Bioinformatics Faculty Publications

Faculty Publications

11-18-2016

Genomes of Gardnerella Strains Reveal an Abundance of Prophages within the Bladder Microbiome

Kema Malki
Loyola University Chicago

Jason W. Shapiro
Loyola University Chicago

Travis Kyle Price
Loyola University Chicago

Evann Elizabeth Hilt
Loyola University Chicago

Krystal Thomas-White
Loyola University Chicago

Follow this and additional works at: https://ecommons.luc.edu/bioinformatics_facpub

Part of the Bacteriology Commons, Bioinformatics Commons, Genomics Commons, and the Virology Commons

See next page for additional authors

Recommended Citation

Authors

Kema Malki, Jason W. Shapiro, Travis Kyle Price, Evann Elizabeth Hilt, Krystal Thomas-White, Trina Sircar, Amy B. Rosenfeld, Michael J. Zilliox, Alan J. Wolfe, and Catherine Putonti

Genomes of *Gardnerella* Strains Reveal an Abundance of Prophages within the Bladder Microbiome

Kema Malki[1], Jason W. Shapiro[1,2], Travis K. Price[3], Evann E. Hilt[3], Krystal Thomas-White[3], Trina Sircar[2], Amy B. Rosenfeld[4], Gina Kuffel[4], Michael J. Zilliox[5], Alan J. Wolfe[3], Catherine Putonti[1,2,6]*

1 Department of Biology, Loyola University Chicago, Chicago, Illinois, United States of America,
2 Bioinformatics Program, Loyola University Chicago, Chicago, Illinois, United States of America,
3 Department of Microbiology and Immunology, Stritch School of Medicine, Loyola University Chicago, Maywood, Illinois, United States of America, 4 Center for Biomedical Informatics, Loyola Genomics Facility, Loyola University Chicago, Maywood, Illinois, United States of America, 5 Department of Public Health Sciences, Stritch School of Medicine, Loyola University Chicago, Maywood, Illinois, United States of America, 6 Department of Computer Science, Loyola University Chicago, Chicago, Illinois, United States of America

* cputonti@luc.edu

Abstract

Bacterial surveys of the vaginal and bladder human microbiota have revealed an abundance of many similar bacterial taxa. As the bladder was once thought to be sterile, the complex interactions between microbes within the bladder have yet to be characterized. To initiate this process, we have begun sequencing isolates, including the clinically relevant genus *Gardnerella*. Herein, we present the genomic sequences of four *Gardnerella* strains isolated from the bladders of women with symptoms of urgency urinary incontinence; these are the first *Gardnerella* genomes produced from this niche. Congruent to genomic characterization of *Gardnerella* isolates from the reproductive tract, isolates from the bladder reveal a large pangenome, as well as evidence of high frequency horizontal gene transfer. Prophage gene sequences were found to be abundant amongst the strains isolated from the bladder, as well as amongst publicly available *Gardnerella* genomes from the vagina and endometrium, motivating an in depth examination of these sequences. Amongst the 39 *Gardnerella* strains examined here, there were more than 400 annotated prophage gene sequences that we could cluster into 95 homologous groups; 49 of these groups were unique to a single strain. While many of these prophages exhibited no sequence similarity to any lytic phage genome, estimation of the rate of phage acquisition suggests both vertical and horizontal acquisition. Furthermore, bioinformatic evidence indicates that prophage acquisition is ongoing within both vaginal and bladder *Gardnerella* populations. The abundance of prophage sequences within the strains examined here suggests that phages could play an important role in the species' evolutionary history and in its interactions within the complex communities found in the female urinary and reproductive tracts.

Background

Gardnerella, a member of the Bifidobacteriaceae family, is a genus of facultative anaerobes within the vaginal microbiota of both healthy women and those diagnosed with bacterial vaginosis (BV) [1–4]. Similarly, our group found *Gardnerella* in urine collected from adult female bladders by transurethral catheter [5–8]. This corresponds with microbiome studies of voided urine: *Gardnerella* was present regardless of sex or symptom status [9–13]. Furthermore, the bladders of healthy individuals include other bacterial taxa also detected within the vaginal microbiota [5–10,12–14]. To date, thirty-nine *G. vaginalis* isolates from the vagina or endometrium have been sequenced [15–19], including four complete genomes; the remaining genomes are represented as scaffolds or contigs. Analyses of *G. vaginalis* genomes found evidence of a large pangenome that consists of a modestly sized core genome in addition to strain-specific genes [15,20].

Prior investigations of *G. vaginalis* genomes from the reproductive tract have uncovered indications of substantial horizontal gene transfer (HGT), including the acquisition of genes from other human-associated taxa [15,20–21]. In addition to natural competence, *G. vaginalis* may also include phage-mediated gene transfer, as coding regions of bacteriophage (phage) origin are ubiquitous within these genomes [20]. Similarly, genomic sequences from other bacterial taxa within the vaginal microbiota also contain parts of or entire temperate phage genomes [22–23]. Prior studies have posited that lysogenic lactobacilli phages could contribute to a shift in the vaginal microbiota leading to BV (for a review, see [4]). Phages have been found to play a crucial role in the structuring of microbial communities, including those residing within the human body [24], driving bacterial genetic diversity [25] and adaptation to changes in the environment [26]. Although several phages induced from vaginal lactobacilli have been identified [27–28], currently no phages have been characterized for *Gardnerella*. While evidence suggests that phages are likely contributors to HGT in commensal communities (e.g. [22,26] and review [29]), the extent of their effect on the human microbiome is just now being explored (e.g., [30–32]). Within many other microbiota, including the bladder, the virome remains largely unexplored.

Herein, we present the genomic sequences of four *Gardnerella* strains isolated from the bladders of adult women with symptoms of urgency urinary incontinence (UUI). Comparative genomics between these strains and publicly available isolates revealed a highly conserved core genome across the bladder and vaginal isolates. Analogous to prior observations for this species, the *Gardnerella* strains isolated from the bladder also contain a large number of prophage gene sequences. The pervasiveness of prophage sequences in *Gardnerella* genomes from both the female urinary and reproductive tracts motivated our thorough bioinformatic investigation. A comprehensive interrogation of the over 400 annotated prophage gene sequences identified here provides insight into the adaptive ability of *Gardnerella*, as well as the larger community of phages within the female urinary community.

Results and Discussion

Genome Features

Four *Gardnerella* strains were previously isolated from the bladders of four different female patients with symptoms of urgency urinary incontinence [8,33]. While MALDI-TOF identified two of these strains as *G. vaginalis* (henceforth denoted as Gv18-4 and Gv23-12), it did not resolve the other two strains to the species level (G26-12 and G30-4). 16S rRNA gene sequencing for each isolate confirmed the strains belong to the genus *Gardnerella* (S1 Fig). All four isolates were sequenced, assembled, and annotated (see Methods) and are publicly available via NCBI. The genomic sequences varied slightly in size from 1.48 to 1.62 Mbp (Table 1).

Table 1. Genomic characteristics of four *Gardnerella* strains isolated from female bladders.

