Prediction of prophages and their host ranges in pathogenic and commensal *Neisseria* species

Giulia Orazi [a], Alan J. Collins[a], Rachel J. Whitaker[a,b]
[a] Carl R. Woese Institute for Genomic Biology
[b] Department of Microbiology
University of Illinois at Urbana-Champaign
Urbana, IL 61801

Corresponding authors
Giulia Orazi, gorazi@illinois.edu
Rachel J. Whitaker, rwhitakr@illinois.edu

ABSTRACT

The genus *Neisseria* includes two pathogenic species, *N. gonorrhoeae* and *N. meningitidis*, and numerous commensal species. *Neisseria* species frequently exchange DNA with one other, primarily via transformation and homologous recombination, and via multiple types of mobile genetic elements (MGEs). Few *Neisseria* bacteriophages (phages) have been identified and their impact on bacterial physiology is poorly understood. Furthermore, little is known about the range of species that *Neisseria* phages can infect. In this study, we used three virus prediction tools to scan 248 genomes of 21 different *Neisseria* species and identified 1302 unique predicted prophages. Using comparative genomics, we found that many predictions are dissimilar from other prophages and MGEs previously described to infect *Neisseria* species. We also identified similar predicted prophages in genomes of different *Neisseria* species. Additionally, we examined CRISPR-Cas targeting of each *Neisseria* genome and predicted prophage. While CRISPR targeting of chromosomal DNA appears to be common among several *Neisseria* species, we found that 20% of the prophages we predicted are targeted significantly more than the rest of the bacterial genome in which they were identified (i.e., backbone). Furthermore, many predicted prophages are targeted by CRISPR spacers encoded by other species. We then used these results to infer additional host species of known *Neisseria* prophages and predictions that are highly targeted relative to the backbone. Together, our results suggest that we have identified novel *Neisseria* prophages, several of which may infect multiple *Neisseria* species. These findings have important implications for understanding horizontal gene transfer between members of this genus.

IMPORTANCE

Drug-resistant *Neisseria gonorrhoeae* is a major threat to human health. Commensal *Neisseria* species are thought to serve as reservoirs of antibiotic resistance and virulence genes for the pathogenic species *N. gonorrhoeae* and *N. meningitidis*. Therefore, it is important to understand both the diversity of mobile genetic elements (MGEs) that can mediate horizontal gene transfer within this genus, and the breadth of species these MGEs can infect. In particular, few bacteriophages (phages) have been identified and characterized in *Neisseria* species. In this study, we identified a large number of candidate phages integrated within the genomes of commensal and pathogenic *Neisseria* species, many of which appear to be novel phages. Importantly, we discovered extensive interspecies targeting of predicted phages by *Neisseria* CRISPR-Cas systems, which may reflect their movement between different species. Uncovering the diversity and host range of phages is essential for understanding how they influence the evolution of their microbial hosts.

INTRODUCTION

The genus *Neisseria* includes the human pathogens *N. gonorrhoeae* and *N. meningitidis* as well as a multitude of highly diverse commensal species that colonize mucosal surfaces of humans and animals (1). Because of the extensive spread of antibiotic resistance (AR) among strains of *N. gonorrhoeae*, infections caused by this pathogen are becoming increasingly difficult to treat (2). Consequently, the WHO and CDC consider *N. gonorrhoeae* a high-priority and urgent threat among antibiotic-resistant pathogens (3, 4).

Neisseria species are naturally competent and frequently exchange DNA with one other via transformation and homologous recombination (5–8). MGEs, such as plasmids, genetic islands, and bacteriophages (phages) can also mobilize genetic material and are powerful forces in shaping bacterial evolution (9–14). Phages are incredibly abundant and can deeply influence the fitness and virulence of their bacterial hosts, particularly when integrated into the bacterial chromosome (15–20). For example, the filamentous prophage MDAΦ promotes attachment of *N. meningitidis* to epithelial cell monolayers (18) and is associated with the ability of this pathogen to cause invasive disease (17).

