Lancet Glob Health*. 2022;10(8):e1115-e1127. doi:10.1016/S2214-109X(22)00241-8

47. Huang L, Guo R, Li S, et al. A multi-kingdom collection of 33,804 reference genomes for the human vaginal microbiome. *Nat Microbiol*. 2024;9(8):2185-2200. doi:10.1038/s41564-024-01751-5

48. O'Leary NA, Cox E, Holmes JB, et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci Data*. 2024;11(1):732. doi:10.1038/s41597-024-03571-y

49. Team G, Riviere M, Pathak S, et al. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv*. Preprint posted online October 2, 2024:arXiv:2408.00118. doi:10.48550/arXiv.2408.00118

50. Shen W, Ren H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *J Genet Genomics*. 2021;48(9):844-850. doi:10.1016/j.jgg.2021.03.006

51. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinforma Oxf Engl*. 2014;30(14):2068-2069. doi:10.1093/bioinformatics/btu153

52. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*. 2016;4:e2584. doi:10.7717/peerj.2584

53. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150-3152. doi:10.1093/bioinformatics/bts565

54. Huerta-Cepas J, Szklarczyk D, Heller D, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. 2019;47(D1):D309-D314. doi:10.1093/nar/gky1085

55. Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FM. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microb Genomics*. 2022;8(1):000748. doi:10.1099/mgen.0.000748

56. Camargo AP, Roux S, Schulz F, et al. Identification of mobile genetic elements with geNomad. *Nat Biotechnol*. 2024;42(8):1303-1312. doi:10.1038/s41587-023-01953-y

57. Johansson MHK, Bortolaia V, Tansirichaiya S, Aarestrup FM, Roberts AP, Petersen TN. Detection of mobile genetic elements associated with antibiotic resistance in Salmonella enterica using a newly developed web tool: MobileElementFinder. *J Antimicrob Chemother*. 2021;76(1):101-109. doi:10.1093/jac/dkaa390

58. Minot SS, Barry KC, Kasman C, Golob JL, Willis AD. geneshot: gene-level metagenomics identifies genome islands associated with immunotherapy response. *Genome Biol*. 2021;22(1):135. doi:10.1186/s13059-021-02355-6

59. Krueger F, James F, Ewels P, et al. FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. Published online February 2, 2023. doi:10.5281/zenodo.7598955

60. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324

61. Twelve years of SAMtools and BCFtools | GigaScience | Oxford Academic. Accessed August 27, 2025. https://academic.oup.com/gigascience/article/10/2/giab008/6137722?login=false

62. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B Methodol*. 1995;57(1):289-300.

63. Gold EB, Bromberger J, Crawford S, et al. Factors Associated with Age at Natural Menopause in a Multiethnic Sample of Midlife Women. *Am J Epidemiol*. 2001;153(9):865-874. doi:10.1093/aje/153.9.865

64. Efron B, Turnbull B, Narasimhan B, Strimmer K. locfdr: Computes Local False Discovery Rates. Published online July 15, 2015. Accessed August 27, 2025. https://cran.r-project.org/web/packages/locfdr/index.html

65. Shimoyama Y. pyGenomeViz: A genome visualization python package for comparative genomics. Published online May 2024. Accessed August 27, 2025. https://github.com/moshi4/pyGenomeViz

66. Blanco-Míguez A, Beghini F, Cumbo F, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol*. 2023;41(11):1633-1644. doi:10.1038/s41587-023-01688-w

67. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res*. 2017;27(4):626-638. doi:10.1101/gr.216242.116

68. Walter J, O'Toole PW. Microbe Profile: The Lactobacillaceae. *Microbiology*. 2023;169(12):001414. doi:10.1099/mic.0.001414

69. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res*. 2024;52(W1):W78-W82. doi:10.1093/nar/gkae268

FIGURE LEGENDS

Figure 1. Construction of a custom pangenome database. A. Schematic for Nextflow pipeline for constructing a custom pangenome database from genomes of a given species. ANI = average nucleotide identity, AMR = antimicrobial resistance. **B.** Number of genomes vs. number of 90% ANI gene clusters for each species with at least 5 genomes in VMGC. Each point represents a species, and points are colored by average genome size for the species. Marginal histograms show the distributions of each variable: number of genomes per species (top) and number of 90% ANI gene clusters in species pangenome (right).