Genome	G26-12	G30-4	Gv18-4	Gv23-12
Length	1,484,647	1,590,395	1,624,172	1,595,255
GC%	42.5	41.9	40.9	41.2
# Protein coding regions	1172	1242	1240	1243
# RNAs	48	48	45	46
# CRISPR arrays	5	8	2	4
# Scaffolds	175	183	324	301
Max Scaffold Length	57,529	42,968	33,481	45,660
N50 Length	16,568	20,497	11,232	11,376
Average Coverage	526	714	197	631
INSDC ID	LWSR00000000	LXJL00000000	LWSP00000000	LWSQ00000000

doi:10.1371/journal.pone.0166757.t001

Homologs, as well as strain-specific coding regions, were identified within the genomes of these bladder *Gardnerella* strains (Fig 1A). Although Gv18-4 and Gv23-12 were indistinguishable on the basis of their 16S rRNA sequences (S1 Fig), instances of both strain-specific gene acquisition and loss within their genomes were detected (S1 Table). Annotation also revealed the type I CRISPR/Cas system in all four genomes, with multiple CRISPR loci per genome (Table 1). Comparison between the two strains identified by MALDI-TOF analysis to the species level (Gv18-4 and Gv23-12) and those identified to the genus level (G26-12 and G30-4) revealed disparity in the number of genes with the annotated functionalities of carbohydrate metabolism, membrane transport, and cell wall and capsule synthesis (Fig 1B); however, the genetic variation observed between the Gv and G strains is statistically significant only for loci that regulate carbohydrate metabolism. Consistent with previous studies that illustrated the ability of different *Gardnerella* strains to ferment and use different carbohydrates [Harwich et al. 2010], annotation of the genome for G26-12 and G30-4 suggest that they lack coding regions with the following functionalities: chitin and N-acetylglucosamine utilization,
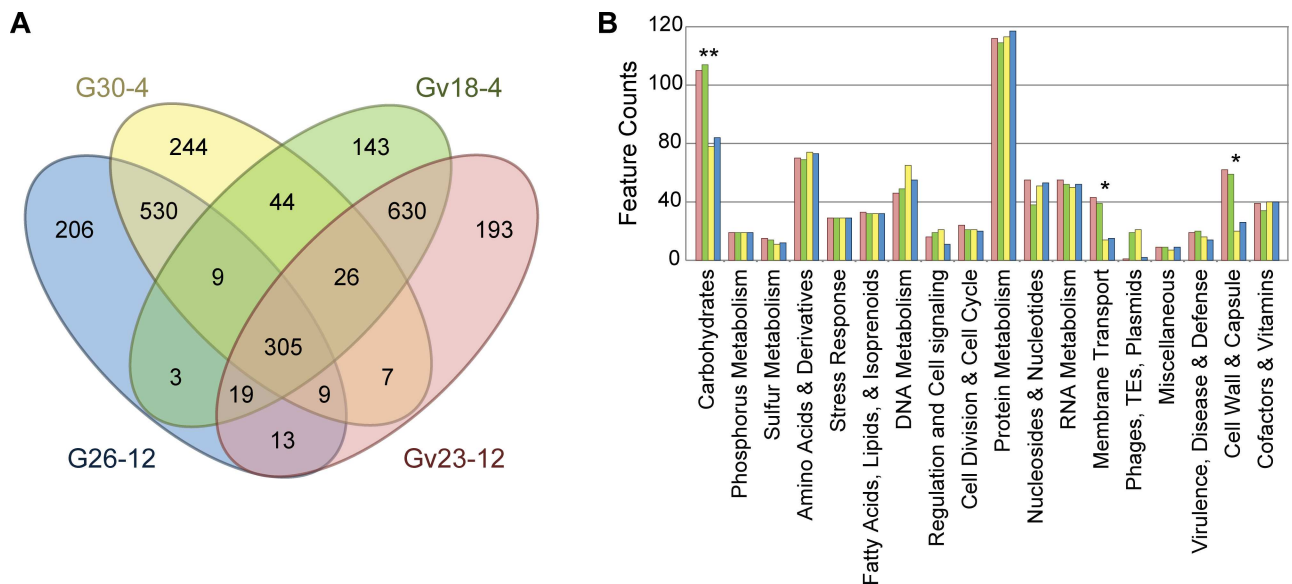
Fig 1. Characterization of predicted coding regions. (A) Homology and **(B)** annotated functionality within the four bladder *Gardnerella* strains. (*******p* value <0.05 between the two predicted *Gardnerella* strains and the two predicted *G. vaginalis* strains; *p value < 0.1). Strains are represented using the same colors in both panels.

doi:10.1371/journal.pone.0166757.g001

trehalose biosynthesis, trehalose uptake and utilization, lactose and galactose uptake and utilization, lactose utilization, formaldehyde assimilation, xylose utilization, deoxyribose and deoxynucleoside catabolism, D-gluconate and ketogluconates metabolism. The absence of these coding regions within the G26-12 and G30-4 genome sequences was confirmed experimentally via PCR (see Methods).

The genomes of these bladder strains were further examined with respect to the presence/absence of coding regions associated with virulence, as previously defined by Yeoman *et al.* [20] (S2 Table). Genes associated with biofilm formation (glycosylases and glycosyltransferases) and epithelial adhesion (fimbria/pili) were identified in all four genomes; however, type-1 fimbrial precursors were found only within the genomes of G26-12 and G30-4. Genes associated with antibiotic/antimicrobial resistance, including those that encode the ABC-type multidrug transport system and the DedA protein (which has been shown to be required for drug resistance in *E. coli* [34]) were present within all four genomes. However, tetracycline resistance proteins were encoded within the Gv18-4, Gv23-12, and G26-12 strains, but not in the G30-4 strain. Genes associated with protection or evasion from the immune response (alkyl hydroperoxide reductase and Rib-family surface protein) were found within all four of the bladder strains. Furthermore, the coding region for the vaginolysin (VLY) gene, which is selective for human cells [20,35], was highly conserved within all four strains, in spite of the two amino acid substitutions that separate the Gv18-4 strain from the type strain for the genus ATCC 14019 (S2 Fig). Finally, the *Gardnerella* strains G26-12 and G30-4 contained an annotated rRNA methyltransferase associated with *Gardnerella* cytotoxicity [20] and with haemolytic activity in other bacterial species [36]. Interestingly, only the Gv18-4 and Gv23-12 strains contained genes that encode sialidase, which has been experimentally proven to contribute to mucin degradation in BV [16], although presence of the gene is not predictive of actual sialidase activity [37] and thus warrants further investigation.

Phylogenetic Analysis and the *Gardnerella* Pangenome

In an effort to assess the similarity/difference between *Gardnerella* strains isolated from the female bladder and those isolated from the female reproductive system, all publicly available sequenced strains isolated from the vagina or endometrium were retrieved from NCBI (S3 Table). This set includes 35 complete, scaffold, or contig genome sequences. Extending beyond the 16S rRNA gene marker, the evolutionary history of this genus was considered by investigating sequence homologies within the "core" *Gardnerella* genome (Methods; S4 Table). In total, 183 genes were identified as belonging to this core gene set, less than identified in prior studies of far fewer genomes [15,20]. As shown in Fig 2, the strains isolated from the bladder are not monophyletic, establishing that the *Gardnerella* core genome does not correlate with a single isolation location. Examination of all coding sequences for all 39 investigated genomes revealed no gene(s) exclusive to the bladder strains. While the vast majority of the genes within the core set demonstrated an evolutionary history identical to that of the species tree (Fig 2), genes indicative of intragenic recombination, such as VLY (S2 Fig), were also identified.

