**Clinical Chemistry** 65:11
000–000 (2019)

Infectious Disease

# In Silico Analysis of *Gardnerella* Genomospecies Detected in the Setting of Bacterial Vaginosis

Robert F. Potter,[1*] Carey-Ann D. Burnham,[2,3,4] and Gautam Dantas[1,2,4,5]

**BACKGROUND:** *Gardnerella vaginalis* is implicated as 1 of the causative agents of bacterial vaginosis, but it can also be isolated from the vagina of healthy women. Previous efforts to study *G. vaginalis* identified 4 to 6 clades, but average nucleotide identity analysis indicates that *G. vaginalis* may be multiple species. Recently, *Gardnerella* was determined to be 13 genomospecies, with *Gardnerella piottii*, *Gardnerella leopoldii*, and *Gardnerella swidsinkii* delineated as separate species.

**METHODS:** We accessed 103 publicly available genomes annotated as *G. vaginalis*. We performed comprehensive taxonomic and phylogenomic analysis to quantify the number of species called *G. vaginalis*, the similarity of their core genes, and their burden of their accessory genes. We additionally analyzed publicly available metatranscriptomic data sets of bacterial vaginosis to determine whether the newly delineated genomospecies are present, and to identify putative conserved features of *Gardnerella* pathogenesis.

**RESULTS:** *Gardnerella* could be classified into 8 to 14 genomospecies depending on the in silico classification tools used. Consensus classification identified 9 different *Gardnerella* genomospecies, here annotated as GS01 through GS09. The genomospecies could be readily distinguished by the phylogeny of their shared genes and burden of accessory genes. All of the new genomospecies were identified in metatranscriptomes of bacterial vaginosis.

**CONCLUSIONS:** Multiple *Gardnerella* genomospecies operating in isolation or in concert with one another may be responsible for bacterial vaginosis. These results have important implications for future efforts to understand the evolution of the *Gardnerella* genomospecies, host–pathogen interactions of the genomospecies during bacterial vaginosis, diagnostic assay development for bacterial vaginosis, and metagenomic investigations of the vaginal microbiota.



Bacterial vaginosis (BV)[6] is a common infectious disease of women, often caused by *Gardnerella vaginalis (1, 2)*. However, *G. vaginalis* has also been identified in healthy women without BV *(3, 4)*. One explanation is that certain strains of *G. vaginalis* are more pathogenic than others. Genome-based taxonomic methods, which have delineated novel species in other genera, have scarcely been applied to *G. vaginalis*. Importantly, 1 recent investigation found that average nucleotide identity (ANI) values between different *G. vaginalis* subgroups were below the species cutoff of 96%, indicating *G. vaginalis* may be multiple species *(5)*. Recently, using ANI and digital DNA-DNA hybridization assays, it was found that 13 different *Gardnerella* genomospecies may currently be annotated as *G. vaginalis (6)*. Three of these species were fully elucidated using phenotypic assays and termed *Gardnerella piotii*, *Gardnerella leopoldii*, and *Gardnerella swidsinkii (7)*.

Historically, delineation of new bacteria taxa has relied on phenotypic differences between strains, such as chemical analysis or biochemical utilization characteristics, with the laborious DNA-DNA hybridization assay representing the gold standard analysis for species-level determination *(7)*. Using 16S rRNA gene sequencing, a cutoff of 97% is typically used to delineate bacteria, but given that recognized different species can have values >97% similarity, whole-genome data such as ANI are often also used *(7, 8)*. ANI values ≥96% are used as thresholds for species-level designations *(9)*. In the absence of an isolated organism, putative novel species can be determined by genetic content alone but are termed genomospecies. In addition, recognized species can also

[1] The Edison Family Center for Genome Sciences and Systems Biology, Washington University in St. Louis School of Medicine, St. Louis, MO; [2] Department of Pathology and Immunology, Washington University in St. Louis School of Medicine, St. Louis, MO; [3] Department of Pediatrics, Washington University in St. Louis School of Medicine, St. Louis, MO; [4] Department of Molecular Microbiology, Washington University in St. Louis School of Medicine, St. Louis, MO; [5] Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO.

* Address correspondence to this author at: E-mail rfpotter@wustl.edu.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Received April 3, 2019; accepted July 25, 2019.
Previously published online at DOI: 10.1373/clinchem.2019.305474



[6] Nonstandard abbreviations: BV, bacterial vaginosis; ANI, average nucleotide identity; NCBI, National Center for Biotechnology Information; AAI, average amino acid identity; COG, cluster of orthologous group; PanPhlAn, pangenome-based phylogenomic analysis.

be reclassified such as *Escherichia hermannii* and *Salmonella subterranea* to *Atlantibacter hermannii* and *Atlantibacter subterranea*, respectively *(10)*. Given the previous analysis on *Gardnerella*, we used multiple in silico taxonomic classification tools to bin publicly available *Gardnerella* genomes into different genomospecies and then performed comparative analysis between the genomospecies.

To address the knowledge gap regarding the taxonomic diversity and relatedness within genomes currently classified as *G. vaginalis*, we performed a retrospective comparative analysis using 103 publicly available genomes as well as BV metatranscriptomes (see Tables 1 and 2 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol65/issue11). Based on the observation that multiple *G. vaginalis* genomes may be related by ANI values less than the species cutoff of 96%, we hypothesized that multiple distinct genomospecies have been collapsed into a single *G. vaginalis* species annotation. Further, we hypothesized that these genomospecies could be distinguished by the relatedness of their shared genes and the differential burden of their accessory genes.

## Materials and Methods

### PUBLICLY AVAILABLE GENOMES AND METATRANSCRIPTOME READS

Four genomes annotated as *Gardnerella* unclassified and 99 genomes annotated as *G. vaginalis* were retrieved from National Center for Biotechnology Information (NCBI) genomes in October 2018. The assembly file containing chromosome and plasmid components for all genomes was used for analysis. All genomes were reannotated for open reading frames using Prokka *(11)*. Paired-end 2 × 100-bp Illumina reads from a metatranscriptomic investigation of BV (BioProject accession number PRJEB21446) were retrieved from the Sequence Read Archive in October 2018 *(12)*. All Linux commands are included in Table 3 of the online Data Supplement.

### IN SILICO TAXONOMIC ANALYSIS

We initially used pyANI (https://github.com/widdowquinn/pyani) to obtain pairwise ANI values with the mummer nucleotide alignment method on all 103 genomes obtained from the NCBI. Representative genomes for the 14 different genomospecies were uploaded to JSpeciesWS in January 2019 and annotated with default conditions for the ANIb and tetranucleotide frequency analysis (http://jspecies.ribohost.com/jspeciesws/#home) *(13, 14)*. The same 14 genomes were uploaded to the ANI matrix software from the Kostas laboratory (http://enve-omics.ce.gatech.edu/g-matrix/index). The faa file from Prokka, containing protein sequences for identified open reading frames, was uploaded to the Kostas laboratory average amino acid identity (AAI) matrix software in January 2019 *(14)*. For the purposes of our investigation, to create genomospecies bins for downstream analysis, we adopted a conservative consensus approach. Thereby, if ≥2 of the tools indicated that the genomes represent the same genomospecies, we then counted them as the same genomospecies.

### CORE GENOME ANALYSIS

Roary was used to cluster the open reading frames in the *Gardnerella* cohort to identify the core genome and accessory genome at 70% identity *(15)*. The 200 core genes were aligned using PRANK *(16)*. The core genome alignment was converted into an approximate maximum likelihood tree with FastTree, and lineages were identified using BAPS within FastGear *(17–19)*. The Newick file from FastTree was viewed as a midpoint rooted tree in iTOL with bootstrap support values as branch labels *(20)*. To construct the nearest neighbor network, the core genome alignment file was uploaded to SplitsTrees *(21)*.

### ACCESSORY GENOME ANALYSIS

The gene presence/absence file constructed by Roary was removed of core genes and analyzed for principal components in RStudio using prcomp *(22)*. The elucidated genomospecies were overlaid onto the genomes. To identify genes responsible for the distinct clustering pattern observed, we used Scoary to identify genes in the Roary pan-genome that are strongly associated with the 9 different genomospecies *(23)*. The presence/absence matrix for genes annotated as sialidases, glycoside hydrolases, ATP-binding import proteins, or allantoin utilization was viewed as a binary matrix in iTOL. Counts for the number of these genes within the different genomospecies were computed and viewed in Prism V8.