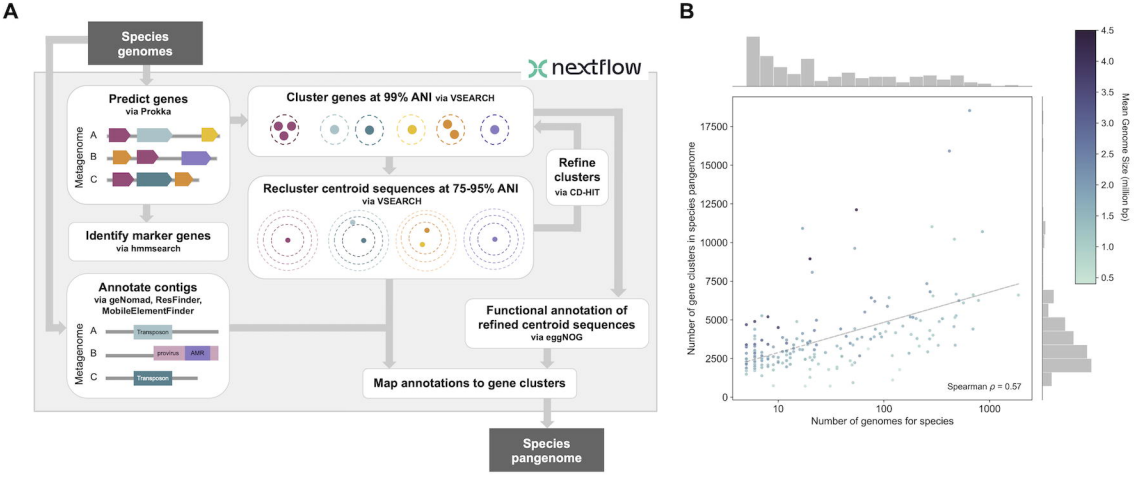
Figure 2. Comparison of pangenome databases constructed from VMGC and GTDB genomes. A. Comparison of the number of 90% ANI gene clusters in the pangenome among shared species in databases constructed from VMGC and GTDB genomes. Species with relative abundance greater than 5% in at least 5% of samples are labeled in bold. Species detected in at least one of 352 samples from Norenhag, et al. (2024) are depicted with filled circles, and species not detected in any of these samples are depicted with an unfilled circle. **B.** Comparison of the number of genomes and their origins for six prevalent vaginal microbiome species. *Bifidobacterium vaginale* is a member of the clade referred to as *Gardnerella vaginalis* in NCBI taxonomy, and *Fannyhessea vaginae* was formerly known as *Atopobium vaginae*. See Figure S4 for proportional representation of genome sources by database. **C-D.** Rarefaction of pangenome size for six selected species using subsampled genomes from VMGC (C) and GTDB (D). The number of sampled genomes was decreased by 10% at each subsampling point, with ten replicates at each subsampling point. The generated data points (number of sampled genomes vs. pangenome size) were fit to power functions.