Nevertheless, the species tree (Fig 2) we derived largely concurs with the tree produced by Ahmed et al. [15]. In comparison to the phylogeny of Ahmed et al. [15], which included 17 genomes and a core genome of 473 genes, their Group 3 and 4 are both contained within Clade A in Fig 2. Our Clade B corresponds to their Group 2 and our Clade C is within their Group 1; only two of the genomes within Clade C were included in the prior analysis of Ahmed et al. [15]. As expected, each clade had significantly more homologous genes (404, 799, 762 and 635 for Clades A through D, respectively) (Fig 2A), a likely residual of undersampling, biased sampling, and/or clade-specific functional conservation.
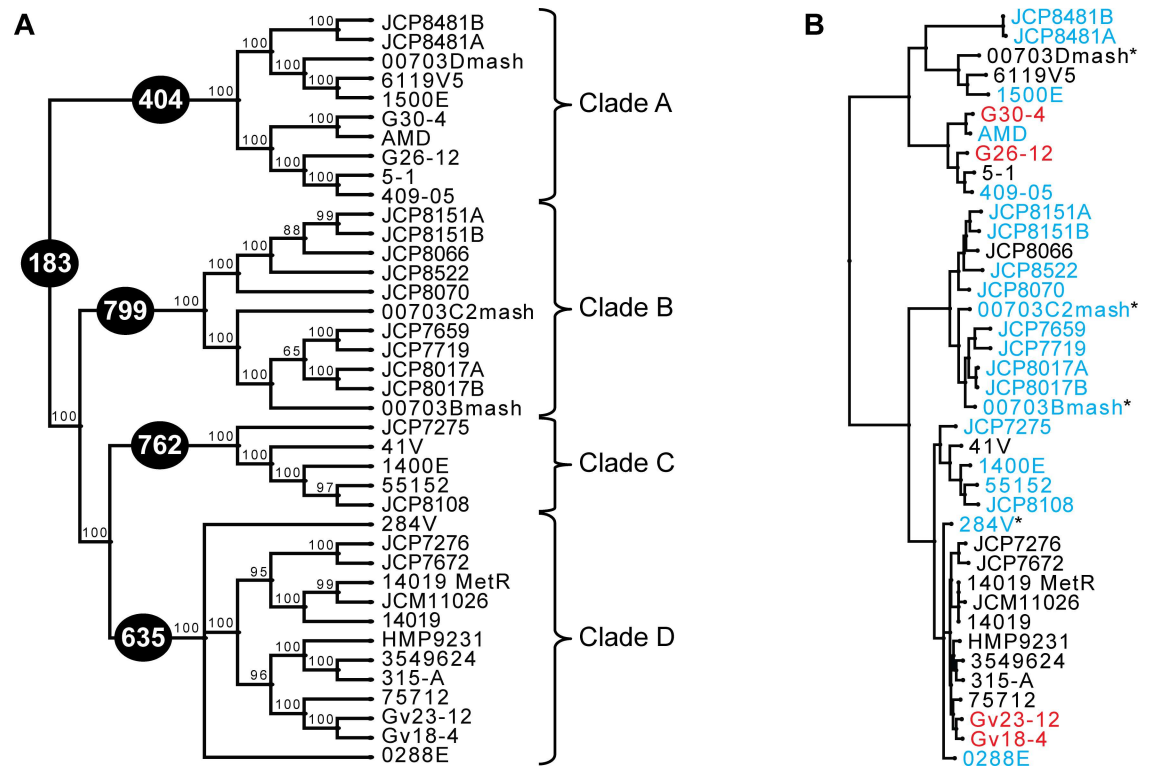
Fig 2. Phylogenetic Analysis of *Gardnerella* Strains. Maximum Likelihood species tree of *Gardnerella* strains based upon sequence homology within the core gene set. **(A)** Phylogenetic tree listing branch supports and distinction of the four clades within the tree. Numbers within black circles indicate the number of homologous genes within each clade and the core *Gardnerella* genome. **(B)** Maximum Likelihood tree including branch lengths and isolation information with respect to location and diagnosed symptom. Strains isolated from the bladder and sequenced in this study are labeled in red; strains listed in green were isolated from the vagina/endometrium of BV+ patients; strains in light blue were isolated the vagina/endometrium of STD+ patients. All remaining strains (indicated in black font) were isolated from the vagina or endometrium.

doi:10.1371/journal.pone.0166757.g002

The bladder strains G26-12 and G30-4 were determined to belong to Clade A, which also includes three strains isolated from BV+ patients (Fig 2B). While the phylogeny based upon the core gene set does not correspond with isolation location, some clades appear to have a higher incidence of strains isolated from patients diagnosed as BV+. Because genes associated with virulence are not exclusively present within genomes isolated from symptomatic patients [15,20], it is not surprising that several are included within the core gene set identified here. The lack of correspondence between phylogenetic history and symptoms is supported by prior studies, which hypothesized that gene expression variation within *G. vaginalis* strains may trigger BV development [38] and single point mutations may result in greater potential for cytotoxicity [39]. Thus, there appears to be no single gene that correlates with BV symptoms. This is, however, not surprising given that *G. vaginalis* strains are present within both the BV- and BV+ vagina [40], and it is but one of the taxa that coincides with BV symptoms [3].

Mobile Elements

The four bladder *Gardnerella* genomes varied in their number of ORFs predicted to be viral (bacteriophage) in origin (see Methods). The number of prophage gene sequences per genome had no correspondence to evolutionary history (Fig 2). This is true for the four *Gardnerella* strains isolated from the bladder, as well as for the 35 strains isolated from the reproductive

system. For example, the *G. vaginalis* strain JCP8522, a vaginal BV+ strain [16], has only one ORF annotated as a phage gene. In contrast, the Gv18-4 strain sequenced as part of this study contains the most, 33, phage-like gene sequences within a *Gardnerella* genome to date. S5 Table lists the 442 annotated prophage genes within the 39 genomes examined. Due to the incomplete (scaffold/contig) status of the majority of the genomes included in the analysis here, we evaluated predicted prophage genes (median length 939 bp) individually.

To identify homologs, the nucleotide sequences for all annotated prophage genes across all *Gardnerella* genomes were examined. Based upon the combination of sequence identity and query coverage (see Methods), 104 clusters of orthologous prophage genes were identified (S1 File). Within larger clusters, prophage homolog sequences were highly conserved between genomes. Forty-nine clusters, however, included only a single prophage gene sequence, indicative of frequent independent acquisition of viral sequences. This is further supported by the variation in prophage genes identified within strains isolated from the same patient. While the genomes of JCP8481A and JCP8481B (both isolated from a single patient as denoted by A and B [16]) contain the same set of prophage gene sequences, the genomes of JCP8151A/B and JCP8017A/B do not (Fig 3; S5 Table); this captures the likely prevalence of lysogenic phage excision, as well as gene loss, events. Likewise, the genomes of 00703Bmash, 00703C2mash, and 00703Dmash, which are from sequential isolates from the same patient [15], vary in their prophage gene content (Fig 3; S5 Table). This diversity, even within a single patient, suggests intra-host prophage gene gain/loss.

Examination of the 37 largest prophage gene clusters suggested that prophage introduction within the *Gardnerella* genus has occurred by both vertical and lateral inheritance. These larger clusters include orthologous prophage gene sequences present within five or more *Gardnerella* genomes; the most pervasive prophage gene sequences were present within 21 of the 39 *Gardnerella* genomes examined here. As shown in Fig 3, several of the larger prophage gene clusters have representatives across the entire phylogenetic tree and some sets of clusters appeared to have been lost within particular lineages. Half of the 104 prophage gene clusters identified exhibit little to no resemblance to any sequence within the current GenBank nr/nt nucleotide database of characterized phage sequences; this includes several of the prophage genes within the larger clusters (represented as open circles in Fig 3). For those prophage gene sequences that exhibit homology to characterized phage genes, hits were frequently identified to the genomes of *Bacillus*-, *Mycobacterium*- and *Staphylococcus*-infecting phages. Previous targeted gene surveys of the bladder have routinely found *Staphylococcus* within the community [6,7,33].