### CLUSTER OF ORTHOLOGOUS GROUPS AND GENE OF INTEREST QUANTIFICATION

We uploaded the pan-genome reference file from Roary, which contains a representative gene for the 7402 genes in the pan-genome database, to EggNOG 4.5.1 in November 2018 to identify functional categories for all possible genes *(24)*. Normalized cluster of orthologous group (COG) counts for each genomospecies were determined by dividing the number of genes for each individual COG annotation by the total number of genes that had any COG assigned. To identify all the genes with a putative role in carbohydrate metabolism, we uploaded the pan-genome reference file from Roary to dbCAN, which used HMMER and DIAMOND to compare our query with the CAZy database *(25)*.

### TAXONOMIC METATRANSCRIPTOME ANALYSIS

To determine the presence of the *Gardnerella* genomospecies within the metatranscriptome samples, we used

the short-read classification program Centrifuge *(26)*. Initially, we made a custom database by assigning the 5971 total contigs from the downloaded FASTA files within the *Gardnerella* cohort to a specific genomospecies using our previously described conservative consensus approach. Therefore, our database contained all open reading frames and intergenic regions. Our classification scheme was designed to ignore the other members of the vaginal microbiota, so each read could be assigned as mapping to ≥1 of the *Gardnerella* genomospecies, mapping uniquely to just 1 genomospecies, or not mapping to any of the genomospecies. For the 20 samples used in our investigation, we then computed the percentage of *Gardnerella*-specific reads that uniquely mapped to an individual genomospecies by quantifying the number of unique reads per genomospecies divided by the total sum of unique reads that mapped to all genomospecies. The percentage matrix created by this analysis was hierarchically clustered using SciPy and viewed as a clustermap in seaborn. Additionally, the percentage values were viewed as a stacked bar plot in Matplotlib.

#### METATRANSCRIPTOME FUNCTIONAL ANALYSIS

We built a pan-genome database with pangenome-based phylogenomic analysis (PanPhlAn) using the presence/absence matrix previously identified by Roary (panphlan_pangenome_generation.py) *(27)*. We mapped the *Gardnerella*-specific transcriptome reads and quantified the coverage amount of every gene in the pan-genome for each metatranscriptome sample (http://panphlan_map.py). We used the mean coverage value for each gene across the 20 metatranscriptome samples. The top 224 expressed genes, defined as the mean plus SD of the coverage level, were analyzed to identify any enriched COGs. To determine whether the COGs were enriched within the metatranscriptome data sets, we computed an enrichment index using the below formula *(28)*:

$$\frac{(\text{Number of top expressed genes with } COG_X / \text{Number of top expressed genes with a COG})}{(\text{Number of total genes with } COG_X / \text{Number of total genes with a COG})} \quad (1)$$

If the index was >1, then it indicated that COG was enriched within the top expressed genes. Additionally, we used a quantitative assessment of coverage values to identify genes within the *Gardnerella* pan-genome that were expressed at a statistically meaningful percentage within the data set. To accomplish this, we viewed each individual gene in the pan-genome by plotting on the *x* axis the number of isolates that harbor the gene identified by Roary and on the *y* axis the mean coverage value from PanPhlAn analysis of the 20 metatranscriptome samples.

#### STATISTICAL ANALYSIS

ANOVA for number of sialidase, glycoside hydrolase, ATP-binding import proteins, allantoin utilization genes, and CAZy database hits in the different genomospecies was performed in GraphPad Prism V8. Paired Student *t*-tests between select groups displayed in Fig. 3C were performed in GraphPad Prism V8. Permutational multivariate analysis of variance (PERMANOVA) was performed on the gene_presence_absence_matrix from Roary in RStudio using the adonis2 command from the Vegan (https://cran.r-project.org/web/packages/vegan/vegan.pdf) package.

## Results

#### IN SILICO TOOL-DEPENDENT CLASSIFICATION OF *G. VAGINALIS* INTO 8 TO 14 GENOMOSPECIES

We began analysis by using pyANI with the mummer nucleotide alignment method to determine pairwise ANI values between the 103 genomes obtained from NCBI (see Fig. 1 and Table 4 in the online Data Supplement). This analysis indicated that in addition to the 13 genomospecies delineated by Vaneechoutte et al. *(6)*, strain NR010 may represent a 14th genomospecies, as it did not have any ANI values ≥96% to any other genome. From this we found that *Gardnerella* may contain a maximum of 14 genomospecies. We then used multiple publicly available tools for verification. Therefore, to further assess whether genomes annotated as the species *G. vaginalis* represent multiple genomospecies, we used 2 additional different ANI platforms—tetranucleotide frequency and AAI tools—to delineate the genomes into genomospecies *(9, 14, 29, 30)*. We chose 13 genomes from the recent delineation of *G. vaginalis*, including the type genomes for *G. vaginalis*, *G. piotii*, *G. leopoldii*, and *G. swidsinkii*, as well as NR010 *(6)* (Table 1). Our results showed that the number of annotated genomospecies ranged from 8 to 14, depending on the classification tools used (Fig. 1). ANI with the BLAST nucleotide alignment method (ANIb) from JSpeciesWS indicated that these genomes represented 14 unique genomospecies (Fig. 1A here and see Table 5 in the online Data Supplement). In comparison, classification by tetranucleotide frequency from JSpeciesWS found that the 14 genomes are 9 genomospecies (Fig. 1B here and Table 5 in the online Data Supplement). Tetranucleotide frequency-based classification found that the type strains *G. leopoldii* UGENT 06.41 (T) and *G. swidsinkii* GS 9838–1 (T) may be the same genomospecies. A separate ANI classifier (ANI calculator from the Kostas lab) indicated that the 14 genomes instead represented 12 genomospecies, and again that the type strains *G. leopoldii* UGENT 06.41 (T) and *G. swidsinkii* GS 9838–1 (T) may be the same genomospecies (Fig. 1C here and Table 6 in the online Data Supplement). Finally, an AAI classifier (AAI calculator

**Table 1. Summary of in silico taxonomic findings for the 14 representative genomospecies from Fig. 1.[a]**

| | | | Method | | | | |
|---|---|---|---|---|---|---|---|
| Number | Species/strain | Assembly | Jspecies-ANIb | Jspecies-Tetra | Kostas-ANI | Kostas-AAI | Conservative consensus |
| 01 | *G. vaginalis* ATCC 14018 (T) | GCA_000178355.1 | 01 | 01 = 02 | 01 | 01 = 02 | GS01 (01 = 02) |
| 02 | JCP8108 | GCA_000414525.1 | 02 | | 02 | | |
| 03 | JCP8017A | GCA_000414605.1 | 03 | 03 = 04 | 03 | 03 = 04 = 11 | GS02 (03 = 04) |
| 04 | *G. piotii* UGENT 18.01 (T) | GCA_003397585.1 | 04 | | 04 | | |
| 05 | *G. leopoldii* UGENT 06.41 (T) | GCA_003293675.1 | 05 | 05 = 06 | 05 = 06 | 05 = 06 | GS03 (05 = 06) |
| 06 | *G. swidsinkii* 9838-1 (T) | GCA_003397705.1 | 06 | | | | |
| 07 | JCP8481A | GCA_000414465.1 | 07 | 07 | 07 | 07 | GS04 (07) |
| 08 | UMB1686 | GCA_002884775.1 | 08 | 08 = 09 = 10 | 08 | 08 = 09 = 10 | GS05 (08 = 09 = 10) |
| 09 | 6119V5 | GCA_000263655.1 | 09 | | 09 = 10 | | |
| 10 | 1500E | GCA_000263595.1 | 10 | | | | |
| 11 | GED7760B | GCA_001546455.1 | 11 | 11 | 11 | | GS06 (11) |
| 12 | CMW7778B | GCA_001563665.1 | 12 | 12 | 12 | 12 | GS07 (12) |
| 13 | KA00225 | GCA_002896555.1 | 13 | 13 | 13 | 13 | GS08 (13) |
| 14 | NR010 | GCA_003408845.1 | 14 | 14 | 14 | 14 | GS09 (14) |
| | Number of species | | 14 | 9 | 12 | 8 | 9 |

[a] The number corresponds to the depiction in Fig. 1. Cases where an analysis indicates that 2 or 3 strains are the same species are denoted with an equal sign. The conservative consensus approach was performed by concatenating strains into genomospecies if ≥2 programs agreed.

from the Kostas lab) produced the most conservative estimate of the number of genomospecies, identifying 8 genomospecies from the 14 genomes (Fig. 1D here and Table 6 in the online Data Supplement). The AAI classifications were concordant with the JSpeciesWS tetranucleotide frequency-based classifications in that *G. vaginalis* ATCC 14018 (T)/JCP8108, *G. piotii* UGENT 18.01 (T)/JCP8017A, and UMB1686/6119V5/1500E were the same genomospecies, respectively. The AAI calculator was the only tool that considered GED7760B the same genomospecies as *G. piotii* UGENT 18.01 (T)/JCP8017A.