In contrast to many highly studied Gammaproteobacteria prophages, few have been identified and characterized in the Betaproteobacteria (21–23). *Neisseria* prophages have been identified primarily in *N. gonorrhoeae* and *N. meningitidis* and consist of a small number of filamentous (24–27) and double-stranded DNA (dsDNA) prophages (28), including Mu-like prophages (21, 29–32). With the exception of MDAΦ, the impact of phages on *Neisseria* biology and pathogenicity remains poorly understood (13). Furthermore, few studies have investigated the host ranges of *Neisseria* phages (33, 34).

Microbes can defend themselves against phages and other MGEs using a variety of systems. One such system is CRISPR-Cas, which is composed of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) arrays and CRISPR-associated (Cas) proteins. Importantly, sequence identity between the spacer and the MGE is required for immunity, which means that CRISPR arrays contain a record of previous encounters with MGEs within the sequences of their spacers. Therefore, this historical record can be used to infer the bacterial hosts of viruses (35–39). Approximately 40% of *N. meningitidis* genomes encode Type II-C CRISPR arrays (40, 41), and multiple putative CRISPR systems have been identified in several commensal species (41–43). In contrast, only degenerate CRISPR arrays have been found in *N. gonorrhoeae* (41, 42).

In this study, we sought to uncover novel *Neisseria* phage diversity. We used bioinformatic virus prediction tools to scan publicly available genomes of pathogenic and commensal *Neisseria* species for prophages. Using comparative genomics, we found that many of these predictions are dissimilar from previously identified *Neisseria* MGEs and are potential targets of CRISPR-Cas systems. Finally, we used interspecies CRISPR targeting of known and predicted prophages to infer whether they may infect multiple different *Neisseria* species.

RESULTS

Predicting prophages within genomes of pathogenic and commensal *Neisseria* species

To search for prophages, we compiled a dataset of 248 publicly available high-quality *Neisseria* genome assemblies from GenBank (44, 45) that includes *N. gonorrhoeae*, *N. meningitidis*, and 19 commensal species (see Methods and **Table S1A** for a description of the dataset). The relationships between the genomes in this dataset are shown in a phylogenetic tree based on ribosomal gene sequences and a heatmap of the average nucleotide identity (ANI) between each pair of genomes (**Fig. 1**). The phylogeny presented here is consistent with previously reported relationships between *Neisseria* species (46).

To increase the likelihood of identifying novel phages, we selected 3 tools that use different approaches to predict prophages within *Neisseria* genomes. PhiSpy uses machine learning to search for characteristics that are unique to prophages (47), VirSorter2 combines alignment and machine learning-based approaches to identify microbial viruses (48, 49), and Seeker uses deep learning to detect phages without relying on any sequence features (50).

We assessed whether these tools could identify 9 known *Neisseria* prophages (**Table S1B**) in *N. gonorrhoeae* FA 1090 (24, 26, 28) and 4 in *N. meningitidis* Z2491 (24, 25, 29). Combined, these three tools predicted 6/9 known, intact prophages in FA 1090 (**Fig. S1A**), and 2/4 in Z2491 (**Fig. S1B**). None of these tools correctly identified known *Neisseria* filamentous prophages; they were either missed entirely or combined with an adjacent dsDNA prophage into a single prediction (NgoΦ2 with NgoΦ6; NgoΦ3 with NgoΦ9) (**Fig. S1A**). These results are consistent with previous observations that several tools (including VirSorter2) have difficulty predicting *Neisseria* prophages (51).

In total, we obtained 2050 predictions (**Table S2A**), which were reduced to 1302 unique predictions after dereplication at 95% length aligned (see Methods and **Table S2B**). No phages identified in different bacterial species were found to be similar at ≥95% length aligned (**Table S2B**). The distribution of lengths of dereplicated prophages predicted by each tool is shown in **Fig. S2**.