Although *Gardnerella*-infecting phages have yet to be isolated, the abundance of prophage gene sequences in the genomes (S5 Table) and the presence of the CRISPR/Cas-system (Table 1) suggests that phages capable of infecting *Gardnerella* spp. exist. This assumption is corroborated by prior analyses of spacer sequences within 21 *G. vaginalis* genomes, in which 70.7% of the spacers showed no homology to sequences within the GenBank database [21]. While many of the clusters exhibiting homology to sequences in GenBank identified lytic phages annotated as infecting an array of different bacterial genera, analysis of eleven of the 104 clusters resulted in the description of phages thought to infect a single taxon. These included several of the larger clusters shown in Fig 3: clusters 12, 15, and 30 were found to have homology to *Bacillus*-infecting phages; clusters 22 and 35 were determined to be similar to phages infecting *Mycobacterium* species; while cluster 29 exhibited a likeness to *Staphylococcus*-infecting phages (S5 Table). Yet, all of these phages are of the family *Siphoviridae*. Homologies identified for these prophage sequences provide a foundation for future work in the isolation of *Gardnerella*-specific phages. More broadly, these prophage gene sequences provide insight into the phage community within the human microbiota.
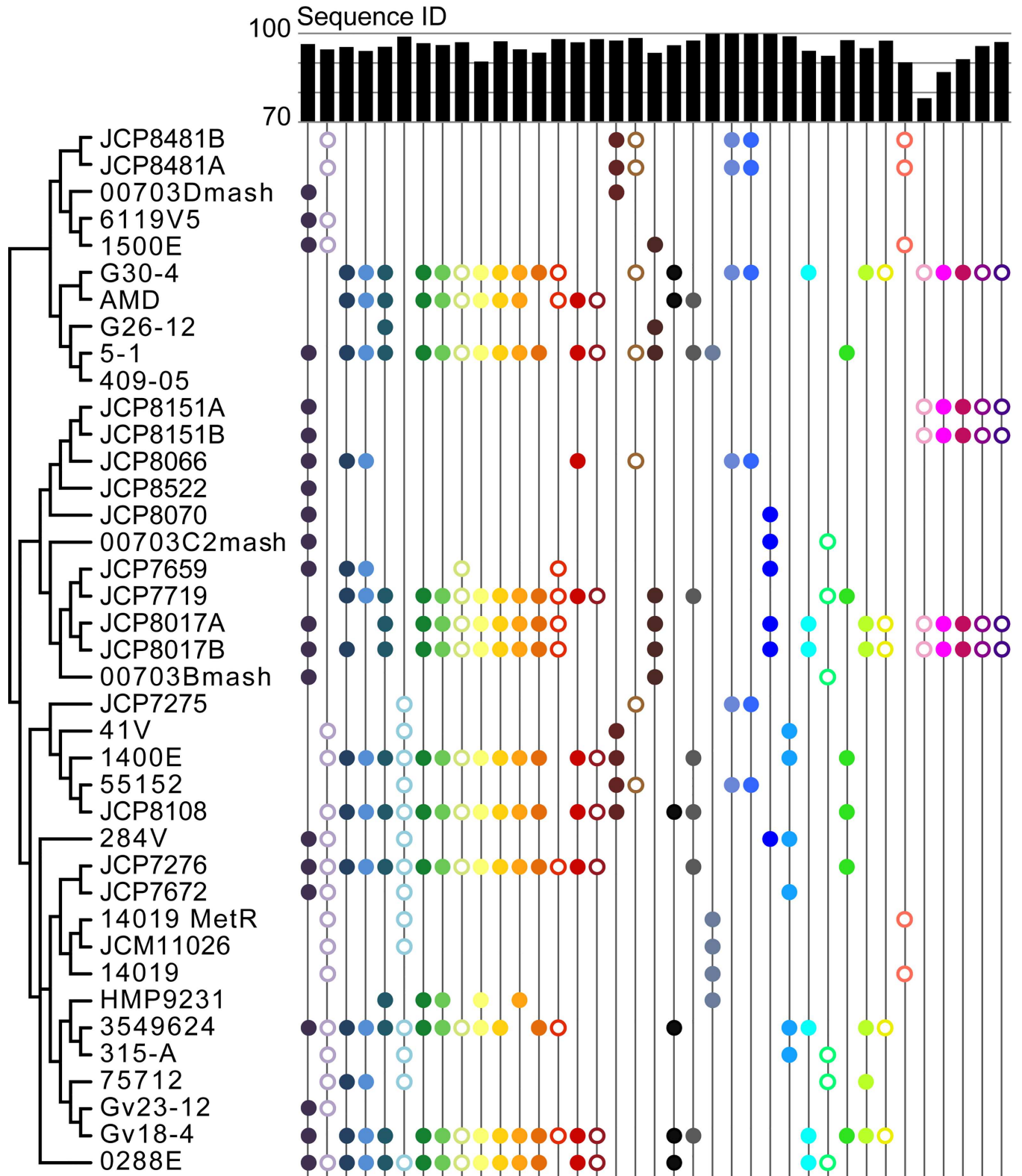
Fig 3. Presence/absence of prophage gene sequences from larger clusters. Clusters of orthologous prophage genes found within five or more of the 39 *Gardnerella* genomes are shown using a distinct color. Open circles indicate that the prophage sequences belonging to the particular cluster show no/poor sequence homology to characterized lytic phages. Closed circles indicate moderate (query coverage ≥80%, nucleotide sequence identity ≥35%) to high homology to a phage genome sequence record in NCBI. Nucleotide sequence identity for the prophage genes in each cluster is indicated by the bar chart at the top of the figure.

Lateral gene acquisition is ubiquitous; however, there is a strong discordance between the evolutionary tree (S1 Fig; Fig 2) and trees derived from strain:prophage gene presence/absence profiles (results not shown). To distinguish clusters acquired via HGT from those linearly inherited from a common ancestor, the genomes of 39 *Gardnerella* strains were examined. Clusters were compared given the number of occurrences and the rate of evolution of viral infection (Fig 4). From this analysis, two groups can be observed. The phage gene clusters in the top right quadrant appear to have evolved quickly relative to their prevalence across strains and have likely been integrated into the *Gardnerella* genomes via HGT. In contrast, the four clusters in the bottom right quadrant appear to have been acquired by vertical transmission. Thus, we hypothesize that prophage integration has been occurring over a long time scale. Given this evidence of phage-host interactions and the results of recent studies illuminating the vital contributions of viruses within other human microbiota (e.g. [31,32,41]), one can surmise that phages are playing a significant role within both the bladder and reproductive system communities.

As the bladder microbiota have only recently been discovered and subsequently surveyed [5–8,10,14,33,42], their microbiome remains largely uncharacterized. The presence of the CRISPR/Cas system within the four *Gardnerella* strains sequenced here suggests phages are present and prolific within the bladder. Given the existence of both *Gardnerella* and *Lactobacillus* within both the female bladder [6–8,33] and vaginal [43–45] microbiota, some insight can be gleaned from the latter. It is likely that phages play an important role in bacterial genome evolution and potentially disease in both niches. Evidence of phages has previously been found within the vaginal microbiota [27,46]. The incidence of Lactobacilliphages may have medical significance, as vaginal lactobacilli may be culled or repressed by phages within the microbiota leading to BV [27] (see review [4]). Increased prophage numbers within *L. crispatus* genomes from the human vagina relative to avian isolates suggests high frequency of phage within the human microbiota [23]. Phage-like sequences within both the bladder and vaginal bacterial microbiomes–prophage sequences as well as CRISPR spacer sequences–often show little to no homology to characterized sequenced phage species, insinuating that numerous genetically diverse phages have yet to be discovered.