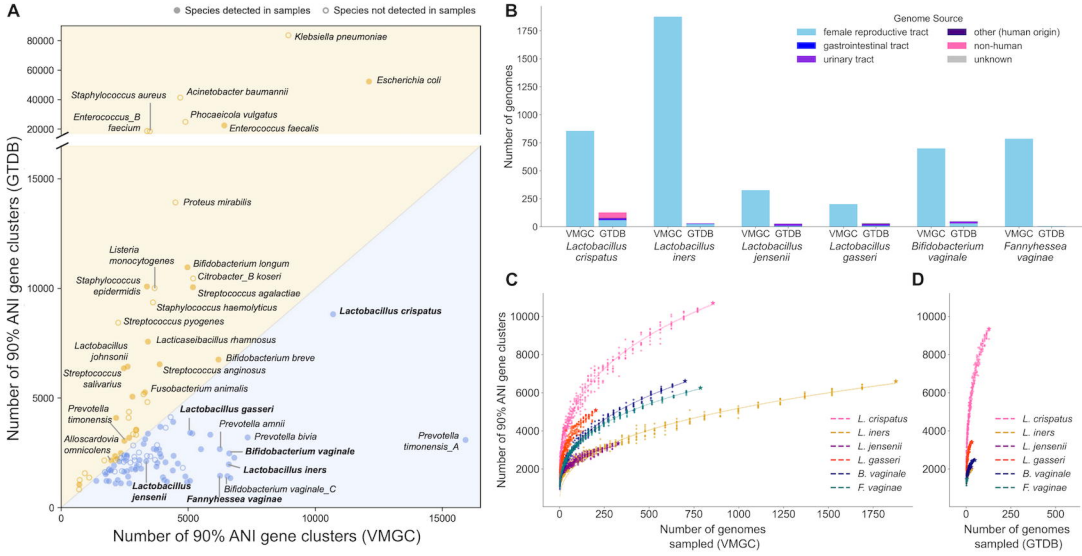
Figure 3. Detection of shared and unique gene clusters in pangenomes constructed from VMGC and GTDB genomes. Each row corresponds to one of six prevalent vaginal microbiome species (labeled at left). **A**. Venn diagrams showing overlap between 90% ANI gene clusters in VMGC and GTDB pangenomes. The number of gene clusters within each group are overlaid on the circles. **B.** Proportion of gene clusters detected (darker color) or not detected (lighter color) in samples from Norenhag et al. (2024), stratified by cluster classification. **C.** Frequency of detection of gene clusters detected in at least one sample, stratified by cluster classification.

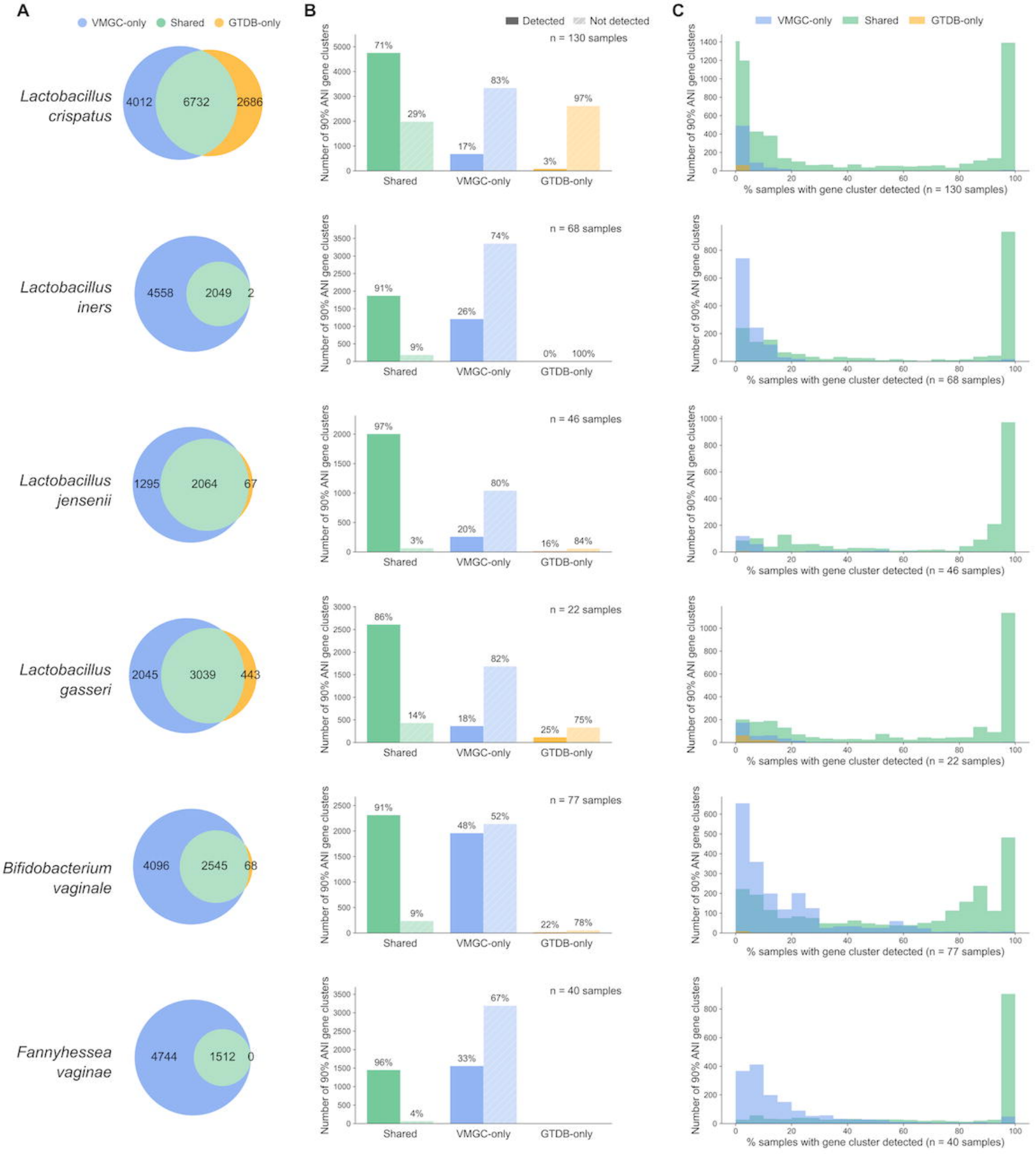
Figure 4. Comparison of number of predicted viral regions among *L. crispatus* genomes, stratified by database category (A), host (B), and human body site (C). Viral regions in genome assemblies were predicted with geNomad. The number of genomes in each group is indicated on the x-axes. Pairwise comparisons within plots were made via a two-sided Mann Whitney U test; pairs with significant differences in number of viral regions are indicated with brackets. **** $p \leq 0.0001$; *** $p \leq 0.001$; ** $p \leq 0.01$; * $p \leq 0.05$. FRT = female reproductive tract.