Subsequently, we present analyses on predictions made by all 3 tools. For the sake of space and clarity, analyses of PhiSpy predictions are presented in the main text, while analyses of VirSorter2 and Seeker predictions are included in the Supplemental Material. We focus on a single tool to avoid the issue of reconciling overlapping predictions between tools and selected PhiSpy because it more accurately predicted the boundaries of known *Neisseria* prophages (**Fig. S1**).

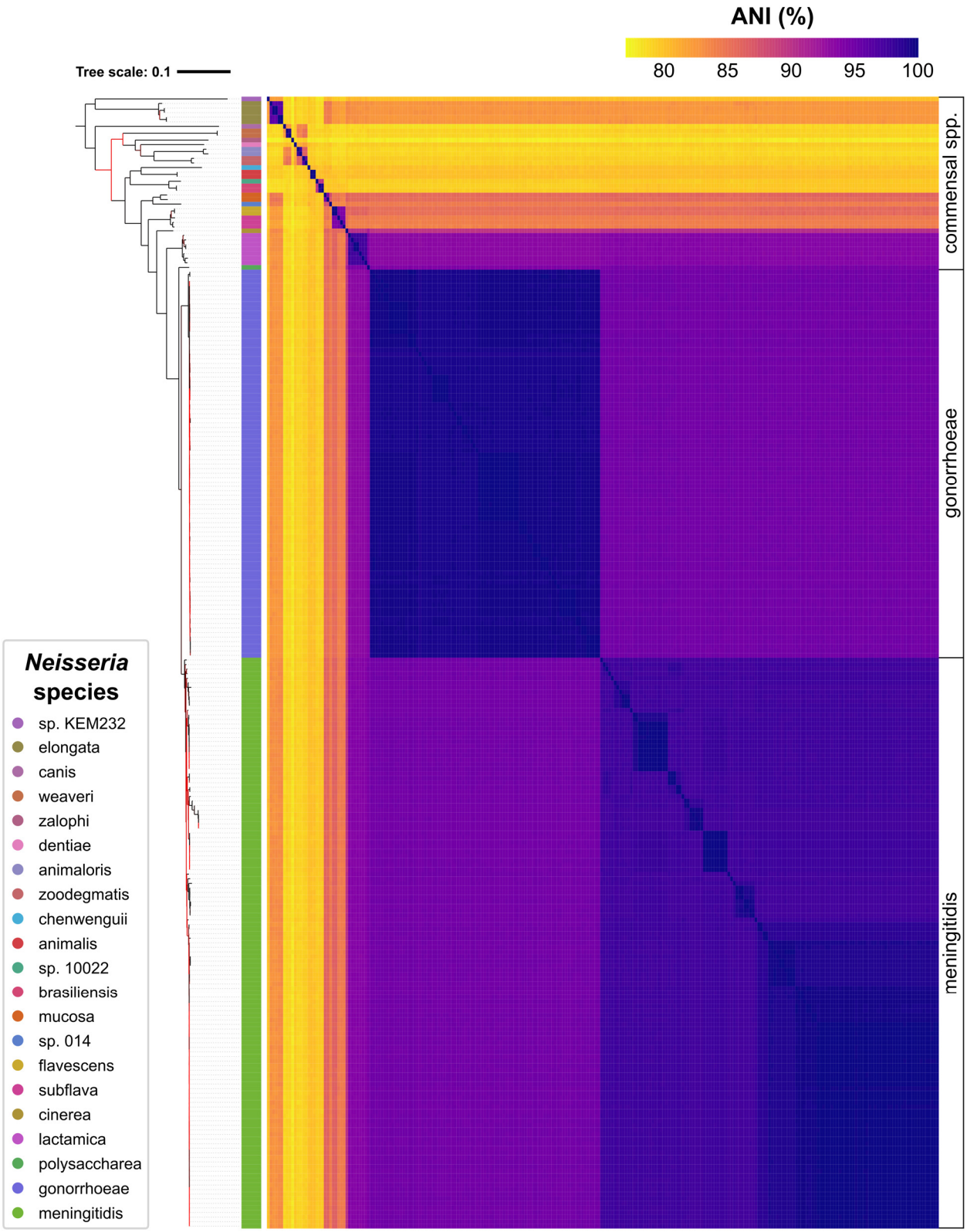

ANI (%)
80
85
90
95
100
Tree scale: 0.1
commensal spp.
gonorrhoeae
meningitidis
Neisseria species
sp. KEM232
elongata
canis
weaveri
zalophi
dentiae
animaloris
zoodegmatis
chenwenguii
animalis
sp. 10022
brasiliensis
mucosa
sp. 014
flavescens
subflava
cinerea
lactamica
polysaccharea
gonorrhoeae
meningitidis

Fig. 1. Relationships between bacterial genomes included in this study. Maximum-likelihood ribosomal MLST (rMLST) tree of the smaller set of high-quality *Neisseria* genomes (described in Table S1A) and a heatmap of pairwise average nucleotide identity (ANI) values between each genome. Ribosomal gene sequences of each genome were identified and concatenated using PubMLST and then used to create a phylogenetic tree using RAxML. The tree was visualized with iTOL and rooted using midpoint rooting. Each colored bar represents a different *Neisseria* species as defined in the key; the order of species in the tree is the same as the order shown in the key. Bootstrap support is indicated by the color of each branch, where red indicates low support. The tree file is provided in Data Set S1, Tab 1. ANI values were calculated using FastANI and represented as a color gradient as indicated in the key.

Few predictions are similar to *Neisseria* plasmids and the Gonococcal Genetic Island

In this study, we used prediction tools that search for viruses. However, because VirSorter2 has been reported to have difficulty distinguishing plasmids from viral sequences (49, 52, 53), we wanted to address the possibility that predictions from any of the tools may resemble other types of *Neisseria* MGEs.

Specifically, we compared our predictions to known *Neisseria* plasmids and the Gonococcal Genetic Island (GGI). To perform this analysis, we performed hierarchical clustering based on percent length aligned of dereplicated predictions and the nucleotide sequences of plasmids and GGI obtained from GenBank (44, 45) (**Table S1C**).