Conclusions

Comparison of four new bladder-associated *Gardnerella* genomes to genomes from the reproductive system identifies strain-specific and lineage-specific genes, suggesting a large *Gardnerella* pangenome may exist. There is, however, no distinct difference between strains isolated from a particular niche. Of particular interest is the high incidence of prophages within the *Gardnerella* genomes and the variability in the number per strain, as well as their putative origin. While prior studies into the prophages of vaginal lactobacilli propose that phage may play a significant role in community dynamics within the vagina (see review [4]), this proposal has yet to be empirically tested. The *Gardnerella* genome analyses conducted here find evidence of ancient, as well as contemporary, phage infection; the fact that isolates from the same individual vary in their prophage gene sequences supports the latter. Bioinformatic inspections of prophage and CRISPR spacer sequences find little to no correspondence with characterized phage sequences, suggesting that *Gardnerella*-infecting phages exist, although they have yet to be isolated in the laboratory. Nonetheless, evidence presented here suggests that phages play a role within the complex microbial communities of both the reproductive tract and the bladder, warranting future exploration of their viromes. The continued isolation and empirical characterization of *Gardnerella* species from the human microbiota is necessary to learn whether microbiome-virome interactions help to establish and maintain ones' bladder health and, if so, how perturbations of this equilibrium result in bladder pathologies.
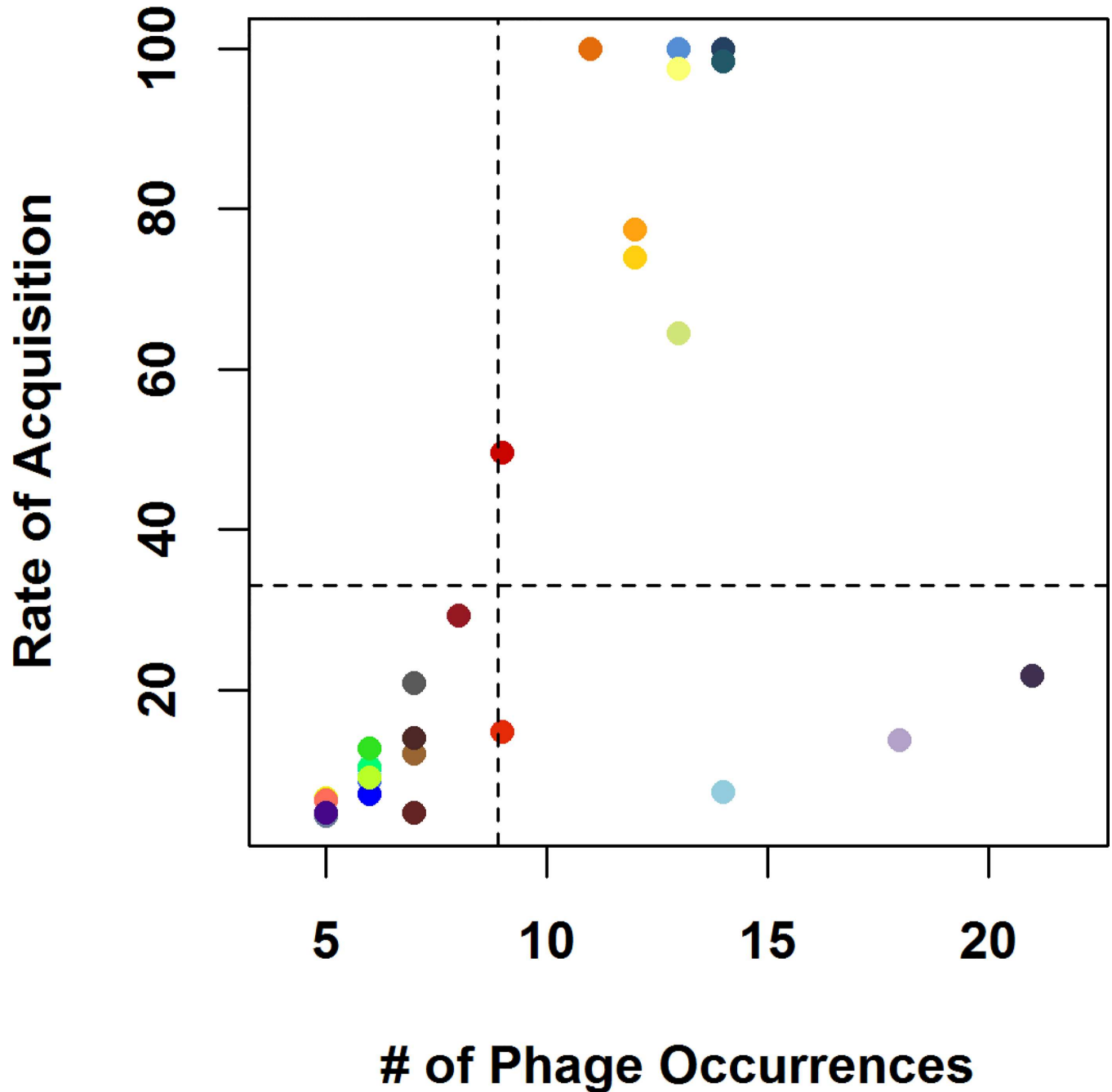
Fig 4. Determining mechanisms of prophage gene acquisition. Colors correspond with the cluster colors in Fig 3. Dashed lines are the mean value for each axis.

Materials and Methods

Strain Isolation and DNA Extraction

The *Gardnerella* and *Gardnerella vaginalis* isolates were isolated from transurethral catheter-ized urine specimens of adult women with UUI [8] using the previously described Enhanced

Quantitative Urine Culture (EQUC) protocol [6]. Microbial identification was determined using a Matrix-Assisted Laser Desorption/Ionization-Time-of-Flight Mass Spectrometer (MALDI-TOF MS, Bruker Daltonics, Billerica, MA) as described [6]. Pure cultures were stored at -80°C in a 2 ml CryoSaver Brucella Broth with 10% Glycerol, no beads, Cryovial, for preservation (Hardy Diagnostics).

The preserved pure culture isolates were grown on CDC Anaerobic 5% sheep blood (Anaerobic BAP) agar (BD BBL™ Prepared Plated Media) under anaerobic conditions at 35°C for 48 hours. MALDI-TOF MS was performed for species/genus verification. An isolated colony was transferred to 5mL tryptic soy broth (TSB) supplemented with 10% fetal bovine serum (FBS) and grown under anaerobic conditions at 35°C for 48 hours. 1 mL of culture was collected, and cells were resuspended in 1 mL of buffered saline peptone (PBS).

Genomic DNA extraction was performed using a phenol-chloroform extraction protocol. Briefly, cells were resuspended in 0.5 mL DNA Extraction Buffer (20 mM Tris-Cl, 2 mM EDTA, 1.2% Triton X-100, pH 8) followed by addition of 50 μL Lysozyme (20mg/mL), 30 μL Mutanolysin, and 5 μL RNase (10 mg/mL). After a 1-hour incubation at 37°C, 80 μL 10% SDS, and 20 μL Proteinase K were added followed by a 2-hour incubation at 55°C. 210 μL of 6 M NaCl and 700 μL phenol-chloroform were then added. After a 30-minute incubation with rotation, the solutions were centrifuged at 13,500 RPM for 10 minutes, and the aqueous phase was extracted. An equivalent volume of Isopropanol was then added, and solution was centrifuged at 13,500 RPM for 10 minutes after a 10-minute incubation. The supernatant was decanted and the DNA pellet was precipitated using 600 μL 70% Ethanol.