We adopted a conservative consensus approach for genomospecies classification for the remainder of our analysis. Specifically, if ≥2 of the aforementioned tools indicated that the genomes represent the same genomospecies, then we counted them as the same. This method had exact concordance with the tetranucleotide frequency tool classification and yielded 9 *Gardnerella* genomospecies (GS01–GS09) (Table 1). All comparative analyses and biological conclusions hereafter are based on these 9 genomospecies.

### CORE GENOME ALIGNMENT SUPPORT RELATEDNESS OF THE GENOMOSPECIES INTO 8 CLADES

To gain further insight into the taxonomic structure of the *Gardnerella* genus, we determined the 200 core genes (the loci present in 100% of strains) at 70% nucleotide identity with the pan-genome tool Roary and aligned these genes with PRANK to create a core gene alignment (see Table 7 in the online Data Supplement). We used FastTree to construct an approximate maximum likelihood tree from the core genome alignment, which depicted the evolutionary relationship between all genomes analyzed and provided a confidence value for every branch point (Fig. 2A). The tree had 100% bootstrap support values at the major branch points, indicating a high degree of confidence on the relatedness of the *Gardnerella* genomospecies to one another. Midpoint rooting of the tree in iTOL depicted a major spit within the genus between GS01/GS02/GS06 and GS03/GS04/GS05/GS07/GS08/GS09. Lineage identification using FastGear/BAPS on the core genome alignment identified 8 major lineages that had almost exact concordance with the consensus delineation into genomospecies, except that GS02 and the single genome GED7760B (GS06) were determined to be in the same lineage (Fig. 2A). FastGear initially assigns clusters with the BAPS software and then uses an additional allele comparison to produce more refined groups. The single genomes/genomospecies KA00225 (GS08) and NR010 (GS09) were counted as their own lineages, indicating that the allele frequencies between GS06 and GS02 may be similar enough compared with the background comparisons that they are linked into the same lineage.

As a second method to view the relatedness of the genomospecies, we visualized the core genome alignment file as a nearest neighbor network in SplitsTrees (Fig.

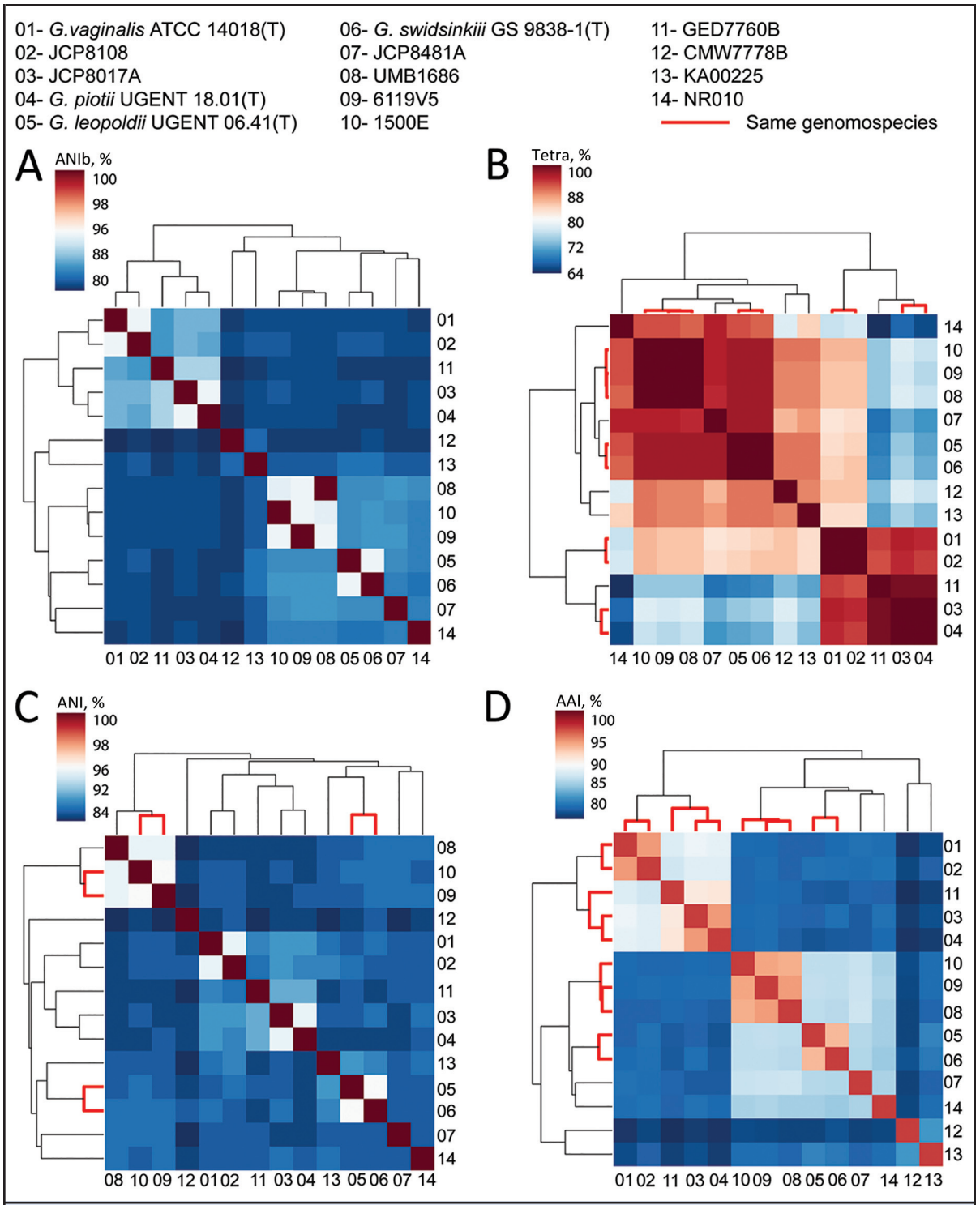


**Fig. 1. Different in silico taxonomic tools produce 8 to 14 *Gardnerella* genomospecies.**

Heatmaps with hierarchal clustering of the pairwise values for JSpeciesWS ANIb (A), JSpeciesWS tetranucleotide frequency (B), Kostas lab ANI (C), and Kostas lab AAI (D) for the 14 representative *Gardnerella* strains. All strains are assigned a numerical identifier. Red lines in the clustering pattern indicate groups of ≥2 genomes that represent the same genomospecies for that tool. The ANI (A and C) and AAI (D) plots are colored so that the taxonomic cutoff of 96% or 90%, respectively, is white. Pairwise values above this cutoff in the red spectrum and values below in the blue spectrum.