Only 14 unique predictions cluster with *Neisseria* plasmids and the GGI based on nucleotide sequence (**Fig. S3, Data Set S1**). Of these 14 predictions, 2 of them (both predicted by VirSorter2) cluster with known *Neisseria* plasmids (**Fig. S3A**), and 12 predictions (all predicted by Seeker) cluster with the GGI (**Fig. S3B**).

Because our study is focused on phages, we excluded these 14 predictions from our subsequent analyses. The results above indicate that the majority of predictions in this study are dissimilar from these other types of known *Neisseria* MGEs.

Comparing predictions to known phages using gene-sharing networks

Classifying phages is challenging due to their high genomic diversity, extensive mosaicism, and lack of universally shared genes (54–57). Therefore, gene-sharing networks are commonly used to compare novel phages to previously identified phages (35, 36, 57–61). Here, we used vConTACT v.2.0 (60, 62) to assess whether the prophages we predicted are similar to known *Neisseria* phages (**Table S1B**) or phages that infect other bacterial taxa (i.e., other reference viruses; see Methods for details).

vConTACT generates a similarity score between each pair of viruses based on the protein clusters they share. If two viruses are significantly similar to one another (i.e., the pair has a score ≥1), then they are connected by an edge. Groups of viruses that are highly similar are placed within the same subcluster and are likely members of the same viral genus (60). We used vConTACT to separately analyze dereplicated predictions from each tool, resulting in 3 distinct networks (**Fig. 2, Fig. S4A and S4B**).

First, we examined whether PhiSpy predictions are significantly similar to known *Neisseria* phages (i.e., connected by an edge in the network). While 83% of PhiSpy predictions (229/277) are *connected to* known *Neisseria* phages, only 52% of PhiSpy predictions (144/277) *cluster with Neisseria* phages (**Table S2C**). These 144 predictions belong to the following 4 subclusters:

181_0 (dark blue circle), 1139_0, 1401_0, 1318_0 (pink circles, **Fig. 2**). Thus, only half of PhiSpy predictions are likely members of the same viral genus as known *Neisseria* phages.

