A Syntenic Pangenome for *Gardnerella* Reveals Taxonomic Boundaries and Stratification of Metabolic and Virulence Potential across Species.

Heather Bouzek[1,2], Sujatha Srinivasan[2], Dakota S. Jones[2], Elsa F. McMahon[2], Susan M. Strenk[2], Tina L. Fiedler[2], David N. Fredricks[2#], Christopher D. Johnston[,1,2#]

[1]*Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, TX, USA*
[2] *Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, Washington, USA*

[#]**Address correspondence to**:

David N. Fredricks, MD, dfredric@fredhutch.org & Christopher D. Johnston, PhD, cdjohnston@mdanderson.org

Abstract

Bacterial vaginosis (BV) is a prevalent condition associated with an imbalance in the vaginal microbiota, often involving species of *Gardnerella*. The taxonomic complexity and inconsistent nomenclature of *Gardnerella* have impeded progress in understanding the role of specific species in health and disease. In this study, we conducted a comprehensive genomic and pangenomic analysis to resolve taxonomic ambiguities and elucidate metabolic and virulence potential across *Gardnerella* species. We obtained complete, closed genomes for 42 *Gardnerella* isolates from women with BV and curated publicly available genome sequences (n = 291). Average nucleotide identity (ANI) analysis, digital DNA-DNA hybridization (dDDH), and the *cpn60* gene sequences identified nine species and eleven subspecies within *Gardnerella*, for which we refined species and subspecies boundaries and proposed updated nomenclature. Pangenome analysis revealed species-specific gene clusters linked to metabolic pathways, virulence factors, and niche adaptations, distinguishing species specialized for mucin degradation in the vaginal environment from those potentially adapted to urinary tract colonization. Notably, we identified lineage-specific evolutionary divergence in gene clusters associated with biofilm formation, carbohydrate metabolism, and antimicrobial resistance. We further discovered the first cryptic plasmids naturally present within the *Gardnerella* genus. Our findings provide a unified framework for *Gardnerella* taxonomy and nomenclature, and enhance our understanding of species-specific functional capabilities, with implications for *Gardnerella* research, diagnostics, and targeted therapeutics in BV.

Introduction

Bacterial vaginosis (BV) is a common and global condition characterized by an imbalance in the vaginal microbiota that affects women around the world. This imbalance typically involves a reduction in beneficial *Lactobacillus* species and an overgrowth of obligate and facultative anaerobes[1], leading to symptoms such as unusual discharge, odor, and discomfort[2]. BV is associated with significant health risks, including increased susceptibility to sexually transmitted infections[3] and adverse pregnancy outcomes such as preterm birth and low birth weight[4,5].

Gardnerella vaginalis was the first bacterium isolated from BV-affected women[6], and *G. vaginalis* was long thought to be the sole causative agent of this condition[7]. This assumption is not baseless, as *Gardnerella* species exhibit a high virulence potential, adhere to vaginal epithelium[8], produce cytotoxins[9], and form biofilms[10], and are present in 95–100% of BV cases[11-13]. Moreover, *Gardnerella* species thrive in the vaginal environment, even at the low pH typical of the *Lactobacillus*-dominant genital tract microbiota[14]. *Gardnerella* species are also present in asymptomatic women, suggesting the potential for both commensal and pathogenic lineages or states, but vaginal concentrations of *Gardnerella* are greater in those with BV[11]. Our previous research revealed that while *G. vaginalis* is correlated with a greater risk ratio for BV than other *Gardnerella* species are, no single *Gardnerella* species is specific to BV. In fact, the high diversity of *Gardnerella* species is a risk factor the development of BV according to a longitudinal study[11]. With the diversity of *Gardnerella* species becoming more relevant in the context of BV, two significant challenges hinder rigorous genetic interrogations of species and subspecies attributes within this genus.

First, despite laudable efforts from multiple research groups, the *Gardnerella* field is stymied by an inconsistent and overly complex nomenclature[15] (**Figure 1**). Piot *et al*.[16], reported that the first biotypes of *Gardnerella* were based on different sugar hydrolysis phenotypes, which, when proven to be unable to capture the diversity of *Gardnerella*, were then expanded to 17 biotypes by the addition of three fermentation targets[17] (**Figure 1**). Early genomic methods used to identify *Gardnerella* species included

RFLP[18] and amplified ribosomal DNA restriction analysis (ARDRA)[19]. The use of the *cpn60* gene (encoding a 60 kDa chaperonin protein) currently provides the most reliable difference between established *Gardnerella* types among single-gene sequences over more traditional 16S rRNA typing, which is not useful for *Gardnerella* speciation[20,21]. Whole-genome comparisons (**Figure 1**) revealed another naming scheme seeking to encompass all strains, where "A", "B", "C", and "D"[22] were proposed designations for a newer condensed species assignment[15,21]. Another method was proposed[23], which uses "1", "2", "3", and "4"[14] for the same clades, leading to a disjointed hybrid designation (C/1, B/2, A/4, and D/3). The species were then renamed as reassorted ecotypes[24], using yet another nomenclature, one that incorporated the genomic context of the distinct types. After whole-genome sequencing became widely available, three more competing standards resulted—each based on average nucleotide identity (ANI), with thirteen genomospecies (Gsp)[25], nine genomospecies (GS)[26], and nine different genotypes (GGtypes)[27] being proposed (**Figure 1**). Formal naming of a subset of these species has added to, rather than alleviated, the confusion, as the groupings do not always align with currently accepted species/subspecies delineations. There are currently six named species within *Gardnerella*[25,28], and each is aligned with a genomospecies. However, according to the ANI standards by which they were designed[29], each of the genomospecies is a subspecies, rather than a distinct species.

A second challenge lies in the lack of comprehensive and accurate genomic and metabolic data for *Gardnerella* species, which hinders our ability to clearly define their metabolic potential and link these characteristics to specific species or subgroups. This challenge is exacerbated by issues with publicly available genomic databases, including strain duplication (same strain listed under multiple different names) and contamination (low-quality sequences or misassigned data), leading to the inclusion of non-*Gardnerella* sequences in genomic and pangenomic studies. If not addressed, these issues could impede progress in understanding *Gardnerella* species and their roles in health and disease states.

In this study, we used multiple genomic and pangenomic approaches to reevaluate the relationships among *Gardnerella* species and provide a single naming scheme for all currently known *Gardnerella* species. To address the challenges associated with current nomenclature strategies, we obtained complete

and closed genomes, with accessory methylomes, for 42 *Gardnerella* isolates from women with BV. By acknowledging previous efforts, we curated publicly available genome sequence data assigned to the *Gardnerella* genus (n=291 as of November 1st, 2024), determining that approximately half of these sequences (n =137) passed acceptable standards for inclusion in genome-based or taxonomic studies. Using ANI thresholds with polished genomes, validated through digital DNA/DNA hybridization and *cpn60* gene typing, we identified nine distinguishable species, encompassing eleven subspecies, within *Gardnerella*. Furthermore, pangenome analysis revealed discrete predicted metabolic features associated with each of these species and subspecies. We subsequently propose an updated nomenclature based on these pangenomic data, which includes refined species boundaries and suggested names for species and subspecies where appropriate. Our work aims to enhance the understanding of genetic characteristics across *Gardnerella* species and subspecies, addressing gaps in knowledge regarding functional capabilities.

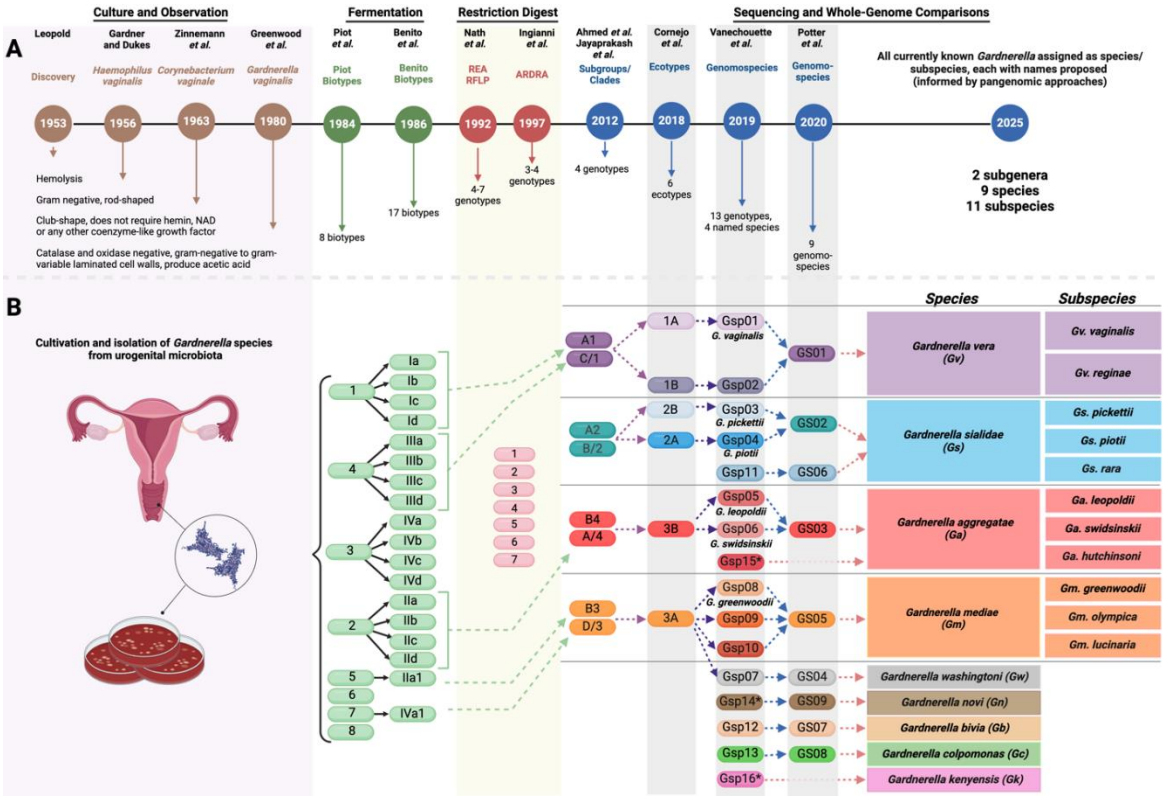

Figure 1: Historical context of *Gardnerella* nomenclature. (**A**) Identification method by type, publication first author, names given to *Gardnerella* or naming framework and dates for different naming schemes and the methods for naming. (**B**) *Gardnerella* subtyping by identification method type. Each cell in a column represents a subset of *Gardnerella* named by the publication in (A). The strains from different publications are connected by arrows. The final two columns contain the new names proposed within this publication.

Results

Clarification of the taxonomy of *Gardnerella*

We incorporated all publicly available *Gardnerella* genome sequences (333 within the NCBI database as of November 2024) with additional sequencing of 42 human *Gardnerella* isolates (complete and closed genomes with associated methylomes) from the vagina, to establish a curated set of *Gardnerella* genomes for pangenomic investigation. ANI and dDDH heatmaps (**Figure 2 and Supplementary Figures 1-4, 16**) from this set of curated genomes show the bifurcation of *Gardnerella* subgenera, as well as distinct speciation and subspeciation within the *Gardnerella* genus, in agreement with and expansion of multiple previous efforts (**Figure 1**)[22,24-26,30,31]. Two new groups were also found, which we tentatively designated as Genomospecies 15 (Gsp15) and Genomospecies 16 (Gsp16). Gsp15 forms a distinct subgroup that is related to Gsp05 and Gsp06, whereas Gsp16 forms its own group, which often clusters near Gsp08 and Gsp15 (**Figure 2**).

Genera: Some classification methods for *Gardnerella* have placed this group within the *Bifidobacterium* genus[27,32]. However, *Gardnerella* is distinct from *Bifidobacterium*, and these two genera should not be considered synonymous. *Bifidobacterium* is a polyphyletic group, consisting of at least six distinct clades[33,34], and its members exhibit a wide range of genetic diversity. Typically, the genomic G+C content does not vary by more than 10% within a genus[35] or by more than 1% within species of the same genus[36]. In *Bifidobacterium*, the genomic G+C content is consistently high (range: 50.4%-67.5%, median of 60.4%) In contrast, *Gardnerella* has lower G+C percentages (range: 37.8%-45.7%, median of 42.0%, $p<0.001$, Wilcoxon post-hoc test) (**Supplementary Figure 5**). In *Bifidobacterium*, genomic prevoteall is consistently high (range: 1.29-3.35 Mb, median of 2.29 Mb). In contrast, *Gardnerella* presented a significantly shorter genomic length that did not overlap (range: 1.24-1.94 Mb, median of 1.60 Mb, $p<0.001$, Wilcoxon post-hoc) (**Supplementary Figure 6**).

Moreover, an ANI cutoff of 80% is typically used to define intra-genus relatedness.[29] A comparison of ANI values between our *Gardnerella* species collection and a collection of 98 strains from the *Bifidobacterium* and closely related *Cardovia* genera (e.g., *Alloscardovia, Aeroscardovia*, etc.) revealed that ANI values between *Bifidobacterium* and *Gardnerella* do not exceed 78% (**Supplementary Figure 7**). In addition, the digital DNA-DNA hybridization (dDDH) of any *Gardnerella* to any *Bifidobacterium* is less than 30% which is the cutoff to delineate genera (formula1_{max} = 13.3%, formula2_{max} = 29.3%, formula3_{max} = 13.6%). The average amino acid identity (AAI) of 24 out of 38 universal orthologous genes revealed differential and nonoverlapping clustering between *Gardnerella* and *Bifidobacterium* (*CpsG*, *DnaN*, *Exo*, *Ffh*, *FtsY*, *FusA, GroL*, *HisS*, *InfB*, *KsgA*, *LeuS*, *MetG*, *NrdA*, *NusG*, *PheS*, *PheP*, *PheT*, *ProS*, *PyrG*, *RecA*, *RplA*, and *Rpl C-E*[37], found in **Supplementary Figures 8-9**).

Subgenera: There are currently two subgenera within *Gardnerella,* which were previously referred to as Set A and Set B[22]. Set A contains the species within the historical C/1 (Gsp01/02) and B/2 (Gsp 03/04). Set B includes the remaining described genomospecies: Gsp05 (*G. leopoldii*), Gsp06 (*G. swidsinskii*), Gsp07, Gsp08 (*G. greenwoodii*), Gsp09, Gsp10, Gsp13, Gsp14, and the two novel genomospecies, Gsp15 and Gsp16, identified in this work.

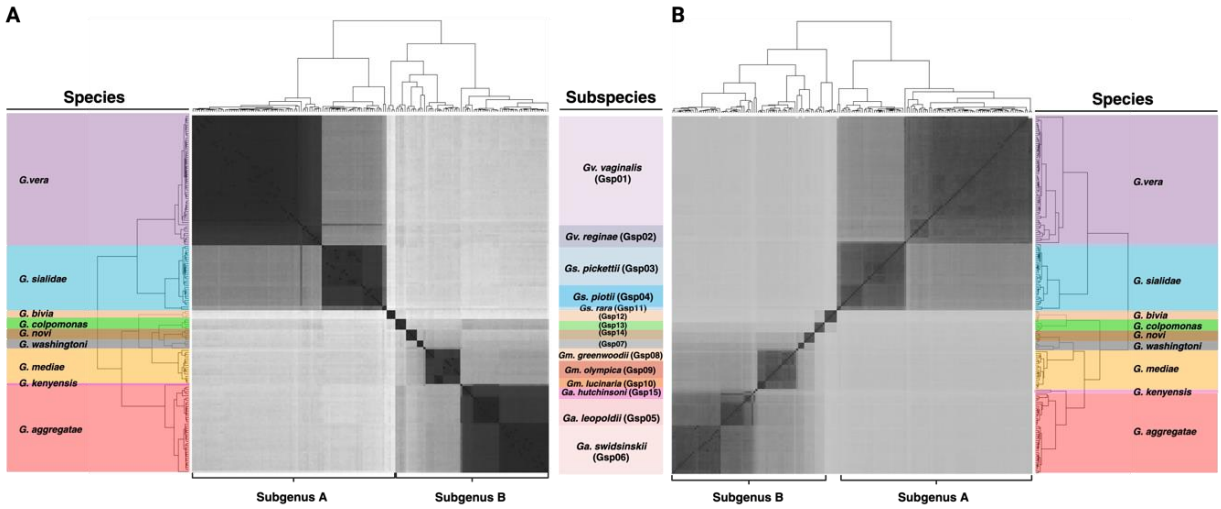

Figure 2: Whole-genome ANI and dDDH similarity heatmaps. (**A**) Average nucleotide identity (ANI) and (**B**) digital DNA–DNA hybridization (dDDH; calculated using Formula 3 length[38]; other formulas in Supplementary Figures) are displayed as heatmaps for curated genomes (including new complete and closed genomes). Horizontal color bands indicate *Gardnerella* species (left and right margins) and subspecies (middle column) assignments. The dendrograms at the top and sides of each heatmap show clustering for each method, with each row and column representing a *Gardnerella* strain. The intensity of color in the heatmaps at the intersection of rows and columns reflects the homology between sequences: darker shades indicate greater similarity, with the identity diagonal (top-left to bottom-right for (**A**) ANI; bottom-left to top-right for (**B**) dDDH) representing comparisons of sequences to themselves. Zero homology is white, and similarity increases with gray intensity, reaching black for full identity. Taxonomic relationships are inferred from squares along the identity diagonal, with genera relationships indicated by the lightest shades (cutoff: 83% ANI; all comparisons above this level). Subgenera are noted in brackets at the bottom of each plot. Species are highlighted by colored bands at the left and right extremes, corresponding to species-level homology (>95% ANI and >70% dDDH, both cutoffs for species-level delineation). Subspecies bands, when present, are marked in variations of the species color and annotated in the center between the heatmaps (>98% ANI).

Species and Subspecies: Notably, the species designations identified here are congruent with the AAI, ANI, and tetranucleotide frequency results reported by Potter *et al.*[26] and were confirmed by our own ANI (**Figure 2 and Supplementary Figure 1**), dDDH (**Figure 2 and Supplementary Figures 2-4)**, and kSNP (**Supplementary Figures 16-18**) results. These groupings are further reinforced by pangenomic approaches (described below) and horizontal gene transfer events that are unique within each grouping.