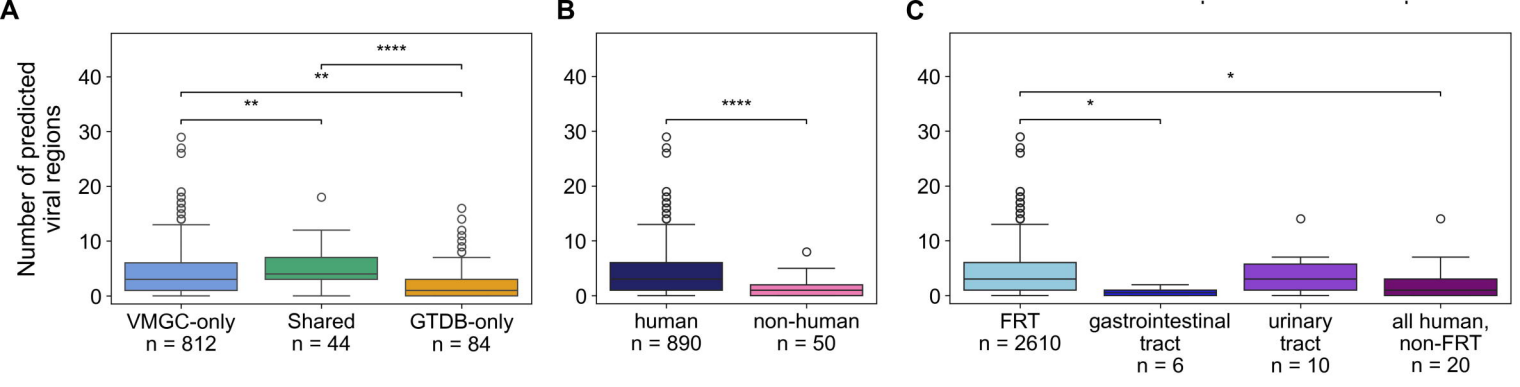
Figure 5. *L. crispatus* accessory genes associated with cervical dysplasia. A. Effect size and significance of 832 accessory genes used as input into association analysis. Genes with locFDR < 0.2 are annotated. **B.** Prevalence of significant genes across samples; significant genes are detected in a larger proportion of cervical dysplasia samples compared to healthy controls. **C.** Presence and absence of significant genes across samples, clustered by Jaccard similarity of gene presence/absence vectors. Genes in red brackets have pairwise Jaccard similarity greater than or equal to 0.8 for all genes in the group.