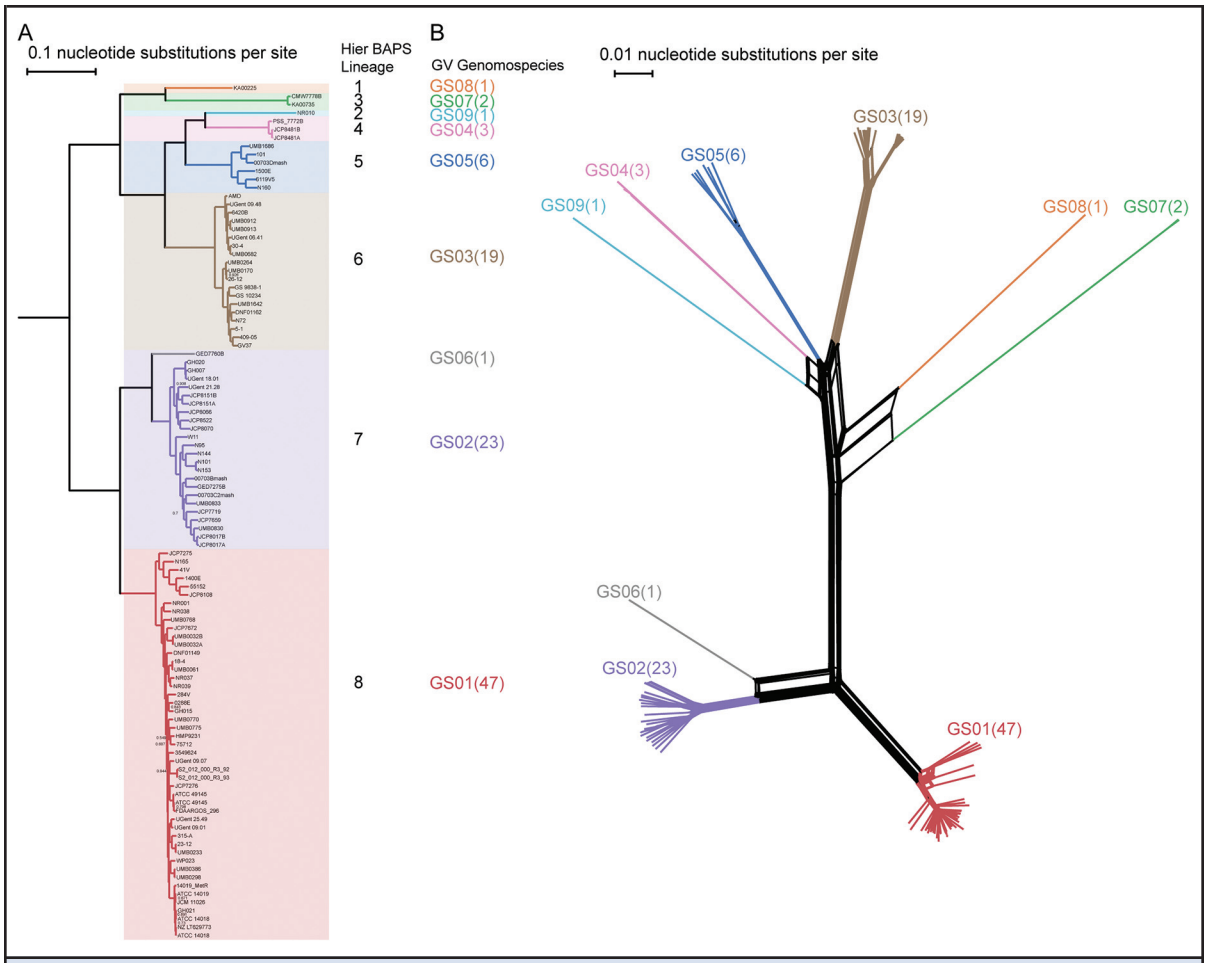


**Fig. 2. Core genome phylogenetic analysis shows the genomospecies fall into 9 distinct clusters.**

(A), Approximate maximum likelihood phylogenetic tree from PRANK alignment of the 200 core genes identified by Roary with FastGear/BAP lineages annotated adjacent to the tree. Groups of genomes that represent the same genomospecies identified by the conservative consensus approach are colored. (B), Nearest neighbor network of the core genome alignment with genomospecies annotated as tip labels.

2B). The clustering pattern of the isolates was visually concordant with the maximum likelihood tree. Importantly, the isolates from GS07 and GS08 deviated away from the center of the network, providing additional evidence on their separation from the group containing GS03, GS04, GS05, and GS09 (Fig. 2B).

### *GARDNERELLA* GENOMOSPECIES HAVE DISTINCT ACCESSORY GENE REPERTOIRES

To understand the differential burden of accessory genes that may contribute to niche adaptation within the vaginal microenvironment and/or to BV pathology, we performed a principal component analysis on the presence/absence matrix of noncore genes identified by Roary (see Table 8 in the online Data Supplement). PERMANOVA using adonis2 from the vegan package in RStudio indicated that genomospecies and accessory genome composition were significantly ($P < 0.00001$) linked. Superposition of the genomospecies classification onto the principal component analysis plot demonstrated that the accessory gene content for GS04, GS05, and GS09 was remarkably similar but that there were large differences between the other genomospecies (Fig. 3A). In particular, the 3 major genomospecies—GS01, GS02, and GS03—were situated in the periphery of the plot, demonstrating disparities between the gene repertoire within these genomospecies (Fig. 3A).

To gain insight into which genes may be driving this clustering pattern, we queried the pan-genome using the pan-genome association tool, Scoary, for genes that were differentially enriched within GS01, GS02, and GS03 (Fig. 3B here and see Table 8 in the online Data Supplement). For genes with putative function as sialidases, glycoside hydrolases, carbohydrate ATP-binding import