Efforts toward a unified nomenclature for *Gardnerella*

Following the resolution of the *Gardnerella* species and subspecies boundaries, we applied this updated taxonomy to refine the *Gardnerella* nomenclature. Given the extensive history of conflicting classification schemes (**Figure 1**), we sought to standardize species and subspecies names for consistency in future

research (**Figure 2**, and **Table 1**). To establish our proposed nomenclature (see **Protologues**), we reassessed previously named species and subspecies, integrating them within newly delineated lineages to create a more consistent naming framework. For instance, the historically recognized '*G. vaginalis'* is now proposed to be designated '*G. vera* subspecies *vaginalis'*, reflecting its phylogenetic placement while preserving historical continuity (**Table 1**). Similarly, we propose distinct species names for lineages with well-supported genomic and phenotypic divergence, including *G. sialidae*, *G. aggregatae*, and *G. mediae*, among others. These refined classifications are further detailed in the **Protologues**, where we provide the rationale for each nomenclatural revision, including species-specific functional attributes and ecological isolation sources. In addition, to prevent ambiguity when referring to *Gardnerella* species in broader studies, each new species name was assigned to enable a unique two-letter shorthand notation (*Gv*, *Gs*, *Ga*, *Gm*, *Gb*, *Gc*, *Gw*, *Gn* and *Gk*). Individual subspecies are assigned unique three-letter abbreviations (**Table 1**) (except for *Gs. pickettii* and *Gs. piotii,* which requires four letters, *Gspk* and *Gspt,* to distinguish between these already named subspecies).

Table 1: Current and proposed naming schemes for *Gardnerella* species and subspecies

Current Species Name	Genomo -species	Ahmed Set Group	Ecotype	GTDB taxa (genus and species)	Proposed Species	Proposed Subspecies
G. vaginalis	01	C/1	1A, 1B	*Bifidobacterium vaginale*	*G. vera* (***Gv***)	*Gv. vaginalis* (***Gvv***)
N/A	02	C/1	1B	*Bifidobacterium vaginale*		*Gv. reginae* (***Gvr***)
G. pickettii	03	B/2	2B	*Bifidobacterium piotii*	*G. sialidae* (***Gs***)	*Gs. pickettii* (***Gspk***)
G. piotii	04	B/2	2A	*Bifidobacterium piotii*		*Gs. piotii* (***Gspt***)
N/A	11	B/2	N/A	*Bifidobacterium vaginale_B*		*Gs. rara* (***Gsr***)
G. leopoldii	05	A/4	3B	*Bifidobacterium leopoldii*	*G. aggregatae* (***Ga***)	*Ga. leopoldii* (***Gal***)
G. swidsinskii	06	A/4	3B	*Bifidobacterium swidsinskii*		*Ga. swidsinskii* (***Gas***)
N/A	15	N/A	N/A	*Bifidobacterium vaginale_C*		*Ga. hutchinsoni* (***Gah***)
G. greenwoodii	08	D/3	3A	*Bifidobacterium vaginale_C/E*	*G. mediae* (***Gm***)	*Gm. greenwoodii* (***Gmg***)
N/A	09	D/3	3A	*Bifidobacterium vaginale_E*		*Gm. lucinaria* (***Gml***)
N/A	10	D/3	3A	*Bifidobacterium vaginale_A*		*Gm. olympica* (***Gmo***)
N/A	12	N/A	N/A	*Bifidobacterium vaginale_F*	*G. bivia* (***Gb***)	*n/a*
N/A	13	N/A	N/A	*Bifidobacterium vaginale_D*	*G. colpomonas* (***Gc***)	*n/a*
N/A	07	N/A	3A	*Bifidobacterium vaginale_H*	*G. washingtoni* (***Gw***)	*n/a*
N/A	14	N/A	N/A	*Bifidobacterium* sp003585735	*G. novi* (***Gn***)	*n/a*
N/A	16	N/A	N/A	*Bifidobacterium vaginale*	*G. kenyensis* (***Gk***)	*n/a*

Pangenomic stratification reveals species-specific gene content

Pangenomes provide a comprehensive view of the genetic repertoire within a group of bacteria, encompassing all genes found across multiple strains or species. These analyses are crucial for understanding genetic diversity, adaptation, and evolutionary relationships. Pangenome construction begins by grouping similar genes into gene clusters, which represent functionally or evolutionarily related groups of genes. These gene clusters are represented as individual circular nodes in **Figures 3-10**. The gene order within individual genomes is conveyed by the connecting lines between these nodes, which provides the syntenic context, enabling us to trace conserved arrangements of genes across multiple strains within the pangenome. This approach highlights regions of conservation, which may correspond to functionally important elements, as well as regions of variability, which drive phenotypic diversity.

One outcome of this analysis is the identification of modules (**Figures 4** and **5**), which consist of gene clusters and their synteny. These modules often represent functional units or clusters of coregulated genes involved in specific biological processes, such as operons. Traditionally, pangenomes have been stratified into two distinct categories: core gene clusters, which are shared by most genomes, and accessory gene clusters, which are present in only some strains[39]. While useful, this binary model often fails to capture intermediate levels of gene variability. To address this in the *Gardnerella* pangenome, we applied PPanGGOLiN[40], which applies a more nuanced approach to split the accessory group into a "cloud" partition (representing low-frequency gene clusters rarely found in the pangenome) and several "shell" partitions (representing gene clusters with intermediate frequency). Shell partitions reflect subgroup-specific variations (**Figures 3** and **4**) and offer a more detailed view of gene dynamics and their role in population diversity. By integrating these concepts into the *Gardnerella* pangenome, we aimed to leverage the combination of modules and shell strata to better understand the functional and evolutionary dynamics driving strain- and species-level diversity. The PPanGGOLiN[40]-derived Gephi[41] plot (**Figures 3–5**) illustrates the syntenic arrangement and order of gene clusters across the *Gardnerella* pangenome, highlighting the stratification among shell partitions. Persistent/core gene clusters (**Figure 3A** and **3B**, nodes in orange) dominate the map, representing highly conserved genomic modules shared by >95% of strains (each cluster is determined by the multivariate Bernoulli mixture model used by PPanGGOLiN[40] which generally follows the 95% threshold). Periodically, conserved regions are interrupted by shell-associated clusters (**Figure 3A** and **3C**, nodes in distinct colors, branching from orange nodes), which mark points of evolutionary divergence. These shell partitions warrant further investigation as they represent alternate evolutionary pathways, often associated with horizontal gene transfer (HGT) events. Indeed, stratification of the shell partitions uncovered lineage-specific gene clusters (**Supplementary Table 1)**, suggesting that gene flow and evolutionary trajectories drive species-specific adaptations in *Gardnerella*.

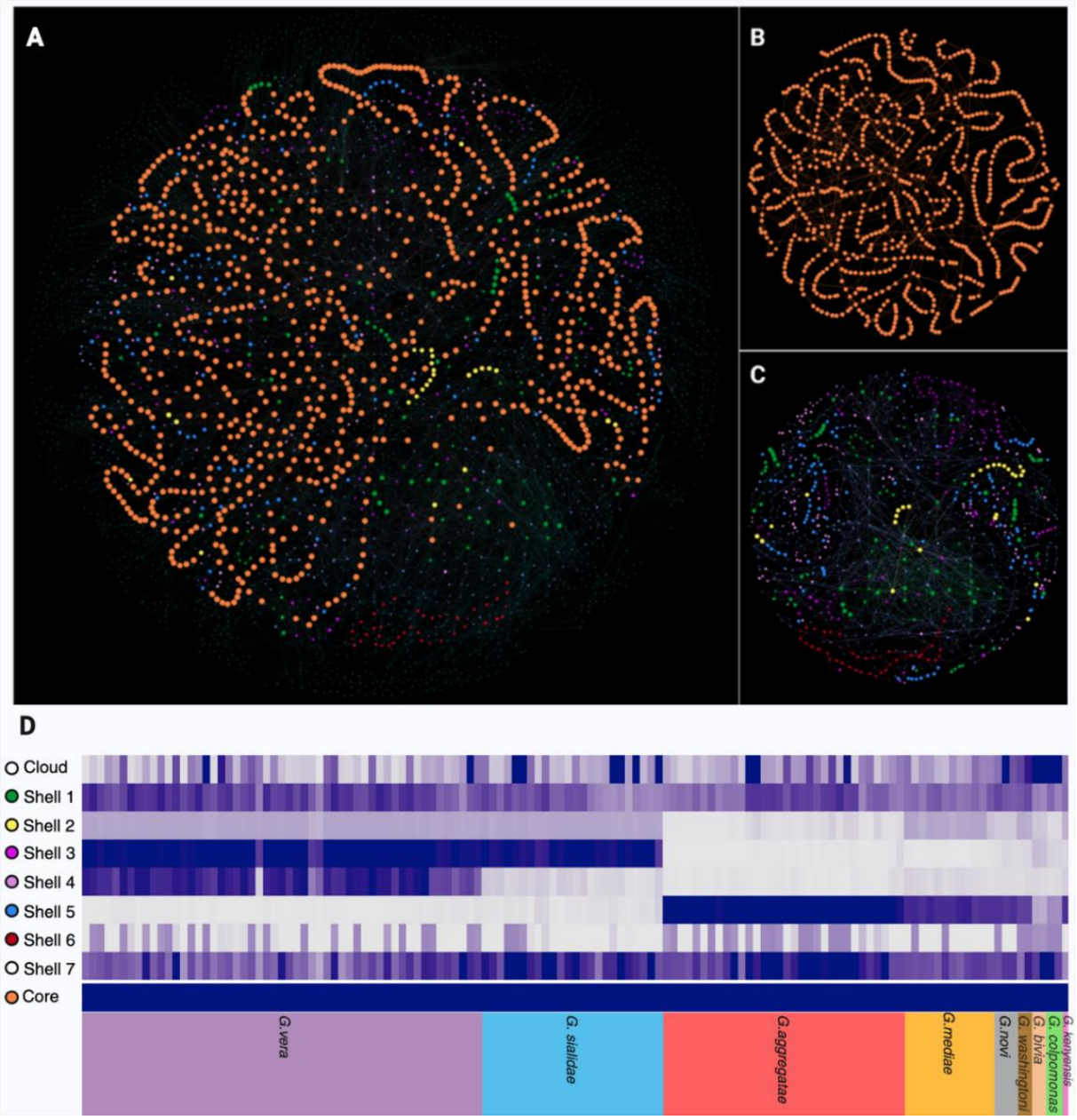

Figure 3: Syntenic pangenome of *Gardnerella*. Node-and-edge maps illustrating the gene clusters within the *Gardnerella* pangenome. The core genes are represented by orange clusters, whereas the light blue cluster denotes the cloud genes. Each shell partition is uniquely colored: Shell 1 (dark green), Shell 2 (yellow), Shell 3 (dark blue), Shell 4 (violet), Shell 5 (pink), and Shell 6 (red). Shell 7 is not shown due to its complexity and small node size. (**A**) The complete pangenome is shown, with a sparsely populated region of core genes at the bottom, identified as a major defense island. The red loop highlights the location of a prophage. (**B**) A detailed view of the core genes reveals long stretches of essential genes in *Gardnerella*. (**C**) The shell regions are displayed without core or shell partitioning. Noteworthy chains in the shells expose syntenic blocks that differentiate the subgenera and species of *Gardnerella*. For example, the operon responsible for removing terminal sialic acid from mucin, along with related transport and degradation enzymes, is shown as a yellow curl in the upper right. The red loop at the bottom highlights Shell 6, which is composed entirely of prophage (vb_Gva_AB1), adjacent to a complex network of genomic rearrangements and genetic elements from multiple infections, predominantly involving Shell 1 (green) and Shell 2 (yellow). (**D**) A heatmap displays the presence of genes in each partition, with color overlay corresponding to the node-and-edge maps. Darker regions indicate absence, whereas lighter or white regions indicate gene presence. Species groups associated with each shell are colored at the bottom, highlighting the relative presence and absence of genes within shells across different *Gardnerella* species.

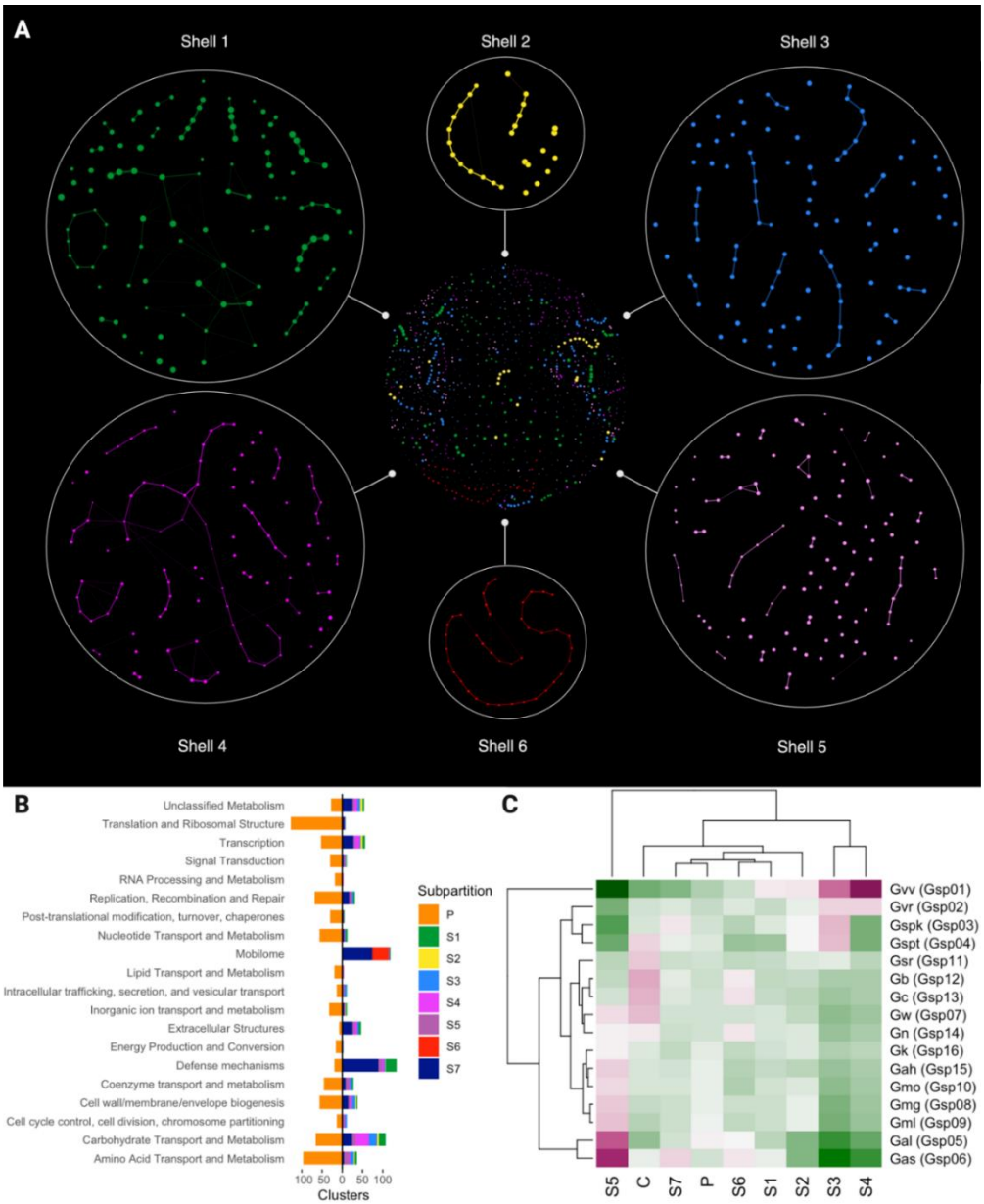

Figure 4: Deconvolution of shell partitions within the *Gardnerella* pangenome. (**A**) Node-and-edge plots of gene clusters within the shell. The entire shell partition is center (Figure 3C). The surrounding plots show single gene clusters and syntenic modules (nodes with edges) within each shell, where strings of nodes are indicative of syntenic gene clusters. (**B**) Bar plot of the Clusters of Orthologous Genes (COG) Categories (y-axis, unknown not shown) versus the numbers of gene clusters annotated with each category (x-axis). Persistent/core clusters (orange, abbreviated "P") are plotted to the left and the shell clusters (abbreviated with "S" and a number and shown in various colors. Cloud partition is not shown.) are plotted to the right. (**C**) The residuals from a chi-square test (df = 120, p-value < 0.001, Supplementary Data) plotted as a heatmap with column and row clustering to show which associations between *Gardnerella* species and subspecies (rows) and the shell partition (columns – P represents the persistent partition, "S" and a number represents the different shell partitions, and C represents the cloud partition) are significantly greater (red) or lower (green) than would be expected by chance. Significantly greater associations are observed between *G. vera* (*Gvv*/*Gvr*) and *G. sialidae* (*Gspk*/*Gspt*) and shell 3 (S3). *G. vera* (*Gvv*/*Gvr*) alone is highly associated with shell 4 (S4), and *G. aggregatae* (*Gal*/*Gas*) is strongly associated with shell 5 (S5). Significantly lower associations are observed for *G. vera vaginalis* (*Gvv*) and shell 5 (S5), as well as for *G. aggregatae* (*Gal*/*Gas*) and shells 3 and 4 (S3/S4).

Persistent Partition: Cellular machinery, DNA replication, transcription and repair

Gene clusters within the persistent partition, also known as the "core" genome, are shared across all *Gardnerella* strains and species, and are present in ~95% of the genomes. Approximately 20% of the gene clusters in the *Gardnerella* pangenome fall within the persistent (core) partition. Like other taxa[40,42], this core partition is enriched for essential cellular functions, including translation and ribosomal structure, transcription, and RNA processing (**Figure 4B**). Additional functions that are for the most part retained across *Gardnerella* species include signal transduction, energy production; and nucleotide, inorganic ion, and lipid metabolism (**Figure 4B**). As the foundational genetic framework essential for cellular viability, persistent partitioning is shaped by evolutionary constraints that preserve genes critical for survival, whereas more flexible, niche-specific functions are partitioned into the shell genome. For example, many of the gene clusters related to amino acid metabolism are persistent and shared across *Gardnerella*, but those related to carbohydrate metabolism are diverse and within shell partitions (**Figures 4B, 5-9**).