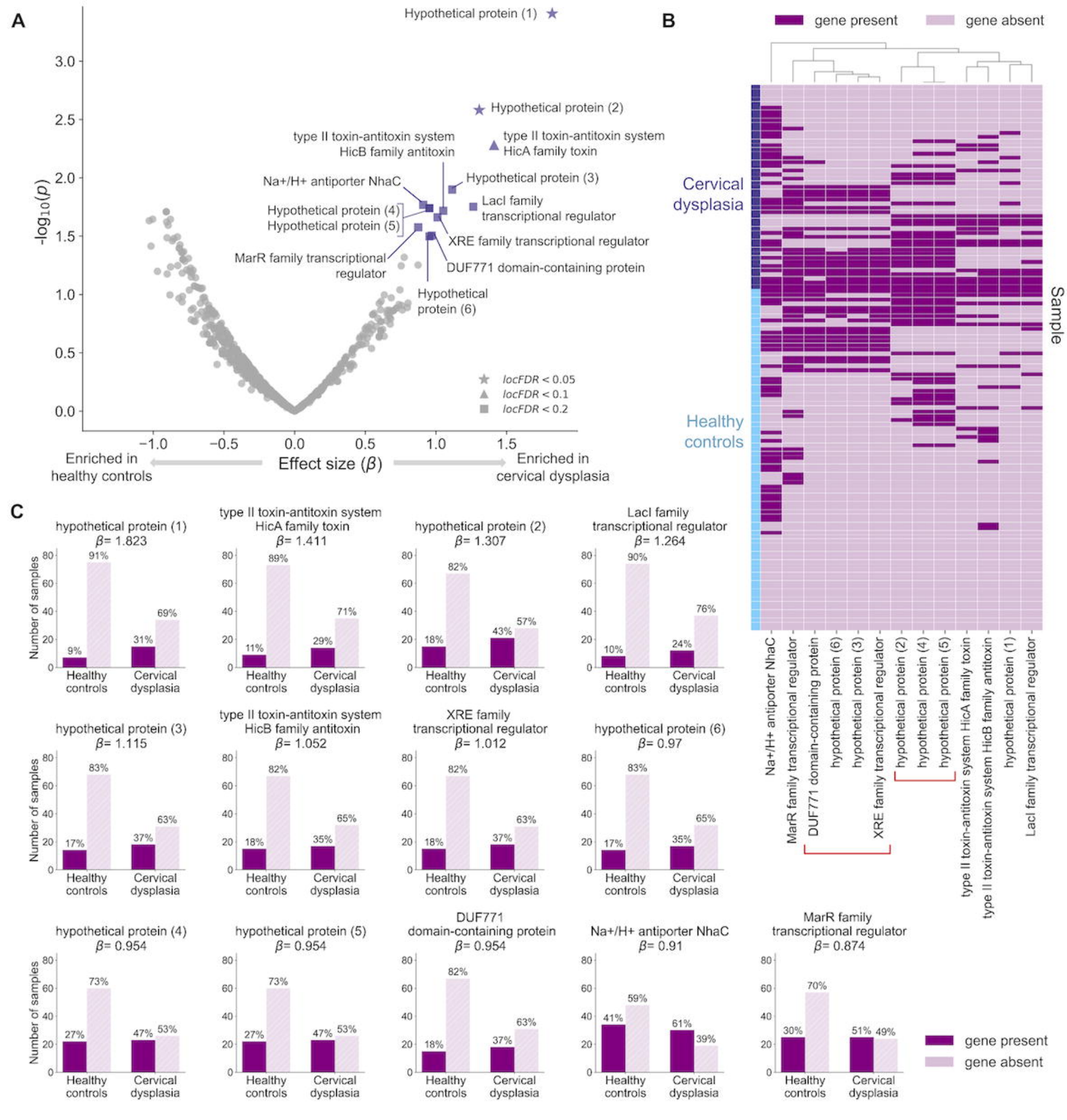
Figure 6. Detection and genomic context of cervical dysplasia-associated genes in NCBI *L. crispatus* isolates. A. Presence/absence of cervical dysplasia-associated genes among 22 *L. crispatus* genome assemblies from single-isolate sequencing. The phylogenetic tree on the right side of the plot was constructed via StrainPhlAn using the complete genome assemblies and depicts the branching topology only (branch lengths are arbitrary). The phylogenetic tree with branch lengths scaled to genetic distance is depicted in Figure S7. **B-G.** 5kb genomic context windows around significant genes among NCBI isolates with the given gene(s) detected. Gray connections across isolates indicate genes are reciprocal best hits as identified by MMseqs2, with sequence identity of at least 75% across at least 75% of both sequences. Brown segments indicate operons predicted by Operon-mapper.

