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

Heather Bouzek[* 1,2], Martha A. Zepeda-Rivera[* 1,2], Sujatha Srinivasan[2], Elliot M. Lee[2], Dakota S. Jones[2], Elsa F. McMahon[2], Susan M. Strenk[2], Tina L. Fiedler[2], Michael France[3,4], Jacques Ravel[3,4], 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

[3] Center for Advanced Microbiome Research and Innovation, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, USA

[4]Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, Maryland, USA

[#]**Address correspondence to**:

David N. Fredricks, MD, dfredric@fredhutch.org

Christopher D. Johnston, PhD, cdjohnston@mdanderson.org

[*] These authors contributed equally to this work.

Abstract

Bacterial vaginosis (BV) is a prevalent condition characterized by an imbalance in the vaginal microbiota, often involving the presence of *Gardnerella* species. The taxonomic complexity and inconsistent nomenclature of *Gardnerella* have impeded progress in understanding the role of this genus in health and disease. In this study, we obtained genomes from 392 *Gardnerella* isolates from asymptomatic individuals and those with BV. With stringent curation of these and additional publicly-available genomes, we conducted a comprehensive genomic analysis (final dataset: n=312). We resolve taxonomic ambiguities using current species thresholds and propose updated *Gardnerella* nomenclature to facilitate formal naming of these twenty-one identified lineages (eleven species, fifteen subspecies). Syntenic pangenome analysis revealed lineage-specific genetic factors and putative operons associated with virulence, metabolism, and genetic defense mechanisms. We also report on the first cryptic plasmids naturally present within the *Gardnerella* genus, which should serve to advance mechanistic studies within these species. Our findings provide a unified framework for *Gardnerella* taxonomy and nomenclature, and advance 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. 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

23 increased susceptibility to sexually transmitted infections[3] and adverse pregnancy outcomes such
24 as preterm birth and low birth weight[4,5].

25 While molecular methods have associated many bacterial species with BV[6-12], *Gardnerella*
26 *vaginalis* was the first bacterium isolated from individuals affected with BV[13] and was long
27 thought to be the sole causative agent of this condition[14]. This assumption was based on
28 *Gardnerella* species exhibiting a high virulence potential, including adherence to the vaginal
29 epithelium[15], production of cytotoxins[16], formation of biofilms[17], and its presence in 95–100% of
30 BV cases[18-20]. Moreover, *Gardnerella* species thrive in the vaginal environment, even at the low
31 pH typical of the *Lactobacillus*-dominant genital tract microbiota[21]. Although vaginal
32 concentrations of *Gardnerella* are greater in those with BV[18], *Gardnerella* species are also present
33 in asymptomatic individuals, suggesting the potential for both commensal and pathogenic lineages
34 or states. Our previous research revealed that while *Gardnerella vaginalis* is correlated with a
35 greater risk ratio for BV than other validly named *Gardnerella* species, no single *Gardnerella*
36 species is specific for BV based on a *cpn60* qPCR. In fact, the presence of a high *Gardnerella*
37 species diversity is a risk factor for the development of BV in our longitudinal study[18]. However,
38 the genetic capabilities of individual *Gardnerella* lineages remain unclear, partially due to
39 inconsistent taxonomic assignments and overly complex *Gardnerella* nomenclature[22], despite
40 laudable efforts from multiple research groups[23-31]. In this study, we used various comparative
41 genomic approaches to taxonomically identify *Gardnerella* lineages and propose a single naming
42 scheme for all currently known species that is reflective of phylogenetic placement and historical
43 context. We obtained genomes from 392 *Gardnerella* isolates from asymptomatic individuals and
44 those affected with BV. We stringently curated additional publicly available *Gardnerella* genomes
45 for a final genome set (n=312), which was analyzed via *cpn60* gene typing, average nucleotide

46 identity, digital DNA/DNA hybridization, and phylogenetic approaches. Our analyses identified
47 eleven distinguishable *Gardnerella* species, encompassing fifteen subspecies. Here, in addition to
48 proposing an updated nomenclature scheme for twenty-one distinct *Gardnerella* lineages, we
49 identify genomic features associated with each through pangenome analysis. Our work aims to
50 advance knowledge on the genetic characteristics and predicted functional capabilities of lineages
51 within the *Gardnerella* genus.
52

53 **Results**

54 ***Gardnerella* and *Bifidobacterium* are distinct genera**

55 We obtained genomes for 392 *Gardnerella* isolates (Fredricks and Ravel bacterial isolate
56 collections) from asymptomatic individuals and those with BV. We curated these and additional
57 publicly available *Gardnerella* genomes (n=387 within the NCBI database as of May 2025) using
58 stringent quality metrics to remove duplicate and low-quality genomes (**Methods,**
59 **Supplementary Table 1**). This resulted in a final curated set of 312 genomes for comparative
60 genomic analyses (**Table 1**, **Supplementary Table 2**). Analysis of the *cpn60* gene (which encodes
61 a 60 kDa chaperonin) currently provides the most reliable method for *Gardnerella* species typing.
62 These species cannot be distinguished by traditional 16S rRNA gene sequence typing[32,33]. BLAST
63 analysis of the *cpn60* gene of these isolates supported their identification as *Gardnerella* species
64 (**Supplementary Table 2**).

65 Despite the classification of these genomes as *Gardnerella* against and within the NCBI
66 database, initial speciation against the Genome Taxonomy Database (GTDB) assigned many of
67 these genomes to *Bifidobacterium* (**Supplementary Table 2**). However, *Bifidobacterium* is known
68 to be a polyphyletic group, consisting of at least six distinct clades[34,35], with its members exhibiting

high genetic diversity. The separation of *Gardnerella* from *Bifidobacterium* has previously been demonstrated through comparisons of genome length, G+C percentage, and overall genome composition.[28,36] Comparisons of genome length and G+C percentage between our *Gardnerella* genome set, *Bifidobacterium,* and closely related *Cardovia* (e.g., *Alloscardovia, Aeroscardovia*, etc.) genomes (n=227) indicate that *Gardnerella* genomes are significantly shorter ($p<0.001$, Wilcoxon *post hoc* test) and have lower G+C content ($p<0.001$, Wilcoxon *post hoc* test) than other *Bifidobacterium* members (**Supplementary Figure 1, Supplementary Table 3**). Additionally, average nucleotide identity (ANI) values fell below generally accepted genus thresholds of 73-75% (**Supplementary Figure 1**). These results confirmed that our final curated genome set included only *Gardnerella* species, further supporting previous evidence that *Gardnerella* and *Bifidobacterium* represent distinct genera within the *Bifidobacteriaceae* family.

Lineage	Abbreviation	No. Genomes	Genome Length	GC %
G. aggregata	*Ga*	n=86	1.37-1.72 Mbp	41.64%-42.50%
G. aggregata subsp. *aggregataa*	*Gaa*	n=5	1.48-1.53 Mbp	42.11%-42.50%
G. aggregata subsp. *hutchinsoni*	*Gah*	n=11	1.52-1.66 Mbp	42.03%-42.43%
G. aggregata subsp. *leopoldii*	*Gal*	n=24	1.37-1.67 Mbp	42.07%-42.34%
G. aggregata subsp. *swidsinskii*	*Gas*	n=46	1.52-1.72 Mbp	41.64%-42.29%
G. bivia	*Gb*	n=6	1.55-1.64 Mbp	37.90%-38.31%
G. coleohominis	*Gc*	n=6	1.64-1.72 Mbp	40.53%-40.76%
G. insolita	*Gi*	n=3	1.59-1.66 Mbp	41.27%-43.80%
G. insolita subsp. *insolita*	*Gii*	n=2	1.61-1.66 Mbp	41.27%-43.80%
G. insolita subsp. *libera*	*Gil*	n=1	1.59 Mbp	43.60%
G. kenyensis	*Gk*	n=1	1.66 Mbp	43.62%
G. media	*Gm*	n=26	1.49-1.65 Mbp	42.90%-43.50%
G. media subsp. *greenwoodii*	*Gmg*	n=13	1.49-1.65 Mbp	42.90%-43.44%
G. media subsp. *media*	*Gmm*	n=2	1.53-1.55 Mbp	43.04%-43.05%
G. media subsp. *pittsburghi*	*Gmp*	n=11	1.50-1.64 Mbp	43.05%-43.50%
G. nova	*Gn*	n=2	1.62-1.70 Mbp	45.35%-45.46%
G. rara	*Gr*	n=6	1.47-1.58 Mbp	43.01%-43.29%
G. sialicivorans	*Gs*	n=60	1.50-1.68 Mbp	42.00%-42.74%
G. sialicivorans subsp. *picketti*	*Gspk*	n=40	1.51-1.68 Mbp	42.00%-42.74%
G. sialicivorans subsp. *piotii*	*Gspt*	n=17	1.50-1.59 Mbp	42.22%-42.61%
G. sialicivorans subsp. *sialicivorans*	*Gss*	n=3	1.50-1.53 Mbp	42.39%-42.48%
G. vaginalis	*Gv*	n=111	1.57-1.78 Mbp	40.99%-41.82%
G. vaginalis subsp. *regina*	*Gvr*	n=15	1.62-1.72 Mbp	41.11%-41.42%
G. vaginalis subsp. *transversa*	*Gvt*	n=17	1.57-1.67 Mbp	41.25%-41.62%
G. vaginalis subsp. *vaginalis*	*Gvv*	n=79	1.58-1.78 Mbp	40.99%-41.82%
G. washingtoni	*Gw*	n=5	1.56-1.65 Mbp	42.89%-43.11%

Table 1: *Gardnerella* lineages referenced in this study

Table 1 shows the *Gardnerella* lineages identified and referenced in this study along with their proposed names and unique abbreviations. Summary statistics, including the number of genomes in this collection, the range of genome length and the G+C content are shown for each lineage.

Identification and proposed revised nomenclature for *Gardnerella* lineages

Given the importance of *Gardnerella* species diversity in BV[18], we wanted to identify the number of lineages present in our genome collection. Although different *Gardnerella* species have been identified through various approaches[27-30,37,38], the naming of these groups has been a

90 consistent challenge for the *Gardnerella* field despite extensive and laudable efforts from multiple
91 research groups (**Figure 1**). In summary, Piot *et al*.[23] initially identified *Gardnerella* biotypes
92 based on sugar hydrolysis phenotypes. When this approach proved insufficient to capture the
93 diversity of the genus, it was expanded to 17 biotypes by adding three fermentation targets[24]. The
94 use of early genomic methods, including restriction fragment length polymorphism (RFLP)[25] and
95 amplified ribosomal DNA restriction analysis (ARDRA)[26], identified distinct *Gardnerella*
96 genotypes. Whole-genome comparisons brought about another naming scheme seeking to
97 encompass all strains, where "A", "B", "C", and "D"[27] were proposed designations for a newer
98 condensed species assignment[22,33]. Nearly simultaneously, another method was proposed[39], using
99 "1", "2", "3", and "4"[21] for the same clades, which led to a disjointed hybrid designation (C/1, B/2,
100 A/4, and D/3). The species were then renamed as reassorted ecotypes[28], using yet another
101 nomenclature, one that incorporated the genomic context of the distinct types. After whole-genome
102 sequencing became widely available, three more competing standards based on ANI emerged,
103 describing thirteen genomic species (Gsp)[29], nine genomospecies (GS)[30], and nine different
104 genotypes (GGtypes)[36]. Formal naming of a limited subset of these species, including *G.*
105 *vaginalis*[13], *G. pickettii*[40], *G. piotii*[29], *G. leopoldii*[29], *G. swidsinskii*[29], and *G. greenwoodii*[40], has
106 further compounded this complex nomenclature (**Figure 1**). Despite the use of different
107 classification schemes, prior studies consistently identified two distinct *Gardnerella* subgenera
108 (Set A and Set B), each encompassing specific lineages[27].