Shell 1: Mobile Genetic Elements, Virulence, and Defense

Shell 1 contains gene clusters found frequently across all species (**Figure 4**, **Supplementary Table 2**) and is enriched with gene clusters consistent with metabolism, genomic defense, and adhesion. These gene clusters do not appear to be lineage specific (**Figure 3D**). The modules of Shell 1 appear to be related to key metabolic processes, including glutamate import (M005)[43], a GH13 α-maltosidase (M013)[44,45], thiamine production and salvage (M030), ribose catabolism via the pentose phosphate pathway (M046)[44,46], and two ABC transporters with unknown target molecules (M018, M026)[47] (**Figure 5**). Additional defense-related gene clusters include a hyperpromiscuous antitoxin[48], a Type I CRISPR–Cas system (M020)[49], a Type I restriction modification (RM) system (M013)[30], and numerous toxin–antitoxin pairs[50]. Several gene clusters consistent with extracellular proteins involved in attachment and virulence including Class C sortase and SpaH/EbpB family LPXTG-anchored major pilins[47,50-52], GA module proteins[50], and internalin B[47,52]. These genes are noteworthy not only for their potential roles in virulence but also for their substantial size: *Gardnerella* pilin genes are 1.6k amino acids (aa), GA module genes range from 5.8k to 6.3k aa, and

internalin B genes span 2.0-2.4k aa. The Shell 1 gene clusters correspond to the *Gardnerella* genes identified by Garcia *et al*., which undergo phase variation[53], a key survival strategy that enables bacteria to dynamically alter surface properties and virulence factors. Phase variation, driven by mobile genetic elements, facilitates immune evasion, pathogenicity, and adaptation to changing environments, allowing bacteria to continuously fine-tune their defenses and offensive capabilities.

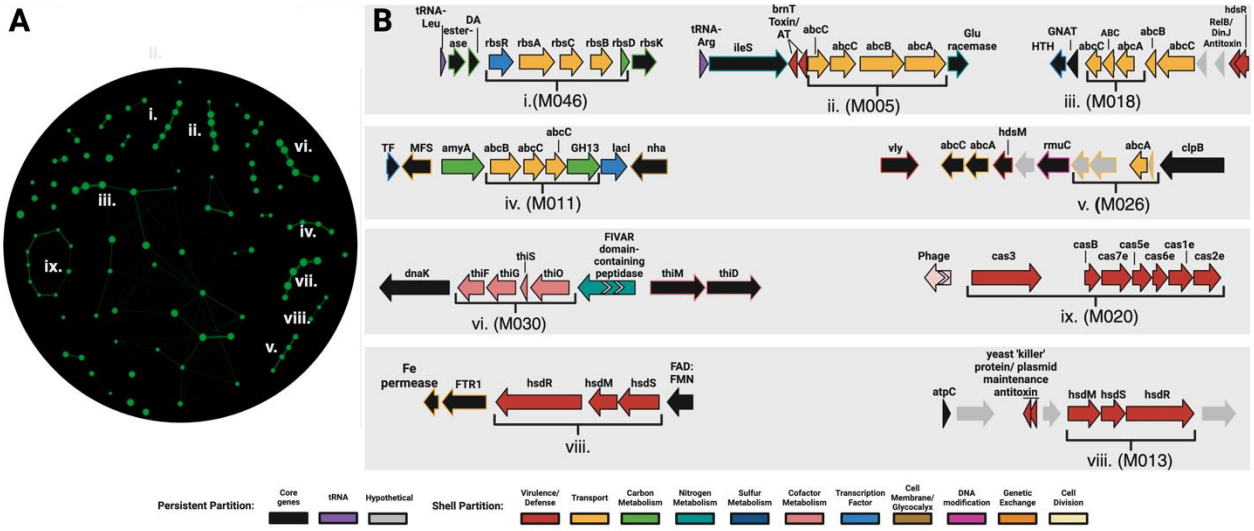

Figure 5: Shell 1 associated modules and gene clusters. (**A**) Node-and-edge plot for Shell 1 which is not associated with any subgroup within *Gardnerella*. Unique modules are labeled (**i** - **ix**). (**B)** Gene arrow plots of the modules labeled in (**A)** with the flanking genes of the module The core genes are colored black for general genes (outlined in the shell-specific color if function is known), purple for tRNA genes, and light gray for hypothetical genes. The specific shell genes are colored by function for Shell 1: red for virulence/defense, yellow for transport, green for carbon metabolism, aqua for nitrogen metabolism, pink for cofactor metabolism, and blue for transcription factors. Gene arrows with chevrons indicate that the size of the protein (or bacteriophage) has been cropped to fit the image. The modules of Shell 1 feature five ATP-binding cassette transporters (ABC); two are glucose-repressed regulons for sugar import (**i**) Ribose *rbsABCDKR* (M046) and (**iv**) α-Maltosidase (GH13)/α-Amylase (*amyA*) (M013) and two modules are proximal to toxin/antitoxin pairs (**ii)** Glutamate import **(**M005) and (**iii**) an import module with an unknown substrate (M018), (**v)** the final ABC-containing module contains two different ABC, one within the shell partition, and one in the core, both upstream of the core vaginolysin (*vly*) cluster **(**M026). (**vi**) Thiazole metabolism (*thiDFGMOS*) (M030) is interrupted by a long extracellular peptidase with a FIVAR domain. The final three are defense related modules—two type I RM systems (**vii)** (M013) and (**viii),** and (**ix**) a circular CRISPR-Cas system (M020).

<u>Shell 2</u>: Sialidase pathway and *Gardnerella* species-specific gene clusters

Shell 2, the smallest shell partition with only two modules (**Figure 4**, **Supplementary Table 2**), contains species-specific gene clusters in *Gardnerella*, which are associated primarily with *G. vera* and *G. sialidae* (**Figure 3D**). These modules are also present in *G. mediae*, *G. novi*, and *G. kenyensis* (**Figure 3D**)

but are largely absent in *G. aggregatae*, *G. colpomonas*, and *G. bivia*. Each module represents a distinct regulon (**Figure 6B**). The first module (M019) is putatively repressed by a response regulator and includes an ABC sugar transport gene cluster with a two-component system, featuring a putative chemosensor (HAMP histidine kinase) and a winged helix transcriptional regulator. The second module (M009) harbors gene clusters related to a key virulence factor in *Gardnerella*, namely a transcriptionally repressed sialidase catabolism operon[54,55]. In addition to these modules, Shell 2 also contains several independent gene clusters that are not linked to a specific module, such as a *lemA* and a glycosyltransferase 2 (GT2) gene cluster, both of which play putative roles in host adhesion and biofilm formation[56,57].

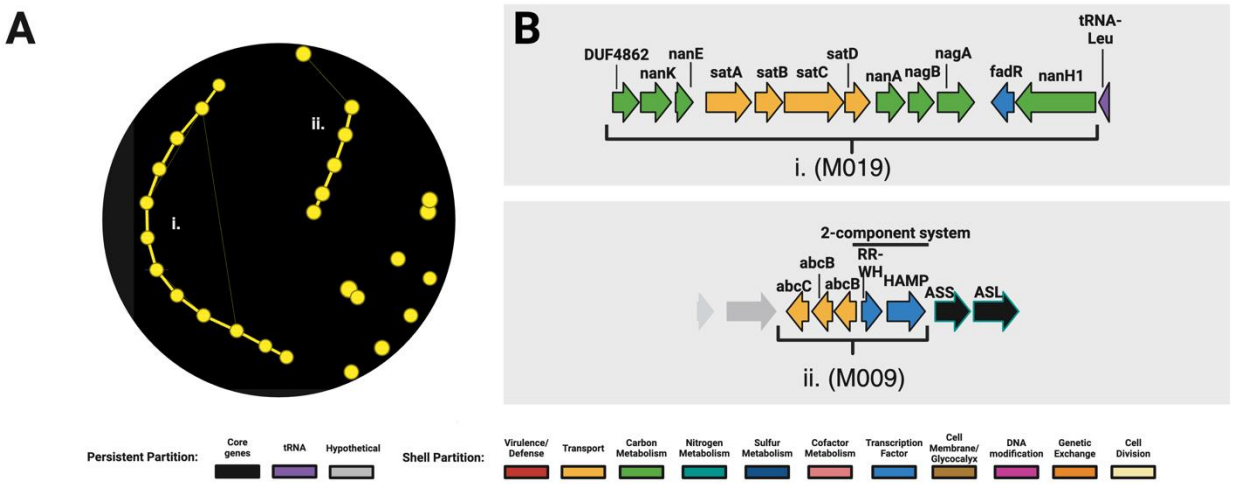

Figure 6: Shell 2 associated modules and gene clusters. (**A**) Node-and-edge plot for Shell 2 which is associated with *G. vera*, *G. sialidae*, *G. mediae*, *G. novi*, and *G. kenyensis*. Unique modules are labeled (**i.** and **ii**.). (**B)** Gene arrow plots of the modules labeled in (**A)** with the flanking genes of the module. The core genes are colored black for general genes (outlined in the shell-specific color if function is known), purple for tRNA genes, and light gray for hypothetical genes. The specific shell genes are colored by function: yellow for transport genes, green for carbon metabolism genes, and blue for transcription factors. The two modules of this shell are both regulons featuring ABC import mechanisms, neither of which are repressed in the presence of glucose. The sialidase import regulon (**i**) features *nanH* (some strains contain a second repressed *nanH*, as shown in brackets), *nanAKE* and *nagAB* for the degradation of sialic acid, and the sialic acid ABC transport system *satABCD* (M019). There is a single gene cluster with unknown function (DUF4862) which is found in the sialidase operons of *Bifidobacteria*. The second (**ii**) is a chemosensor (HAMP histidine kinase) linked to a transcriptional repressor (RR-WH for winged-helix response regulator) in a two-component system with another ABC transport system (M009). The two flanking core genes are arginosuccinate synthase (ASS) and arginosuccinate lyase (ASL).

216 Shell 3: Modules associated predominantly with *G. vera* and *G. sialidae*

217 Shell 3 contains gene clusters associated with *G. vera* and *G. sialidae* (**Figure 4**, **Supplementary**
218 **Table 2**), which are largely absent in all other *Gardnerella* species. However, two distinct populations are
219 present within Shell 2. The first population consists of modules with putative functions that are unique to
220 *G. vera* and *G. sialidae*. In contrast, the second population consists of gene clusters that, while initially
221 appearing uniquely present in these two species, instead have sets of homolog gene clusters in other
222 *Gardnerella* species with the same putative function but that have deviated in their underlying DNA
223 sequences. As a result, PPanGGOLiN no longer groups these sequences together across species when
224 clustering thresholds are used. These latter gene clusters indicate vertical evolutionary divergence between
225 *Gardnerella* species, where across species they share the same genomic loci (flanked by the same core gene
226 clusters) and putative functional annotations within identical metabolic pathways but differ sufficiently at
227 the DNA level to cluster separately.

228 Among the Shell 3 modules (**Figure 4**) that are unique to *G. vera* and *G. sialidae* are those involved
229 in altered cell wall biosynthesis, cell division, and lipopolysaccharide (LPS) production. A unique module
230 putatively responsible for synthesizing an O-polysaccharide with a rhamnolipid backbone (M000)
231 potentially assists in promoting adhesion and immune evasion, similar to the mechanisms studied in
232 Streptococci[58-60]. Additionally, is a putative module harboring a β-1,6-N-acetylglucosamine operon (M016)
233 is present, which may contribute to biofilm production[15]. A putative cobalt uptake ABC transport system
234 (M001) is associated with *lemA* from Shell 2. Other unique modules, such as a myristoyl transferase (*nmt*1),
235 which is known to aid in protein modification of cell membrane components are also consistent with
236 potential immune modulation. Furthermore, we identified a gene cluster associated with aminoglycoside
237 phosphotransferase resistance, as well as cell division and DNA repair proteins, such as *divIVA* and *recO*
238 (M004), related to the maintenance of cellular integrity under stress. Other modules in Shell 3 relate to
239 sugar and amino acid import, including a module potentially involved in fucose metabolism (M006)[14,61].
240 Additionally, isolated α-mannosidase[62] and galactose-1-phosphate uridyltransferase gene clusters are
241 present, encoding key enzymes in the Leloir pathway for galactose metabolism which are implicated in

mucin catabolism following deprotection by fucosidases or sialidases[63]. Another module (M021) contains gene clusters that represent the oxidative branch of the pentose phosphate pathway, which generates reducing equivalents for biosynthesis, whereas its nonoxidative counterpart supports alternative sugar metabolic pathways. The remaining modules are consistent with methionine salvage (M014) and nucleic acid uptake (M003).

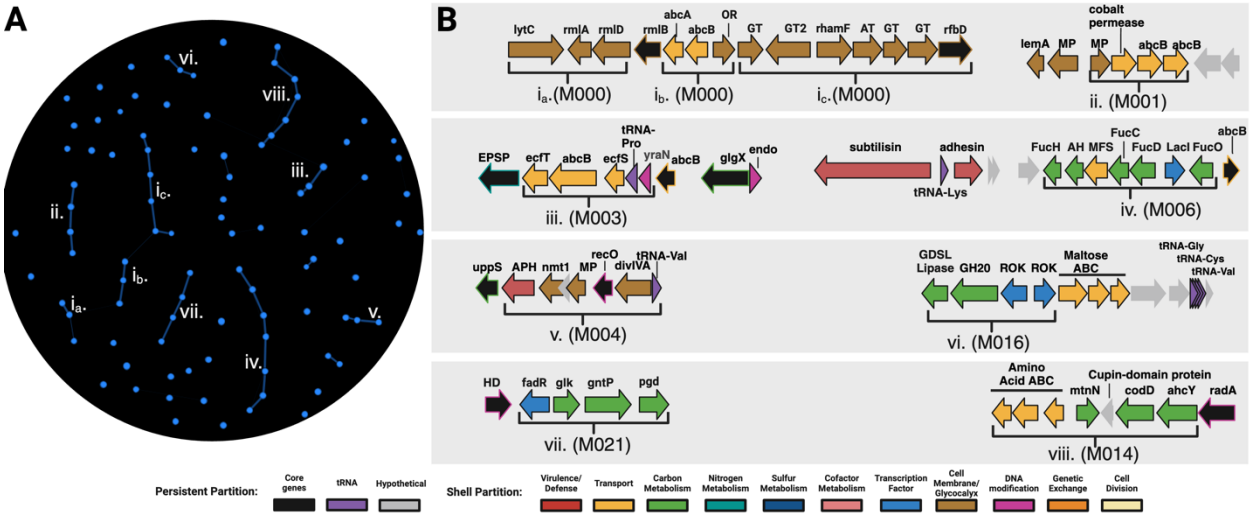

Figure 7: Shell 3 associated modules and gene clusters. (**A**) Node-and-edge plot for Shell 3 which is associated with *G. vera* and *G. sialidae* and the subgenera they form (Set A). Unique modules are labeled (**i-ix**). (**B**) Gene arrow plots of the modules labeled in (**A**) with the flanking genes of the module. Genes are colored according to their function. *G. vera* and *G. sialidae* contain a module that is involved in the synthesis of lipopolysaccharide (LPS) (M000). Cell walls are imbedded with rhamnose via three submodules: (**i_a**) rhamnose biosynthesis genes *rmlABD* with *lytC* for autolysis of the cell wall, (**i_b**) extrusion the LPS with *abcAB* and a NAD(P)-dependent oxidoreductase (OR), and (**i_c**) LPS sugar modification with glucotransferases (GT), as well as an acetyltransferase (AT**)**. A second membrane and import module (**ii**) contains membrane proteins (MP) and possible cobalt ABC imports (M001). (**iii**) DNA/RNA uptake module (M003) involves two energy-coupled transfer subunits, *ecfST* and two ABC transporters with *yraN* and an endonuclease (endo). The fucose catabolism operon (**iv**) is glucose-repressed (*LacI*) with an entire fucose catabolism *fucCDHO* operon (M006) coupled to a major facilitator superfamily (MFS) import. A probable defense island (**v** M004) with cell immune modulation via a myristoyl transferase (*nmt*1), an MP defense with aminoglycoside phosphotransferase (APH, which may provide antibiotic resistance), and cell division and DNA repair proteins (*divIVA* and *recO*). (**vi**) β-1,6-N-acetylglucosamine (GH20) has a module with GDSL lipase and two repressor kinases (ROK) (M016) are in close proximity to an ABC maltose uptake module. Part of the oxidative pentose phosphate pathway (**vii**, M021) is the *fadR*-repressed module ,which contains glucokinase (*glk*), gluconate permease (*gntP*) and 6-phosphogluconate dehydrogenase (*pgd*). Finally, (**viii**) amino acid salvage of methionine and purine metabolism through an ABC transport mechanism (M014), 5-methylthioadenosine nucleosidase (*mtnN*), cytosine deaminase (*codD*), and adenosylhomocysteinase (*ahcY*).

Shell 4: *G. vera* Specific Modules Associated with Sugar Import

Shell 4 contains gene clusters that are specific to *G. vera* (**Figure 4**, **Supplementary Table 2**). Ten distinct sugar-related modules are present, nine of which are consistent with regulons for carbohydrate

use[64]. These include galactose degradation (M035/M036, M052) and sialic acid catabolism (M019) operons. Galactose is commonly linked to GalNAc in host mucins, and its breakdown is a crucial step in mucin deprotection, whereas sialic acid (Neu5Ac, Nan) and a fucose (Fuc) degradation and import module (M040, in addition to the Shell 3 module M006)[22,24] are the negatively-charged sugar moieties found at the terminal ends of many glycan chains and are used to protect the mucin from degradation. Two glucose-repressed modules putatively targeting α-glucans (M034) are present in *G. vera*, in addition to a separate module related to glutamine metabolism (M044). *G. vera* also contains modules involved in L-arabinose metabolism (M033), which include a ROK family repressor. ROK repression also factors into modules for putative xylose/fructose[24] degradation (M043), and galactose import (M036). Another unspecified helix–turn–helix repressor is adjacent to a separate α/β-galactosidase[24,62] module associated with the Leloir pathway (M035). In addition to sugar degradation, The ability of *G. vera* to adhere to host tissues may be facilitated by biofilm-related gene clusters and modules in Shell 4. These include gene cluster orthologs of internalin B[65], a pilin, a conjugal transfer protein, and other adhesion-related lone gene clusters such as a discoidin-domain protein (M040).