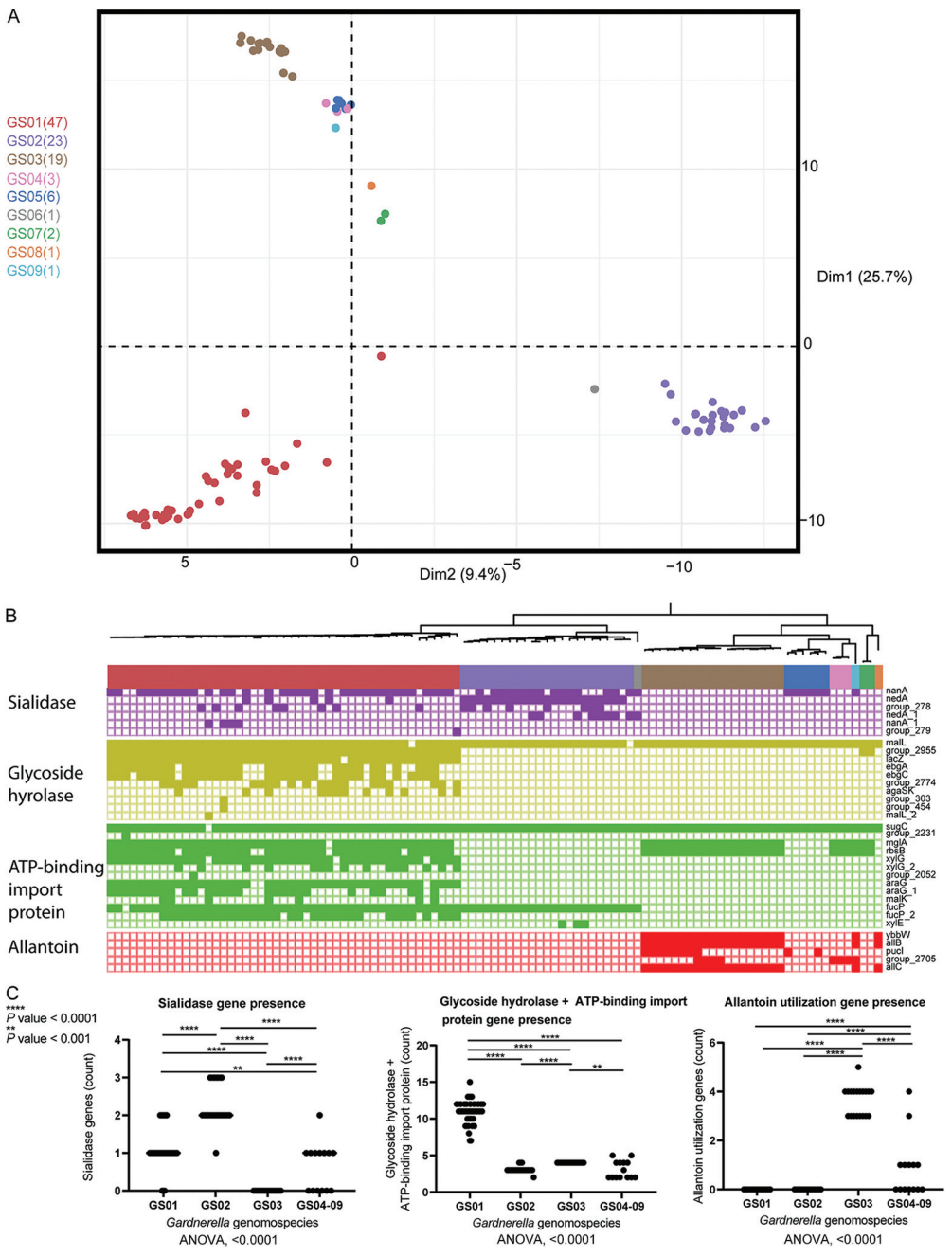


**Fig. 3. Accessory gene burden is different between the major genomospecies.**

(A), Principal component analysis of the accessory gene presence/absence matrix created by Roary with the genomes colored by their genomospecies. Each individual point represents the accessory gene content of a *Gardnerella* genome. Points closer to one another have similar accessory gene content. (B), Presence/absence matrix for accessory genes with putative sialidase, glycoside hydrolase, carbohydrate ATP-binding, and allantoin utilization roles adjacent to the phylogenetic tree from Fig. 2A (distances not to scale) with genomospecies identity. Each filled square represents the presence of a given gene, and a blank square represents the absence of that gene. Squares are colored based off predicted function. (C), Total counts for the number of genes identified with suspected function as sialidases, glycoside hydrolases/carbohydrate ATP-binding, and allantoin utilization abilities. Paired Student *t*-test results are shown for all significantly different gene burdens between GS01, GS02, GS03, and GS04–09.

protein, and allantoin metabolism, we viewed the presence/absence of the clusters in iTOL and quantified the gene burden in Prism. Interestingly, we found that genes annotated as sialidases were significantly ($P < 0.0001$) enriched in GS02 over GS01 and GS04 through GS09 (Fig. 3C). These genes were completely absent from GS03 genomes. Similarly, genes annotated as glycoside hydrolases and carbohydrate ATP-binding import proteins were significantly ($P < 0.0001$) enriched in GS01 (Fig. 3C). Lastly, we found that genes involved in the uptake and usage of allantoin were enriched ($P < 0.0001$) in GS03, absent in GS01 and GS02, and sparsely present in the other genomospecies (Fig. 3C). To understand overall differences in metabolic potential between the genomospecies, we submitted the pan-genome reference FASTA to EggNOG for COG annotation and quantified the number of COGs present in each genomospecies (see Table 9 in the online Data Supplement). The results were remarkably similar across all COGs except a notable increase in genes related to "carbohydrate transport and metabolism" in GS01 (see Fig. 2A in the online Data Supplement). Similarly, GS01 had a significantly ($P < 0.0001$) higher amount of carbohydrate utilization genes annotated by the CAZy database compared with the other genomospecies (see Fig. 2B in the online Data Supplement). In summary, we found that the different *Gardnerella* genomospecies could largely be distinguished by the presence/absence of their accessory genes and that accessory genes with specific functions were enriched or absent in certain genomospecies. These results indicated that different *Gardnerella* genomospecies had distinct gene repertories, which may lead to niche separation within the vaginal environment.

#### TAXONOMIC SIGNATURES OF NOVEL GENOMOSPECIES DURING BV

Metatranscriptomes are the genes that are expressed by a community of bacteria in any given environment. Given our improved resolution of *Gardnerella* into 9 genomospecies, we wanted to investigate whether any of the newly elucidated genomospecies could be identified in the metatranscriptomes of BV samples. To accomplish this, we used the short-read classifier Centrifuge on metatranscriptome sequencing reads from BV samples of women before and after receipt of metronidazole therapy *(26, 27)*.

We used Centrifuge to identify the percentage of *Gardnerella* reads that uniquely map to just 1 of the genomospecies within BV metatranscriptome samples for all of the genomospecies (see Table 2 in the online Data Supplement) *(26, 27)*. Metatranscriptome reads mapping uniquely to only a single genomospecies were identified for all genomospecies, but GS04, GS06, GS08, and GS09 had a mean presence of 1.74%, 1.0%, 0.49%, and 0.41% across all samples (see Table 2 in the online Data Supplement). As our classification scheme was designed specifically to focus only on the *Gardnerella* genomospecies, in 3 of 20 samples, the largest value of unique reads was unmapped. We found 08_003_a had the largest number of *Gardnerella*-specific unique reads, with only 15.01% unmapped (see Table 2 in the online Data Supplement). Hierarchical clustering of the genomospecies percent values showed that clustering was primarily driven by taxonomic signatures, rather than treatment outcomes to metronidazole therapy (Fig. 4A). Visual interpretation of the heatmap showed 4 primary clusters of similar taxonomic profiles.

The first cluster containing only 08_003_a is largely dominated by a single genomospecies, as 90.56% of the *Gardnerella*-specific reads uniquely map to GS07 (Fig. 4B). The second cluster contains 06_006_a, 06_006_b, and 06_001_a, and is notable for their abundance of GS05 because it composes 91.79%, 84.20%, and 83.89%, respectively, of the *Gardnerella*-specific reads in these samples (Fig. 4B). The third cluster contains 17_004_b, 08_006_a, and 05_012_a, and has high levels of GS02 metatranscriptome unique reads at 79.95%, 70.93%, and 46.29%, respectively (Fig. 4B). The fourth cluster containing 13 of 20 of the samples is notable for having the highest mean percent of GS01 (30.51%) and GS03 (46.18%) (see Table 2 in the online Data Supplement). Unique reads for GS04 had the greatest prevalence within this cluster, as sample 13_019_a contains 29.15% of GS04 unique reads (Fig. 4B). These results indicate that the newly elucidated *Gardnerella* genomospecies can be identified as major contributors to *Gardnerella*-specific metatranscriptome reads during BV.

#### EXPRESSION OF TRANSLATION MACHINERY AND PUTATIVE VIRULENCE FACTORS BY GARDNERELLA DURING BV

Because metatranscriptomes provide a snapshot of the genes that are being transcribed, we finally wanted to investigate conserved features of *Gardnerella* gene expression during BV. To accomplish this, we used PanPhlAn to quantify gene coverage values for every gene in the pan-genome matrix created by Roary and the EggNOG COG annotation to identify enriched functions within highly expressed genes (1 SD above mean coverage levels across all samples) (see Table 9 in the online Data Supplement). Of these 224 highly expressed genes, 194 had a COG annotation. Sixteen of 26 COGs were identified in these 194 genes (Fig. 5A). Seven of 16 of these COGs had enrichment indices >1, indicating that they were disproportionally found among the highly expressed genes. The propagation and maintenance of proteins was found to be especially important, as COGs for transcription (K), translation (J), and protein turnover (O) were all enriched within the highly expressed genes (Fig. 5A). The other evident pattern was that COGs for carbohydrate