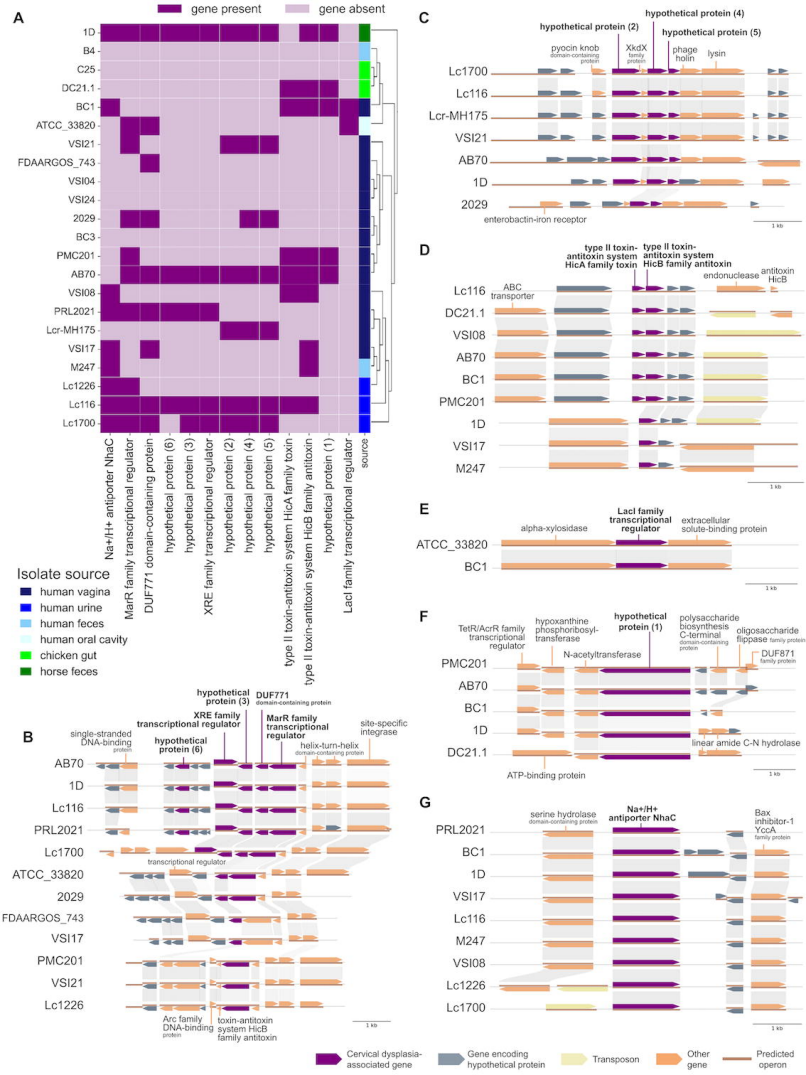

A
Species genomes
nextflow
Predict genes
via Prokka
Metagenome
A
B
C
Cluster genes at 99% ANI via VSEARCH
Refine clusters
via CD-HIT
Recluster centroid sequences at 75-95% ANI
via VSEARCH
Identify marker genes
via hmmsearch
Annotate contigs
via geNomad, ResFinder, MobileElementFinder
Metagenome
A
B
C
Transposon
provirus
AMR
Transposon
Functional annotation of refined centroid sequences
via eggNOG
Map annotations to gene clusters
Species pangenome
B
Number of gene clusters in species pangenome
17500
15000
12500
10000
7500
5000
2500
0
10
100
1000
Number of genomes for species
Spearman ρ = 0.57
Mean Genome Size (million bp)
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5