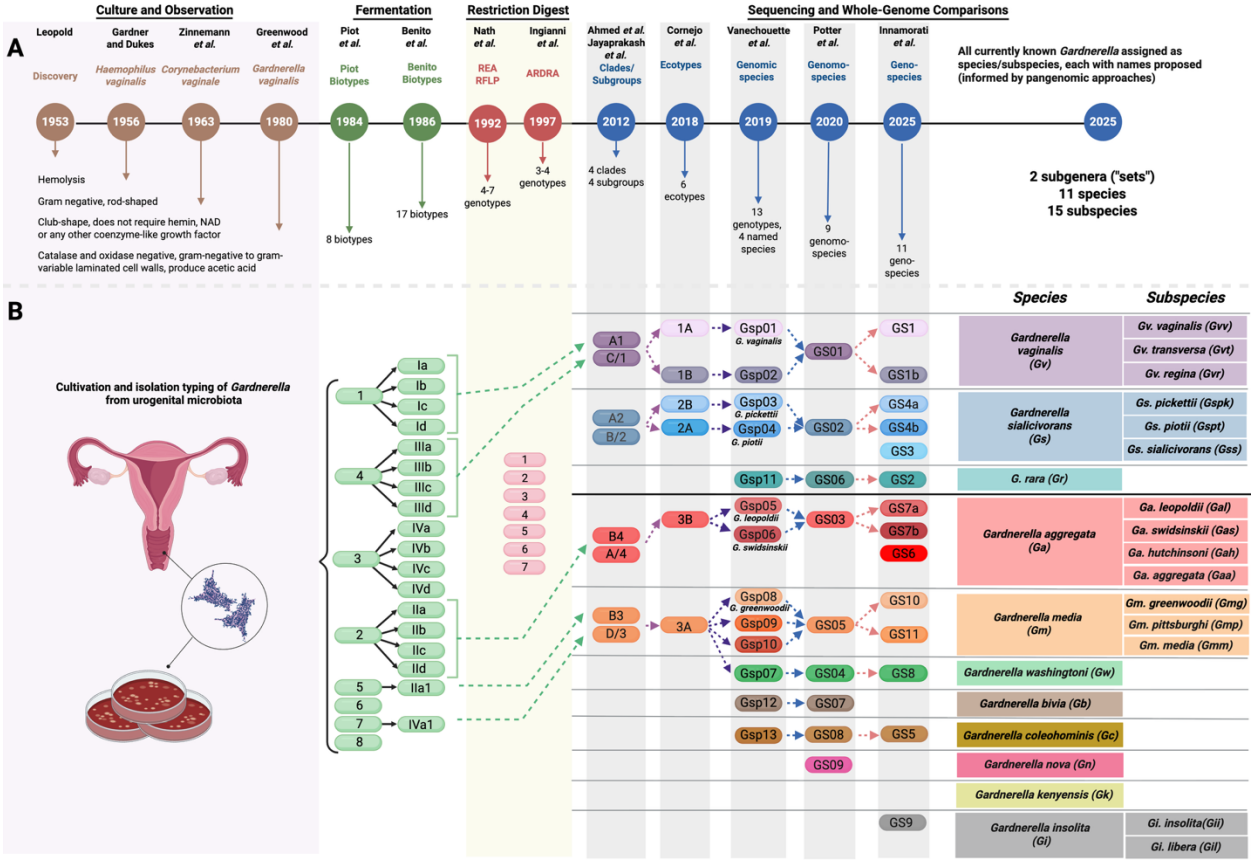
109 Consistent with this, *cpn60* gene sequences and reference-free whole-genome phylogenies
110 from our collection clearly separated all genomes into Set A or Set B (**Figure 2**, **Supplementary**
111 **Figure 2**). To determine the number of *Gardnerella* species represented, we compared the
112 genome-wide ANI values and identified a 94% threshold that delineated intra-species comparisons

113 (**Supplementary Figure 3**), slightly below the prokaryotic species ANI threshold of 95 ± 0.5 %
114 reported by Jain *et al.*[41]. ANI and digital DNA-DNA hybridization (dDDH) values were co-linear
115 (**Supplementary Figure 3**) and as such dDDH did not provide additional information. Using a
116 94% ANI threshold, we identified eleven *Gardnerella* species within our dataset. Whole-genome
117 phylogenies combined with ANI values further highlighted subspecies groups in five of these
118 species, resulting in twenty-one distinct *Gardnerella* lineages (**Figure 2, Supplementary Table**
119 **4**). These species designations were congruent with the classification proposed by Potter *et al.*[30].
120 Given the extensive history of conflicting classification schemes (**Figure 1**), we next sought to
121 standardize species and subspecies names across all lineages identified (**Figure 1**, **Figure 2**, **Table**
122 **1**, **Supplementary Table 2**).

123 Details on the rationale for each proposed nomenclatural revision, including species-
124 specific functional attributes and ecological isolation sources, are provided in the **Protologues**.
125 Our aim was to integrate previously used species and subspecies names to maintain a consistent
126 framework. Historically recognized "*G. vaginalis*" corresponds to a subspecies group that we now
127 propose to designate as *G. vaginalis* subspecies *vaginalis*, preserving historical continuity while
128 reflecting its phylogenetic placement (**Figure 1**, **Table 1**, **Supplementary Table 1**). Similarly, we
129 propose distinct species names for lineages with well-supported genomic and phenotypic
130 divergence, including *G. sialicivorans*, *G. aggregata*, and *G. media*, among others. To minimize
131 ambiguity in referring to *Gardnerella* species in broader studies, we developed guidelines for
132 assigning unique shorthand notation for each lineage (**Methods**). Species names are to be
133 abbreviated with two letters (*Ga*, *Gb*, *Gc*, *Gi*, *Gk, Gm*, *Gn, Gr*, *Gs*, *Gv* and *Gw*), while individual
134 subspecies should be designated with three-letter abbreviations (except for *Gs* subsp. *pickettii* and

135 *Gs* subsp. *piotii,* which require four letters, *Gspk* and *Gspt,* to distinguish them from other named

136 subspecies) (**Figure 1**, **Table 1**, **Methods**).



137

138 **Figure 1: Historical context of *Gardnerella* nomenclature**

139 **A**) From top to bottom schematic shows the identification method by type, publication first author,

140 names given to *Gardnerella* or naming framework, date of publication, and resulting conclusions

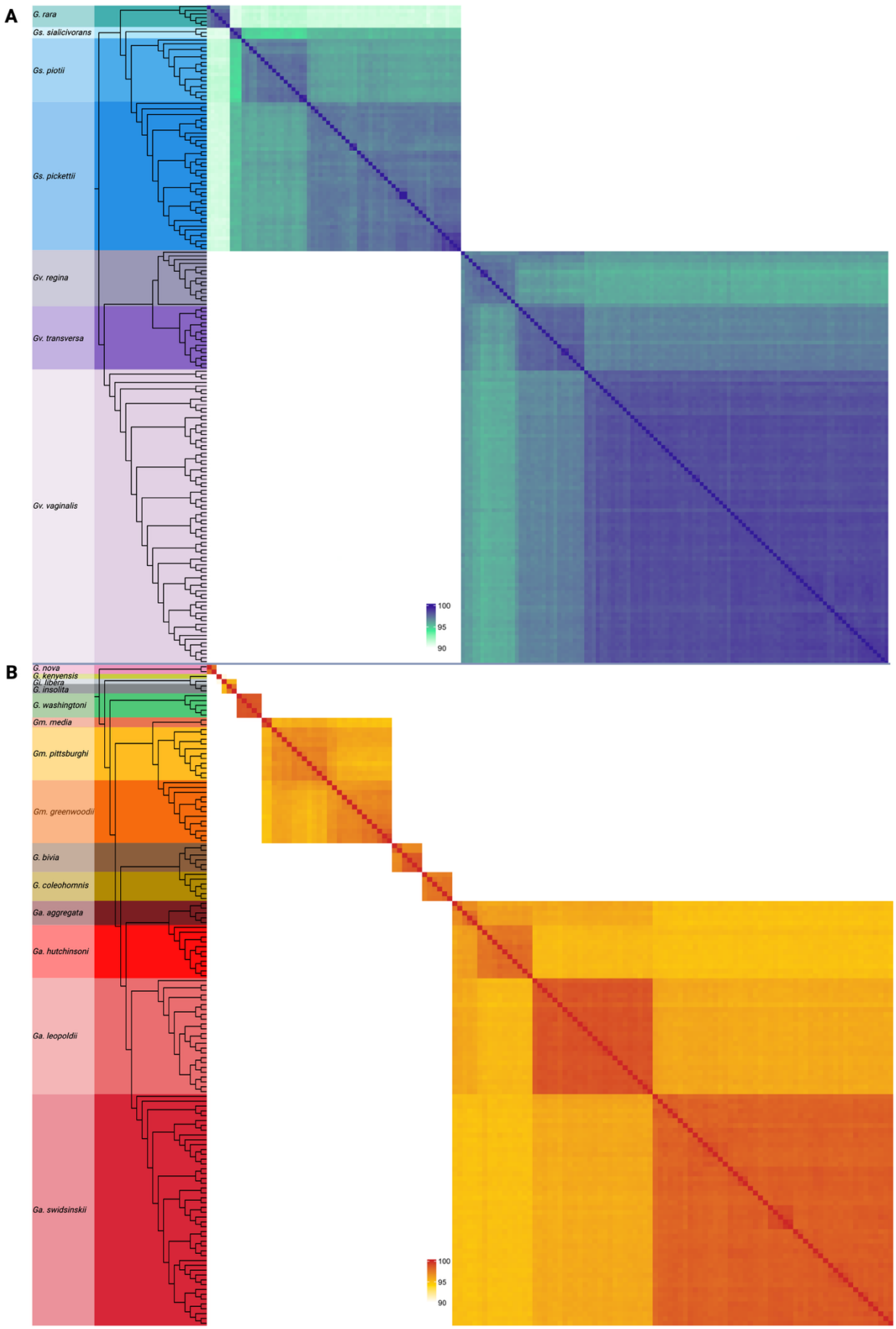
141 from naming schemes. **B**) *Gardnerella* subtyping resulting from the publication in (A) where each

142 column represents a subset of *Gardnerella* named. Lineages between different publications are

143 connected by arrows. The final two columns contain the updated lineages with corresponding

144 species and subspecies names proposed within this study.

145


A
G. rara
Gs. sialicivorans
Gs. piotii
Gs. pickettii
Gv. regina
Gv. transversa
Gv. vaginalis
100
95
90
B
G. nova
G. kenyensis
G. libera
G. insolita
G. washingtoni
Gm. media
Gm. pittsburghi
Gm. greenwoodii
G. bivia
G. coleohomnis
Ga. aggregata
Ga. hutchinsoni
Ga. leopoldii
Ga. swidsinskii
100
95
90

146

147 **Figure 2: Average nucleotide identity values between *Gardnerella* genomes**

148 Average nucleotide identity values between *Gardnerella* genomes in historical **A**) Set A and **B**) Set
149 B. Key for each panel indicates corresponding values to each color. White represents values below
150 90%. Genomes are organized based on a kSNP4[42]-generated whole-genome reference-free
151 phylogeny.