Genome Sequencing, Assembly, and Annotation

DNA samples were diluted in water to a concentration of 0.2 ng/μl as measured by a fluorometric-based method (Life Technologies) and 5 μl was used to obtain a total of 1 ng of input DNA. Library preparation was performed using the Nextera XT DNA Library Preparation Kit (Illumina) according to manufacturer's instructions. The isolates were barcoded, pooled and each isolate was sequenced twice, on two separate runs, using the Illumina MiSeq platform and the MiSeq Reagent Kit v2 (300-cycles) to produce 150 bp paired-end reads. Sequencing reads were parsed into individual folders according to the respective barcodes.

The following protocol produced an assembly with the least number of scaffolds and the highest overall coverage (Table 1). Reads were paired using Geneious (Biomatters Ltd., Auckland, New Zealand) for each isolate for each sequencing run. *De novo* assembly was performed, combining the two runs per isolate, using the Geneious plug-in for Velvet [47] (k = 99). Sequence contigs were then extended and scaffolds were constructed using the tool SSPACE [48]. Resulting contigs were again assembled using the Geneious *de novo* assembler at the Medium-Low sensitivity setting. Annotations were performed for each of the contigs using the RAST annotation pipeline [49], as well as the BASys bacterial annotation system [50]. CRISPR arrays were predicted using CRISPRdb [51]. A local nucleotide database was created for each strain using the protein coding regions predicted by RAST. Each predicted coding region was then reciprocally BLASTed (blastn). Genes were determined to be homologs if the query coverage and the sequence identity were both greater than 70%.

Raw sequencing reads as well as assembled contigs are available through NCBI: Gv18-4 (SRA: SRX1688291, WGS: LWSP00000000), Gv23-12 (SRA: SRX1688198, WGS: LWSQ00000000), G26-12 (SRA: SRX1688301, WGS: LWSR00000000), and G30-4 (SRA: SRX1688300, WGS: LXJL00000000). S1 Table lists the annotated protein functionalities within the four strains; annotations are available and can be queried through NCBI.

Phylogenetic Analysis

Single gene trees. 16S rRNA gene sequence analysis was performed by excising the full length 16S sequence from the assembled genomes and querying each against the NCBI nr/nt database via blastn. Sequences producing full-length hits were collected and aligned using MUSCLE [52]. The phylogenetic tree was constructed using RAxML [53] and visualized using PhyloWidget [54]. Phylogenetic analysis of the VLY gene sequences was aligned using ClustalW [55]. As before, RAxML [53] and PhyloWidget [54] were used to derive and visualize the tree, respectively.

Core genome tree. While 43 strains are presently publicly available through NCBI for *G. vaginalis*, this study considered only those that were (1) isolated from the vagina or endometrium and (2) are documented within the literature [15–19,39]. S3 Table lists the 35 publicly available strains meeting these criteria. The species tree was derived by first identifying the core set of genes within the 39 *Gardnerella* strain sequences (genome, scaffold, or contig collections). Incomplete genomes (scaffold or contig status) were retrieved from NCBI. Their sequences were individually submitted to the RAST server and annotated [49]. For the three complete genomes, G. *vaginalis* 409–05 (NC_013721), 14019 (NC_014644), and HMP9231 (NC_017456), the ffn format files were retrieved from NCBI's FTP site. The core gene set was determined by first creating a local nucleotide BLAST database with the coding regions annotated for the *G. vaginalis* 14019 strain. Annotated coding regions in other sequences were BLASTed locally (using BLASTn, e-value$<10^{-5}$), returning the top hit only. Genes producing hits in all other 38 genomes were identified as members of the core genome. The core gene set contains 183 genes (S4 Table), a significantly smaller group than previously used when considering smaller numbers of strains [15,20]. Each core gene and its orthologs were aligned using ClustalW [55]. Alignments were concatenated producing a concatenated gene alignment (super-gene alignment) using an in-house script in R (available upon request). The Maximum-likelihood tree was derived using RAxML [53]. Trees were visualized using PhyloWidget [54].

Assessing Presence/Absence of Coding Regions of Interest

The absence of coding regions identified with key carbohydrate metabolism functionalities within the G26-12 and G30-4 genome analyses was experimentally verified via PCR. Coding regions within the Gv18-4 genomic sequence with the following functional annotations were retrieved: (1) chitin and N-acetylglucosamine utilization, (2) deoxyribose and deoxynucleoside catabolism, (3) D-gluconate and ketogluconates metabolism, (4) formaldehyde assimilation, (5) lactose and galactose uptake and utilization, (6) trehalose biosynthesis, uptake and utilization, and (7) xylose utilization. In total, 22 genes were identified. Each was BLASTed against the nr/nt database, specifying the genus *Gardnerella*, to retrieve orthologs from other strains. All annotations within the GenBank files retrieved were manually inspected verifying similar confirmed/predicted protein functions. Primers were designed for each gene, targeting conserved regions amongst all orthologs, and obtained from Eurofins MWG Operon (Huntsville, AL). In total, 66 PCR reactions were conducted, testing each of the primer pairs against: Gv18-4 (serving as a positive control), G30-4, and nuclease-free water (serving as a negative control). All Gv18-4 reactions produced amplicons of the expected sizes. G30-4 and the negative controls did not produce amplicons. Primer sequences are listed in S6 Table.

Prophage Identification

Genes annotated as phage or viral in origin were extracted from each of the genome sequence annotations. Local BLAST databases were used for both identifying the orthologous clusters of

prophage sequences, as well as the putative origin of prophage clusters. First, a local nucleotide database was created, including all of the sequences predicted as phage from all 39 *Gardnerella* strains. Each individual putative phage coding region sequence was then BLASTed (blastn) against this local database (maximum 100 results), detecting identity to itself, as well as similarity to other sequences. Each hit was further qualified; hits with a sequence identity and query coverage greater than or equal to 80% were considered homologous. Once clusters were identified, the sequences within the cluster were aligned using ClustalW [55] and manually inspected to guarantee correct clustering (S1 File). Identification for the origin of each cluster was performed similar to above; the local database used for conducting blastn searches was the complete collection of viral RefSeq coding regions. The all.fna.tar.gz file was retrieved from ftp:// ftp.ncbi.nlm.nih.gov/genomes/Viruses/. Homologs were called when query coverage was ≥80%. The sequence identity threshold was considerably lower ≥35%. A cluster was considered to show no homology to any RefSeq coding region if it did not meet both of these conditions.

Estimating Rates of Phage Acquisition

Rates of phage acquisition were estimated using the R package corHMM [56], which implements a Hidden Markov Model to estimate the rate of evolution of traits occurring across a phylogeny. Each phage was encoded as a binary trait according to its presence/absence as visualized in Fig 3. Using this method to estimate the "rate of evolution," therefore, provides a meaningful proxy for the rate of phage acquisition across the tree. For any given phage, the rate of acquisition is expected to scale with its total number of occurrences in the tree, and horizontally transmitted phages are expected to have faster relative rates than vertically transmitted phages (as seen in Fig 4).