A
Species detected in samples
Species not detected in samples
Klebsiella pneumoniae
Escherichia coli
Staphylococcus aureus
Acinetobacter baumannii
Enterococcus_B faecium
Phocaeicola vulgatus
Enterococcus faecalis
Proteus mirabilis
Listeria monocytogenes
Bifidobacterium longum
Staphylococcus epidermidis
Citrobacter_B koseri
Streptococcus agalactiae
Staphylococcus haemolyticus
Lactobacillus crispatus
Streptococcus pyogenes
Lacticaseibacillus rhamnosus
Lactobacillus johnsonii
Bifidobacterium breve
Streptococcus salivarius
Streptococcus anginosus
Fusobacterium animalis
Lactobacillus gasseri
Prevotella timonensis
Prevotella amnii
Prevotella bivia
Prevotella timonensis_A
Alloscardovia omnicolens
Bifidobacterium vaginale
Lactobacillus iners
Lactobacillus jensenii
Bifidobacterium vaginale_C
Fannyhessea vaginae
Number of 90% ANI gene clusters (GTDB)
Number of 90% ANI gene clusters (VMGC)
B
Genome Source
female reproductive tract
gastrointestinal tract
urinary tract
other (human origin)
non-human
unknown
Number of genomes
VMGC GTDB
Lactobacillus crispatus
Lactobacillus iners
Lactobacillus jensenii
Lactobacillus gasseri
Bifidobacterium vaginale
Fannyhessea vaginae
C
Number of 90% ANI gene clusters
Number of genomes sampled (VMGC)
L. crispatus
L. iners
L. jensenii
L. gasseri
B. vaginale
F. vaginae
D
Number of genomes sampled (GTDB)


A
VMGC-only
Shared
GTDB-only
Lactobacillus crispatus
4012
6732
2686
Lactobacillus iners
4558
2049
2
Lactobacillus jensenii
1295
2064
67
Lactobacillus gasseri
2045
3039
443
Bifidobacterium vaginale
4096
2545
68
Fannyhessea vaginae
4744
1512
0
B
Detected
Not detected
n = 130 samples
71%
29%
17%
83%
3%
97%
n = 68 samples
91%
9%
26%
74%
0%
100%
n = 46 samples
97%
3%
20%
80%
16%
84%
n = 22 samples
86%
14%
18%
82%
25%
75%
n = 77 samples
91%
9%
48%
52%
22%
78%
n = 40 samples
96%
4%
33%
67%
Shared
VMGC-only
GTDB-only
Number of 90% ANI gene clusters
C
VMGC-only
Shared
GTDB-only
Number of 90% ANI gene clusters
% samples with gene cluster detected (n = 130 samples)
% samples with gene cluster detected (n = 68 samples)
% samples with gene cluster detected (n = 46 samples)
% samples with gene cluster detected (n = 22 samples)
% samples with gene cluster detected (n = 77 samples)
% samples with gene cluster detected (n = 40 samples)


A
B
C
Number of predicted viral regions
0
10
20
30
40
**
**
****
****
*
*
VMGC-only
n = 812
Shared
n = 44
GTDB-only
n = 84
human
n = 890
non-human
n = 50
FRT
n = 2610
gastrointestinal tract
n = 6
urinary tract
n = 10
all human, non-FRT
n = 20