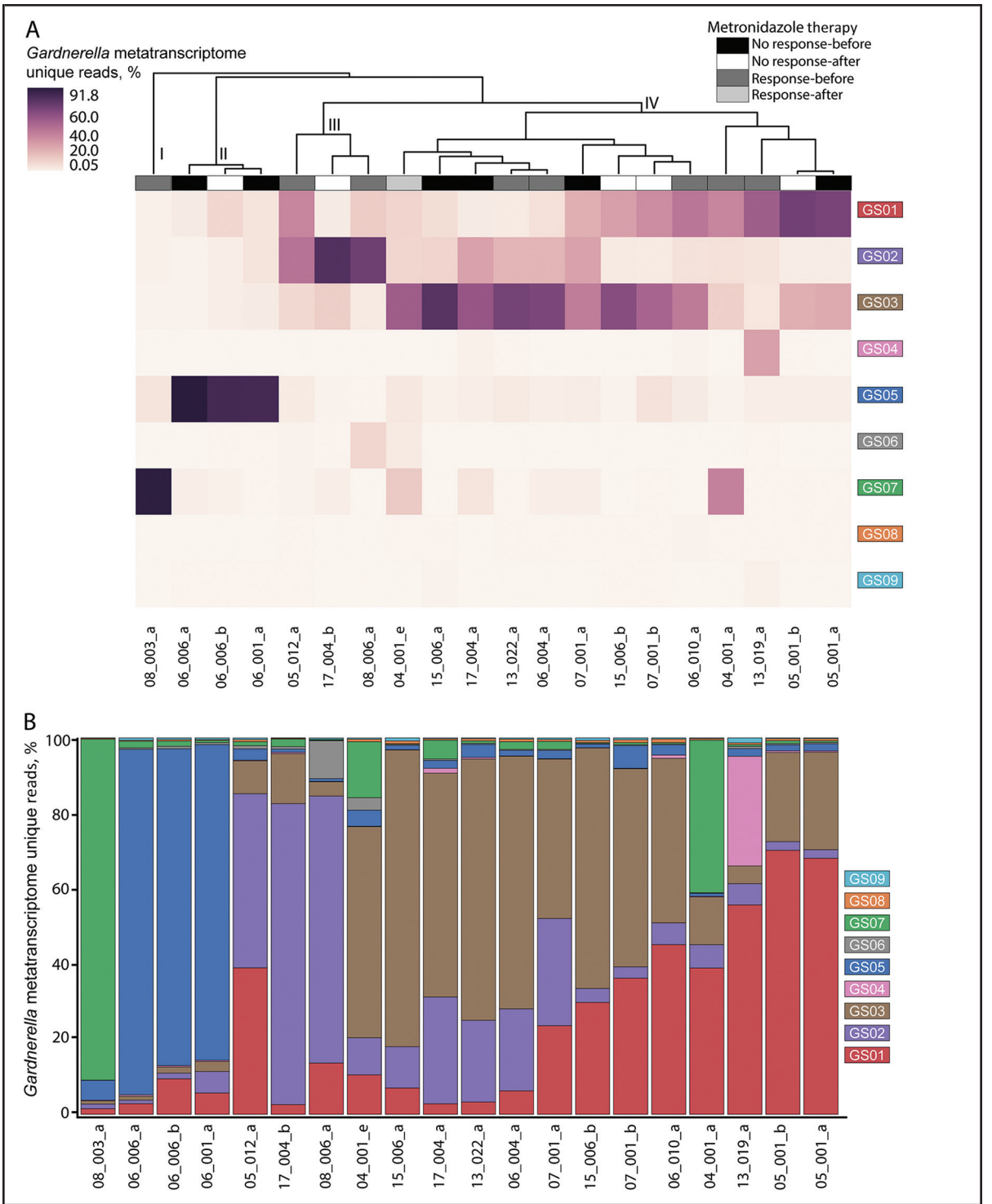


**Fig. 4. Newly elucidated genomospecies are identified in BV metatranscriptome samples.**

(A), Heatmap with hierarchal clustering from Centrifuge depicting the percentage of unique *Gardnerella* genomospecies-specific reads for each patient sample. Samples are labeled for clinical information related to efficacy of metronidazole treatment. Clusters of similar taxonomic profiles mentioned in the Results are labeled using roman numerals at the root of each cluster. (B), Stacked bar plots from Centrifuge showing the percentage of unique *Gardnerella* reads for each genomospecies within the metatranscriptome samples.

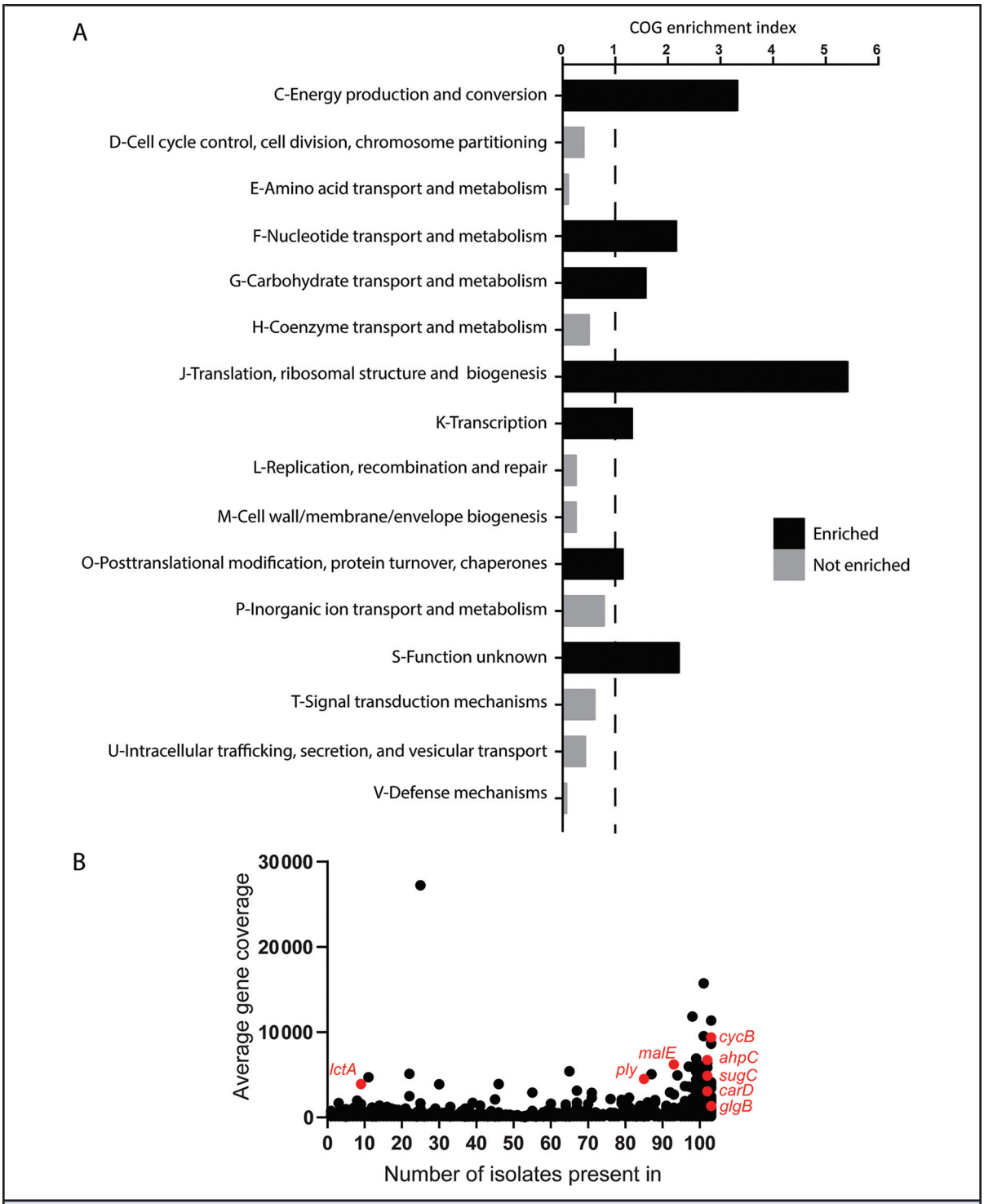


**Fig. 5.** ***Gardnerella*** **translation machinery and vaginolysin expression during BV.**

(A), Enrichment index scores for the COGs from EggNOG of the top 224 expressed *Gardnerella* pan-genome genes. Scores >1 are considered significantly enriched and are depicted in black; nonenriched COGs identified within the top 224 expressed genes are in gray. (B), Pan-genome plot depicting each of the 7402 genes as the number of isolates that harbor that gene (*x* value) and the PanPhlAn gene coverage value (*y* axis). Genes mentioned in article text are shown in red.

transport/metabolism (G) and energy production/conversion (C) were also enriched.