152

153 **A syntenic pangenome for *Gardnerella***

154 After stringently identifying the *Gardnerella* species and subspecies lineages in our
155 genome set and assigning each a standardized name, we were able to accurately investigate the
156 genetic features associated with each. Pangenome analysis provides a comprehensive view of the
157 genetic repertoire within a group of bacteria by identifying genetic features that are well-conserved
158 across most (>95%) genomes ("core genome") versus those that are only present in a subset of
159 genomes ("accessory genome"). While the core genome often contains genes involved in essential
160 cellular processes[43,44], analysis of the accessory genomes can provide key insights about genetic
161 diversity, adaptation, and evolutionary relationships. For example, previous pangenome efforts in
162 *Gardnerella* have expanded taxonomic groupings, identified metabolically essential genes, and
163 uncovered virulence and niche adaptation features[31,45,46]. However, bacterial genes often function
164 in co-regulated units or clusters involved in specific biological processes (e.g. operons). To identify
165 putative operons within the *Gardnerella* pangenome, we applied a syntenic approach that accounts
166 for gene order by using the Partitioned PanGenome Graph of Linked Neighbors (PPanGGOLiN)
167 tool[43] (**Figure 3**, **Figure 4, Supplementary Table 5**). Briefly, the coding sequences of genes are
168 grouped into "nodes" based on amino acid similarity and visualized as circular nodes, with node
169 size reflecting the relative presence of that gene group across the genome set (**Figure 4A-C**).

170 Nodes representing neighboring genes are linked into syntenic units called "modules"
171 (**Supplementary Figure 4**). This analysis clearly distinguished between the core ("persistent"
172 partition) and accessory genomes, and enabled identification of accessory genetic elements present
173 at intermediate frequency ("shell" partitions) or very low frequency ("cloud" partition) (**Figure 3**).
174 By leveraging modules and pangenome partitions, we can trace conserved arrangements of genes
175 across multiple genomes and simultaneously gain insight into the functional and evolutionary
176 dynamics driving diversity at the genus, species, and strain levels.

177 The core genome in our *Gardnerella* dataset reflected approximately 16% of identified
178 nodes, enriched in essential cellular functions including translation and ribosomal structure,
179 transcription, RNA processing, signal transduction, energy production, and metabolism of
180 nucleotides, amino acids, inorganic ions, and lipids (**Figure 3, Supplementary Figure 4**).
181 Periodically, these conserved nodes are interrupted by shell-associated nodes (**Figure 4A,** nodes
182 in distinct colors, branching from orange nodes), which mark points of evolutionary divergence.
183 These shell nodes often represent alternate evolutionary pathways, commonly associated with
184 horizontal gene transfer (HGT) events. We noted node differences between shell partitions that
185 were consistent with virulence, metabolism, and defense systems, warranting further investigation.

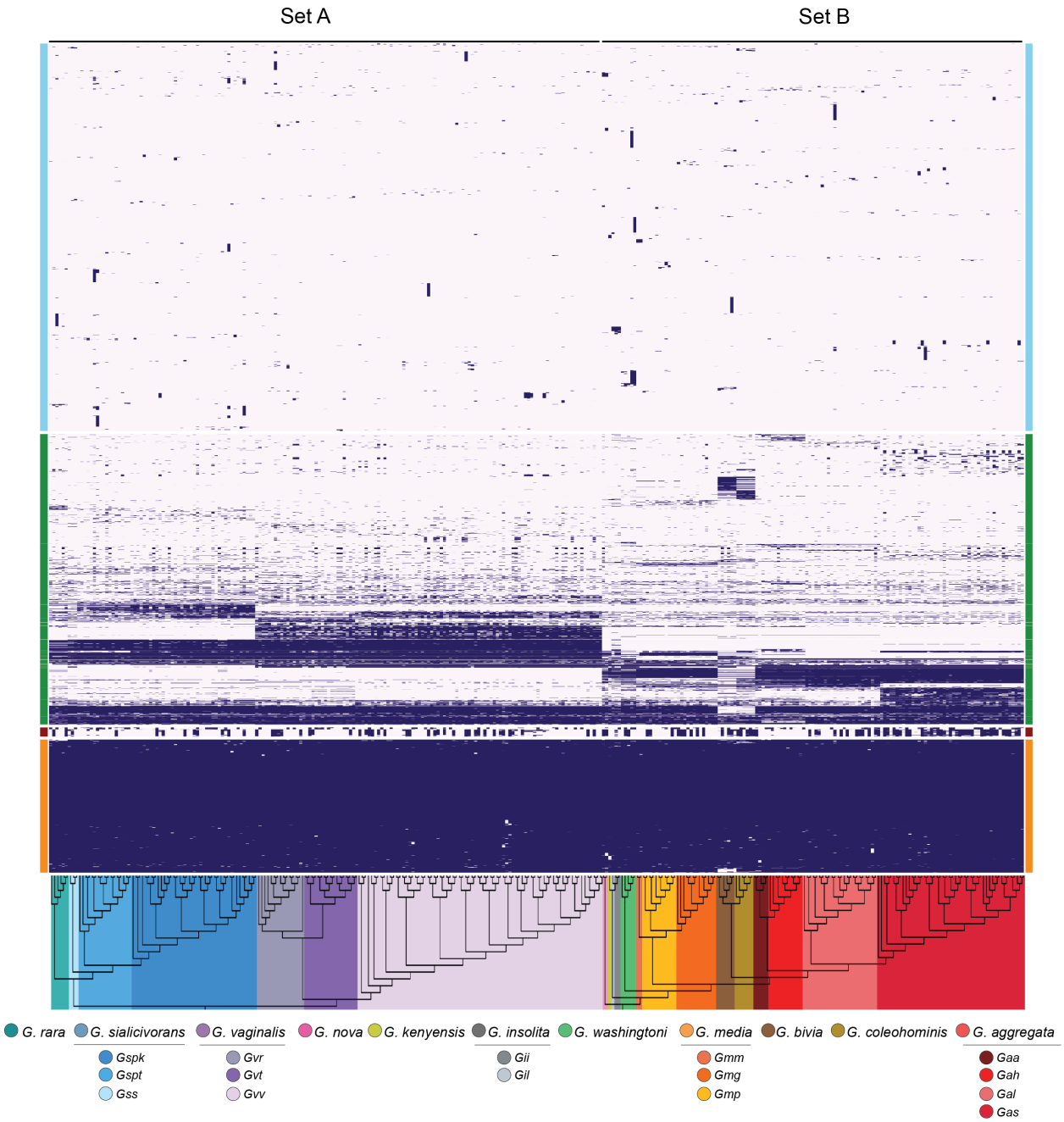
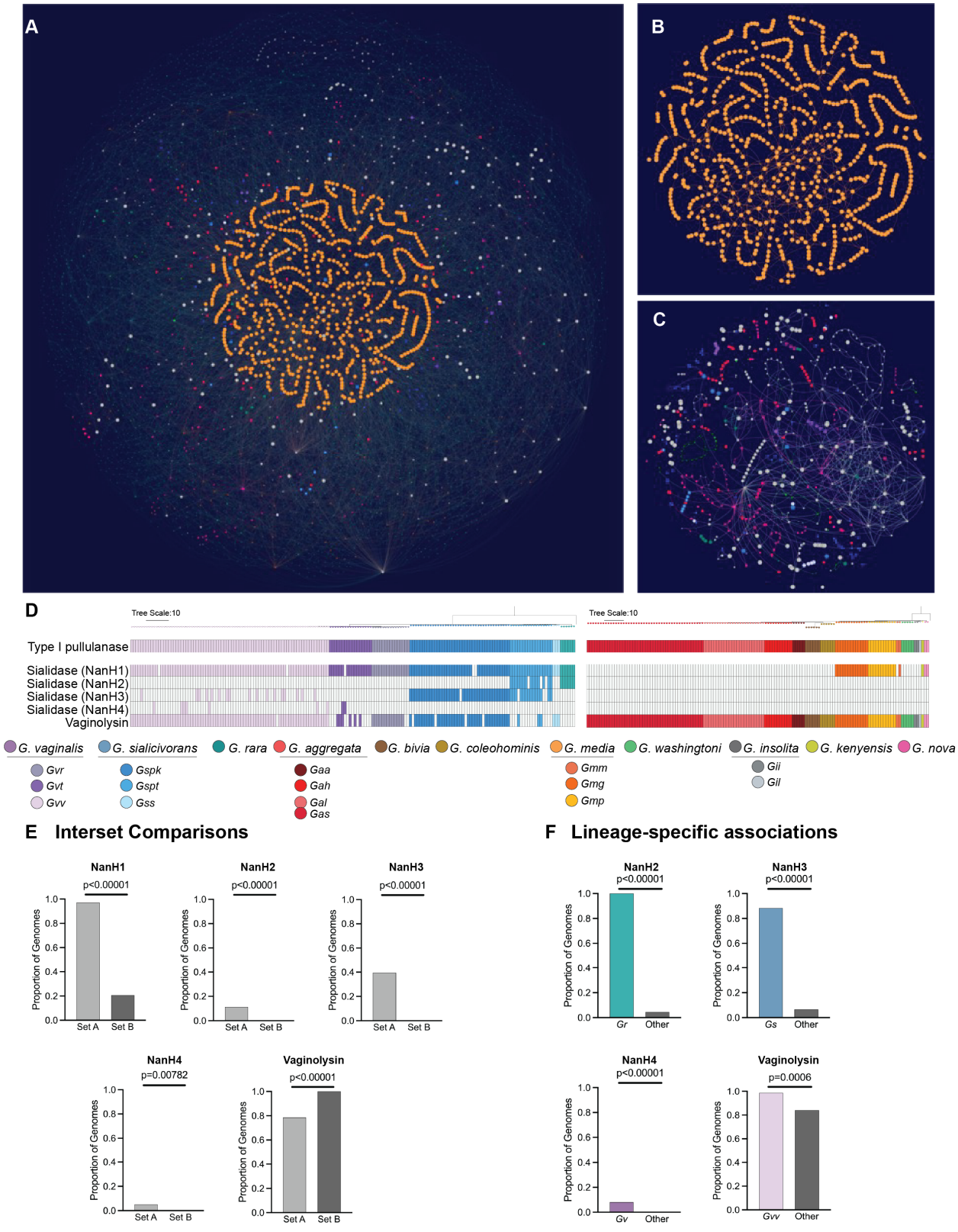


Figure 3: *Gardnerella* pangenome partitions

Heatmap displays the presence and absence of nodes in each partition where each row represents a node, and each column corresponds to a *Gardnerella* genome. Colors on the side indicate the core/persistent (orange), phage (red), shells (green), and cloud (blue) partitions. Genomes are ordered by kSNP4[42] phylogenies of Set A and Set B, with colors on dendrogram indicating lineage.


A
B
C
D
Tree Scale:10
Tree Scale:10
Type I pullulanase
Sialidase (NanH1)
Sialidase (NanH2)
Sialidase (NanH3)
Sialidase (NanH4)
Vaginolysin
G. vaginalis
Gvr
Gvt
Gvv
G. sialicivorans
Gspk
Gspt
Gss
G. rara
G. aggregata
Gaa
Gah
Gal
Gas
G. bivia
G. coleohominis
G. media
Gmm
Gmg
Gmp
G. washingtoni
G. insolita
Gii
Gil
G. kenyensis
G. nova
E Interset Comparisons
NanH1
p<0.00001
NanH2
p<0.00001
NanH3
p<0.00001
NanH4
p=0.00782
Vaginolysin
p<0.00001
Proportion of Genomes
Set A
Set B
F Lineage-specific associations
NanH2
p<0.00001
NanH3
p<0.00001
NanH4
p<0.00001
Vaginolysin
p=0.0006
Gr
Gs
Gv
Gvv
Other

192

193

194 **Figure 4: Syntenic maps of the *Gardnerella* pangenome**

195 **A-C**) Visualizations of the PPanGGOLiN[43]-generated *Gardnerella* pangenome. Each node
196 represents a gene group, with syntenic nodes representing neighboring genes. Size of node
197 indicates relative presence across *Gardnerella* genomes. Colors indicate core (orange) or shell
198 (other colors) partitions. Core and shell partitions are shown individually in **B**) and **C**),
199 respectively. **D**) Presence-absence heatmap of canonical *Gardnerella* virulence factors and the
200 well-conserved pullulanase across each *Gardnerella* lineage. Heatmap is organized by kSNP4[42]
201 phylogenies of Set A and Set B. Colors indicate lineage and presence/absence as denoted in key.
202 **E-F**) Graphs show the virulence factor presence compared by genome proportion between
203 *Gardnerella* **E**) Set A and Set B and **E**) specific lineages and all others. Statistical analysis was
204 carried out using a two-sample z-test, two-tailed.