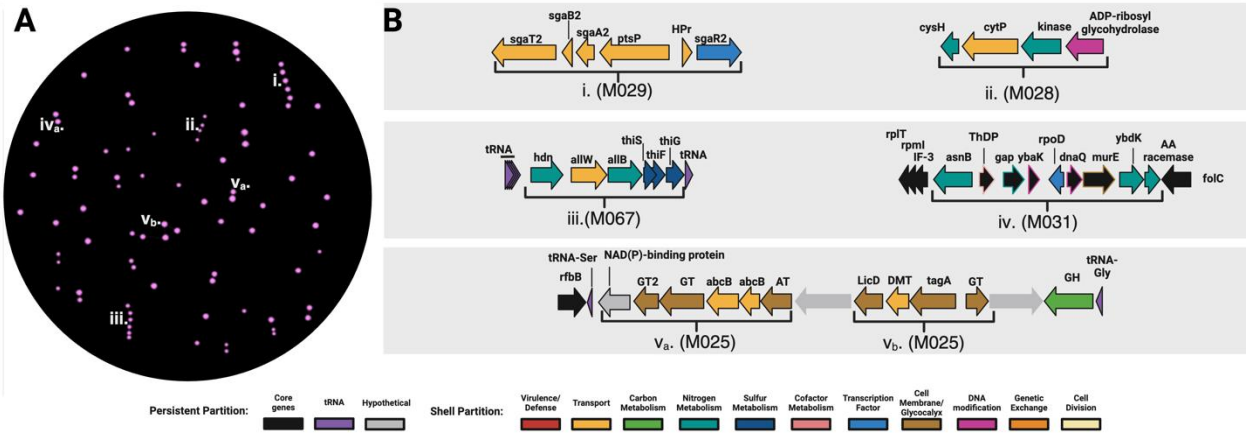

Figure 8: Shell 4 associated modules and gene clusters. (**A**) Node-and-edge plot for Shell 4, which is associated with *G. vera*. Unique modules are labeled (**i-x**). (**B**) Gene arrow plots of the modules labeled in (**A**) with the flanking genes of the module. Genes are colored according to their function. The modules most closely associated with *G. vera* are almost all associated with import of sugars for probable carbon catabolism. Many are specifically repressed by the presence of glucose (*lacI*) or N-acetylglucosamine (*nagC*). The largest (**i**) is an arabinose import (*abcABC*) and metabolism regulon (*ara1ABD*) with the accessory genes CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase (*pgsA*) and biotin--[acetyl-CoA-carboxylase] ligase (*birA*). In close proximity is (**ii**) Xylose import (*abcABC*) and metabolism (*xylAB*) (M043). Lactose repression is probable for (**iii**) β-galactosidase (β-gal)/GH65, β-phosphoglucomutase (*pgmB*) and their associated import genes (*abcABC*) (M052), and (**vii**) an α-glucosidase (*aglA*) with an import mechanism (*abcABC*) (M034). (**viii**) Lactose repression is coupled with the synthesis of glutamate (*glnE*) a bile salt hydrolase (*bsh*) and a drug/metabolite transporter (DMT) (M044). The combined modules (M035 and

M036) (**iv.** and **v.**, respectively) together constitute a putative galactose regulon, *galEKMT,* and import (*abcABC*) has two potential repressors, the helix–turn–helix protein and the ROK repressor. The modules also include a zeta toxin (*bzt*), phosphotransferase IIC (pts-IIC) and a hypothetical protein (DUF4838). (**viii**) The potential nucleoside import module contains energy-coupled transport *ecfST* and *abcAC* with the pyrimidine hydrolase *rihA* (M044). (**x**) An import module (*abcCBB*) with a GH2 glucoside hydrolase is located in a region associated with a core glutaredoxin (*glx*) and a *G. vera*-specific discoidolysin (*dly*) and fucose isomerase (*fucI*). (**vi**) This module has the environmentally sensitive *marR* for transcriptional repression and imports fucose (*fucP*) and is coupled to ribose metabolism genes (*rbsK*, A1E, and *deoC*).

Shell 5: Species-specific amino acid catabolism

Shell 5 contains gene clusters that are specific to *G. aggregatae, G. mediae, G. novi, G. washingtoni, G. bivia, G. colpomonas,* and *G. kenyensis* (**Figure 4**, **Supplementary Table 2**) and is consistent with various genetic strategies to survive in the urinary tract environment. The import and utilization of operons for allantoin/hydantoin and ascorbate (M067 and M029) may allow their use as carbon sources in urine. Genes responsible for the synthesis of essential compounds, such as cytosine (M028), asparagine (M031), and folate (*folP*, *folk* as lone gene clusters in the shell), are also present. A cell membrane/biofilm structural module (M025) is also present, which may influence the adherence, aggregation, and biofilm formation of these species.

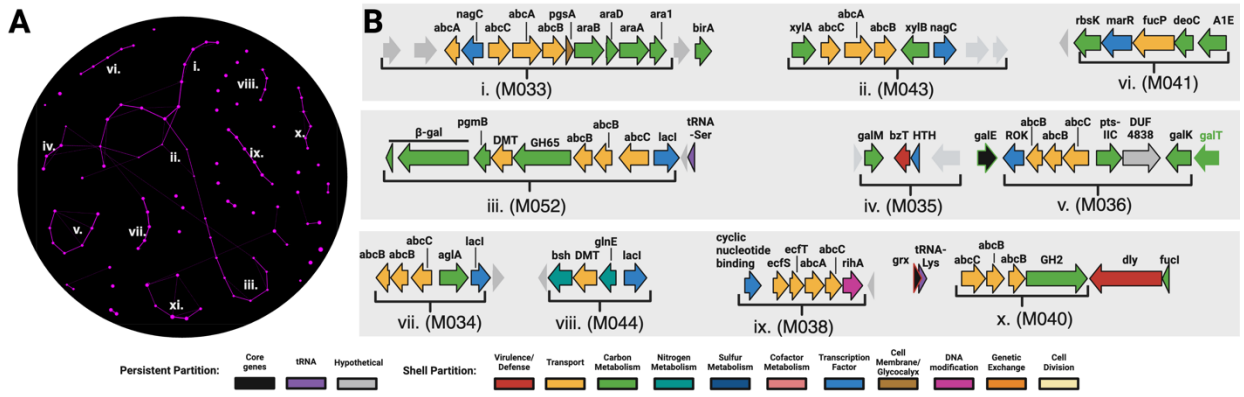

Figure 9: Shell 5 associated modules and gene clusters. (**A**) Node-and-edge plot for Shell 5, which is associated with the subgenus that contains *G. aggregatae, G. mediae, G. novi, G. washingtoni, G. colpomonas,* and *G. bivia*. (**B**) Gene arrow plots of the modules labeled in (**A**) with the flanking genes of the module. Genes are colored according to their function. The genes in the other modules are colored according to function. Unique modules are labeled (**i-v**). (**i**) Ascorbate import gene clusters (*sgaT2,B2,A2, ptsP* and *HPr*) are transcriptionally repressed (*sgaR2*) (M029) (**ii**) ADP-ribosylglucohydrolase, cytosine permease (*cytP)*, cysteine hydrolase (*cysH*) and a kinase (M028) (**iii**) Allantoin (*allBW*)/ hydantoin (*hyd*)/ thiamine (*thiFGS*) import and utilization module (M067). (**iv**) Synthesis of the amino acid asparagine (M031). (**v**) O-antigen production consists of glycosyl transferases (GT), acetyl transferases (AT), DMT/ABC transferases for extrusion (DMT, *abcBB*), and strengthening and protecting cell membranes (*licD*, *tagA*) (M025).

Shell 3 and Shell 5 Homolog Gene Clusters: Vertical Transmission within *Gardnerella* species

We noted in Shell 3 that there was a population of gene clusters that appeared uniquely present in certain species but instead had highly related "paired" gene clusters in Shell 5 that had deviated in their underlying DNA sequence (**Figure 10A**). The affected gene clusters in Shell 3 are associated with *G. vera* and *G. sialidae,* whereas their partner gene clusters in Shell 5 are associated with *G. mediae*, *G. aggregatae*, *G. novi*, and *G. washingtoni*. An accumulation of point mutations within these gene clusters appears to have driven this "split" along phylogenetic lines, shaped by vertical transmission and individual genes evolving under distinct selective pressures while remaining reproductively isolated. These diverging gene clusters are interspersed among core genes, and no regions of genomic plasticity were identified in these areas. The genes involved are related to functions such as the competence machinery (for DNA uptake) and peptidoglycan remodeling[30], and their divergence between subgenera suggests evolutionary pressure to modify the genes involved in the maintenance and remodeling of the bacterial cell wall and envelope.

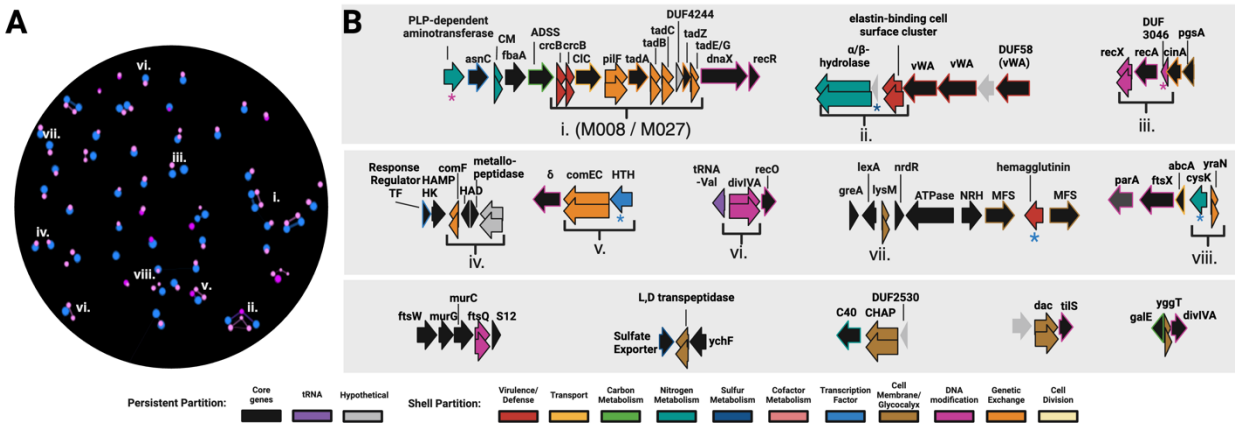

Figure 10. Gene clusters under vertical evolutionary divergence related to DNA competence and peptidoglycan remodeling in *Gardnerella* **(A)** Node-and-edge plot for the homologs in Shells 3 and 5. **(B)** Homologous genes are found in operons related to competence and genetic exchange Genes are colored according to their function. The genes in the other modules are colored according to function. Duplicated clusters are shown with double arrows. Gene clusters with members in only one shell are given an asterisk. The asterisk is colored pink if it is absent in Shell 3 and blue if it is absent in Shell 5. **(i)** Split module with type IV competence pili (*pilF*), attachment and competence (*tadABCZE/G*), toxicity pump (*crcB*), choirmate mutase (CM), and PLP-dependent aminotransferase (M008 in Shell 3, M027 in Shell 5). **(ii)** Cell membrane/host interface with Von Willibrand (vWA) domains, elastin-binding domains and alpha/beta hydrolase. **(iii)** Competence (*cinA*) and DNA recombination and repair (*recXA*). **(iv)** Competence (*comF*) and a hypothetical cluster **(v)** Competence (*comEC*) and helix-hairpin-helix (HTH), (**vi**) Cell division (*divIVA*), (**vii**) Peptidoglycan-binding (*lysM*) and an unpaired hemagglutinin, (**viii**) Cysteine kinase (*cysK*) and natural transformation (*yraN*). Single sets of pairs include peptidoglycan binding (L,D-transpeptidase, CHAP, and D-alanyl-D-alanine carboxypeptidase - *dac*), and the cell division gene, *ftsQ*.

285 Shell 6: A *Gardnerella* Prophage

286 Shell 6 contains gene clusters that are not specific to any one species (**Figure 4**). There is a well-
287 described prophage from the *Siphoviridae* family in *Gardnerella* known as vB_gva_AB1 (or
288 "phage_cluster_1")[66,67]. Shell 6 is composed entirely of this phage, which is scattered across distinct
289 *Gardnerella* species groups. There are additional variant gene clusters related to this prophage in Shell 7
290 and within the cloud partition, consistent with previous reports of heterogeneity in this prophage genome[30].
291 vB_gva_AB1 is sometimes reported as several different prophages, yet the discrete clustering and synteny
292 within our PPanGGOLiN pangenome suggest that these represent variants of a single bacteriophage.

293 Shell 7: Transition from the Pangenome Cloud Partition

294 Shell 7 is the lowest frequency shell and contains clusters from across the pangenome, comprising
295 highly dynamic gene clusters subject to frequent HGT events[68]. These gene clusters are often exchanged
296 between strains and include clusters from species with too few genomes to form their own distinct shells.
297 As a result, Shell 7 serves as a transitional zone between the pangenome cloud and more structured shell
298 partitions. For example, although lacking its own dedicated shell, *G. sialidae*, which is closely related to
299 *G. vera*, has an additional sialidase variant (NanH3)[69]. Unlike NanH1, NanH3 does not appear to reside
300 within a regulon. This variant is present in multiple copies across strains of *G. sialidae* subspecies (*Gs.*
301 *pickettii* has 2.0 copies per genome and *Gs. piotii* has an average of 2.7). Shell 7 also harbors *Ga.*
302 *swidsinskii*-specific gene clusters consistent with a bacteriocin (*latA*)[70] and related-machinery (M072) as
303 well as a type II restriction-modification system (M066) putatively recognizing GA^{m}GGC sequence
304 motifs[71].

305 With limited sequenced genomes available, *G. bivia* stands apart from other *Gardnerella* species
306 (see genomic G+C content and genome length) within Shell 7. This species appears to specialize in mucin
307 degradation, encoding multiple enzymes that target the terminal sugars of glycoproteins. These include two
308 α-L-fucosidases: a GH95, which is consistent with the removal of fucose-linked α-1-2 to galactose (via an
309 inverting mechanism[72]), and a GH29 which is associated with the removal of α1-3/4/6-linked fucose from

310 GlcNAc[73]. Additional gene clusters for mucolytic enzymes include an N-acetyl-β-D-glucosaminase
311 (GlcNAc - GH3), and a GH101 that removes the α-1,3-linkage between N-acetyl galactosamine (GalNAc)
312 and serine / threonine residues within O-linked glycoproteins.

313 Cloud partition: Dispersed or strain-specific accessory gene clusters

314 The cloud genome partition encompasses dispersed and strain-specific accessory gene clusters,
315 representing emergent genetic elements of *Gardnerella* that are too infrequent to cluster into shell partitions.
316 Collectively, this partition comprises 56.91% of all gene clusters in the *Gardnerella* pangenome, and
317 includes a diverse array of gene clusters (**Figure 4C)** including isolated gene modules (*Gw*, *Gb*, *Gc,* and
318 *Gn* cells in the cloud partition column representing modules M054, M063, M080-M083, M086, M088,
319 M092, M103, M112-M115), phage-related genes (M060, M077, M091, M104), defense modules (M002,
320 M058, M064, M070, M079, M082, M084, M089, M093, M099, M101, M102, M106, M107) and more
321 isolated gene clusters.

322 Among these cloud-associated modules, we discovered a small cryptic plasmid in *G. sialidae*
323 subspecies *pickettii* strain DNF1199S. This plasmid was first identified in PacBio-sequenced genomes as
324 an independent circular contig, with high coverage, forming a second replicon within an otherwise closed
325 genome sequence. Subsequent plasmid isolation and validation confirmed its presence and sequence,
326 marking the first cryptic plasmid identified within the *Gardnerella* genus. Using structural sequence
327 insights from this plasmid, we identified two additional putative plasmid sequences within the cloud
328 partition: one in the *G. vera vaginalis* strain 0756_546_1_1_BHK74[74], which was isolated from
329 cervicovaginal mucus in South Africa, and another in *G. kenyensis* strain NR04775[75], isolated from a
330 vaginal swab in Nairobi, Kenya. Both plasmids share similar sequence identities and ORF to those of the
331 DNF1199S plasmid, and their replication initiation genes clustered together in our pangenome analysis,
332 suggesting a potentially conserved replication mechanism for cryptic plasmids within the *Gardnerella*
333 genus.

Discussion

The study of *Gardnerella* has been hindered by taxonomic inconsistencies and confusing nomenclature that complicate comparative genomic analyses, making it inherently difficult to delineate the functional roles of specific species or subspecies in health and disease states within the urogenital tract. Our work seeks to resolve these ambiguities by integrating whole-genome sequencing, ANI clustering, dDDH analysis, and pangenomic stratification, to produce a more refined taxonomic classification system for *Gardnerella*. Therefore, we further propose an updated and unified *Gardnerella* nomenclature. Proposing nomenclature changes, particularly for well-established organisms such as '*G. vaginalis',* (updated here to *G. vera* subspecies *vaginalis*), can often be contentious and may face resistance within the scientific community. However, updating *Gardnerella* lineages to appropriate species and subspecies designations using accepted criteria is necessary to prevent further taxonomic confusion. Moreover, as additional *Gardnerella* species and subspecies are isolated and genome-sequenced, this updated classification provides a robust framework onto which new taxa can be systematically integrated. Ultimately, our naming efforts seek to increase clarity and ensure that the functional role of specific species in BV can be accurately interpreted, while preserving previous names where possible.

Additionally, we addressed misclassifications in public genome databases, confirming that *Gardnerella* is a distinct lineage separate from *Bifidobacterium*, but within the same *Bifidobacteriaceae* family. Public genome repositories currently contain numerous low-quality or misclassified *Gardnerella* sequences, with one major consequence being the misclassification of *Gardnerella* as *Bifidobacterium* within the Genome Taxonomy Database (GTDB). We addressed these challenges by curating a high-confidence dataset of complete genomes (**Supplementary Tables 9-12**), establishing a reference set for genomic and functional studies. This work revealed that the incorrect assignment stems from contamination of *Gardnerella* genome sequences with other genera rather than true phylogenetic relatedness. When contaminating sequences are removed, *Gardnerella* forms a distinct lineage, separate from *Bifidobacterium*,

358 which is supported by genomic G+C content, genome size, ANI, dDDH, and phylogenetic analyses (**Figure**
359 **1**,**2, Supplementary Figures 1-9**, **16** and **18**).

360 Our pangenome findings suggest that *Gardnerella* likely evolved from an ancestral
361 *Bifidobacteraceae* capable of broad carbohydrate metabolism. While the ancestral *Bifidobacteraceae* are
362 metabolic generalists, selective pressures likely drove genomic reduction and specialization in mucin
363 degradation as *Gardnerella* transitioned into the vaginal niche from elsewhere in the microbiota. After
364 adapting to the vagina, *Gardnerella* bifurcated into two major clades (historically Set A and Set B), one
365 retaining an apparent mucolytic strategy (*G. vera*, *G. sialidae*) and the other adapting to transient urinary
366 colonization (*G. aggregatae*, *G. mediae*) suggesting this divergence was driven by ecological constraints.
367 The vaginal mucin sugar-adapted clade (*G. vera*, *G. sialidae*) is the most widely studied. In the human
368 vaginal niche, mucin is composed of O-linked glycosylations with N-acetylgalactosamine (GalNAc)
369 serving as the core sugar. GalNAc is linked to serine or threonine residues in the mucin protein backbone,
370 forming sugar chains (glycans) that are vital for protecting and lubricating epithelial surfaces. The four of
371 the most common sugars attached to GalNAc are sialic acid, fucose, mannose, and galactose. In the
372 *Gardnerella* pangenome, modules associated with the utilization of these specific sugars are tightly
373 associated with both *G. vera* and *G. sialidae* (Shells 2 and 3, **Figure 3B**). The machinery to break down
374 three out of four of these sugars is present in both of these species: sialidase[15,24,42,62,73-7] in (M019),
375 fucose[24,62,75] in complete operons (M009), and a lone mannosidase[62]. *G. vera*, alone, has multiple operons
376 to metabolize the fourth sugar, β-galactose[24,62] in (M035/M036 and M052).