Supporting Information

S1 Fig. Phylogenetic tree based on 16S rRNA gene. Maximum-Likelihood phylogenetic tree for the 16S rRNA gene. Strains isolated from the bladder are indicated in red.
(TIF)

S2 Fig. Phylogenetic analysis of VLY gene. (A) Amino acid sequence alignment of four bladder *Gardnerella* isolates and representatives from other clades; clades are indicated to the left of each strain/isolate name. Mismatches within the alignment are highlighted (red/blue text).
(B) Maximum-Likelihood phylogenetic tree for the VLY gene. Branch supports are indicated.
(TIF)

S1 Table. Presence/absence of gene functionality within the four bladder isolates.
(XLSX)

S2 Table. Presence/absence of virulence genes within the four bladder isolates.
(XLSX)

S3 Table. List of 35 genomes retrieved from NCBI for genome comparisons.
(XLSX)

S4 Table. Core genes identified for the 39 *Gardnerella* genomes examined.
(XLSX)

S5 Table. Prophage sequences, clusters, and putative origins within the 39 *Gardnerella* genomes.
(XLSX)

S6 Table. PCR primers for amplification of Gv-specific carbohydrate metabolism coding regions.
(XLSX)

S1 File. FASTA sequences for the individual prophage gene clusters.
(ZIP)

Acknowledgments

Author Contributions

Conceptualization: AJW CP.

Data curation: KM CP.

Formal analysis: KM JWS TS.

Funding acquisition: AJW CP.

Investigation: TKP EEH KT-W ABR GK MJZ.

Methodology: AJW CP.

Resources: AJW CP.

Software: JWS CP.

Validation: KT-W TS.

Writing – original draft: AJW CP.

Writing – review & editing: KM JWS TKP EEH KT-W TS ABR GK MJZ AJW CP.

References

1. Fredricks DN, Fiedler TL, Marrazzo JM. Molecular identification of bacteria associated with bacterial vaginosis. N Engl J Med. 2005; 353:1899–1911. doi: 10.1056/NEJMoa043802 PMID: 16267321

2. Hyman RW, Fukushima M, Diamond L, Kumm J, Giudice LC, Davis RW. Microbes on the human vaginal epithelium. Proc Natl Acad Sci USA. 2005; 102:7952–7957. doi: 10.1073/pnas.0503236102 PMID: 15911771

3. Srinivasan S, Hoffman NG, Morgan MT, Matsen FA, Fiedler TL, Hall RW, et al. Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. PLoS One. 2012; 7:e37818. doi: 10.1371/journal.pone.0037818 PMID: 22719852

4. Turovskiy Y, Sutyak Noll K, Chikindas ML. The aetiology of bacterial vaginosis. J Appl Microbiol. 2011; 5:1105–1128.

5. Wolfe AJ, Toh E, Shibata N, Rong R, Kenton K, Fitzgerald M, et al. Evidence of uncultivated bacteria in the adult female bladder. J Clin Microbiol 2012; 50:1376–1383. doi: 10.1128/JCM.05852-11 PMID: 22278835

6. Hilt EE, McKinley K, Pearce MM, Rosenfeld AB, Zilliox MJ, Mueller ER, et al. Urine is not sterile: use of enhanced urine culture techniques to detect resident bacterial flora in the adult female bladder. J Clin Microbiol. 2014; 52:871–876. doi: 10.1128/JCM.02876-13 PMID: 24371246

7.   Pearce MM, Zilliox MJ, Rosenfeld AB, Thomas-White KJ, Richter HE, Nager CW, et al. Pelvic Floor Dis-
     orders Network. The female urinary microbiome in urgency urinary incontinence. Am J Obstet Gynecol.
     2015; 213:347.e1-11. doi: 10.1016/j.ajog.2015.07.009 PMID: 26210757

8.   Thomas-White KJ, Hilt EE, Fok C, Pearce MM, Mueller ER, Kliethermes S, et al. Incontinence medica-
     tion response relates to the female urinary microbiota. Int Urogynecol J. 2015; 27:723–33. doi: 10.1007/
     s00192-015-2847-x PMID: 26423260

9.   Siddiqui H, Nederbragt AJ, Lagesen K, Jeansson SL, Jakobsen KS. Assessing diversity of the female
     urine microbiota by high throughput sequencing of 16S rDNA amplicons. BMC Microbiology 2011;
     11:244. doi: 10.1186/1471-2180-11-244 PMID: 22047020

10. Fouts DE, Pieper R, Szpakowski S, Pohl H, Knoblach S, Suh MJ, et al. Integrated next-generation
     sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymp-
     tomatic bacteriuria in neuropathic bladder associated with spinal cord injury. J. Transl. Med. 2012;
     10:174. doi: 10.1186/1479-5876-10-174 PMID: 22929533

11. Nelson DE, Dong Q, Van Der Pol B, Toh E, Fan B, Katz BP, et al. Bacterial communities of the coronal
     sulcus and distal urethra of adolescent males. PLoS One. 2012; 7:e36298. doi: 10.1371/journal.pone.
     0036298 PMID: 22606251

12. Lewis DA, Brown R, Williams J, White P, Jacobson SK, Marchesi JR, et al. The human urinary micro-
     biome; bacterial DNA in voided urine of asymptomatic adults. Front Cell Infect Microbiol. 2013; 3:41.
     doi: 10.3389/fcimb.2013.00041 PMID: 23967406

13. Xu W, Yang L, Lee P, Huang WC, Nossa C, Ma Y, et al. Mini-review: perspective of the microbiome in
     the pathogenesis of urothelial carcinoma. Am J Clin Exp Urol. 2014; 2:57–61. PMID: 25126590

14. Khasriya R, Sathiananthamoorthy S, Ismail S, Kelsey M, Wilson M, Rohn JL, et al. Spectrum of bacterial
     colonization associated with urothelial cells from patients with chronic lower urinary tract symptoms. J
     Clin Microbiol. 2013; 51:2054–2062. doi: 10.1128/JCM.03314-12 PMID: 23596238

15. 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 con-
     sistent with subspeciation into genovars. J Bacteriol. 2012; 194:3922–3937. doi: 10.1128/JB.00056-12
     PMID: 22609915

16. Lewis WG, Robinson LS, Gilbert NM, Perry JC, Lewis AL. Degradation, foraging, and depletion of
     mucus sialoglycans by the vagina-adapted Actinobacterium *Gardnerella vaginalis*. J Biol Chem. 2013;
     288:12067–12079. doi: 10.1074/jbc.M113.453654 PMID: 23479734

17. Oshima K, Hisamatsu S, Toh H, Nakano A, Kiuchi M, Kuroyanagi H, et al. Complete genome sequence
     of *Gardnerella vaginalis* strain JCM 11026(T), isolated from vaginal tracts of women. Genome Announc.
     2015; 3:e00286–15. doi: 10.1128/genomeA.00286-15 PMID: 25858849

18. Schuyler JA, Chadwick SG, Mordechai E, Adelson ME, Gygax SE, Hilbert DW. Draft genome sequence
     of a Metronidazole-resistant *Gardnerella vaginalis* isolate. Genome Announc. 2015; 3:e00992–15. doi:
     10.1128/genomeA.00992-15 PMID: 26337887

19. Schuyler JA, Mordechai E, Adelson ME, Gygax SE, Hilbert DW. Draft genome sequence of a Metroni-
     dazole-resistant *Gardnerella vaginalis* strain ATCC 14019. Genome Announc. 2015; 3:e01345–15. doi:
     10.1128/genomeA.01345-15 PMID: 26564054

20. 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. doi: 10.1371/journal.pone.0012411 PMID: 20865041