205

206 **Virulence factors and predicted pathogenicity across *Gardnerella* lineages**

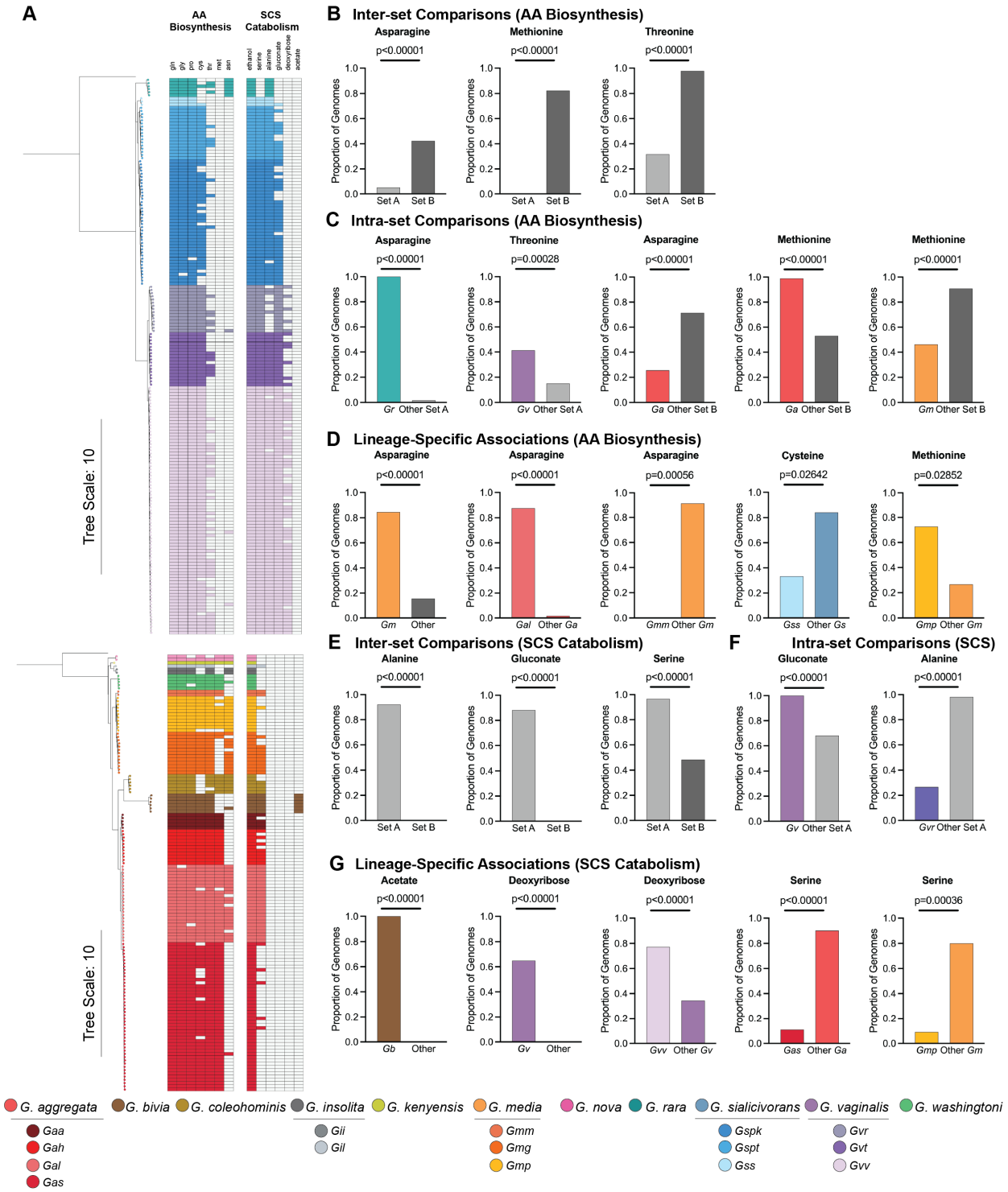
207 *Gardnerella* colonization of the vaginal niche, as well as its pathogenesis, is dependent on
208 mucin degradation, adherence to the vaginal epithelium, and the formation of stable biofilms[15].
209 Canonical *Gardnerella* virulence factors include four sialidase variants[47] (NanH1, NanH2, NanH3,
210 NanH4) that aid in mucin degradation, and vaginolysin[16,48], which binds and lyses host cells in the
211 vaginal mucosa. In our pangenome dataset we identified specific nodes associated with each of
212 these virulence factors (**Supplementary Table 5**).

213 The Type I pullulanase protein was well conserved (present in 100% of *Gardnerella*
214 genomes), in agreement with previous reports (**Figure 4D**). Comparisons between *Gardnerella* Set
215 A and Set B showed that the four sialidase proteins (NanH1-NanH4) are significantly enriched in
216 Set A genomes, while vaginolysin is significantly enriched in Set B genomes (**Figure 4E**).

217 Lineage-specific associations were also observed for three of the sialidase variants as well as
218 vaginolysin (NanH2-*Gr*, NanH3-*Gs*, NanH4-*Gv*, vaginolysin-*Gvv*) (**Figure 4F**). Our results
219 support the hypothesis that variable lineage-dependent pathogenic potential exists across
220 *Gardnerella*.
221
222 **Predicted metabolic capabilities across *Gardnerella* lineages**
223 Given the observed difference in metabolism-related nodes, we bioinformatically predicted
224 the biosynthesis of amino acids and the catabolism of small carbon sources across our *Gardnerella*
225 genome dataset (**Figure 5A, Supplementary Table 6**). The biosynthesis of glutamine, glycine,
226 and proline appeared to be well conserved (>99% of genomes) while a high proportion of genomes
227 (89%) harbored the capacity for cysteine biosynthesis (**Figure 5A**). Compared to Set A, genomes
228 across Set B were significantly enriched for asparagine, threonine, and methionine biosynthesis
229 (**Figure 5B**). Within Set B, asparagine biosynthesis was significantly depleted in *Ga*, threonine
230 biosynthesis was significantly depleted in *Gn,* and methionine biosynthesis was significantly
231 depleted in *Gm* and *Gb* and significantly enriched in *Ga* (**Figure 5C**). Although asparagine and
232 threonine biosynthesis were significantly enriched in Set B, we noted their species-specific
233 enrichments within Set A. Compared to all other Set A genomes, asparagine and threonine
234 biosynthesis were significantly enriched in *Gr* and *Gv*, respectively (**Figure 5C**). *Gardnerella*
235 species comparisons indicated that asparagine biosynthesis was significantly enriched in *Gm,*
236 cysteine biosynthesis was significantly depleted in *Gc* and *Gr*, and proline biosynthesis was
237 significantly depleted in *Gn* (**Figure 5D**). Lineage comparisons within species indicated that
238 asparagine biosynthesis was significantly enriched in *Gal*, methionine biosynthesis was

239 significantly enriched in *Gmp,* asparagine biosynthesis was significantly depleted in *Gmm*, and
240 cysteine biosynthesis was significantly depleted in *Gss* (**Figure 5D**).

241 Genes for the catabolism of ethanol appear to be a well-conserved feature of *Gardnerella*
242 genomes (**Figure 5A**). Genomes across Set A were further significantly enriched for the catabolism
243 of alanine, gluconate, and serine (**Figure 5E**). Notably, lineage-specific associations within Set A
244 were observed for the catabolism of each of these small carbon sources. While gluconate
245 catabolism is significantly enriched in *Gv*, alanine catabolism is significantly depleted in *Gvr*
246 (**Figure 5F**). Serine catabolism was completely absent in *Gr* and within Set B genomes was
247 significantly depleted in *Gmp* and *Gas* compared to other lineages within their respective species
248 (**Figure 5G**). Interestingly, the catabolism of acetate and deoxyribose appeared to be lineage-
249 specific and unique to *Gb* and *Gv*, respectively (**Figure 5G**). In the case of deoxyribose, within *Gv*
250 its catabolism was significantly enriched in subspecies *Gvv* (**Figure 5G**). Our results indicate that
251 Set A and Set B genomes are distinguished by their relative enrichment of pathways for catabolism
252 of small carbon sources or synthesis of distinct amino acids with lineage-specific associations.



253

254 **Figure 5: Predictions for metabolic capabilities across *Gardnerella* lineages**

255 **A**) Presence-absence heatmap for GapMind-predicted[49-51] pathways to synthesize amino acids

256 (AA) or catabolize small carbon sources (SCS) where each row represents a *Gardnerella* genome.

257 Genomes are ordered by kSNP4[42] phylogenies of Set A and Set B. Presence by proportion of
258 genomes between Set A and Set B (**B** and **E**), within Set A or Set B (**C** and **F**), and between specific
259 lineages (**D** and **G**) are shown. Statistical analysis was carried out using a two-sample z-test, two-
260 tailed. Colors represent each lineage.

261

262 **Phage and antiphage defense mechanisms across *Gardnerella* genomes**

263 In our pangenome data set, we noted that multiple nodes were related to phage-associated
264 proteins that encompassed two distinct *Siphoviridae* family *Gardnerella* prophage, including the
265 previously described vB_Gva_AB1[52,53] and a novel phage we have tentatively designated as
266 "phage_gamma" (**Supplementary Table 7**). Although vB_Gva_AB1 is sometimes reported as
267 several different prophages, the discrete synteny within our PPanGGOLiN pangenome suggests
268 that these variants represent a single phage (**Supplementary Table 5**). It is notable that some
269 strains appear to have remnant phage genes, although these appeared to be non-syntenic and
270 scattered throughout the genome. An arbitrary cutoff of greater than ten phage_gamma genes and
271 greater than twenty-five vB_Gva_AB1 genes was considered for phage presence (**Supplementary**
272 **Figure 5**). While both vB_Gva_AB1 and phage_gamma were identified in Set A and Set B
273 genomes, vB_Gva_AB1 was significantly enriched in Set B genomes (**Supplementary Figure 5**).

274 To counteract phages, bacteria have evolved an array of defense systems[54], with the most
275 widespread and well-characterized being restriction-modification (RM) and CRISPR-Cas
276 mechanisms[55]. Across our *Gardnerella* genome dataset we identified the presence of sixty different
277 systems, with an average of twelve defenses per genome (**Figure 6A-B**, **Supplementary Figure**
278 **6, Supplementary Table 8**). Of the defense systems with described mechanisms, dXTPase
279 systems[56] were well conserved (99% genomes). Type I RM systems were prevalent (56% of
280 genomes) and did not differ significantly between *Gardnerella* sets. Set A genomes were

281 significantly enriched for some Abi proteins (AbiD[57] and AbiL[58]), CBASS defenses[59,60], and DRT
282 mechanisms[61,62], while Set B showed enrichment for a larger number of defenses, including those
283 with Abi-mechanisms (AbiE[63], Gabija[64,65], Lamassu[64,66,67]), CRISPR-Cas (CRISPR arrays and Cas
284 Type IE proteins), and Type II RM defenses (**Supplementary Figure 7**). Across *Gardnerella*
285 lineages, *Gas* showed an enrichment for DNA-modification system (DMS) proteins, while *Gvv*
286 showed an enrichment for AbiL and DRT defenses (**Supplementary Figure 7**).