377 The association of the sialidase activity of *Gardnerella* and BV is well documented[15,43,63,76-79], as is
378 the association with a specific operon and pathway for the import and catabolism of sialic acids[55,63].
379 However, the sialidase operon has not previously been definitively linked to any clade or lineage. Two
380 separate species emerged from the mucin sugar clade: *G. vera*, an 'archivist' of sorts, retaining a series of
381 repressed operons associated with sugar-degradation, and *G. sialidae*, which seems to have specialized
382 gene-clusters for breaking down the outer defense sugar moieties of cervico-vaginal mucin. *G. vera*
383 (formally *G. vaginalis*) has developed a sophisticated system to interact with human mucin and alter its

384 sugar composition, potentially contributing to both its virulence and survival in the vaginal environment.
385 *G. vera* is both an ancient reproductive tract pathogen[80] and one of the two most frequently detected species
386 of *Gardnerella* in vaginal secretions today (alongside *G. aggregatae*)[11,14,23,81]. It comprises the largest
387 percentage of our *Gardnerella* genomes, allowing more detailed scrutiny in pangenome analysis and
388 gaining its own shell partition of gene clusters, Shell 4. Shell 4 contains two β-galactose operons–one is
389 putatively inhibited by the presence of glucose through LacI repression and channels the sugar into the
390 nonoxidative pentose phosphate pathway (PPP) (M052), and the other contains gene clusters common in
391 the Leloir pathway for galactose metabolism (M035/M036). Historically, fermentation of the pentose sugars
392 arabinose and xylose was used to categorize *Gardnerella* subtypes[17,24,82]. The modules for both are in *G.*
393 *vera* and, like galactose, are linked to PPP. The pentose, ribose, is an operon that has a gene cluster
394 consistent with the import of fucose (fucP), which is intriguing as fucose fermentation was previously
395 shown to stimulate ribose utilization[83]. The other species within this clade is *G. sialidae*. As specialists, it
396 harbors up to three different sialidases, often in multiple copies in the genome (hence the name of this
397 species)[24,77,84,85]. *G. sialidae* lacks the regulons for wide sugar utilization that are available to *G. vera*. It
398 also has a smaller overall genome (**Supplementary Figure 12**). This finding suggests that while *G. vera*
399 maintains metabolic versatility, *G. sialidae* has undergone reductive evolution toward a more specialized
400 niche.

401 The second clade of *Gardnerella* lacks the mucolytic traits of the first but contains gene clusters
402 that may support transient urinary tract colonization. More genetically diverse, this clade has operons linked
403 to nutrient scarcity, the use of amino acids as carbon sources, and molecules that are abundant in urine.
404 Bacteria surviving in the urinary tract must adjust to conditions such as low carbon availability, increased
405 salinity, low-to-moderate oxygen levels, an abundance of nitrogen, and limited iron. Shell 5, linked to the
406 species *G. aggregate*, *G. mediae*, G. *novi*, and *G. washingtoni*, suggests that these strains may be well suited
407 to urinary environments. In humans, allantoin, a stable product of uric acid oxidation, is found primarily in
408 urine. Ascorbate (vitamin C) is present in the urine when its intake exceeds the body's absorption capacity.

These species contain gene clusters consistent with their utilization, akin to other opportunistic uropathogens known to utilize ascorbate as a carbon source.

As the *Gardnerella* clades evolved and diverged from one another, certain protein-coding genes, particularly those involved in genetic exchange and host interactions, appeared to have undergone genetic diversification (**Figure 10**). Proteins responsible for HGT, such as those mediating conjugation, transduction, and competence, exhibit increased variation between the *Gardnerella* clades. This diversification likely reflects lineage-specific adaptations to distinct ecological niches or selective pressures. However, it remains unclear whether this divergence in HGT-related machinery was a cause or a consequence of *Gardnerella* clade bifurcation (historically Set A and Set B). Further investigation of these factors as potential drivers of *Gardnerella* speciation is warranted.

Finally, we report the first isolation of a cryptic plasmid from the *Gardnerella* genus (**Supplementary Table 14**). Previously, *Gardnerella* was thought to lack such mobile genetic elements, which are inherently valuable as vector-chassis backbones for genetic engineering. Recent efforts to adapt genetic tools from *Bifidobacterium* for engineering *Gardnerella* strains showed promising results, achieving reproducible transformation, albeit with low efficiency[86]. Our discovery of this extrachromosomal and self-replicative genetic element, along with the identification of additional homologous sequences in public datasets, now provides the opportunity to develop genetic tools for *Gardnerella*. Overcoming genetic intractability within *Gardnerella* should be prioritized, as this will greatly advance our ability to investigate the role of *Gardnerella* species in health and disease states.

Conclusions

Our refined *Gardnerella* taxonomy and pangenome analysis highlight the functional genetic diversity among *Gardnerella* species, with direct implications for clinical diagnostics and therapeutic strategies. Accurate taxonomy and resolution of analysis are crucial for diagnostics, yet current clinical tests fail to differentiate *Gardnerella* species, leading to a broad association of this genus with BV. The presence of

433 several species with distinct metabolic capabilities suggests that individual *Gardnerella* lineages occupy
434 specialized ecological niches within the host. For example, the mucin-degrading capabilities of *G. vera* and
435 *G. sialidae* indicate adaptation to vaginal colonization, whereas the biofilm-associated traits of *G.*
436 *aggregatae* and *G. mediae* raise questions about their role in persistent infections and BV recurrence. Our
437 complete genomes, curated datasets, refined taxonomy and updated nomenclature provide a more robust
438 foundation for *Gardnerella* researchers and provide a framework for understanding the contributions of
439 each species to the high diversity of *Gardnerella* species found in BV.

440 **Protologues**

441 ***i)*** **Description of the species *Gardnerella vera* comb. nov.**

442 With the use of current species delineation thresholds (greater than 95% ANI[29], **Figure 2**, and greater than
443 70% dDDH, **Figure 2**), genomospecies 01 (Gsp01) and genomospecies 02 (Gsp02) should be classified as
444 subspecies within a single species. This proposal is consistent with previous studies that have referred to
445 the species as C/1[21,22], Ecotype1[24], GGtype09[27], GS01[26], and *Bifidobacterium vaginale*[87] (**Figure 1**), and is
446 supported by phylogenetic data from kSNP[88] and *cpn60*[21] (**Supplementary Figures 16-19**). We propose
447 renaming this species as *Gardnerella vera* (IPA: /ˈvɪə.ɹə/), From the Latin nominative feminine singular of
448 *verus* meaning "true" or "genuine." The original strains of the genus fall within this species, including those
449 described in the original definitions of *Gardnerella vaginalis*[6,89] and *Bifidobacterium vaginale* [87].

450 This modest nomenclature change, shifted from the current *G. vaginalis* (*Gv*) to the new *Gv.*
451 *vaginalis* (*Gvv*), is intended to minimize disruption and confusion while facilitating ease of broader
452 adoption within the field. While these proposed changes, especially the reclassification of *G. vaginalis*, may
453 prompt discussion among researchers, updating these taxa to subspecies status is warranted under current
454 standards and aligns with the evolving understanding of *Gardnerella* taxonomy.

455 There are 52 genomes from cultivated strains and 11 metagenomes from the species *Gardnerella*
456 *vera*. Representative cultures have been isolated within the United States, Belgium, China, Russia, and

Denmark. These strains were isolated from the endocervical secretions, urine, catheter urine, endometrium, and skin (for one metagenome) of human subjects (**Supplementary Table 9**). The median genome length of the *Gardnerella vera* genome is 1.67 Mb. The median G+C percentage content is 41.36. The type strain for *Gardnerella* subspecies *vaginalis* is the same as that for the genus: ATCC 14018^{T}. (This strain is the same as Dukes 594^{T}, DSM 4944[90], CCUG 3717^{T}, JCM 11026[91], NCTC 10915^{T}, and NCTC 10287^{T}) The whole-genome sequence of ATCC 14018^{T} is available from NCBI under assembly accession number GCA_003397685.1.

<u>Emended description of *Gardnerella vera* subsp. *vaginalis*</u>: The strains formerly classified as Gsp01[25], *G. vaginalis*, and Ecotype1A[24] should be reclassified as *Gardnerella vera* subspecies *vaginalis*. All publicly available *Gardnerella vera* subsp. *vaginalis* strains have been isolated from the urogenital tract of women or identified via gut metagenomes. Our curated collection included 43 genomes from cultivated strains and 10 metagenomes for this subspecies (**Supplementary Table 9**). The mean genome length of this collection is 1.67 Mb. The median % G+C content of the DNA is 41.26. The type strain for the *Gardnerella vera* subspecies *vaginalis* should remain ATCC 14018^{T} (also known as Dukes 594^{T}, DSM 4944^{T}, CCUG 3717^{T,} JCM 11026^{T}, NCTC 10915^{T}, or NCTC 10287^{T}). The complete genome sequence of the type strain is available from NCBI under assembly accession number GCA_003397685.1.

<u>Description of *Gardnerella vera* subsp. *reginae*</u>: These strains were previously designated as genomospecies 02. This species is isolated from the urogenital tract of women, with eight representative genomes from cultured isolates (55152[22], 1400E[22], 41V[92], JCP8108[78], N165[75], 0756_546_1_1_BHK10[74], DNF00354, and DNF00622A) and one metagenome (GCA_947253645.1) for this subspecies. (**Supplementary Table 9**). The median genome length in this collection is 1.64 Mb. The median % genomic G+C content of the DNA is 41.31. We suggest that strain JCP8108 be assigned as the type strain for *Gardnerella vera* subsp. *reginae*. This strain is currently a reference genome for the Human Microbiome Project[93] (HMP ID 1114) and is available through ATCC/BEI.

481 **ii) <u>Description of the species *Gardnerella sialidae* comb. nov.</u>**

482 Genomic analyses support the establishment of a new species that includes genomospecies 03 (Gsp03, *G.*
483 *pickettii*) and genomospecies 04 (Gsp04, *G. piotii*), which are currently collectively referred to as B/2[21,22],
484 Ecotype 2[24], Gsp03/04/11[25], GGtype08[27], GS02[26], and *Bifidobacterium piotii*[87]. We propose naming this
485 new species *Gardnerella sialidae* (IPA: /ˈsaɪ.æl.ə.deɪ/, abbreviated *Gs*). The name is derived from the Greek
486 word σίαλον (*sialon*), which is the root of sialic acid, an abundant human glycoprotein found in mucin. The
487 species is named for the intact sialidase catabolism operon and one or more sialidase genes present in the
488 genome of each strain. The suffix "-*idae*" is the Latin feminine genitive singular.

489 As part of this revision, we suggest the following nomenclature: Gsp03 would be reclassified as *G.*
490 *sialidae* subspecies *pickettii* (abbreviated *Gs. pickettii* and *Gspk*) and Gsp04 can be reclassified as *G.*
491 *sialidae* subspecies *piotii* (abbreviated as *Gs. piotii* and *Gspt*). While abbreviating these names presents
492 challenges and requires four-letter designations, we have maintained the established nomenclature (i.e.,
493 *pickettii* and *piotii*) for consistency. Additionally, Gsp11 (GS06[26] , GGtype08[27], or *Bifidobacterium vaginale*
494 B[87]) clusters within the *G. sialidae* species. Although only a single cultured strain (GED7760B[94]) is
495 currently available, we propose recognizing this strain as a distinct subspecies, which would be reclassified
496 into the *G. sialidae* subspecies *rara* (abbreviated *Gs. rara* and *Gsr*).

497 Our curated collection included 26 genomes from cultured isolates and seven metagenomes (MAG)
498 that passed quality thresholds (**Supplementary Table 10**). These isolates were found in the endocervical
499 secretions, urine, and bladders of human subjects. Representative cultures have been isolated within the
500 United States, Canada, China, Kenya, South Africa, Portugal, and Belgium. The median genome length is
501 1.55 Mb. The median % genomic G+C content of DNA is 42.37.

502 <u>Emended description of *Gardnerella sialidae* subsp. *pickettii*</u>

503 Previously designated *Gardnerella pickettii*[28], Ecotype 2B[95], and genomospecies 03 (Gsp03)[25], *Gardnerella*
504 *sialidae* subsp. *pickettii* (IPA: /ˈpɪk.ɛt.iː.iː/, abbreviated *Gspk*) is given the Latin masculine genitive singular,
505 meaning "belonging to Pickett". Sousa *et al*. originally coined this name. This subspecies was named in

honor of M.J. Pickett, who reclassified and renamed *Gardnerella vaginalis* from *Haemophilus vaginalis*[25,89]. There are 16 cultivated *Gardnerella sialidae* subsp. *pickettii* strains in our collection and four metagenomes (GCA_947089705.1, GCA_946999255.1, GCA_963512395.1, and GCA_963511315.1) that passed quality thresholds. The median genome length is 1.56 Mb. The median % genomic G+C content is 42.28. The type strain is retained as c17Ua_112^{T}, which was isolated from the urine of a human female in Porto, Portugal, in October 2017[28]. The complete genome sequence of this strain is available from NCBI under assembly accession number GCA_029226345.1.

Emended description of *Gardnerella sialidae* subsp. *piotii*

Previously designated as *Gardnerella piotii*, genomospecies 04 (Gsp04)[25], or Ecotype2A[24], *Gardnerella sialidae* subsp. *piotii* (IPA: /piːˈɒt.iː.iː/, abbreviated *Gspt*) is given the Latin masculine genitive singular of the family name for Piot, meaning “belonging to Piot”. It was named in honor of infectious disease researcher Peter Piot, who authored an article distinguishing sugar fermentation types within *Gardnerella*[25,96]. Our curated collection included nine cultivated strains and two metagenomes (GCA_947041305.1 and GCA_947254105.1) that passed quality thresholds. The median genome length is 1.54 Mb. The median % genomic G+C content of DNA is 42.47. The type strain is UGent 18.01[25] (also known as LMG 30818^{T} and CCUG 72427^{T}), which was isolated from human cervical secretions in Ghent, Belgium, in October 2012. The complete genome sequence of the type strain is available from NCBI under assembly accession number GCA_003397585.1.

Description of *Gardnerella sialidae* subsp. *rara*

Previously designated as Gsp11 within the NCBI database, *Gardnerella sialidae* subspecies *rara* (IPA: /ˈrɛə.rə/, abbreviated *Gsr*) bears the Latin feminine nominative singular of *rarus,* meaning "rare". This subspecies contains one known cultivated strain, GED7760B[94] and one metagenome (GCA_947090325.1). These strains were both isolated from human cervical secretions. The median length is 1.39 Mb, and the median % genomic G+C content of the type strain is 43.45.

530 **iii) <u>Description of the species *Gardnerella aggregatae* comb. nov.</u>**

531 Another species group is formed by genomic thresholds and contains the species currently designated as
532 *Gardnerella leopoldii* (Gsp05), *Gardnerella swidsinskii* (Gsp06), and a collection of genomes that could be
533 tentatively considered as genomospecies 15 (Gsp15). These three groups are collectively known as A/4[21,22],
534 Ecotype 3B[24], Gsp05/06[25], GGtype04[27], GS03[26], and *Bifidobacterium leopoldii/swidsinskii*[87]. Thus, we
535 propose to name this species *Gardnerella aggregatae* (IPA: /agˈɾɛːgaː.teː/, abbreviated *Ga*), from the Latin
536 nominative feminine singular of *aggregare,* meaning "to flock together". *Gardnerella aggregatae* differ
537 from *G. vera* and *G. sialidae* owing to the presence of genes which are content consistent with adherence
538 and biofilm formation (see the pangenome section below). Within *Gardnerella aggregatae*, we propose
539 reclassification of the three subspecies as *G. aggregatae* subspecies *leopoldii*, *G. aggregatae* subspecies
540 *swidsinskii*, and the Gsp15 group as *G. aggregatae* subspecies *hutchinsoni*.

541 There are 34 genomes from cultivated strains and seven metagenomes that pass quality thresholds
542 for the species *G. aggregatae* (**Supplementary Table 11**). Representative cultures have been isolated in the
543 United States, Canada, China, Russia, South Africa, Kenya, Denmark, Belgium, France, and Italy. These
544 were isolated from endocervical secretions, the gut microbiota, urine, catheter urine, and blood from human
545 subjects. The median genome length for *G. aggregatae* is 1.58 Mb. The median % genomic G+C content
546 of DNA is 42.05. We suggest that the type strain for the *G. aggregatae* species should be *Gardnerella*
547 *aggregatae* subsp. *swidsinskii*, GS 9838-1[25] (also known as LMG 30812^{T} and CCUG 72429^{T}), which was
548 isolated from human cervical secretions in St. Petersburg, Russia in October 2008. The whole-genome
549 sequence of the type strain is available from NCBI under assembly accession number GCA_003293675.1.

550 <u>Emended description of *Gardnerella aggregatae* subsp. *leopoldii*</u>

551 Previously designated *Gardnerella leopoldii* or genomospecies 05 (Gsp05)[25], *Gardnerella aggregatae*
552 subsp. *leopoldii* (IPA: /ˈliː.əpoʊld.iː.iː/, abbreviated as *Gal*) was given the Latin genitive masculine singular
553 for the family name for Leopold, meaning "belonging to Leopold". Vaneechouette *et al*. originally coined
554 this name in honor of Sidney Leopold, who described the first strains of the *Gardnerella* genus[25,97]. Within

our curated collection are 13 cultivated strains, but no currently available MAG pass quality thresholds (**Supplementary Table 11**). The median genome length is 1.55 Mb, and the DNA G+C content is 42.17%. The type strain is UGent 06.41[25] (also known as LMG 30814^T and CCUG 72425^T), which was isolated from human endocervical secretions in Ghent, Belgium in October 2012. The complete genome sequence of the type strain is available from NCBI under assembly accession number GCA_003293675.1.