Analysis of individual genes among the highly expressed group identified several candidates involved in BV pathogenesis, including the known vaginolysin toxin gene (*ply*)[7] (Fig. 5B) *(31)*. Arguably the best studied pathogen in the Actinobacteria is *Mycobacterium tuberculosis*, and although *M. tuberculosis* and *G. vaginalis* exist in different human-associated environments, several of the highly expressed genes in BV transcriptomes are known virulence factors in the actinobacterial pathogen *M. tuberculosis*, suggesting a conserved importance in *G. vaginalis* (Fig. 5B) *(32–35)*. These include the transcriptional regulator *carD*, the trehalose import protein *sugC*, the glycan-branching enzyme *glgB*, and the oxidative stress response gene *ahpC. cycB*, a gene involved in maltose binding, was also found to be highly expressed, consistent with the importance of carbohydrate metabolism in the COG enrichment analysis *(36)*. Additionally, we found that the lantibiotic lacticin gene, *lctA*, was not strongly conserved in the *Gardnerella* pan-genome but was expressed at a high level *(37)*. These data identify several genetic loci that may be a conserved feature of *Gardnerella* pathogenesis in BV.

## Discussion

In microbial taxonomy, phylogenomic methods are being used with increasing frequency to specifically and accurately delineate new bacterial species from several previously known genera, including commensal and pathogenic members of *Klebsiella* and *Propionibacterium (22, 38)*. These delineations can have important implications in understanding the biology and clinical significance of these closely related organisms, and their potential differential contributions to health and disease in various hosts. For instance, although previously believed to be a benign environmental species, *Klebsiella variicola* strains can cause higher bladder infection titers in a mouse model of urinary tract infections compared with the canonical pathogen *Klebsiella pneumoniae (39)*. Similarly, presence of gene clusters encoding ABC transporters and phosphotransferase systems were differentially present within different genera of cutaneous propionibacteria, which may enable adaptation of *Cutibacterium*, *Pseudopropionibacterium*, and *Acidipropionibacterium* to different skin niches *(38)*. Our goal was to use initial binning of *Gardnerella* genomes into genomospecies using a variety of available tools and then explore differences in phylogeny, gene content, and metatranscriptome presence to provide insights into the biology of *Gardnerella* during BV.

Several previous reports have used whole-genome sequencing to compare pathogenic and commensal *Gardnerella* strains but did not systematically use taxonomic tools to define clear genomospecies *(5, 40–42)*. An early analysis between strains 409–05 (GS03), ATCC 14018 (GS01), and ATCC 14019 (GS01) found that 409–05 lacked mucin-degrading sialidases *(40)*. Our analysis corroborates this finding, as we did not identify any sialidases in GS03 genomes. A broader analysis of the core genome similarity between 17 strains found that they could be classified into 4 separate clades *(42)*. By applying taxonomic methods to compare 103 publicly available genomes annotated as *G. vaginalis*, we found exact concordance between these initial clades and our genomospecies, as Group 1 corresponds to GS01, Group 2 corresponds to GS02, Group 3 corresponds to GS05, and Group 4 corresponds to GS03 *(42)*. The differentiation of *G. vaginalis* into 4 clades had been recapitulated by alignment of just the cpn60 locus *(5)*. Again, we found complete concordance between these earlier delineations and our genomospecies as the subgroup A isolates corresponding to GS03, subgroup B were GS02, subgroup C were GS01, and subgroup D were GS05. Importantly, this latter study used pairwise ANI analysis to determine that ANI values ≤95% were found between the cpn60 group designations, suggesting that they constituted separate genomospecies *(5)*. One report on comparative analysis of 37 *Gardnerella* isolates identified 6 clades based off of conserved gene similarity and distinct gene presence *(43)*. This study, however, indicated that JCP8481A/JCP8481B (GS04) and 6119V4/00703Dmash (GS05) were both in clade 3A *(43)*. Conversely, they suggested that JCP775/ATCC 14019 and 00703C2mash/JCP8070, GS01 and GS02, respectively, were in separate clades *(43)*.

Recently, 1 analysis of 81 *Gardnerella* strains found that they represented 13 genomospecies, 3 of which were elucidated to be the new species *G. piotii*, *G. leopoldii*, and *G. swidsinkii*, using a combination of in silico tools and phenotypic assays *(6)*. This report did not include NR010, which was consistently annotated as a separate genomospecies in the 4 taxonomic tools that we used, making the current maximum number of *Gardnerella* genomospecies as 14. However, we show that for the type strains within *G. vaginalis*, *G. piotii*, *G. leopoldii*, *G. swidsinkii*, and representatives of the other 10 genomospecies, conflicting taxonomic information can arise from ANI vs tetranucleotide frequency vs AAI tool use *(9, 29, 30)*. The greatest discrepancies were between the JSpeciesWS ANIb method (14 species) and the Kostas laboratory AAI tool (8 genomospecies). For the purposes of our study, we took the conservative consensus for the number of genomospecies across the 4 tools, which had

[7] Genes: *ply*, pneumolysin; *carD*, G-protein-coupled receptor; *sugC*, *sugar ABC transporter ATP-binding protein SugC*; *glgB*, *1,4-alpha-glucan branching enzyme*; *ahpC*, alkyl hydroperoxide reductase, AhpC component; *cycB*, cyclin B; *lctA*, •••.

exact concordance with the tetranucleotide frequency analysis in JSpeciesWS. Strikingly, in 3 of 4 of the tools tested, the type strains *G. leopoldii* UGENT 06.41 (T) and *G. swidsinkii* GS 9838–1 (T) were annotated as being the same genomospecies, conflicting the phenotypic results *(6)*. It is possible that these 2 strains may represent different subspecies of the same *Gardnerella* species. The approximate maximum likelihood tree and nearest neighbor network both showed that the genomospecies can be readily distinguished by the similarity of their 200 core genes. Similarly, the genomospecies had vastly different repertories of accessory genomes, except the group that contained GS05, GS04, and GS09. These differences may be important for adaptation to the vaginal microenvironment or BV pathology because some of the genes driving this difference include those known for virulence (e.g., sialidases) *(44, 45)*.

Given that GS07 and GS04 through GS09 were unknown in these prior genomic studies and that there was occasionally ambiguity between previous group determinations, we used metatranscriptome sequencing reads from BV samples from women before and after metronidazole, to determine whether the *Gardnerella* genomospecies could be implicated in BV pathogenesis *(12)*. Similar to the original study, our results did not find any association between the presence of specific *Gardnerella* genomospecies and resistance to metronidazole, even with the improved taxonomic classification *(12)*. However, we detected unique transcripts to all genomospecies in every sample. When we combined the taxonomic information to identify conserved features of *Gardnerella* pathogenesis in BV, we found enrichment for COGs involved in carbohydrate transport and conversion into energy as well as propagation of protein machinery. Importantly, the known virulence factor *ply* and several genes implicated in the pathogenesis of the Actinobacteria *M. tuberculosis* were some of the genes with the highest coverage values.

The major limitation of this investigation is that as a retrospective genomic analysis, we do not have immediate access to the isolates for species characterization. Comprehensive analysis of differences in membrane lipids and biochemical utilizations would be necessary to correctly classify these 9 genomospecies into species, with proper Latin nomenclature *(6)*. Another limitation relates to the fact that our analysis includes metatranscriptome reads rather than metagenomic reads; thus, we are unable to accurately quantify absolute abundance of the different genomospecies within the BV samples because it is possible that a genomospecies could compose a smaller overall fraction but express a large number of genes.

In summary, we performed a taxonomic investigation of >100 *Gardnerella* genomes. Our consensus analysis using several different tools found that *Gardnerella* was composed of 9 different genomospecies that could be readily distinguished from one another by the similarity of their core genes and by the presence/absence of their accessory genes. Although there was a high level of similarity in overall COG presence, certain features were enriched within specific genomospecies, supporting differentiation into specific niches within the vaginal microenvironment and possibly alternative mechanisms for BV pathology. We found evidence of the presence of these genomospecies in publicly available BV metatranscriptome reads and determined that translation machinery and putative virulence factors constitute a conserved *Gardnerella* transcriptome during BV. This work may enable future epidemiological investigations on the presence of the different genomospecies during health and disease states, mechanistic insight into their individual or combined roles in causing BV, and development of precision treatments tailored for specific genomospecies.

---

**Author Contributions:** *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 4 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved.*

R.F. Potter performed the in silico analysis and drafted the manuscript. C.-A.D. Burnham and G. Dantas advised the analysis and edited the manuscript.