287 Although defense mechanisms evolved primarily to protect against bacteriophages, their
288 activity extends against other horizontally transferred genetic elements, as well as plasmid vectors
289 used during *in vitro* transformation. RM defenses, which are dependent upon detection and
290 destruction of inappropriately methylated foreign DNA, are one of the most prevalent across
291 bacteria. Further, it is notable that approaches to overcome these RM systems during genetic
292 engineering have been shown to significantly improve transformation efficiency across multiple
293 bacterial species[68-71]. Thus, we sought to characterize RM defenses across a subset of our
294 *Gardnerella* genome dataset. We obtained and analyzed methylomes for 42 *Gardnerella* isolates
295 across thirteen lineages within the Fredricks isolate collection (**Supplementary Table 2**). A total
296 of 39 methylated motifs were detected and assigned to Type I (25 motifs), Type II (11 motifs), or
297 Type III (3 motifs) RM systems (**Figure 6C**, **Supplementary Table 9**).

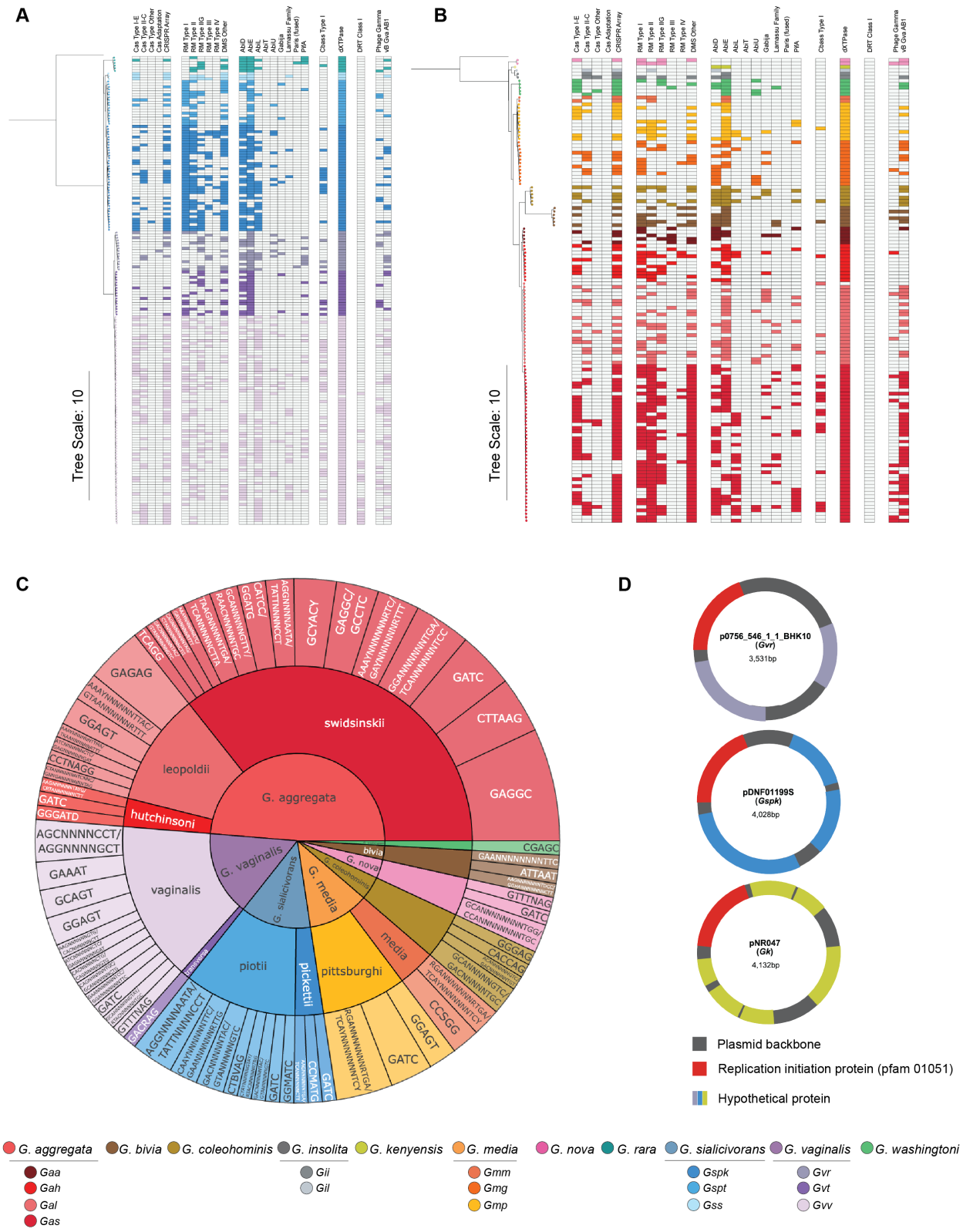


Figure 6: Defense systems and cryptic plasmids across *Gardnerella* genomes

300 **A-B**) Presence-absence heatmaps for PADLOC-predicted[66,72] defense mechanisms. Genomes are
301 ordered by kSNP4[42] phylogenies of **A**) Set A and **B**) Set B. **C**) Sunburst plot shows the sequence
302 of each methyl-modified motif detected within each *Gardnerella* lineage. The size of the section
303 represents the prevalence within each lineage. **D**) Schematics depict plasmid annotations for
304 pDNF1199S-01, p0756_546_1_1_BHK10-01, and pNR047-01.
305
306 ***Identification of a native Gardnerella cryptic plasmid***
307 There is renewed interest in developing genetic tools for engineering *Gardnerell*a
308 species[73,74]. Because *Gardnerella* were long thought to lack mobile genetic elements, such as
309 native plasmids, most efforts have focused on adapting non-*Gardnerella* genetic tools for use in
310 this genus. Recent studies have been extensive and show promising results, achieving targeted
311 chromosomal integrations and deletions using *E. coli*-based suicide vectors[73], as well as
312 reproducible transformation with a *Bifidobacterium-E.coli* shuttle vector, albeit with low
313 efficiency[74]. During genome sequencing of the Fredricks clinical isolate collection, we observed
314 an independent circular contig, with high coverage, forming a second replicon within the otherwise
315 closed genome sequence of *Gspk* strain DNF01199S (**Figure 6D, Supplementary Table 10**).
316 Annotation of this contig showed three coding sequences for hypothetical proteins and a single
317 putative replication initiation protein (**Supplementary Table 11**). Subsequent plasmid isolation
318 and validation confirmed the episomal nature of this contig, marking pDNF01199S-01 as the first
319 cryptic plasmid identified within the *Gardnerella* genus.
320 Motivated by this discovery, we assessed whether this or additional cryptic plasmids were
321 present in our *Gardnerella* pangenome. We reasoned that the current limited knowledge of such
322 plasmids reflects their rarity. As such, we focused on the cloud partition of the accessory genome,

323 which highlights strain-specific genetic factors and collectively comprises 47.23% of the
324 *Gardnerella* pangenome. Using structural sequence insights from pDNF01199S-01, we identified
325 two additional putative plasmid sequences within the cloud partition: one in *Gvv* strain
326 0756_546_1_1_BHK10[75] (p0756_546_1_1_BHK10-01), which was isolated from cervicovaginal
327 mucus in South Africa, and another in *Gk* strain NR047 (pNR047-01)[76], isolated from a vaginal
328 swab in Nairobi, Kenya. All three plasmids are approximately 4 kb in size, contain coding
329 sequences for putative hypothetical proteins, and share a highly similar replication initiation
330 protein (84.48%-95.26% similar), suggesting a potentially conserved replication mechanism for
331 cryptic plasmids within the *Gardnerella* genus (**Figure 6D**, **Supplementary Tables 10-11**).

332 Given that pDNF01199S-01 stably replicates, we next sought to engineer a *Gardnerella-*
333 *E.coli* shuttle vector that could be used for the transformation of other *Gardnerella* isolates. To
334 establish an antibiotic selectable marker, we noted that within our pangenome dataset, 22% of
335 genomes contained the presence of *tetM*, including *Gspk* DNF01199S. This was consistent with
336 previous reports on the variable tetracycline resistance phenotypes across *Gardnerella* strains[77].
337 Thus, we assessed the ability of both *tetM*+ and *tetM*- isolates to grow in the presence of
338 supplemented tetracycline. As predicted, *tetM*+ isolates were resistant to tetracycline while *tetM*-
339 isolates showed susceptibility, indicating that the *tetM* cassette confers tetracycline resistance *in*
340 *vitro*, and could serve as a selectable marker for this vector. To enable the propagation of our
341 *Gardnerella-E.coli* shuttle vector in *E. coli* strains, we combined pDNF01199S-01, *tetM* cloned
342 from DNF01199S, and the *E. coli* minicircle plasmid backbone (pMC) into a single parental
343 plasmid (herein designated pG01-pp) (**Supplementary Figure 8A-C**). As previously reported[68],
344 we induced and purified the pDNF01199S-tetM-MC minicircle vector (pG01-MC). We
345 subsequently attempted to transform this minicircle into the *tetM*- strain, *Gvv* ATCC 14018

346 (**Supplementary Figure 8D**). While this yielded no transformants, we predicted that overcoming
347 RM defenses present in these strains would increase transformation efficiency. A previously
348 reported RM system modifying "GGCC" sites has been described in ATCC 14018[78], and bypassing
349 this system was shown to improve transformation efficiency with suicide-based vectors[73]. Thus,
350 we used commercially available *HaeIII* methyltransferase (MTase) to methyl-modify these sites
351 in pG01-MC prior to transformation. In combination with further optimization of growth and
352 transformation conditions, transformation efficiency was substantially increased through these
353 approaches from a minimum of 2.30×10^0 transformants per µg of DNA to 6.85×10^2
354 (**Supplementary Figure 8D**). Collectively, these results show that pG01 is the first replicative
355 shuttle vector based upon a native *Gardnerella* cryptic plasmid. Together with recent work by the
356 Dillard and Hill groups[73,74], these efforts should establish a robust platform for more mechanistic
357 investigations of *Gardnerella* biology and pathogenesis.
358
359
360

361 **Discussion**

362 Taxonomic inconsistencies and fluctuating nomenclature have made it inherently
363 challenging to delineate the roles of *Gardnerella* lineages in health and disease states. Our work
364 seeks to resolve these ambiguities by integrating whole-genome analyses to produce a more refined
365 *Gardnerella* taxonomic classification system. We further propose an updated and unified
366 *Gardnerella* nomenclature that is reflective of these lineages. Proposing nomenclatural changes,
367 particularly for well-established organisms such as '*G. vaginalis',* (updated here to *G. vaginalis*
368 subsp. *vaginalis*), can often be contentious and may face resistance within the scientific
369 community. However, updating *Gardnerella* lineages to appropriate species and subspecies
370 designations, using established criteria, is necessary to prevent further taxonomic confusion. We
371 noted that while the *cpn60* gene, often used for *Gardnerella* typing, can largely distinguish between
372 *Gardnerella* species, its resolution is insufficient for subspecies assignments. Accurate subspecies
373 assignments require whole-genome approaches through either phylogenetic assessment or average
374 nucleotide identity comparisons. When additional *Gardnerella* lineages are isolated, whole-
375 genome sequencing efforts should allow for continuous integration into this updated taxonomic
376 and nomenclature scheme. Ultimately, our naming efforts seek to increase clarity and ensure that
377 the functional role of specific species in the genital tract can be accurately interpreted, while
378 preserving previous naming efforts, where possible.