Emended description of *Gardnerella aggregatae* subsp. *swidsinskii*

Previously described as *Gardnerella swidsinskii,* genomospecies 06 (Gsp06)[25]*, Gardnerella aggregatae* subsp. *swidsinskii* (IPA: /ˈɛfˠid͡zin.ski.iː/., abbreviated as *Gas*) has a name derived from the Latin masculine singular from the family name for Swidsinski, meaning "belonging to Swidsinski." Named in honor of Alexander Swidsinski by Vaneechouette *et al*.[25] because he recognized that clue cells in the Gram stains of vaginal swabs of women with BV are indicative of *Gardnerella* biofilm formation[25,98]. Our curated collection included 19 cultured isolates and five metagenomes that passed quality thresholds (**Supplementary Table 11**). The median genome length is 1.62 Mb. The median % genomic G+C content of DNA is 41.95. The type strain is GS 9838-1^T (LMG 30812^T and CCUG 72429^T), which was isolated from endocervical secretions in St. Petersburg, Russia in October 2008. The complete genome sequence of the type strain is available from NCBI under assembly accession number GCA_003293675.1.

Description of *Gardnerella aggregatae* subsp. *hutchinsoni* comb. nov.

Designated here as genomospecies 15 (Gsp15) and with a MAG previously assigned as *Bifidobacterium* sp003585845[87], *Gardnerella aggregatae* subsp. *hutchinsoni* (IPA: /ˈhʌtʃɪnˈsoʊn.aɪ /, abbreviated *Gah*) is named with the Latinized genitive masculine singular for the Fred Hutchinson Cancer Center, where this strain was first cultivated. There are two cultivated members (DNF00983 and DNF01192) and two MAG (GCA_946998015.1 and GCA_947252625.1) that pass quality thresholds in this group. The genome length is 1.44 Mb, while the % genomic G+C content of DNA is 42.25.

578 **iv) <u>Description of the species *Gardnerella mediae* comb. nov.</u>**

579 Another species group is formed by genomic thresholds and contains the species previously known as
580 D/3[21,22], Ecotype 3B[24], Gsp08/09/10[25], GGtype04[27], GS05[26], or *Bifidobacterium vaginale*_C/E[87].
581 Pangenome analysis revealed that this species shares biologically relevant operons with the Set A subgenus
582 containing *G. vera* and *G. sialidae*, including a sialidase operon. By ANI and dDDH thresholds, this species
583 also falls within the Set B subgenera along with *G. aggregatae*, *G. washingtoni*, *G. bivia*, *G. novi*, and *G.*
584 *kenyensis*, suggesting that the species is an intermediate between Set A and Set B. Thus, we propose naming
585 this species *Gardnerella mediae* (IPA: /ˈmeː.dɪ.ae/, abbreviated *Gm*), from the Latin nominative feminine
586 plural of *medius*, meaning "the middle ones" or "the central ones," for the phylogenetic position of this
587 species. Within *Gardnerella mediae*, we propose reclassification of three subspecies as *G. mediae*
588 subspecies *greenwoodii*, *G. mediae* subspecies *olympica*, and *G. mediae* subspecies *lucinaria*.

589 There are 14 genomes from cultivated strains and five metagenomes that pass quality thresholds
590 for the species *G. mediae* (**Supplementary Table 12**). The median genome length is 1.52 Mb. The median
591 % genomic G+C content of DNA is 43.24. The type strain is c31Ua_26[28], which was isolated from human
592 urine in Porto, Portugal, in 2019[28].

593 <u>Emended description of *Gardnerella mediae* subsp. *greenwoodii*</u>

594 Previously described as *G. greenwoodii* (or genomospecies 08)[25,28], *Gardnerella mediae* subsp. *greenwoodii*
595 (IPA: /griːnwʊdiːiː/, abbreviated *Gmg*) is derived from the Latinized genitive masculine singular of
596 Greenwood. Sousa[99] *et al*. originally proposed this name in honor of J. R. Greenwood, who reclassified and
597 renamed *Gardnerella vaginalis* from *Haemophilus vaginalis*[25,89]. Our curated collection included four
598 isolates (101[100], 00703Dmash[22], c31Ua_26[28], and KC1[101]) and one metagenome that passed quality
599 thresholds (GCA_946892205.1). The median genome length is 1.47 Mb. The median % genomic G+C
600 content is 43.36. The type strain is c31Ua_26^{T}, which was isolated from human urine in Porto, Portugal, in
601 2019[28]. The complete genome sequence of the type strain is available from NCBI under accession number
602 GCA_029207615.1.

Description of *Gardnerella mediae* subsp. *lucinaria* comb. nov.

Previously described as genomospecies 09 (Gsp09)[25], *Gardnerella mediae* subsp. *lucinaria* (IPA: /luː.kɪˈnaː.ri.a/, abbreviated *Gml*) stems from the Latin feminine nominative plural *lux* + adjective -*inus*, meaning "lighted," and -*aria*, meaning "one who brings forth," and is named for Lucina (Roman goddess of midwives, childbirth, and newborns). Our curated collection included six isolates (1683_1263_1_CBA4[74], 6119V5[22], DNF00172, KA00288, N160[75], and NR044[75]) and three metagenomes (GCA_946891375.1, GCA_947090285.1, and GCA_963511645.1) that passed quality thresholds. The median genome length is 1.51 Mb. The median % genomic G+C content of DNA is 48.71.

Description of *Gardnerella mediae* subsp. *olympica*

Previously described as genomospecies 10 (Gsp10)[25], *Gardnerella mediae* subsp. *olympica* (IPA: /oʊˈlɪm.pɪ.kə/, abbreviated *Gmo*) is from the Greek Ὄλυμπος (*olympos*), the mountainous home of the gods, and the Greek feminine diminutive ending -ίκα (-*ica*). This subspecies is named for Washington State's Olympic Mountains, which are near Seattle, where several of these strains were cultivated. Within our curated collection, four unique isolates were available (1500E[22], DNF01141, KA00603, and WP012[75]) and one metagenome that passed quality thresholds (GCA_946997205.1). The median genome length is 1.56 Mb. The median % genomic G+C content of DNA is 43.02.

v) Description of the species *Gardnerella washingtoni* comb. nov.

The next species formed by genomic thresholds contains the groups previously described as genomospecies 07 (Gsp07)[25], Ecotype3A[24], GGtype06[27], GS04[26], and *Bifidobacterium vaginale*_D[87]. This species encompasses several cultivated strains isolated from urine and endocervical secretions, originally by the Lewis group at Washington University in Missouri, USA[94,102], along with a strain isolated by our group in Washington state (current study, DNF00536). Thus, we propose naming this species *G. washingtoni* (IPA: /ˌwɑːʃɪŋˈtoʊnaɪ/, abbreviated *Gw*), a Latin masculine genitive singular toponym for the two Washingtons involved in the cultivation of this species. Our curated collection included three unique isolates (DNF00536, JCP8481B[78], PSS 7772B[103], **Supplementary Table 13**). The median genome length is 1.61 Mb. The median

% genomic G+C content is 42.95. We suggest that the type strain be JCP8481B as this strain is currently a reference genome for the Human Microbiome Project (HMP ID 1118) and is available through ATCC/BEI.

vi) <u>Description of the species *Gardnerella bivia* comb. nov.</u>

The next species formed by genomic thresholds contains the groups previously described as genomospecies 12 (Gsp12)[25], GGtype02[27], GS07[26], and *Bifidobacterium vaginale*_A[87]. This species forms a unique cluster that falls outside of both *Gardnerella* subgenera according to ANI thresholds. However, by dDDH, this species is included within subgenera that contain *G. aggregatae* and *G. mediae*. These species have genes associated with two major pathways for the degradation of mucins. We propose that this species be named *Gardnerella bivia* (IPA: /ˈbɪ.vi.ə/, abbreviated as *Gb*), from the Latin feminine nominative plural -- *bivia,* meaning "two roads". There are currently two unique isolates (CMW7778B[94] and DNF00502) and four metagenomes (GCA_943912865.1, GCA_946891935.1, GCA_946997805.1, GCA_947280595.1) that pass quality thresholds within our collection (**Supplementary Table 13**). The median genome length is 1.62 Mb. The median % genomic G+C content is 38.01, which is considerably lower than that of other *Gardnerella* species. We suggest that the type strain be CMW7778B as this strain is currently a reference genome for the Human Microbiome Project (HMP ID 3230) and is available through ATCC/BEI.

vii) <u>Description of the species *Gardnerella colpomonas* comb. nov.</u>

The next species formed by genomic thresholds contains the groups previously described as genomospecies 13 (Gsp13)[31], GGtype01[27], GS08[26], and *Bifidobacterium vaginale*_F[87]. It often closely clusters with *G. bivia* during comparative genome analyses but nevertheless remains distinct by accepted thresholds. We propose that this species be named *Gardnerella colpomonas* (IPA: /ˌkɑːl.pəˈmoʊ.nəs/, abbreviated as *Gc*), from the Greek κόλπος (colpos), meaning “vagina,” and μόνας (monas), meaning "single" or "unit," together meaning a single-celled organism cultured from the vagina. Within our curated collection are two unique isolates (KA00225 and KA00255) and three metagenomes available through NCBI (GCA_902481265.1, GCA_946891425.1, GCA_947252875.1) that pass quality thresholds (**Supplementary Table 13**). The

652 median genome length of the isolated strains is 1.60 Mb. The median % genomic G+C content of DNA is
653 40.53.

654 **viii) <u>Description of the species *Gardnerella novi* comb. nov.</u>**

655 The next species formed by genomic thresholds contains the groups previously described as genomospecies
656 14 (Gsp14)[5,26,62], GGtype03[27], GS09[26] and *Bifidobacterium vaginale*_H[87], and is part of the A/4 group[31].
657 This species has been cultivated by groups in the United States (KA00747, this study) and Nairobi, Kenya[31],
658 and is detected as a MAG by groups in the United States and France[104]. This was the first "new"
659 genomospecies uncovered after the original partitioning of *Gardnerella* into genomospecies by
660 Vaneechouette *et al*. Thus, we propose that this species be named *Gardnerella novi* (IPA: /ˈnoʊ.vi/,
661 abbreviated as *Gn*) from the Latin neuter nominative neuter plural meaning "the new". Within our curated
662 collection are two unique isolates (NR010[75] and KA00747) and three metagenomes (GCA_940796375.1,
663 GCA_947252425.1, and GCA_946891105.1) within NCBI that pass quality thresholds (**Supplementary**
664 **Table 13**). The median genome length of the isolated strains is 1.65 Mb. The median % genomic G+C
665 content of DNA is 45.41.

666 **ix) <u>Description of the species *Gardnerella kenyensis* comb. nov.</u>**

667 The final species formed by the genomic thresholds included the groups previously described as A/4[31] and
668 *Bifidobacterium* sp003585735[87]. We originally designated these as genomospecies 16 (Gsp16), in
669 accordance with the previous naming schemes based on ANI. However, we propose that this species be
670 named *Gardnerella kenyensis* (IPA: /kɛ.njɛnˈsɪs/, abbreviated as *Gk*), from the name of the country where
671 all strains have been found to date, Kenya, and the Latin adjectival suffix meaning "originating in,"-*ensis*.
672 All strains currently in cultivation were isolated in Nairobi, Kenya[31], and are available through NCBI
673 (NR002[75], NR021[75], and NR047[75]). Only one passed our quality threshold metrics (NR002,
674 **Supplementary Table 13**). The genome size is 1.51 Mb. The % G+C content of the DNA is 43.59.

675 **Acknowledgement**

676 Research reported in this publication was supported the National Human Genome Research Institute (R01
677 HG005816 to D.N.F.) as part of the Human Microbiome Project Initiative focused on technology
678 development. In addition to funds were provided by Fred Hutchinson Cancer Center (to C.D.J.), and the
679 National Institute of Dental and Craniofacial Research (NIDCR)/National Institutes of Health Office of
680 the Director (OD) Common Fund under award number [R01 DE027850] (to C.D.J.).

681 **Methods**

682 ***Collection, isolation, sequencing, and initial typing of Gardnerella species from the human vaginal***
683 ***microbiota.*** *Gardnerella* strains subjected to whole-genome sequencing were cultivated from human study
684 participants with and without BV in Seattle, Washington[11,105]. Bacterial isolates consistent with *Gardnerella*
685 morphotypes on Gram stain were subjected to PCR of the 16S rRNA gene with amplicon sequencing to
686 confirm phylogenetic placement in the *Gardnerella* genus and *cpn60* gene PCR with amplicon sequencing
687 to further specify these bacteria[11]. Phylogenetically diverse *Gardnerella* species were selected for whole-
688 genome sequencing.

689 ***PacBio single-molecule real-time sequencing and genome assembly.*** A total of forty-two strains were
690 selected for PacBio single-molecule real-time sequencing and genome assembly. This was performed via a
691 PacBio Sequel instrument (Pacific Biosciences) and a PacBio Sequel II instrument (Pacific Biosciences).
692 The sequencing reads were processed via Microbial Assembly software within the Pacific Biosciences'
693 SMRTAnalysis pipeline (v.9.0.0.92188). Additional assembly was carried out via Flye assembler v.2.8 as
694 needed[106]. The completeness and contamination of the genomes were assessed via CheckM[107]. The
695 genomes were annotated via RAST[108]. Two sequences were removed from the collection after average
696 nucleotide identity (ANI)[29] which revealed that they shared greater than 99.5% ANI with other isolates
697 (DNF00747 is 99.999% similar to DNF00186; DNF00753 is 99.995% similar to DNF00172). As each of

these duplicates was collected from the same patient as its respective double, it was determined that these duplicates were clonal isolates of the original genome.

Curating publicly available datasets from the National Center for Biotechnology Information (NCBI). We used publicly available genomes both to provide a direct reference for previous studies and to supplement our own sequencing efforts. While these public genomes represent a convenience sample of the true population genome and are thus prone to time, niche, and location clusters, they can still provide valuable insight into the true *Gardnerella* population, especially for rarely encountered species and strains. We searched GenBank assembly databases[109] for the genus *Gardnerella* and family *Bifidobacteraceae* (without genus assignment). Using the Type (Strain) Genome Server (TYGS)[38] to confirm the genus of these genomes, we discovered that several of the unassigned *Bifidobacteraceae* bacteria in NCBI were *Gardnerella* and included them in our analysis (further investigation into the collection revealed 268 genomes or genome-length metagenomes as of November 2024). *Bifidobacteraceae* (without genus assignment) were included if they were collected as part of the same NCBI BioProject which added 20 *Gardnerella* that were not typed further than the *Bifidobacteraceae* family. Additionally, we included three sequences (350_GVAG, 476_GVAG, and 842_GVAG) that were initially assigned to *Gardnerella* but were later renamed to *Alloscardovia* to verify this placement. The initial number of potential genomes was n = 288 (Supplementary Tables 1–7 show the sequences removed at each step below).

To verify the taxonomy of these strains we used the TYGS type strain server[38] to validate the genus. We cross-referenced the NCBI Assembly accession number with the GTDB[87] (n.b. we disagree with the taxonomic assignment of the genus *Gardnerella* in this database) and found that 29 strains did not belong to the genus *Gardnerella* (**Supplementary Table 3**). We began curating these genomes by excluding public genomes if they failed to meet one or more of our criteria. Draft genomes with more than 250 contigs were removed due to inherent unreliability (n=21 genomes were excluded via this threshold; see **Supplementary Table 4** for sequences removed). The genomes that were modified in the laboratory were also removed

722 (14018c[110] and 14019metR[111] have laboratory-induced resistance to clindamycin and metronidazole,
723 respectively; n=2, **Supplementary Table 5**).

724 Next, we screened for high-quality genomes. Genome assemblies with TYGS marker lineages in
725 any family other than *Bifidobacteraceae* and greater than 1% contamination (by CheckM) were excluded
726 (n=52, Supplementary Table 4). With some overlap in exclusion categories, truncated genomes with less
727 than 90% completeness (by CheckM) were removed (n=54, **Supplementary Tables 6 and 7**).

728 We examined type strains to remove duplicates. The most common reference strain in used in
729 *Gardnerella* studies is designated as ATCC 14018[112] in the American Type Culture Collection. There are
730 many duplicate genomes under different names for this strain, as it was deposited in type collections
731 worldwide; it is variously known as CCUG 3717 by the Culture Collection University of Gothenburg
732 (https://www.ccug.se/), JCM 11026 in the Japan Collection of Microorganisms
733 (https://jcm.brc.riken.jp/en/), and NCTC 10287 in the National Collection of Type Cultures
734 (https://www.culturecollections.org.uk/about-us/nctc/). Each of these sequences exists in the NCBI archive
735 as a separate sequence of varying length and quality. The reference strain ATCC 14018 was deposited at
736 the same time as the strain ATCC 14019[50] from the original collection of Gardner and Dukes (then known
737 as strains CV594 and CV317, respectively)[6]. The genomic sequences of these two reference strains are
738 99.97% identical by ANI and are therefore interchangeable. In separate efforts, ATCC 14018 (as both ATCC
739 14018 and NCTC 10287) was sequenced as a reference within two different studies (in addition to several
740 distinct *Gardnerella* isolates in these efforts). The submitted genome sequences were given study-specific
741 names (GH021[75] and KC4[101]) but remained identical to those of ATCC 14018, suggesting inadvertent
742 duplication of the reference sequence during public release. Other reference strains have similar
743 duplications, but to a lesser extent. ATCC 49145[112] (originally designated as the "AmMS 117 isolate" from
744 the Baxter Healthcare Corporation, https://www.atcc.org/products/49145) was chosen by the FDA-ARGOS
745 project[113] to represent *Gardnerella*. This single isolate exists twice within the database (as
746 FDAARGOS_296[113] and FDAARGOS_568[113]). The type strain for *Gardnerella piotii* is "UGent18.01"[25].
747 It appears multiple times in NCBI as GH007[75], GH019[75], and GH020[114]. The recently added type strains for

748 *G. greenwoodii* and *G. pickettii* are greater than 99.5% identical to isolates that existed prior to the
749 publication of these strains. The *G. greenwoodii* type strain (c31Ua_26[28]) is nearly identical to UMB1686[115]
750 (ANI > 99.99%). The *G. pickettii* type strain c17Ua_112[28] is 99.98% similar in ANI to GED7275B[94]. Two
751 sets of sequences in our own collection shared greater than 99.5% ANI with our own isolates (DNF00747
752 is 99.999% similar to DNF00186, DNF00753 is 99.995% similar to DNF00172). As these duplicates were
753 each collected from the same patient as their respective double, we determined that these duplicates were
754 clonal isolates of the original genome.

755 For genomes (other than reference strains) that share greater than 99% ANI, one of the pairs was
756 excluded from the study (52 were removed with the less complete, less contaminated, shorter, or
757 metagenome excluded; see **Supplementary Table 8**). In total, the curated set includes both new PacBio
758 genomes and public genomes (n=177).