**Authors' Disclosures or Potential Conflicts of Interest:** *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

**Employment or Leadership:** None declared.
**Consultant or Advisory Role:** None declared.
**Stock Ownership:** None declared.
**Honoraria:** None declared.
**Research Funding:** R.F. Potter, a National Institute of General Medical Sciences training grant through award T32 GM007067 (PI: James Skeath) and the Monsanto/Bayer Excellence Fund graduate fellowship; G. Dantas, awards through the National Institute of Allergy and Infectious Diseases, and the Eunice Kennedy Shriver National Institute of Child Health & Human Development, of the National Institutes of Health under award numbers R01AI123394 and R01HD092414, respectively.
**Expert Testimony:** None declared.
**Patents:** None declared.

**Role of Sponsor:** The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

**Acknowledgments:** The authors thank members of the Dantas laboratory for insightful discussions of the results and conclusions, especially Alaric W. D'Souza for his assistance with RStudio.

## References


1. Bagnall P, Rizzolo D. Bacterial vaginosis: a practical review. JAAPA 2017;30:15–21.
2. Gardner HL, Dukes CD. Haemophilus vaginalis vaginitis: a newly defined specific infection previously classified non-specific vaginitis. Am J Obstet Gynecol 1955;69: 962–76.
3. Mikamo H, Sato Y, Hayasaki Y, Hua YX, Tamaya T. Vaginal microflora in healthy women with Gardnerella vaginalis. J Infect Chemother 2000;6:173–7.
4. Hickey RJ, Forney LJ. Gardnerella vaginalis does not always cause bacterial vaginosis. J Infect Dis 2014; 210:1682–3.
5. Schellenberg JJ, Paramel Jayaprakash T, Withana Gamage N, Patterson MH, Vaneechoutte M, Hill JE. Gardnerella vaginalis subgroups defined by cpn60 sequencing and sialidase activity in isolates from Canada, Belgium and Kenya. PLoS One 2016;11:e0146510.
6. Vaneechoutte M, Guschin A, Van Simaey L, Gansemans Y, Van Nieuwerburgh F, Cools P. 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 2019;69:679–87.
7. Tindall BJ, Rossello-Mora R, Busse HJ, Ludwig W, Kampfer P. Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst Evol Microbiol 2010;60:249–66.
8. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, et al. Microbiology in the post-genomic era. Nat Rev Microbiol 2008;6:419–30.
9. Ciufo S, Kannan S, Sharma S, Badretdin A, Clark K, Turner S, et al. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. Int J Syst Evol Microbiol 2018;68: 2386–92.
10. Hata H, Natori T, Mizuno T, Kanazawa I, Eldesouky I, Hayashi M, et al. Phylogenetics of family Enterobacteriaceae and proposal to reclassify Escherichia hermannii and Salmonella subterranea as Atlantibacter hermannii and Atlantibacter subterranea gen. nov., comb. nov. Microbiol Immunol 2016;60:303–11.
11. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–9.
12. Deng ZL, Gottschick C, Bhuju S, Masur C, Abels C, Wagner-Dobler I. Metatranscriptome analysis of the vaginal microbiota reveals potential mechanisms for protection against metronidazole in bacterial vaginosis. mSphere 2018;3.
13. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 2009;106:19126–31.
14. Richter M, Rossello-Mora R, Oliver Glockner F, Peplies J. JspeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. Bioinformatics 2016;32:929–31.
15. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;31:3691–3.
16. Loytynoja A. Phylogeny-aware alignment with PRANK. Methods Mol Biol 2014;1079:155–70.
17. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 2010;5:e9490.
18. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol 2013; 30:1224–8.
19. Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. Efficient inference of recent and ancestral recombination within bacterial populations. Mol Biol Evol 2017;34:1167–82.
20. Letunic I, Bork P. Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 2007;23:127–8.
21. Huson DH. Splitstree: analyzing and visualizing evolutionary data. Bioinformatics 1998;14:68–73.
22. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. Proc Natl Acad Sci U S A 2015;112:E3574–81.
23. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol 2016;17:238.
24. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 2019;47:D309–D14.
25. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 2018; 46:W95–W101.
26. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res 2016;26:1721–9.
27. Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods 2016;13:435–8.
28. Murray GL, Tsyganov K, Kostoulias XP, Bulach DM, Powell D, Creek DJ, et al. Global gene expression profile of Acinetobacter baumannii during bacteremia. J Infect Dis 2017;215:S52–S7.
29. Noble PA, Citek RW, Ogunseitan OA. Tetranucleotide frequencies in microbial genomes. Electrophoresis 1998;19:528–35.
30. Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes. J Bacteriol 2005;187: 6258–64.
31. Gelber SE, Aguilar JL, Lewis KL, Ratner AJ. Functional and phylogenetic characterization of vaginolysin, the human-specific cytolysin from Gardnerella vaginalis. J Bacteriol 2008;190:3896–903.
32. Weiss LA, Harrison PG, Nickels BE, Glickman MS, Campbell EA, Darst SA, Stallings CL. Interaction of CarD with RNA polymerase mediates Mycobacterium tuberculosis viability, rifampin resistance, and pathogenesis. J Bacteriol 2012;194:5621–31.
33. Kalscheuer R, Weinrick B, Veeraraghavan U, Besra GS, Jacobs WR Jr. Trehalose-recycling ABC transporter LpqY-SugA-SugB-SugC is essential for virulence of Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 2010;107:21761–6.
34. Dkhar HK, Gopalsamy A, Loharch S, Kaur A, Bhutani I, Saminathan K, et al. Discovery of Mycobacterium tuberculosis alpha-1,4-glucan branching enzyme (GlgB) inhibitors by structure- and ligand-based virtual screening. J Biol Chem 2015;290:76–89.
35. Master SS, Springer B, Sander P, Boettger EC, Deretic V, Timmins GS. Oxidative stress response genes in Mycobacterium tuberculosis: role of ahpC in resistance to peroxynitrite and stage-specific survival in macrophages. Microbiology 2002;148:3139–44.
36. Kamionka A, Dahl MK. Bacillus subtilis contains a cyclodextrin-binding protein which is part of a putative ABC-transporter. FEMS Microbiol Lett 2001;204: 55–60.
37. Furgerson Ihnken LA, Chatterjee C, van der Donk WA. In vitro reconstitution and substrate specificity of a lantibiotic protease. Biochemistry 2008;47:7352–63.
38. Scholz CF, Kilian M. The natural history of cutaneous propionibacteria, and reclassification of selected species within the genus Propionibacterium to the proposed novel genera Acidipropionibacterium gen. nov., Cutibacterium gen. nov. and Pseudopropionibacterium gen. nov. Int J Syst Evol Microbiol 2016;66: 4422–32.
39. Potter RF, Lainhart W, Twentyman J, Wallace MA, Wang B, Burnham CA, et al. Population structure, antibiotic resistance, and uropathogenicity of Klebsiella variicola. mBio 2018;9.
40. Yeoman CJ, Yildirim S, Thomas SM, Durkin AS, Torralba M, Sutton G, et al. Comparative genomics of Gardnerella vaginalis strains reveals substantial differences in metabolic and virulence potential. PLoS One 2010; 5:e12411.
41. Harwich MD Jr, Alves JM, Buck GA, Strauss JF 3rd, Patterson JL, Oki AT, et al. Drawing the line between commensal and pathogenic Gardnerella vaginalis through genome analysis and virulence studies. BMC Genomics 2010;11:375.
42. Ahmed A, Earl J, Retchless A, Hillier SL, Rabe LK, Cherpes TL, 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 2012; 194:3922–37.
43. Cornejo OE, Hickey RJ, Suzuki H, Forney LJ. Focusing the diversity of Gardnerella vaginalis through the lens of ecotypes. Evol Appl 2018;11:312–24.
44. Govinden G, Parker JL, Naylor KL, Frey AM, Anumba DOC, Stafford GP. Inhibition of sialidase activity and cellular invasion by the bacterial vaginosis pathogen Gardnerella vaginalis. Arch Microbiol 2018;200: 1129–33.
45. Hardy L, Jespers V, Van den Bulck M, Buyze J, Mwambarangwe L, Musengamana V, et al. The presence of the putative Gardnerella vaginalis sialidase a gene in vaginal specimens is associated with bacterial vaginosis biofilm. PLoS One 2017;12:e0172522.