379 With this refined taxonomic and nomenclature framework, we probed for genetic features
380 and putative operons associated with each lineage by applying a syntenic pangenome approach.
381 We noted that the vast majority of *Gardnerella* gene content (~84%) is species and strain-variable
382 with many differences associated with virulence, metabolism, and genetic defense systems.
383 *Gardnerella* Set A genomes (*Gr, Gs,* and *Gv*) showed an enrichment for sialidases and an increased

potential for the catabolism of small carbon sources. In contrast, Set B genomes (*Ga, Gb, Gc, Gi, Gk, Gm, Gn,* and *Gw*) were enriched in putative adhesion and virulence proteins, displayed a greater capacity for amino acid biosynthesis, had a higher propensity to harbor the vG_gva_AB1 prophage, and contained a broader repertoire of genetic defense mechanisms.

Mechanistic studies to assess the roles of *Gardnerella* genetic factors, including those identified in this study, are currently hindered by limited genetic tools available for *Gardnerella*. In fact, it has previously been reported that *Gardnerella* species lack mobile genetic elements, such as plasmids[27]. However, our extensive genome sequencing efforts reveal the first *Gardnerella* cryptic plasmid naturally occurring within a single cultivated strain, along with additional homologous sequences in publicly available genomes. Coupled with recent advances in *Gardnerella* species engineering, this discovery of extrachromosomal and self-replicative genetic elements within the *Gardnerella* genus should facilitate the development of advanced genetic tools. Collectively, overcoming genetic intractability within *Gardnerella* species will greatly advance our ability to investigate the role of the genus in health and disease states.

Acknowledgements

Research reported in this publication was supported in part by the National Human Genome Research Institute as part of the Human Microbiome Project Initiative focused on technology development under award number R01HG005816 to D.N.F., the National Institute for Allergy and Infectious Diseases of the National Institutes of Health under award number U19AI084044 to J.R., the National Institute of Dental and Craniofacial Research (NIDCR) of the National Institutes of Health Office of the Director (OD) Common Fund under award number R01DE027850 to C.D.J.

406 and the Gates Foundation under award INV010470 to J.R. In addition, funds were provided by
407 Fred Hutchinson Cancer Center (to C.D.J.).

408 Materials and Methods

409 *Collection, isolation, sequencing, and initial typing of Gardnerella species from the human*
410 *vaginal microbiota*

411 *Gardnerella* strains subjected to whole-genome sequencing were cultivated either from the
412 Fredricks collection, harboring isolates from human study participants with and without BV in
413 Seattle, Washington, USA[18,79] or from the Ravel isolate collection, harboring strains from multiple
414 geographic locations, including USA and Africa (Supplementary Table 2). Bacterial isolates
415 consistent with *Gardnerella* morphotypes on Gram stain were subjected 16S rRNA gene amplicon
416 sequencing to confirm phylogenetic placement in the *Gardnerella* genus and *cpn60* gene PCR with
417 amplicon sequencing to further specify these bacteria[18]. Phylogenetically diverse *Gardnerella*
418 species were selected for whole-genome sequencing. High molecular weight DNA was extracted
419 from each isolate using the MasterPure DNA purification kit (Lucigen), with two
420 phenol/chloroform cleanups prior to DNA precipitation where needed. DNA extraction was
421 quantified on a TapeStation 2200 instrument run with a Genomic DNA tape (Agilent).

422 *Genome sequencing and assembly*

423 *Sequencing on the Pacific Biosciences SeqI/II Platforms*

424 A total of forty-two strains from the Fredricks collection, and seventeen from the Ravel collection,
425 were selected for single-molecule real-time (SMRT) long-read sequencing using either the PacBio
426 SeqI or SeqII instrument (Pacific Biosciences) at the Fred Hutchinson Cancer Center or the

427 Genomic Resource Center of the University Maryland School of Medicine, respectively.
428 Sequencing libraries were prepared using the SMRTBell Template Prep Kit 1.0 or 2.0 and were
429 size-selected using a BluePippin (Sage Science) with a cutoff of 3-5 kb. Sequencing reads were
430 processed via Microbial Assembly software within the Pacific Biosciences' SMRTAnalysis
431 pipeline (v.9.0.0.92188).

432 *Sequencing on the Illumina HiSeq 4000 Platform*

433 A total of 332 isolates were selected for Illumina sequencing on the HiSeq4000 platform. Short
434 read libraries were prepared using the KAPA HyperPlus Kit (Kapa Biosystems) with KAPA
435 Single-Indexed Adapter Kit Set B. A fixed volume (35 µL) of genomic DNA was used as input,
436 and libraries were prepared following the manufacturer's protocol with modifications based on
437 their amount of input DNA. For samples with 0.5 or 0.2 ng input DNA, the fragmentation enzyme
438 was diluted 1:2 or 1:5 with water. All samples were fragmented at 37°C for 5 min. Adapter
439 concentrations varied according to the input, and the adapter ligation was carried out overnight at
440 4°C for all samples. The post-ligation cleanup was performed with 0.8x Ampure XP beads
441 (Beckman Coulter, Indianapolis IN) and 20 µL of sample was used in library amplification. Post-
442 amplification cleanup was performed with 1x Ampure XP beads; libraries with remaining adapter
443 dimer peaks were cleaned a second time. The final elution was in 25 µL of nuclease-free water.
444 Libraries were run on a TapeStation instrument with a D1000 tape (Agilent) to assess quality and
445 concentration. Libraries were sequenced (8 libraries/lane) on an Illumina HiSeq 4000 instrument
446 using the 150 bp paired-end protocol. Reads were assembled using Unicycler.

447

448 ***Publicly available genomes from the National Center for Biotechnology Information (NCBI)***

449 We used publicly available genomes both to provide a direct reference for previous studies and to

450 supplement our own sequencing efforts. We searched GenBank[80] assembly databases for the genus
451 *Gardnerella* and family *Bifidobacteraceae* (without genus assignment). Using the Type (Strain)
452 Genome Server (TYGS)[81] to confirm the genus of these genomes, we discovered that several of
453 the unassigned *Bifidobacteraceae* bacteria in NCBI were *Gardnerella* and included them in our
454 analysis. *Bifidobacteraceae* (without genus assignment) were included if they were collected as
455 part of the same NCBI BioProject which added 20 *Gardnerella* that were not typed further than
456 the *Bifidobacteraceae* family. Additionally, we included three sequences (350_GVAG,
457 476_GVAG, and 842_GVAG) that were initially assigned to *Gardnerella* but were later renamed
458 to *Alloscardovia* to verify this placement. The initial number of potential genomes was n = 779.
459 To verify the taxonomy of these strains we used the TYGS type strain server[81] to validate the genus
460 and removed genomes that were assigned outside of *Gardnerella*. Draft genomes with more than
461 250 contigs were removed due to inherent unreliability and genomes that were modified in the
462 laboratory were also removed.
463 *Bifidobacterium* strains were collected and curated in much the same manner as the
464 *Gardnerella* strains. A total of 98 strains from all known species of the *Bifidobacterium* genus and
465 the closely related *Cardovia* genera—*Alloscardovia, Aeroscardovia*, *Bombiscardovia*,
466 *Pseudoscardovia*, and *Scardovia*—were downloaded from NCBI. Whenever possible, the type
467 strains were collected. Sequences were assessed using CheckM, TYGS, and ANI via cutoffs
468 similar to those used for the *Gardnerella* strains. *Bifidobacteriaceae* groups were assigned
469 according to Alessandri *et al*.[82].
470
471 ***Genome quality assessments***

472 The completeness and contamination of all genomes (sequenced or available on NCBI) were
473 assessed via CheckM[83]. The genomes were annotated via RAST[84]. Genomes with contamination
474 greater than 1% or completeness less than 90% by CheckM were excluded. Genome sequences
475 were marked as duplicates if they shared an average nucleotide identity (ANI) greater than 99.5%
476 with other genomes in the collection (**Supplementary Table 1**). Notes regarding criteria for which
477 a specific representative was chosen can be found in Supplementary Table 1.
478
479 ***Average nucleotide identity and digital DNA-DNA hybridization comparisons***
480 Average nucleotide identity (ANI) metrics were obtained for Gardnerella Set A and Set B
481 genomes, separately, using fastANI[41] and visualized using the R packages ComplexHeatmap[85] and
482 "seriation"[86]. Digital DNA-DNA hybridization (dDDH) [87] values were calculated via the Genome-
483 to-Genome Distance Calculator[81] and visualized in R.
484
485 ***kSNP4 whole-genome phylogeny***
486 kSNP4[42] was used to construct a reference-free phylogenetic tree for *Gardnerella* Set A (kmer=17,
487 FCK=0.167) and *Gardnerella* Set B (kmer=17, FCK=0.797) genomes, separately. *Bombiscardovia*
488 *bombi* strain DSM19703 was included to root both dendrograms. The resulting parsimony trees
489 were visualized iTOL v7[88], with GapMind[49-51] metabolism predictions and PADLOC [66,72] defense
490 mechanism predictions annotated where appropriate.
491
492 ***cpn60 whole-genome phylogeny***

493 RAST-annotated nucleotide sequences for *cpn60* were MUSCLE-aligned and used to generate a
494 maximum-likelihood dendrogram using MEGA v11[89]. The resulting tree was visualized and
495 annotated using iTOL v7[88].
496
497 ***PPanGGOLiN Gardnerella pangenome***
498 We used PPanGGOLiN[43] on our curated genome set to graph and cluster the genes from all strains,
499 find regions of genomic plasticity and synteny[90], and separate the pangenome gene clusters into
500 core, shell, and cloud partitions, with clustering set to 50% coverage and 50% identity. EggNOG[91],
501 InterPro[92] and KEGG Mapper[93] were run separately on each genome to gather and confirm the
502 CAZyme[94], Pfam[95], and KEGG designations of the gene clusters. COG categories[96] were assigned
503 on the basis of the consensus of these designations. We added the annotation "Unclassified
504 Metabolism" to capture gene clusters that are clearly metabolism related, but the metabolite is not
505 defined or could be of multiple metabolic types.
506
507 ***Isolation of pDNF1199S-01 plasmid***
508 *Gvr* strain DNF1199S was grown in New York City III (NYCIII) liquid media in an AS-580
509 anaerobic chamber (Anaerobe Systems, Morgan Hill, CA) in a gas mixture of 5% CO_2, 5% H_2,
510 and 90% N_2 at 37°C. A QIAamp DNA Miniprep kit (Qiagen, Hilden, Germany) was used to isolate
511 pDNF1199S-01 from a 5mL culture of the bacteria. Isolated DNA was digested with *XhoI* (New
512 England Biolabs, Ipswich, MA) and visualized by gel electrophoresis to confirm plasmid presence.
513
514 ***Gardnerella* tetracycline resistance testing**