Next, we compared PhiSpy predictions to viruses that infect bacterial taxa other than *Neisseria* (i.e., other reference viruses). We found that 86% of PhiSpy predictions (239/277) are significantly *connected to* reference viruses (**Table S2C**). However, only 15% of PhiSpy predictions (42/277) *cluster with* other reference viruses (**Table S2C**); these predictions belong to either subcluster 181_0 (dark blue circle) or 239_0 (light blue circle, **Fig. 2**). Below, we explore these 2 subclusters that contain both PhiSpy predictions and other reference viruses.

Subcluster 181_0 includes Mu-like phages that infect *N. meningitidis* (Pnm1-2, MuMenB), *Mannhemia haemolytica* (3927AP2), and *Haemophilus parasuis* (SuMu, shown in bold, **Fig. 3A**). Previously, Pnm1-2 and MuMenB were found to resemble a Mu-like phage that infects *H. influenzae* (21). There is a high degree of synteny between members of 181_0, and the proteins shared between known and predicted *Neisseria* prophages in this subcluster have >50% sequence identity (**Fig. 3A**).

Similarly, subcluster 239_0 contains two other Mu-like phages that infect *Burkholderia cenocepacia* (BcepMu) and *B. thailandensis* (phiE255, **Fig. 3B**). Except for several late phage genes, most predicted proteins are shared between the *Burkholderia* phages and *Neisseria* predictions (at ~30-50% sequence identity, **Fig. 3B**). Together, the results above suggest that several *Neisseria* species may be infected by phages that are similar to those that infect *Haemophilus, Mannheimia,* and *Burkholderia* — microbes that *Neisseria* species may encounter within the respiratory tracts of humans and animals.

We also explored subcluster 7_0 (orange circle, **Fig. 2**). Although this subcluster does not include any reference viruses, its members share many genes with reference viruses (many surrounding dark grey nodes, **Fig. 2**). In particular, a *Neisseria* sp. KEM 232 prediction that is part of 7_0 shares 48% of predicted proteins (29/60) with other members of 7_0 (**Fig. 3C**), and also shares 30% of proteins (18/60) with two *Mannheimia* P2-like phages that do not belong to this subcluster (587AP1 and phiMHaA1, **Fig. 3C**).

While the majority of PhiSpy predictions do not cluster with known phages (**Table S2C**), many predictions were found to have a low degree of similarity to different reference phages (as indicated by the low similarity scores between predictions and reference viruses, **Fig. S5**). For example, the *N. lactamica* and *N. elongata* predictions belonging to 479_0 each have a low degree of similarity to ~20-50 different reference viruses (**Fig. 3D**). Therefore, these findings suggest that several *Neisseria* prophages are distantly related to multiple viruses that infect other bacterial taxa.

Finally, we found differences in how similar predictions from each tool are to known viruses. Specifically, 1) a smaller proportion of Seeker predictions are *connected to* reference phages compared to PhiSpy and VirSorter2 predictions (**Table S2C**); 2) the *degree to which* Seeker predictions are similar to reference viruses is significantly lower compared to the other tools (**Fig. S5**); and 3) zero Seeker predictions *cluster with* reference viruses (**Table S2C**). Thus, Seeker predictions may represent novel phages, other MGEs, or alternatively, regions of the chromosome that were incorrectly called.