A
Hypothetical protein (1)
Hypothetical protein (2)
type II toxin-antitoxin system HicB family antitoxin
type II toxin-antitoxin system HicA family toxin
Na+/H+ antiporter NhaC
Hypothetical protein (3)
Hypothetical protein (4)
Hypothetical protein (5)
LacI family transcriptional regulator
XRE family transcriptional regulator
MarR family transcriptional regulator
DUF771 domain-containing protein
Hypothetical protein (6)
locFDR < 0.05
locFDR < 0.1
locFDR < 0.2
-log10(p)
Effect size (β)
Enriched in healthy controls
Enriched in cervical dysplasia
B
gene present
gene absent
Cervical dysplasia
Healthy controls
Sample
Na+/H+ antiporter NhaC
MarR family transcriptional regulator
DUF771 domain-containing protein
hypothetical protein (6)
hypothetical protein (3)
XRE family transcriptional regulator
hypothetical protein (2)
hypothetical protein (4)
hypothetical protein (5)
type II toxin-antitoxin system HicA family toxin
type II toxin-antitoxin system HicB family antitoxin
hypothetical protein (1)
LacI family transcriptional regulator
C
hypothetical protein (1) β= 1.823
type II toxin-antitoxin system HicA family toxin β= 1.411
hypothetical protein (2) β= 1.307
LacI family transcriptional regulator β= 1.264
hypothetical protein (3) β= 1.115
type II toxin-antitoxin system HicB family antitoxin β= 1.052
XRE family transcriptional regulator β= 1.012
hypothetical protein (6) β= 0.97
hypothetical protein (4) β= 0.954
hypothetical protein (5) β= 0.954
DUF771 domain-containing protein β= 0.954
Na+/H+ antiporter NhaC β= 0.91
MarR family transcriptional regulator β= 0.874
Number of samples
Healthy controls
Cervical dysplasia
gene present
gene absent


A
gene present
gene absent
1D
B4
C25
DC21.1
BC1
ATCC_33820
VSI21
FDAARGOS_743
VSI04
VSI24
2029
BC3
PMC201
AB70
VSI08
PRL2021
Lcr-MH175
VSI17
M247
Lc1226
Lc116
Lc1700
Na+/H+ antiporter NhaC
MarR family transcriptional regulator
DUF771 domain-containing protein
hypothetical protein (6)
hypothetical protein (3)
XRE family transcriptional regulator
hypothetical protein (2)
hypothetical protein (4)
hypothetical protein (5)
type II toxin-antitoxin system HicA family toxin
type II toxin-antitoxin system HicB family antitoxin
hypothetical protein (1)
LacI family transcriptional regulator
source
Isolate source
human vagina
human urine
human feces
human oral cavity
chicken gut
horse feces
B
hypothetical protein (3)
DUF771 domain-containing protein
XRE family transcriptional regulator
MarR family transcriptional regulator
hypothetical protein (6)
single-stranded DNA-binding protein
helix-turn-helix domain-containing protein
site-specific integrase
AB70
1D
Lc116
PRL2021
Lc1700
transcriptional regulator
ATCC_33820
2029
FDAARGOS_743
VSI17
PMC201
VSI21
Lc1226
Arc family DNA-binding protein
toxin-antitoxin system HicB family antitoxin
1 kb
C
hypothetical protein (4)
hypothetical protein (2)
hypothetical protein (5)
pyocin knob domain-containing protein
XkdX family protein
phage holin
lysin
Lc1700
Lc116
Lcr-MH175
VSI21
AB70
1D
2029
enterobactin-iron receptor
1 kb
D
type II toxin-antitoxin system HicA family toxin
type II toxin-antitoxin system HicB family antitoxin
endonuclease
antitoxin HicB
Lc116
ABC transporter
DC21.1
VSI08
AB70
BC1
PMC201
1D
VSI17
M247
1 kb
E
LacI family transcriptional regulator
alpha-xylosidase
extracellular solute-binding protein
ATCC_33820
BC1
1 kb
F
TetR/AcrR family transcriptional regulator
hypoxanthine phosphoribosyltransferase
hypothetical protein (1)
polysaccharide biosynthesis C-terminal domain-containing protein
oligosaccharide flippase family protein
N-acetyltransferase
DUF871 family protein
PMC201
AB70
BC1
1D
DC21.1
linear amide C-N hydrolase
ATP-binding protein
1 kb
G
serine hydrolase domain-containing protein
Na+/H+ antiporter NhaC
Bax inhibitor-1 YccA family protein
PRL2021
BC1
1D
VSI17
Lc116
M247
VSI08
Lc1226
Lc1700
1 kb
Cervical dysplasia-associated gene
Gene encoding hypothetical protein
Transposon
Other gene
Predicted operon