21. Pleckaityte M, Zilnyte M, Zvirbliene A. Insights into the CRISPR/Cas system of *Gardnerella vaginalis*.
     BMC Microbiology. 2012; 12:301. doi: 10.1186/1471-2180-12-301 PMID: 23259527

22. Baugher JL, Durmaz E, Klaenhammer TR. Spontaneously induced prophages in *Lactobacillus gasseri*
     contribute to horizontal gene transfer. Appl Environ Microbiol. 2014; 80:3508–3517. doi: 10.1128/AEM.
     04092-13 PMID: 24682298

23. Ojala T, Kankainen M, Castro J, Cerca N, Edelman S, Westerlund-Wikström B, et al. Comparative
     genomics of *Lactobacillus crispatus* suggests novel mechanisms for the competitive exclusion of *Gard-
     nerella vaginalis*. BMC Genomics. 2014; 15:1070. doi: 10.1186/1471-2164-15-1070 PMID: 25480015

24. Clokie MR, Millard AD, Letarov AV, Heaphy S. Phages in nature. Bacteriophage. 2011; 1:31–45. doi:
     10.4161/bact.1.1.14942 PMID: 21687533

25. Koskella B, Brockhurst MA. Bacteria-phage coevolution as a driver of ecological and evolutionary pro-
     cesses in microbial communities. FEMS Microbiol Rev. 2014; 38:916–931. doi: 10.1111/1574-6976.
     12072 PMID: 24617569

26. Roberts AP, Kreth J. The impact of horizontal gene transfer on the adaptive ability of the human oral
     microbiome. Front Cell Infect Microbiol. 2014; 4:124. doi: 10.3389/fcimb.2014.00124 PMID: 25250243

27. Kiliç AO, Pavlova SI, Alpay S, Kiliç SS, Tao L. Comparative study of vaginal *Lactobacillus* phages isolated from women in the United States and Turkey: Prevalence, morphology, host range, and DNA homology. Clin Vaccine Immunol. 2001; 8:31–39.

28. Martin R, Escobedo S, Suárez JE. Induction, structural characterization, and genome sequence of Lv1, a prophage from a human vaginal *Lactobacillus jensenii* strain. Int Microbiol. 2010; 13:113–121. doi: 10.2436/20.1501.01.116 PMID: 20890845

29. Huddleston JR. Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes. Infect Drug Resist. 2014; 7:167–176. doi: 10.2147/IDR.S48820 PMID: 25018641

30. Abeles SR, Pride DT. Molecular bases and role of viruses in the human microbiome. J Mol Biol. 201; 426:3892–3906. doi: 10.1016/j.jmb.2014.07.002 PMID: 25020228

31. Hannigan GD, Meisel JS, Tyldsley AS, Zheng Q, Hodkinson BP, SanMiguel AJ, et al. The human skin double-stranded DNA virome: Topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. mBio. 2015; 6:e01578–15. doi: 10.1128/mBio.01578-15 PMID: 26489866

32. Ogilvie LA, Jones BV. The human gut virome: a multifaceted majority. Front Microbiol. 2015; 6:918. doi: 10.3389/fmicb.2015.00918 PMID: 26441861

33. Pearce MM, Hilt EE, Rosenfeld AB, Zilliox MJ, Thomas-White K, Fok C, et al. The female urinary microbiome: a comparison of women with and without urgency urinary incontinence. MBio. 2014; 5:e01283–14. doi: 10.1128/mBio.01283-14 PMID: 25006228

34. Kumar S, Doerrler WT. Members of the conserved DedA family are likely membrane transporters and are required for drug resistance in *Escherichia coli*. Antimicrob Agents Chemother. 2014; 58:923–930. doi: 10.1128/AAC.02238-13 PMID: 24277026

35. Gelber SE, Aguilar JL, Lewis KL, Ratner AJ. Functional and phylogenetic characterization of vaginolysin, the human-specific cytolysin from *Gardnerella vaginalis*. J Bacteriol. 2008; 190:3896–3903. doi: 10.1128/JB.01965-07 PMID: 18390664

36. Wren BW, Stabler RA, Das SS, Butcher PD, Mangan JA, Clarke JD, et al. Characterization of a haemolysin from *Mycobacterium tuberculosis* with homology to a virulence factor of *Serpulina hyodysenteriae*. Microbiology. 1998; 144:1205–1211. doi: 10.1099/00221287-144-5-1205 PMID: 9611795

37. 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. doi: 10.1371/journal.pone.0146510 PMID: 26751374

38. Castro J, Alves P, Sousa C, Cereija T, França Â, Jefferson KK, 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. 2015; 5:11640. doi: 10.1038/srep11640 PMID: 26113465

39. 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. doi: 10.1186/1471-2164-11-375 PMID: 20540756

40. Hickey R J, Forney LJ. *Gardnerella vaginalis* does not always cause bacterial vaginosis. J Infect Dis. 2014; 210:1682–1683. doi: 10.1093/infdis/jiu303 PMID: 24855684

41. Edlund A, Santiago-Rodriguez TM, Boehm TK, Pride DT. Bacteriophage and their potential roles in the human oral cavity. J Oral Microbiol. 2015; 7:27423. doi: 10.3402/jom.v7.27423 PMID: 25861745

42. Brubaker L, Nager CW, Richter HE, Visco A, Nygaard I, Barber MD, et al. Urinary bacteria in adult women with urgency urinary incontinence. Int Urogynecol J. 2014; 25:1179–1184. doi: 10.1007/s00192-013-2325-2 PMID: 24515544

43. Zhou X, Brown CJ, Abdo Z, Davis CC, Hansmann MA, Joyce P, et al. Differences in the composition of vaginal microbial communities found in health Caucasian and black women. ISME J. 2007; 1:121–133. doi: 10.1038/ismej.2007.12 PMID: 18043622

44. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, et al. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci USA. 2011; 108:4680–4687. doi: 10.1073/pnas.1002611107 PMID: 20534435

45. Ma B, Forney LJ, Ravel J. Vaginal microbiome: rethinking health and disease. Annu Rev Microbiol. 2012; 66:371–389. doi: 10.1146/annurev-micro-092611-150157 PMID: 22746335

46. Anukam KC, Macklaim JM, Gloor GB, Reid G, Boekhorst J, Renckens B, et al. Genome sequence of *Lactobacillus pentosus* KCA1: Vaginal isolate from a healthy premenopausal woman. PLoS One. 2013; 8:e59239. doi: 10.1371/journal.pone.0059239 PMID: 23527145

47. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. doi: 10.1101/gr.074492.107 PMID: 18349386

48. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011; 27:578–579. doi: 10.1093/bioinformatics/btq683 PMID: 21149342

49. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014; 42:D206–214. doi: 10.1093/nar/gkt1226 PMID: 24293654

50. Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, et al. BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res. 2005; 33:W455–459. doi: 10.1093/nar/gki593 PMID: 15980511

51. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics. 2007; 8:172. doi: 10.1186/1471-2105-8-172 PMID: 17521438

52. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. doi: 10.1093/nar/gkh340 PMID: 15034147

53. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30:1312–1313. doi: 10.1093/bioinformatics/btu033 PMID: 24451623

54. Jordan GE, Piel WH. PhyloWidget: web-based visualizations for the tree of life. Bioinformatics. 2008; 24:1641–1642. doi: 10.1093/bioinformatics/btn235 PMID: 18487241

55. Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics. 2002; 2:2.3.

56. Beaulieu JM, Oliver JC. 2012. corHMM: hidden Markov models in R, version 1. https://r-forge.r-project.org/projects/corhmm/.