759 *Bifidobacterium* strains were collected and curated in much the same manner as the *Gardnerella*
760 strains. A total of 98 strains from all known species of the *Bifidobacterium* genus and the closely related
761 *Cardovia* genera—*Alloscardovia, Aeroscardovia*, *Bombiscardovia*, *Pseudoscardovia*, and *Scardovia*—
762 were downloaded from NCBI. Whenever possible, the type strains were collected. Sequences were assessed
763 using CheckM, TYGS, and ANI via cutoffs similar to those used for the *Gardnerella* strains. The subgenera
764 "groups" of *Bifidobacterium* were assigned according to Alessandri *et al*.[116].

765 ***Integration of additional clinically isolated genomes to enhance pangenomic examinations of***
766 ***Gardnerella***

767 Prior to this work, 20 complete and closed genomes of *Gardnerella* strains were available through public
768 databases. Among these strains, three are the type strains (ATCC 14018) and one is a duplicate (JNFY4[117]
769 = *JNFY13[118]; the one retained in our analyses is marked with an asterisk), yielding 17 unique and complete
770 strains. Most samples were sequenced in 2019 and are believed to have been collected prior to sequencing
771 and submission. Most (16/17) samples were collected from endocervical secretions; one exception was
772 cultured from blood (GV37). Furthermore, these genomes are dominated by *Gardnerella vaginalis* (Gsp01,

773 n=9), with the others distributed among *G. pickettii* (Gsp03, n=1), *G. piotii* (Gsp04, n=2), *G. leopoldii*
774 (Gsp05, n=1), *G. swidsinskii* (Gsp06, n=3), and Gsp09 (n=1).

775 We added 42 complete and closed genomes of *Gardnerella* to NCBI. These strains were typed by
776 *cpn60*, kSNP, TYGS, and ANI. The strains are Gsp01 (n=9), Gsp02 (n=2), Gsp03 (n=2), Gsp04 (n=2),
777 Gsp05 (n=7), Gsp06 (n=10), Gsp07 (n=1), Gsp09 (n=3), Gsp10 (n=2), Gsp12 (n=1), Gsp13 (n=1), Gsp14
778 (n=1), and a novel strain comprising a new genomospecies, Gsp15 (n=2). There were some duplicates by
779 ANI of our strains, although some of the duplicates were from previous sequencing efforts (DNF0172* and
780 DNF0753, DNF0186* and DNF0747, DNF00502* and KA00735 (previously sequenced), DNF1159* and
781 DNF1162 (previously sequenced)). We added 40 unique strains (Supplementary Table 8), which will be
782 combined with the curated public strains (n=177 genomes).

783 ***Building core genome phylogeny***

784 We used the whole-genome comparisons of ANI[29] and digital DNA–DNA hybridization (dDDH)[38],
785 traditional phylogenetic trees for the single gene *cpn60*[21] (*GroEL*), and genome-wide analysis using SNPs
786 from kSNP[88] to determine the phylogeny of *Gardnerella*. The 16S rRNA sequences for the named
787 *Gardnerella* species are nearly identical, and no signature sequences exist to differentiate these strains[23].
788 We extracted *cpn60* from the curated strains and aligned and created a tree using maximum parsimony via
789 iTol[119] (**Supplementary Figure 16**).

790 Whole-genome alignments were performed using kSNP, a tool that identifies SNPs across hundreds
791 of microbial genomes without the need for genome alignment or a reference genome. We used kSNP to
792 create parsimony, maximum likelihood, and neighbor-joining trees (**Supplementary Figures 16-18**).

793 The threshold for genera using ANI has not been standardized, but it ranges from as low as 70%[120]
794 to as high as 83%[121] and may depend on the genera and their history. FastANI was used to determine which
795 species were above the 95% intraspecies threshold; this threshold grouped together several previously
796 named genomospecies that formed subgroups within the species threshold. The ANI matrices were formed

797 in R via ComplexHeatmap[122]. The R package "seriation"[123] was used to apply banded-anti-Robinson
798 dissimilarity rules[124] to each of the matrices.

799 dDDH comparisons were generated via comparisons of FASTA files input into the Genome-to-
800 Genome Distance Calculator[38]. Three measures were used to calculate genomic distance via dDDH: high-
801 scoring segment pairs divided by genome length, identities divided by genome length, and distance by
802 dividing the identities by the high-scoring segment pair length. All three measures were coupled with a
803 generalized linear model (GLM) to calculate the probability that two strains were related[125]. The 79% dDDH
804 was the threshold above which subspecies relationship can be inferred; more than 70% dDDH indicates a
805 species relationship[126]. The data were cleaned, condensed, and formed into matrices via R software.

806 The percentage of guanidine and cytosine in the DNA of each genome (G+C mol %) was calculated
807 and plotted for each strain (**Supplementary Figures 10 and 11**). We used the POGO-DB[37] for universal
808 orthologous gene average amino acid identity comparisons (AAI) (**Supplementary Figures**).

809 ***Constructing and building the Gardnerella pangenome***

810 We used PPanGGOLiN[40] on all the non-metagenomic sequences (n=134) to graph and cluster the genes
811 from all strains, find regions of genomic plasticity[127] and synteny[128], and separate the pangenome gene
812 clusters into core, shell, and cloud partitions (using default parameters except that clustering was set to
813 coverage: 0.8 and identity: 0.5, resulting in the following numbers: total genes = 161977, total genomes =
814 134, families = 4347, edges = 10252, persistent = 866, shell = 1007, cloud = 2474, partitions = 9, S1 =
815 106, S2 = 27, S3 = 102, S4 = 106, S5 = 113, S6 = 40, S7 = 512, modules = 117, families in modules =
816 654). EggNOG[129], InterPro[130] and KEGG Mapper[131] were run separately on each genome to gather and
817 confirm the CAZyme[132], PFam[133] and KEGG designations of the gene clusters. COG categories[134] were
818 assigned on the basis of the consensus of these designations. We added the annotation "Unclassified
819 Metabolism" to capture gene clusters that are clearly metabolism related, but the metabolite is not defined
820 or could be of multiple metabolic types.

Bibliography

1 Fredricks, D. N., Fiedler, T. L. & Marrazzo, J. M. Molecular Identification of Bacteria Associated with Bacterial Vaginosis. *New England Journal of Medicine* **353**, 1899-1911 (2005). https://doi.org/10.1056/nejmoa043802

2 Amsel, R. *et al.* Nonspecific vaginitis. Diagnostic criteria and microbial and epidemiologic associations. *Am J Med* **74**, 14-22 (1983). https://doi.org/10.1016/0002-9343(83)91112-9

3 Alcendor, D. J. Evaluation of Health Disparity in Bacterial Vaginosis and the Implications for <scp>HIV</scp>-1 Acquisition in African American Women. *American Journal of Reproductive Immunology* **76**, 99-107 (2016). https://doi.org/10.1111/aji.12497

4 Klebanoff, M. A. & Brotman, R. M. Treatment of bacterial vaginosis to prevent preterm birth. *Lancet* **392**, 2141-2142 (2018). https://doi.org/10.1016/s0140-6736(18)32115-9

5 Berman, H. L., Goltsman, D. S. A., Anderson, M., Relman, D. A. & Callahan, B. J. Gardnerella diversity and ecology in pregnancy and preterm birth. *mSystems* **9**, e0133923 (2024). https://doi.org/10.1128/msystems.01339-23

6 Gardner, H. L. & Dukes, C. D. Haemophilus vaginalis vaginitis: a newly defined specific infection previously classified non-specific vaginitis. *Am J Obstet Gynecol* **69**, 962-976 (1955).

7 Schwebke, J. R., Muzny, C. A. & Josey, W. E. Role of Gardnerella vaginalis in the Pathogenesis of Bacterial Vaginosis: A Conceptual Model. *The Journal of Infectious Diseases* **210**, 338-343 (2014). https://doi.org/10.1093/infdis/jiu089

8 Patterson, J. L., Stull-Lane, A., Girerd, P. H. & Jefferson, K. K. Analysis of adherence, biofilm formation and cytotoxicity suggests a greater virulence potential of Gardnerella vaginalis relative to other bacterial-vaginosis-associated anaerobes. *Microbiology* **156**, 392-399 (2010). https://doi.org/10.1099/mic.0.034280-0

9 Gelber, S. E., Aguilar, J. L., Lewis, K. L. T. & Ratner, A. J. Functional and Phylogenetic Characterization of Vaginolysin, the Human-Specific Cytolysin fromGardnerella vaginalis. *Journal of Bacteriology* **190**, 3896-3903 (2008). https://doi.org/10.1128/jb.01965-07

10 Castro, J. *et al.* Using an in-vitro biofilm model to assess the virulence potential of Bacterial Vaginosis or non-Bacterial Vaginosis Gardnerella vaginalis isolates. *Scientific Reports* **5**, 11640 (2015). https://doi.org/10.1038/srep11640

11 Munch, M. M. *et al.* Gardnerella Species and Their Association With Bacterial Vaginosis. *J Infect Dis* **230**, e171-e181 (2024). https://doi.org/10.1093/infdis/jiae026

12 Muzny, C. A. *et al.* An Updated Conceptual Model on the Pathogenesis of Bacterial Vaginosis. *The Journal of Infectious Diseases* **220**, 1399-1405 (2019). https://doi.org/10.1093/infdis/jiz342

13 Janulaitiene, M. *et al.* Prevalence and distribution of Gardnerella vaginalis subgroups in women with and without bacterial vaginosis. *BMC Infectious Diseases* **17** (2017). https://doi.org/10.1186/s12879-017-2501-y

14 Balashov, S. V., Mordechai, E., Adelson, M. E. & Gygax, S. E. Identification, quantification and subtyping of Gardnerella vaginalis in noncultured clinical vaginal samples by quantitative PCR. *Journal of Medical Microbiology* **63**, 162-175 (2014). https://doi.org/https://doi.org/10.1099/jmm.0.066407-0

15 Shvartsman, E., Hill, J. E., Sandstrom, P. & Macdonald, K. S. Gardnerella Revisited: Species Heterogeneity, Virulence Factors, Mucosal Immune Responses, and Contributions to Bacterial Vaginosis. *Infection and Immunity* **91** (2023). https://doi.org/10.1128/iai.00390-22

16 Piot, P. *et al.* Biotypes of Gardnerella vaginalis. *J Clin Microbiol* **20**, 677-679 (1984). https://doi.org/10.1128/jcm.20.4.677-679.1984

17 Benito, R., Vazquez, J. A., Berron, S., Fenoll, A. & Saez-Nieto, J. A. A Modified Scheme for Biotyping Gardnerella Vaginalis. *Journal of Medical Microbiology* **21**, 357-359 (1986). https://doi.org/https://doi.org/10.1099/00222615-21-4-357

18 Nath, K., Devlin, D. & Beddoe, A. M. Heterogeneity in restriction patterns of Gardnerella vaginalis isolates from individuals with bacterial vaginosis. *Research in Microbiology* **143**, 199-209 (1992). https://doi.org/https://doi.org/10.1016/0923-2508(92)90009-D

19 Ingianni, A., Petruzzelli, S., Morandotti, G. & Pompei, R. Genotypic differentiation ofGardnerella vaginalisby amplified ribosomal DNA restriction analysis (ARDRA). *FEMS Immunology & Medical Microbiology* **18**, 61-66 (1997). https://doi.org/10.1111/j.1574-695x.1997.tb01028.x

20 Links, M. G., Dumonceaux, T. J., Hemmingsen, S. M. & Hill, J. E. The Chaperonin-60 Universal Target Is a Barcode for Bacteria That Enables De Novo Assembly of Metagenomic Sequence Data. *PLoS ONE* **7**, e49755 (2012). https://doi.org/10.1371/journal.pone.0049755

21 Paramel Jayaprakash, T., Schellenberg, J. J. & Hill, J. E. Resolution and Characterization of Distinct cpn60-Based Subgroups of Gardnerella vaginalis in the Vaginal Microbiota. *PLoS ONE* **7**, e43009 (2012). https://doi.org/10.1371/journal.pone.0043009

22 Ahmed, A. *et al.* Comparative Genomic Analyses of 17 Clinical Isolates of Gardnerella vaginalis Provide Evidence of Multiple Genetically Isolated Clades Consistent with Subspeciation into Genovars. *Journal of Bacteriology* **194**, 3922-3937 (2012). https://doi.org/10.1128/jb.00056-12

23 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. *Infection and Immunity* **87** (2019). https://doi.org/10.1128/iai.00532-19

24 Cornejo, O. E., Hickey, R. J., Suzuki, H. & Forney, L. J. Focusing the diversity of Gardnerella vaginalis through the lens of ecotypes. *Evolutionary Applications* **11**, 312-324 (2018). https://doi.org/10.1111/eva.12555

25 Vaneechoutte, M. *et al.* Emended description of Gardnerella vaginalis and description of Gardnerella leopoldii sp. nov., Gardnerella piotii sp. nov. and Gardnerella swidsinskii sp. nov., with delineation of 13 genomic species within the

genus Gardnerella. *Int J Syst Evol Microbiol* **69**, 679-687 (2019). https://doi.org/10.1099/ijsem.0.003200

26 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). https://doi.org/10.1373/clinchem.2019.305474

27 Tarracchini, C. *et al.* Assessing the Genomic Variability of Gardnerella vaginalis through Comparative Genomic Analyses: Evolutionary and Ecological Implications. *Applied and Environmental Microbiology* **87** (2020). https://doi.org/10.1128/aem.02188-20

28 Sousa, M. *et al.* Gardnerella pickettii sp. nov. (formerly Gardnerella genomic species 3) and Gardnerella greenwoodii sp. nov. (formerly Gardnerella genomic species 8) isolated from female urinary microbiome. *Int J Syst Evol Microbiol* **73** (2023). https://doi.org/10.1099/ijsem.0.006140

29 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9** (2018). https://doi.org/10.1038/s41467-018-07641-9

30 Bohr, L. L., Mortimer, T. D. & Pepperell, C. S. Lateral Gene Transfer Shapes Diversity of Gardnerella spp. *Front Cell Infect Microbiol* **10**, 293 (2020). https://doi.org/10.3389/fcimb.2020.00293

31 Schellenberg, J. J. *et al.* Gardnerella vaginalis Subgroups Defined by cpn60 Sequencing and Sialidase Activity in Isolates from Canada, Belgium and Kenya. *PLOS ONE* **11**, e0146510 (2016). https://doi.org/10.1371/journal.pone.0146510

32 Barisic, V., Abdelhadi, A., Frank, A., Riojas, M. & Hazbon, M. 1 (ASM Microbe 2019 Poster, San Francisco, California, United States, 2019).

33 Zhang, G., Gao, B., Adeolu, M., Khadka, B. & Gupta, R. S. Phylogenomic Analyses and Comparative Studies on Genomes of the Bifidobacteriales: Identification of Molecular Signatures Specific for the Order Bifidobacteriales and Its Different Subclades. *Frontiers in Microbiology* **7** (2016). https://doi.org/10.3389/fmicb.2016.00978

34 Lugli, G. A. *et al.* Comparative genomic and phylogenomic analyses of the Bifidobacteriaceae family. *BMC Genomics* **18** (2017). https://doi.org/10.1186/s12864-017-3955-4

35 Mesbah, N. M., Whitman, W. B. & Mesbah, M. in *Methods in Microbiology* Vol. 38 (eds Fred Rainey & Aharon Oren) 299-324 (Academic Press, 2011).

36 Meier-Kolthoff, J. P., Klenk, H.-P. & Göker, M. Taxonomic use of DNA G+ C content and DNA–DNA hybridization in the genomic age. *International journal of systematic and evolutionary microbiology* **64**, 352-356 (2014).

37 Lan, Y., Morrison, J. C., Hershberg, R. & Rosen, G. L. POGO-DB—a database of pairwise-comparisons of genomes and conserved orthologous genes. *Nucleic Acids Research* **42**, D625-D632 (2014). https://doi.org/10.1093/nar/gkt1094

38 Meier-Kolthoff, J. P. & Göker, M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nature Communications* **10** (2019). https://doi.org/10.1038/s41467-019-10210-3

39 Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae : Implications for the microbial “pan-genome”.

Proceedings of the National Academy of Sciences **102**, 13950-13955 (2005). https://doi.org/10.1073/pnas.0506758102

40 Gautreau, G. *et al.* PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLOS Computational Biology* **16**, e1007732 (2020). https://doi.org/10.1371/journal.pcbi.1007732

41 Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media* **3**, 361-362 (2009). https://doi.org/10.1609/icwsm.v3i1.13937

42 Flores Ramos, S. *et al.* Genomic Stability and Genetic Defense Systems in Dolosigranulum pigrum , a Candidate Beneficial Bacterium from the Human Microbiome. *mSystems* **6** (2021). https://doi.org/10.1128/msystems.00425-21

43 Srinivasan, S. *et al.* Metabolic Signatures of Bacterial Vaginosis. *mBio* **6**, e00204-00215 (2015). https://doi.org/10.1128/mbio.00204-15

44 Catlin, B. W. Gardnerella vaginalis: characteristics, clinical considerations, and controversies. *Clinical Microbiology Reviews* **5**, 213-237 (1992). https://doi.org/10.1128/cmr.5.3.213

45 Bhandari, P., Tingley, J. P., Palmer, D. R. J., Abbott, D. W. & Hill, J. E. Characterization of an α-Glucosidase Enzyme Conserved in Gardnerella spp. Isolated from the Human Vaginal Microbiome. *J Bacteriol* **203**, e0021321 (2021). https://doi.org/10.1128/jb.00213-21

46 Shimada, T., Kori, A. & Ishihama, A. Involvement of the ribose operon repressor RbsR in regulation of purine nucleotide synthesis in Escherichia coli. *FEMS Microbiology Letters* **344**, 159-165 (2013). https://doi.org/10.1111/1574-6968.12172

47 Marín, E. *et al.* Unraveling Gardnerella vaginalis Surface Proteins Using Cell Shaving Proteomics. *Frontiers in Microbiology* **9** (2018). https://doi.org/10.3389/fmicb.2018.00975

48 Kurata, T. *et al.* A hyperpromiscuous antitoxin protein domain for the neutralization of diverse toxin domains. *Proceedings of the National Academy of Sciences* **119**, e2102212119 (2022). https://doi.org/10.1073/pnas.2102212119

49 Pleckaityte, M., Zilnyte, M. & Zvirbliene, A. Insights into the CRISPR/Cas system of Gardnerella vaginalis. *BMC Microbiology* **12**, 301 (2012). https://doi.org/10.1186/1471-2180-12-301

50 Yeoman, C. J. *et al.* Comparative Genomics of Gardnerella vaginalis Strains Reveals Substantial Differences in Metabolic and Virulence Potential. *PLoS ONE* **5**, e12411 (2010). https://doi.org/10.1371/journal.pone.0012411

51 Susmitha, A., Bajaj, H. & Madhavan Nampoothiri, K. The divergent roles of sortase in the biology of Gram-positive bacteria. *The Cell Surface* **7**, 100055 (2021). https://doi.org/https://doi.org/10.1016/j.tcsw.2021.100055