515 *Gardnerella* strains were grown anaerobically in an AS-580 anaerobic chamber (Anaerobe
516 Systems, Morgan Hill, CA) in a gas mixture of 5% CO_2, 5% H_2, and 90% N_2 at 37°C. Bacteria
517 were recovered from frozen stocks on NYCIII agar plates and sub-cultured twice in 2mL of
518 NYCIII liquid media. Culture purity was verified by Gram stain. The final cultures were
519 normalized to 0.5 McFarland turbidity using fresh NYCIII broth and 25µL was spotted onto an
520 NYCIII agar plate containing 16µg/mL tetracycline and an antibiotic-free NYCIII agar plate. 25µL
521 of *L. reuteri* CF48-3A was spotted onto each tetracycline plate as a tetracycline-resistant control[97].
522 Tetracycline resistance was determined according to Clinical and Laboratory Standards Institute
523 antimicrobial susceptibility testing protocols and any growth on tetracycline NYCIII agar plates
524 was considered as resistant to tetracycline.
525
526 ***pMC-pDNF1199S-tetM construction and minicircle preparation***
527 pDNF1199S-01 and *tetM* were both cloned from gDNA isolated from strain DNF1199S. Gene
528 fragments were assembled with pMC by Gibson assembly[98]. The parental construct was
529 transformed into *E. coli* ZYCY10P3S2T (System Biosciences, Palo Alto, CA) by heat shocking
530 according to the SBI Minicircle DNA Technology user manual. Transformant colonies were
531 selected by growing on Luria-Burtani (LB) agar plates containing 50 µg/mL kanamycin. A
532 successfully transformed colony was chosen and stocked, then induction of minicircles was
533 performed according to the SBI Minicircle DNA Technology user manual by growing the bacteria
534 in induction media containing arabinose. pDNF1199S-*tetM* minicircle DNA was then purified
535 from the induced *E. coli* using a Qiagen Plasmid Midi Kit (Qiagen, Hilden, Germany). The
536 sequence of purified plasmid was confirmed against the construct sequence by whole-plasmid
537 sequencing performed by Plasmidsaurus (Plasmidsaurus, Eugene, OR). "GGCC" sites on the

538 plasmid were protected by incubating 1µg of the DNA with 1µL of *HaeIII* MTase (New England
539 Biolabs, Ipswich, MA) at 37°C for 4hrs in the presence of 160µM S-Adenosyl methionine (New
540 England Biolabs, Ipswich, MA). Following enzyme treatment, DNA was purified with Monarch
541 PCR & DNA Cleanup Kit (New England Biolabs, Ipswich, MA).
542
543 **Preparation of competent *Gardnerella* cells**
544 *Gvv* type strain ATCC 14018 was grown in an AS-580 anaerobic chamber (Anaerobe Systems,
545 Morgan Hill, CA) in a gas mixture of 5% CO_2, 5% H_2, and 90% N_2 at 37°C. Cells were recovered
546 from frozen stocks on NYCIII agar plates and subcultured into NYCIII liquid media. 500µL of
547 culture was subcultured into 5mL NYCIII media and grown overnight. 4mL of this culture was
548 inoculated into 40mL of modified de Man-Rogosa-Sharpe media[99] supplemented with 5g/L
549 glucose, 0.4g/L L-cysteine HCl, 100mL/L filter-sterilized horse serum, and 100 mM NaCl when
550 stated. Once the target OD_{600} was reached, cell were harvested by centrifugation at 4°C at 4,052x*g*
551 for 10min. Cells were washed in one of the following three ways: 1) twice in 25mL ice-cold
552 maltose citrate buffer (0.2625g/L citric acid and 0.5M maltose at pH 5.8), 2) twice in 25mL ice-
553 cold autoclaved 10% glycerol, or 3) once in 25mL ice-cold autoclaved water from a Milli-Q
554 IQ7000 pure water system (Millipore Sigma, St. Louis, MO, USA) followed by a second wash in
555 25mL ice-cold maltose citrate buffer. Washed cells were then resuspended in 200µL of maltose
556 citrate buffer, mixed with 200µL 80% glycerol, and stored at -80°C in aliquots.
557
558 ***Gardnerella electrotransformation and verification***
559 Competent cells were mixed with plasmid DNA on ice and transferred into pre-chilled 0.1cm gap
560 Gene Pulser/MicroPulser Electroporation Cuvettes (Bio-Rad, Hercules, CA). Electroporation was

561 then performed on a Gene Pulser Xcell (Bio-Rad, Hercules, CA) at 25µF, 200Ω, and 2,000V.
562 Immediately following electroporation, 950µL of pre-warmed NYCIII broth was added to the cells
563 and they were transferred to an AS-580 anaerobic chamber (Anaerobe Systems, Morgan Hill, CA)
564 in a gas mixture of 5% CO_2, 5% H_2, and 90% N_2 to recover at 37°C for 1 hour. 100µL – 200µL of
565 transformed cells were then plated on NYCIII agar plates with and without 16 µg/mL tetracycline.
566 The plates were incubated anaerobically at 37°C for 48 hours and transformed colonies on the
567 tetracycline plates were counted to determine transformation efficiency. After the first successful
568 transformation, transformant colonies were grown in 5mL NYCIII media with 16 µg/mL
569 tetracycline and pDNF1199S-*tetM* was recovered from them using a QIAamp DNA Miniprep kit
570 (Qiagen, Hilden, Germany). The sequence of the purified plasmid was confirmed by sequencing
571 performed by Plasmidsaurus (Plasmidsaurus, Eugene, OR). Following transformation of
572 *Gardnerella* strain DNF00354 with pDNF1199S-tetM, the transformation reaction was plated onto
573 NYCIII agar plates with and without 16 µg/mL tetracycline. A small amount of growth from a
574 transformant colony on the tetracycline plate or lawn growth from the control plate was
575 resuspended in sterile water and colony PCR was performed to amplify and detect the *tetM* gene.
576 1µL of resuspended bacteria was added to a reaction mix of 12.5µL GoTaq Green master mix
577 (Promega, Madison, WI), 0.5µL of each primer, and 6.5µL H_2O. PCR was carried out for 28 cycles
578 with an annealing temperature of 60°C and an extension time of 2.5 min.
579 Primer sequences were as follows (5' → 3'):
580 Forward: CGCGACCAAATATTGGTACATTATTACAGCTATTTTGTAATCACGTACTC
581 Reverse: GAGACTGTCGTATTGATGTCACGGGCTTAATAGGTCTTATGTCGG
582
583

References

584 1 Fredricks, D. N., Fiedler, T. L. & Marrazzo, J. M. Molecular identification of bacteria
585 associated with bacterial vaginosis. *N Engl J Med* **353**, 1899-1911 (2005).
586 https://doi.org/10.1056/NEJMoa043802
587 2 Amsel, R. *et al.* Nonspecific vaginitis. Diagnostic criteria and microbial and epidemiologic
588 associations. *Am J Med* **74**, 14-22 (1983). https://doi.org/10.1016/0002-9343(83)91112-9
589 3 Alcendor, D. J. Evaluation of Health Disparity in Bacterial Vaginosis and the Implications
590 for HIV-1 Acquisition in African American Women. *Am J Reprod Immunol* **76**, 99-107
591 (2016). https://doi.org/10.1111/aji.12497
592 4 Klebanoff, M. A. & Brotman, R. M. Treatment of bacterial vaginosis to prevent preterm
593 birth. *Lancet* **392**, 2141-2142 (2018). https://doi.org/10.1016/S0140-6736(18)32115-9
594 5 Berman, H. L., Goltsman, D. S. A., Anderson, M., Relman, D. A. & Callahan, B. J.
595 *Gardnerella* diversity and ecology in pregnancy and preterm birth. *mSystems* **9**, e0133923
596 (2024). https://doi.org/10.1128/msystems.01339-23
597 6 Fredricks, D. N., Fiedler, T. L., Thomas, K. K., Mitchell, C. M. & Marrazzo, J. M. Changes
598 in vaginal bacterial concentrations with intravaginal metronidazole therapy for bacterial
599 vaginosis as assessed by quantitative PCR. *J Clin Microbiol* **47**, 721-726 (2009).
600 https://doi.org/10.1128/JCM.01384-08
601 7 Fredricks, D. N., Fiedler, T. L., Thomas, K. K., Oakley, B. B. & Marrazzo, J. M. Targeted
602 PCR for detection of vaginal bacteria associated with bacterial vaginosis. *J Clin Microbiol*
603 **45**, 3270-3276 (2007). https://doi.org/10.1128/JCM.01272-07
604 8 Hummelen, R. *et al.* Deep sequencing of the vaginal microbiota of women with HIV. *PLoS*
605 *One* **5**, e12078 (2010). https://doi.org/10.1371/journal.pone.0012078
606 9 Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A*
607 **108 Suppl 1**, 4680-4687 (2011). https://doi.org/10.1073/pnas.1002611107
608 10 Shipitsyna, E. *et al.* Composition of the vaginal microbiota in women of reproductive age-
609 -sensitive and specific molecular diagnosis of bacterial vaginosis is possible? *PLoS One* **8**,
610 e60670 (2013). https://doi.org/10.1371/journal.pone.0060670
611 11 Srinivasan, S. *et al.* Bacterial communities in women with bacterial vaginosis: high
612 resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria.
613 *PLoS One* **7**, e37818 (2012). https://doi.org/10.1371/journal.pone.0037818
614 12 van de Wijgert, J. H. *et al.* The vaginal microbiota: what have we learned after a decade of
615 molecular characterization? *PLoS One* **9**, e105998 (2014).
616 https://doi.org/10.1371/journal.pone.0105998
617 13 Gardner, H. L. & Dukes, C. D. Haemophilus vaginalis vaginitis: a newly defined specific
618 infection previously classified non-specific vaginitis. *Am J Obstet Gynecol* **69**, 962-976
619 (1955).
620 14 Schwebke, J. R., Muzny, C. A. & Josey, W. E. Role of *Gardnerella vaginalis* in the
621 pathogenesis of bacterial vaginosis: a conceptual model. *J Infect Dis* **210**, 338-343 (2014).
622 https://doi.org/10.1093/infdis/jiu089
623 15 Patterson, J. L., Stull-Lane, A., Girerd, P. H. & Jefferson, K. K. Analysis of adherence,
624 biofilm formation and cytotoxicity suggests a greater virulence potential of *Gardnerella*
625 *vaginalis* relative to other bacterial-vaginosis-associated anaerobes. *Microbiology*
626 *(Reading)* **156**, 392-399 (2010). https://doi.org/10.1099/mic.0.034280-0

16 Gelber, S. E., Aguilar, J. L., Lewis, K. L. & Ratner, A. J. Functional and phylogenetic characterization of Vaginolysin, the human-specific cytolysin from *Gardnerella vaginalis*. *J Bacteriol* **190**, 3896-3903 (2008). https://doi.org/10.1128/JB.01965-07

17 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. *Sci Rep* **5**, 11640 (2015). https://doi.org/10.1038/srep11640

18 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

19 Muzny, C. A. *et al.* An Updated Conceptual Model on the Pathogenesis of Bacterial Vaginosis. *J Infect Dis* **220**, 1399-1405 (2019). https://doi.org/10.1093/infdis/jiz342

20 Janulaitiene, M. *et al.* Prevalence and distribution of Gardnerella vaginalis subgroups in women with and without bacterial vaginosis. *BMC Infect Dis* **17**, 394 (2017). https://doi.org/10.1186/s12879-017-2501-y

21 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. *J Med Microbiol* **63**, 162-175 (2014). https://doi.org/10.1099/jmm.0.066407-0

22 Shvartsman, E., Hill, J. E., Sandstrom, P. & MacDonald, K. S. *Gardnerella* Revisited: Species Heterogeneity, Virulence Factors, Mucosal Immune Responses, and Contributions to Bacterial Vaginosis. *Infect Immun* **91**, e0039022 (2023). https://doi.org/10.1128/iai.00390-22

23 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

24 Benito, R., Vazquez, J. A., Berron, S., Fenoll, A. & Saez-Neito, J. A. A modified scheme for biotyping *Gardnerella vaginalis*. *J Med Microbiol* **21**, 357-359 (1986). https://doi.org/10.1099/00222615-21-4-357

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

26 Ingianni, A., Petruzzelli, S., Morandotti, G. & Pompei, R. Genotypic differentiation of *Gardnerella vaginalis* by amplified ribosomal DNA restriction analysis (ARDRA). *FEMS Immunol Med Microbiol* **18**, 61-66 (1997). https://doi.org/10.1111/j.1574-695X.1997.tb01028.x

27 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. *J Bacteriol* **194**, 3922-3937 (2012). https://doi.org/10.1128/JB.00056-12