52 Garcia, E. M. *et al.* Phase variable colony variants are conserved across Gardnerella spp. and exhibit different virulence-associated phenotypes. *mSphere* **9** (2024). https://doi.org/10.1128/msphere.00450-24

53 Garcia, E. M. *et al.* Sequence Comparison of Vaginolysin from Different Gardnerella Species. *Pathogens* **10** (2021). https://doi.org/10.3390/pathogens10020086

54 Horne, C. R. *et al.* Mechanism of NanR gene repression and allosteric induction of bacterial sialic acid metabolism. *Nature Communications* **12** (2021). https://doi.org/10.1038/s41467-021-22253-6

55 Severi, E. *et al.* Multiple evolutionary origins reflect the importance of sialic acid transporters in the colonization potential of bacterial pathogens and commensals. *Microb Genom* **7** (2021). https://doi.org/10.1099/mgen.0.000614

56 Lenz, L. L., Dere, B. & Bevan, M. J. Identification of an H2-M3-restricted Listeria epitope: implications for antigen presentation by M3. *Immunity* **5**, 63-72 (1996). https://doi.org/10.1016/s1074-7613(00)80310-6

57 Nori, S. R. C. *et al.* Profiling of vaginal Lactobacillus jensenii isolated from preterm and full-term pregnancies reveals strain-specific factors relating to host interaction. *Microbial Genomics* **9** (2023). https://doi.org/https://doi.org/10.1099/mgen.0.001137

58 Shibata, Y., Yamashita, Y., Ozaki, K., Nakano, Y. & Koga, T. Expression and Characterization of StreptococcalrgpGenes Required for Rhamnan Synthesis inEscherichia coli. *Infection and Immunity* **70**, 2891-2898 (2002). https://doi.org/10.1128/iai.70.6.2891-2898.2002

59 Graninger, M., Nidetzky, B., Heinrichs, D. E., Whitfield, C. & Messner, P. Characterization of dTDP-4-dehydrorhamnose 3,5-Epimerase and dTDP-4-dehydrorhamnose Reductase, Required for dTDP-l-rhamnose Biosynthesis in Salmonella enterica Serovar Typhimurium LT2. *Journal of Biological Chemistry* **274**, 25069-25077 (1999). https://doi.org/10.1074/jbc.274.35.25069

60 Kalynych, S., Morona, R. & Cygler, M. Progress in understanding the assembly process of bacterial O-antigen. *FEMS Microbiology Reviews* **38**, 1048-1065 (2014). https://doi.org/10.1111/1574-6976.12070

61 Bhandari, P. & Hill, J. E. Transport and Utilization of Glycogen Breakdown Products by Gardnerella spp. from the Human Vaginal Microbiome. *Microbiol Spectr* **11**, e0443522 (2023). https://doi.org/10.1128/spectrum.04435-22

62 Putonti, C. *et al.* Genome Investigation of Urinary Gardnerella Strains and Their Relationship to Isolates of the Vaginal Microbiota. *mSphere* **6** (2021). https://doi.org/10.1128/msphere.00154-21

63 Agarwal, K. & Lewis, A. L. Vaginal sialoglycan foraging byGardnerella vaginalis: mucus barriers as a meal for unwelcome guests? *Glycobiology* **31**, 667-680 (2021). https://doi.org/10.1093/glycob/cwab024

64 Khoroshkin, M. S., Leyn, S. A., Van Sinderen, D. & Rodionov, D. A. Transcriptional Regulation of Carbohydrate Utilization Pathways in the Bifidobacterium Genus. *Frontiers in Microbiology* **7** (2016). https://doi.org/10.3389/fmicb.2016.00120

65 Quereda, J. J. *et al.* Reassessing the role of internalin B in Listeria monocytogenes virulence using the epidemic strain F2365. *Clinical Microbiology and Infection* **25**, 252.e251-252.e254 (2019). https://doi.org/10.1016/j.cmi.2018.08.022

66 Bordigoni, A., Bouchard, S. & Desnues, C. Isolation and Characterization of Gardnerella Phage vB_Gva_AB1, a Bacteriophage Infecting a Clinical Strain of Gardnerella vaginalis. *Microbiology Resource Announcements* **10** (2021). https://doi.org/10.1128/mra.00053-21

67 Malki, K. *et al.* Genomes of Gardnerella Strains Reveal an Abundance of Prophages within the Bladder Microbiome. *PLOS ONE* **11**, e0166757 (2016). https://doi.org/10.1371/journal.pone.0166757

68 Horesh, G. *et al.* Different evolutionary trends form the twilight zone of the bacterial pan-genome. *Microbial Genomics* **7** (2021). https://doi.org/https://doi.org/10.1099/mgen.0.000670

69 Robinson, L. S., Schwebke, J., Lewis, W. G. & Lewis, A. L. Identification and characterization of NanH2 and NanH3, enzymes responsible for sialidase activity in the vaginal bacterium Gardnerella vaginalis. *J Biol Chem* **294**, 5230-5245 (2019). https://doi.org/10.1074/jbc.RA118.006221

70 Hugerth, L. W. *et al.* Defining Vaginal Community Dynamics: daily microbiome transitions, the role of menstruation, bacteriophages, and bacterial genes. *Microbiome* **12** (2024). https://doi.org/10.1186/s40168-024-01870-5

71 Meinke, G., Phelan, P. J., Harrison, C. J. & Bullock, P. A. Analysis of the Costructure of the Simian Virus 40 T-Antigen Origin Binding Domain with Site I Reveals a Correlation between GAGGC Spacing and Spiral Assembly. *Journal of Virology* **87**, 2923-2934 (2013). https://doi.org/10.1128/jvi.02549-12

72 Katayama, T. *et al.* Molecular Cloning and Characterization of Bifidobacterium bifidum 1,2-α- <scp>I</scp> -Fucosidase (AfcA), a Novel Inverting Glycosidase (Glycoside Hydrolase Family 95). *Journal of Bacteriology* **186**, 4885-4893 (2004). https://doi.org/10.1128/jb.186.15.4885-4893.2004

73 Perna, V. N., Barrett, K., Meyer, A. S. & Zeuner, B. Substrate specificity and transglycosylation capacity of α-L-fucosidases across GH29 assessed by bioinformatics-assisted selection of functional diversity. *Glycobiology* **33**, 396-410 (2023). https://doi.org/10.1093/glycob/cwad029

74 Zhu, M. *et al.* Vaginal Lactobacillus fatty acid response mechanisms reveal a novel strategy for bacterial vaginosis treatment. *bioRxiv*, 2023.2012.2030.573720 (2023). https://doi.org/10.1101/2023.12.30.573720

75 Khan, S., Vancuren, S. J. & Hill, J. E. A Generalist Lifestyle Allows Rare Gardnerella spp. to Persist at Low Levels in the Vaginal Microbiome. *Microbial Ecology* **82**, 1048-1060 (2021). https://doi.org/10.1007/s00248-020-01643-1

76 Lewis, W. G., Robinson, L. S., Gilbert, N. M., Perry, J. C. & Lewis, A. L. Degradation, Foraging, and Depletion of Mucus Sialoglycans by the Vagina-adapted Actinobacterium Gardnerella vaginalis. *Journal of Biological Chemistry* **288**, 12067-12079 (2013). https://doi.org/10.1074/jbc.m113.453654

77 Robinson, L. S., Schwebke, J., Lewis, W. G. & Lewis, A. L. Identification and characterization of NanH2 and NanH3, enzymes responsible for sialidase activity in the vaginal bacterium Gardnerella vaginalis. *Journal of Biological Chemistry* **294**, 5230-5245 (2019). https://doi.org/10.1074/jbc.ra118.006221

78 Agarwal, K. *et al. Resident microbes shape the vaginal epithelial glycan landscape* (Cold Spring Harbor Laboratory, 2022).

79 Kamiya, T. *et al. Resource landscape shapes the composition and stability of the human vaginal microbiota* (Cold Spring Harbor Laboratory, 2024).

80 Devault, A. M. *et al.* A molecular portrait of maternal sepsis from Byzantine Troy. *eLife* **6**, e20983 (2017). https://doi.org/10.7554/eLife.20983

81 Shipitsyna, E. *et al.* Quantitation of all Four Gardnerella vaginalis Clades Detects Abnormal Vaginal Microbiota Characteristic of Bacterial Vaginosis More Accurately than Putative G. vaginalis Sialidase A Gene Count. *Molecular Diagnosis & Therapy* **23**, 139-147 (2019). https://doi.org/10.1007/s40291-019-00382-5

82 van Esbroeck, M. *et al.* Polyphasic approach to the classification and identification of Gardnerella vaginalis and unidentified Gardnerella vaginalis-like coryneforms present in bacterial vaginosis. *Int J Syst Bacteriol* **46**, 675-682 (1996). https://doi.org/10.1099/00207713-46-3-675

83 Autieri, S. M. *et al.* L-fucose stimulates utilization of D-ribose by Escherichia coli MG1655 DeltafucAO and E. coli Nissle 1917 DeltafucAO mutants in the mouse intestine and in M9 minimal medium. *Infect Immun* **75**, 5465-5475 (2007). https://doi.org/10.1128/iai.00822-07

84 Lopes dos Santos Santiago, G. *et al.* Gardnerella vaginalis comprises three distinct genotypes of which only two produce sialidase. *American Journal of Obstetrics & Gynecology* **204**, 450.e451-450.e457 (2011). https://doi.org/10.1016/j.ajog.2010.12.061

85 Novak, J. *et al.* Cervicovaginal Gardnerella sialidase-encoding gene in persistent human papillomavirus infection. *Scientific Reports* **13** (2023). https://doi.org/10.1038/s41598-023-41469-8

86 Kularatne, B. M. D. N. & Hill, J. E. *Transformation ofGardnerella vaginaliswith aBifidobacterium-Escherichia colishuttle vector plasmid* (Cold Spring Harbor Laboratory, 2025).

87 Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* **50**, D785-D794 (2022). https://doi.org/10.1093/nar/gkab776

88 Gardner, S. N., Slezak, T. & Hall, B. G. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* **31**, 2877-2878 (2015). https://doi.org/10.1093/bioinformatics/btv271

89 Greenwood, J. R. & Pickett, M. J. Transfer of Haemophilus vaginalis Gardner and Dukes to a New Genus, Gardnerella: G. vaginalis (Gardner and Dukes) comb. nov. *International Journal of Systematic Bacteriology* **30**, 170-178 (1980). https://doi.org/10.1099/00207713-30-1-170

90 Varghese, N. (2016).

91 Oshima, K. *et al.* Complete Genome Sequence of Gardnerella vaginalis Strain JCM 11026 ^T , Isolated from Vaginal Tracts of Women. *Genome Announcements* **3**, e00286-00215 (2015). https://doi.org/10.1128/genomea.00286-15

92 Alves, J. M. B., G.A. (2010).

93 Nelson, K. E. *et al.* A Catalog of Reference Genomes from the Human Microbiome. *Science* **328**, 994-999 (2010). https://doi.org/10.1126/science.1183605

94 Lewis, A. L. *et al.* Genome Sequences of 11 Human Vaginal Actinobacteria Strains. *Genome Announcements* **4**, e00887-00816 (2016). https://doi.org/10.1128/genomea.00887-16

95 Hardy, L., Cerca, N., Jespers, V., Vaneechoutte, M. & Crucitti, T. Bacterial biofilms in the vagina. *Research in Microbiology* **168**, 865-874 (2017). https://doi.org/https://doi.org/10.1016/j.resmic.2017.02.001

96 Piot, P. *et al.* Biotypes of Gardnerella vaginalis. *Journal of Clinical Microbiology* **20**, 677-679 (1984). https://doi.org/10.1128/jcm.20.4.677-679.1984

97 Leopold, S. Heretofore undescribed organism isolated from the genitourinary system. *U S Armed Forces Med J* **4**, 263-266 (1953).

98 Swidsinski, A. *et al.* Adherent Biofilms in Bacterial Vaginosis. *Obstetrics & Gynecology* **106** (2005).

99 Alves, P., Castro, J., Sousa, C., Cereija, T. B. & Cerca, N. Gardnerella vaginalis Outcompetes 29 Other Bacterial Species Isolated From Patients With Bacterial Vaginosis, Using in an In Vitro Biofilm Formation Model. *The Journal of Infectious Diseases* **210**, 593-596 (2014). https://doi.org/10.1093/infdis/jiu131

100 Alves, J. M. B., G.A. (NCBI, 2010).

101 Horrocks, V. *et al.* Nuclear Magnetic Resonance Metabolomics of Symbioses between Bacterial Vaginosis-Associated Bacteria. *mSphere* **7** (2022). https://doi.org/10.1128/msphere.00166-22

102 Kline, K. A. & Lewis, A. L. Gram-Positive Uropathogens, Polymicrobial Urinary Tract Infection, and the Emerging Microbiota of the Urinary Tract. *Microbiology Spectrum* **4** (2016). https://doi.org/10.1128/microbiolspec.uti-0012-2012

103 Mitreva, M., Pepin,K.H., Mihindukulasuriya,K.A., Fulton,R.,Fronick,C., O'Laughlin,M., Miner,T., Herter,B., Rosa,B.A.,Cordes,M., Tomlinson,C., Wollam,A., Palsikar,V.B., Mardis,E.R. and Wilson,R.K. (2016).

104 Mitchell, A. L. *et al.* EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Research* **46**, D726-D735 (2018). https://doi.org/10.1093/nar/gkx967

105 Srinivasan, S. *et al.* More Easily Cultivated Than Identified: Classical Isolation With Molecular Identification of Vaginal Bacteria. *Journal of Infectious Diseases* **214**, S21-S28 (2016). https://doi.org/10.1093/infdis/jiw192

106 Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**, 540-546 (2019). https://doi.org/10.1038/s41587-019-0072-8

107 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**, 1043-1055 (2015). https://doi.org/10.1101/gr.186072.114

108 Aziz, R. K. *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**, 75 (2008). https://doi.org/10.1186/1471-2164-9-75

109 Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Research* **44**, D67-D72 (2016). https://doi.org/10.1093/nar/gkv1276

110 Weeks, R. M. *et al.* Cationic amphiphiles against Gardnerella vaginalis resistant strains and bacterial vaginosis-associated pathogens. *Pathog Dis* **77** (2019). https://doi.org/10.1093/femspd/ftz059

111 Schuyler, J. A., Mordechai, E., Adelson, M. E., Gygax, S. E. & Hilbert, D. W. Draft Genome Sequence of a Metronidazole-Resistant Derivative of Gardnerella

vaginalis Strain ATCC 14019. *Genome Announcements* **3**, e01345-01315-e01345 (2015). https://doi.org/10.1128/genomea.01345-15

112 Weeks, R. M. *et al.* Cationic amphiphiles against Gardnerella vaginalis resistant strains and bacterial vaginosis-associated pathogens. *Pathogens and Disease* **77** (2019). https://doi.org/10.1093/femspd/ftz059

113 Sichtig, H. *et al.* FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. *Nature Communications* **10** (2019). https://doi.org/10.1038/s41467-019-11306-6

114 Kurukulasuriya, S. P., Patterson, M. H. & Hill, J. E. Slipped-Strand Mispairing in the Gene Encoding Sialidase NanH3 in Gardnerella spp. *Infect Immun* **89** (2021). https://doi.org/10.1128/IAI.00583-20

115 Thomas-White, K. *et al.* Culturing of female bladder bacteria reveals an interconnected urogenital microbiota. *Nature Communications* **9** (2018). https://doi.org/10.1038/s41467-018-03968-5

116 Alessandri, G., van Sinderen, D. & Ventura, M. The genus bifidobacterium: From genomics to functionality of an important component of the mammalian gut microbiota running title: Bifidobacterial adaptation to and interaction with the host. *Comput Struct Biotechnol J* **19**, 1472-1487 (2021). https://doi.org/10.1016/j.csbj.2021.03.006

117 Zhang, K. *et al.* Antibiotic resistance and pathogenicity assessment of various Gardnerella sp. strains in local China. *Frontiers in Microbiology* **13** (2022). https://doi.org/10.3389/fmicb.2022.1009798

118 Iyer, L. M., Zhang, D., Rogozin, I. B. & Aravind, L. Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucleic Acids Research* **39**, 9473-9497 (2011). https://doi.org/10.1093/nar/gkr691

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

120 Barco, R. A. *et al.* A Genus Definition for Bacteria and Archaea Based on a Standard Genome Relatedness Index. *mBio* **11** (2020). https://doi.org/10.1128/mbio.02475-19

121 Lindsey, R. L. *et al.* Rapid identification of enteric bacteria from whole genome sequences using average nucleotide identity metrics. *Frontiers in Microbiology* **14** (2023). https://doi.org/10.3389/fmicb.2023.1225207

122 Gu, Z. Complex heatmap visualization. *iMeta* **1** (2022). https://doi.org/10.1002/imt2.43

123 Hahsler, M., Hornik, K. & Buchta, C. Getting Things in Order: An Introduction to the R Package seriation. *Journal of Statistical Software* **25**, 1 - 34 (2008). https://doi.org/10.18637/jss.v025.i03

124 Robinson, J. An Iterative Method of Solving a Game. *Annals of Mathematics* **54**, 296-301 (1951). https://doi.org/10.2307/1969530

125 Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P. & Göker, M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14**, 60 (2013). https://doi.org/10.1186/1471-2105-14-60

126 Wayne, L. G. *et al.* Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic and Evolutionary Microbiology* **37**, 463-464 (1987). https://doi.org/https://doi.org/10.1099/00207713-37-4-463

127 Bazin, A., Gautreau, G., Médigue, C., Vallenet, D. & Calteau, A. panRGP: a pangenome-based method to predict genomic islands and explore their diversity. *Bioinformatics* **36**, i651-i658 (2020). https://doi.org/10.1093/bioinformatics/btaa792

128 Bazin, A., Medigue, C., Vallenet, D. & Calteau, A. *panModule: detecting conserved modules in the variable regions of a pangenome graph* (Cold Spring Harbor Laboratory, 2021).

129 Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309-D314 (2019). https://doi.org/10.1093/nar/gky1085

130 Blum, M. *et al.* InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research* **53**, D444-D456 (2025). https://doi.org/10.1093/nar/gkae1082

131 Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research* **53**, D672-D677 (2025). https://doi.org/10.1093/nar/gkae909

132 Drula, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research* **50**, D571-D577 (2022). https://doi.org/10.1093/nar/gkab1045

133 Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412-D419 (2021). https://doi.org/10.1093/nar/gkaa913

134 Galperin, M. Y. *et al.* COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research* **49**, D274-D281 (2021). https://doi.org/10.1093/nar/gkaa1018