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

29 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

30 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

31 Innamorati, K. A. *et al.* Metronidazole response profiles of Gardnerella species are congruent with phylogenetic and comparative genomic analyses. *Genome Med* **17**, 28 (2025). https://doi.org/10.1186/s13073-025-01446-4

32 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

33 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

34 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. *Front Microbiol* **7**, 978 (2016). https://doi.org/10.3389/fmicb.2016.00978

35 Lugli, G. A. *et al.* Comparative genomic and phylogenomic analyses of the Bifidobacteriaceae family. *BMC Genomics* **18**, 568 (2017). https://doi.org/10.1186/s12864-017-3955-4

36 Tarracchini, C. *et al.* Assessing the Genomic Variability of Gardnerella vaginalis through Comparative Genomic Analyses: Evolutionary and Ecological Implications. *Appl Environ Microbiol* **87** (2020). https://doi.org/10.1128/AEM.02188-20

37 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

38 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

39 Hill, J. E., Albert, A. Y. K. & Group, V. R. Resolution and Cooccurrence Patterns of *Gardnerella leopoldii, G. swidsinskii, G. piotii,* and *G. vaginalis* within the Vaginal Microbiome. *Infect Immun* **87** (2019). https://doi.org/10.1128/IAI.00532-19

40 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

41 Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114 (2018). https://doi.org/10.1038/s41467-018-07641-9

42 Hall, B. G. & Nisbet, J. Building Phylogenetic Trees From Genome Sequences With kSNP4. *Mol Biol Evol* **40** (2023). https://doi.org/10.1093/molbev/msad235

43 Gautreau, G. *et al.* PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput Biol* **16**, e1007732 (2020). https://doi.org/10.1371/journal.pcbi.1007732

44 Flores Ramos, S. *et al.* Genomic Stability and Genetic Defense Systems in *Dolosigranulum pigrum*, a Candidate Beneficial Bacterium from the Human Microbiome. *mSystems* **6**, e0042521 (2021). https://doi.org/10.1128/mSystems.00425-21

45 Dillard, L. R., Glass, E. M., Lewis, A. L., Thomas-White, K. & Papin, J. A. Metabolic Network Models of the Gardnerella Pangenome Identify Key Interactions with the Vaginal Environment. *mSystems* **8**, e0068922 (2023). https://doi.org/10.1128/msystems.00689-22

46 Teixeira, P., Sousa, M., Gonçalves Ribeiro, T., Grosso F., Peixe L. Unraveling the *Gardnerella* Pangenome: Insights into Genetic Diversity, Virulence, and Antibiotic Resistance. *bioRxiv* **2025.03.03.641253** (2025). https://doi.org/doi: https://doi.org/10.1101/2025.03.03.641253

47 Chen, L., Li, J. & Xiao, B. The role of sialidases in the pathogenesis of bacterial vaginosis and their use as a promising pharmacological target in bacterial vaginosis. *Front Cell Infect Microbiol* **14**, 1367233 (2024). https://doi.org/10.3389/fcimb.2024.1367233

48 Rottini, G. *et al.* Identification and partial characterization of a cytolytic toxin produced by *Gardnerella vaginalis*. *Infect Immun* **58**, 3751-3758 (1990). https://doi.org/10.1128/iai.58.11.3751-3758.1990

49 Price, M. N., Shiver A.L., Day, L.A., Torres, M., Lesea, H.P., Owens, T.K., Deutschbauer, A.M., Arkin, A.P. . Improving the annotation of amino acid biosynthesis pathways: GapMind 2024. *Biorxiv* **2024.10.14.618325** (2024). https://doi.org/https://doi.org/10.1101/2024.10.14.618325

50 Price, M. N., Deutschbauer, A. M. & Arkin, A. P. GapMind: Automated Annotation of Amino Acid Biosynthesis. *mSystems* **5** (2020). https://doi.org/10.1128/mSystems.00291-20

51 Price, M. N., Deutschbauer, A. M. & Arkin, A. P. Filling gaps in bacterial catabolic pathways with computation and high-throughput genetics. *PLoS Genet* **18**, e1010156 (2022). https://doi.org/10.1371/journal.pgen.1010156

52 Bordigoni, A., Bouchard, S. & Desnues, C. Isolation and Characterization of Gardnerella Phage vB_Gva_AB1, a Bacteriophage Infecting a Clinical Strain of *Gardnerella vaginalis*. *Microbiol Resour Announc* **10** (2021). https://doi.org/10.1128/MRA.00053-21

53 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

54 Georjon, H. & Bernheim, A. The highly diverse antiphage defence systems of bacteria. *Nat Rev Microbiol* **21**, 686-700 (2023). https://doi.org/10.1038/s41579-023-00934-x

55 Tesson, F. *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun* **13**, 2561 (2022). https://doi.org/10.1038/s41467-022-30269-9

56 Tal, N. *et al.* Bacteria deplete deoxynucleotides to defend against bacteriophage infection. *Nat Microbiol* **7**, 1200-1209 (2022). https://doi.org/10.1038/s41564-022-01158-0

57 McLandsborough, L. A., Kolaetis, K. M., Requena, T. & McKay, L. L. Cloning and characterization of the abortive infection genetic determinant abiD isolated from pBF61 of Lactococcus lactis subsp. lactis KR5. *Appl Environ Microbiol* **61**, 2023-2026 (1995). https://doi.org/10.1128/aem.61.5.2023-2026.1995

58 Deng, Y. M., Liu, C. Q. & Dunn, N. W. Genetic organization and functional analysis of a novel phage abortive infection system, AbiL, from Lactococcus lactis. *J Biotechnol* **67**, 135-149 (1999). https://doi.org/10.1016/s0168-1656(98)00175-8

59 Cohen, D. *et al.* Cyclic GMP-AMP signalling protects bacteria against viral infection. *Nature* **574**, 691-695 (2019). https://doi.org/10.1038/s41586-019-1605-5

60 Millman, A., Melamed, S., Amitai, G. & Sorek, R. Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. *Nat Microbiol* **5**, 1608-1615 (2020). https://doi.org/10.1038/s41564-020-0777-y

61 Gao, L. *et al.* Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science* **369**, 1077-1084 (2020). https://doi.org/10.1126/science.aba0372

62 Mestre, M. R. *et al.* UG/Abi: a highly diverse family of prokaryotic reverse transcriptases associated with defense functions. *Nucleic Acids Res* **50**, 6084-6101 (2022). https://doi.org/10.1093/nar/gkac467

63 Dy, R. L., Przybilski, R., Semeijn, K., Salmond, G. P. & Fineran, P. C. A widespread bacteriophage abortive infection system functions through a Type IV toxin-antitoxin mechanism. *Nucleic Acids Res* **42**, 4590-4605 (2014). https://doi.org/10.1093/nar/gkt1419

64 Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359** (2018). https://doi.org/10.1126/science.aar4120

65 Cheng, R. *et al.* A nucleotide-sensing endonuclease from the Gabija bacterial defense system. *Nucleic Acids Res* **49**, 5216-5229 (2021). https://doi.org/10.1093/nar/gkab277

66 Payne, L. J. *et al.* Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Res* **49**, 10868-10878 (2021). https://doi.org/10.1093/nar/gkab883

67 Millman, A. *et al.* An expanded arsenal of immune systems that protect bacteria from phages. *Cell Host Microbe* **30**, 1556-1569 e1555 (2022). https://doi.org/10.1016/j.chom.2022.09.017

68 Johnston, C. D. *et al.* Systematic evasion of the restriction-modification barrier in bacteria. *Proc Natl Acad Sci U S A* **116**, 11454-11459 (2019). https://doi.org/10.1073/pnas.1820256116

69 Finn, M. B., Ramsey, K. M., Tolliver, H. J., Dove, S. L. & Wessels, M. R. Improved transformation efficiency of group A Streptococcus by inactivation of a type I restriction modification system. *PLoS One* **16**, e0248201 (2021). https://doi.org/10.1371/journal.pone.0248201

70 Hu, S. *et al.* Altering under-represented DNA sequences elevates bacterial transformation efficiency. *mBio* **14**, e0210523 (2023). https://doi.org/10.1128/mbio.02105-23

71 Chen, C. *et al.* Development of a Pre-Modification Strategy to Overcome Restriction-Modification Barriers and Enhance Genetic Engineering in Lactococcus lactis for Nisin Biosynthesis. *Int J Mol Sci* **26** (2025). https://doi.org/10.3390/ijms26052200

72 Payne, L. J. *et al.* PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. *Nucleic Acids Res* **50**, W541-W550 (2022). https://doi.org/10.1093/nar/gkac400

73 Klimowicz, A. K., Garcia, E.M., Jefferson, K.K., Dillard, J.P. Genetic transformation of *Gardnerella* species and characterization of vaginolysin and sialidase mutants. *bioRxiv* **2025.05.09.653190** (2025). https://doi.org/https://doi.org/10.1101/2025.05.09.653190

74 Kularatne, B. & Hill, J. E. Transformation of Gardnerella vaginalis with a Bifidobacterium-Escherichia coli shuttle vector plasmid. *Microbiol Spectr* **13**, e0048125 (2025). https://doi.org/10.1128/spectrum.00481-25

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

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

77 Kharsany, A. B., Hoosen, A. A. & Van den Ende, J. Antimicrobial susceptibilities of *Gardnerella vaginalis*. *Antimicrob Agents Chemother* **37**, 2733-2735 (1993). https://doi.org/10.1128/AAC.37.12.2733

78 Bulavaite, A., Dalgediene, I., Michailoviene, V. & Pleckaityte, M. Type II Restriction-Modification System from Gardnerella vaginalis ATCC 14018. *Pathogens* **9** (2020). https://doi.org/10.3390/pathogens9090703

79 Srinivasan, S. *et al.* More Easily Cultivated Than Identified: Classical Isolation With Molecular Identification of Vaginal Bacteria. *J Infect Dis* **214 Suppl 1**, S21-28 (2016). https://doi.org/10.1093/infdis/jiw192

80 Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res* **44**, D67-72 (2016). https://doi.org/10.1093/nar/gkv1276

81 Meier-Kolthoff, J. P. & Goker, M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nat Commun* **10**, 2182 (2019). https://doi.org/10.1038/s41467-019-10210-3

82 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

83 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 Res* **25**, 1043-1055 (2015). https://doi.org/10.1101/gr.186072.114

84 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

85 Gu, Z. Complex heatmap visualization. *Imeta* **1**, e43 (2022). https://doi.org/10.1002/imt2.43

86 Hahsler, M., Hornik, K., Buchta, C. . Getting Things in Order: An Introduction to the R Package Seriation. *Journal of Statistical Software* **25(3),1-34** (2008). https://doi.org/doi.org/10.18637/jss.v025.i03

87 Auch, A. F., von Jan, M., Klenk, H. P. & Goker, M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci* **2**, 117-134 (2010). https://doi.org/10.4056/sigs.531120

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

89 Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol Biol Evol* **38**, 3022-3027 (2021). https://doi.org/10.1093/molbev/msab120

90 Bazin, A., Gautreau, G., Medigue, 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

91 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 Res* **47**, D309-D314 (2019). https://doi.org/10.1093/nar/gky1085

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

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

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

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

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

97 Egervarn, M., Roos, S. & Lindmark, H. Identification and characterization of antibiotic resistance genes in *Lactobacillus reuteri* and *Lactobacillus plantarum*. *J Appl Microbiol* **107**, 1658-1668 (2009). https://doi.org/10.1111/j.1365-2672.2009.04352.x

98 Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343-345 (2009). https://doi.org/10.1038/nmeth.1318

99 Douwe van Sinderen, M. V. *Bifidobacteria: Methods and Protocols*. 1 edn, (Humana, 2021).