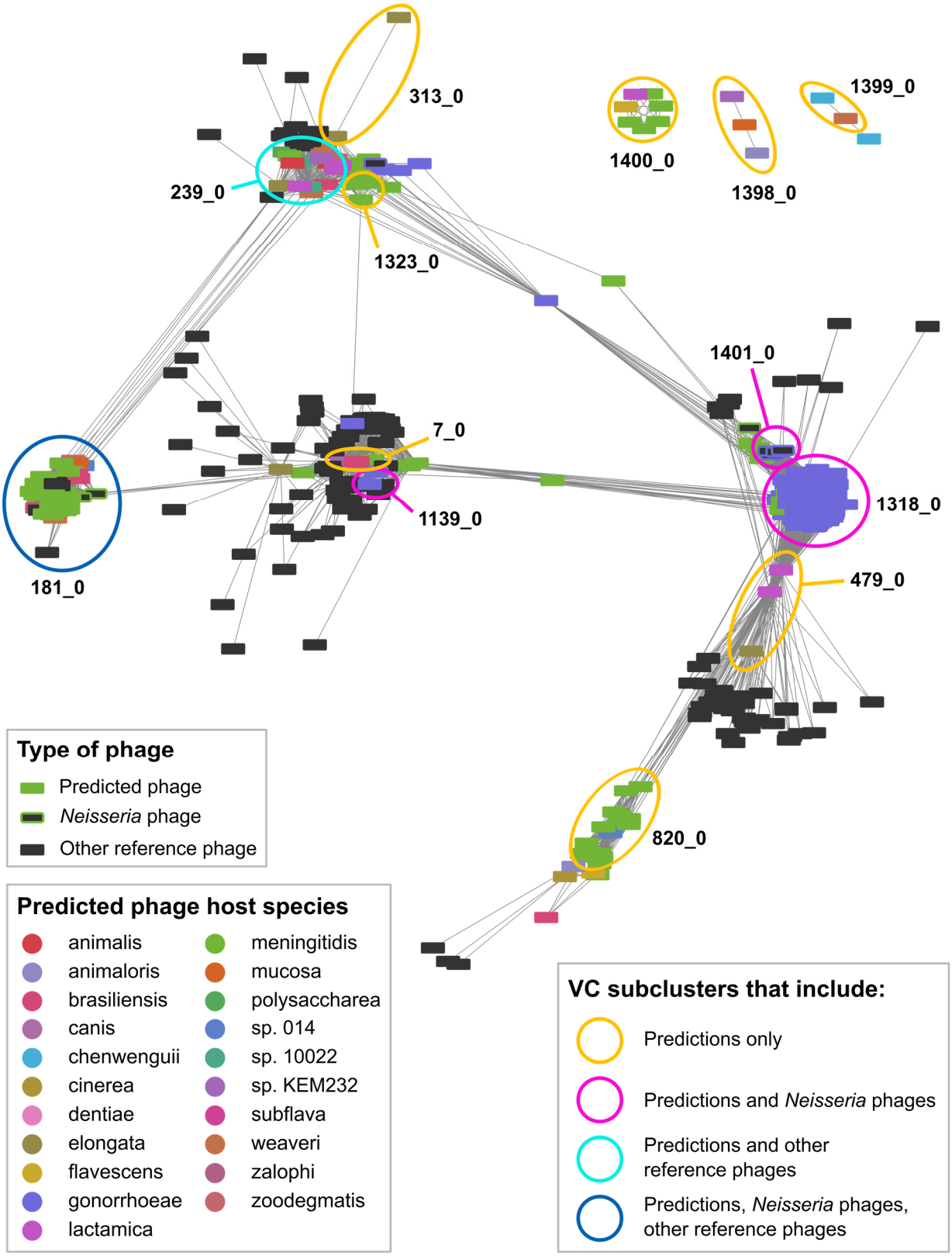

313_0
1400_0
1398_0
1399_0
239_0
1323_0
1401_0
7_0
1139_0
1318_0
181_0
479_0
820_0
Type of phage
Predicted phage
Neisseria phage
Other reference phage
Predicted phage host species
animalis
animaloris
brasiliensis
canis
chenwenguii
cinerea
dentiae
elongata
flavescens
gonorrhoeae
lactamica
meningitidis
mucosa
polysaccharea
sp. 014
sp. 10022
sp. KEM232
subflava
weaveri
zalophi
zoodegmatis
VC subclusters that include:
Predictions only
Predictions and Neisseria phages
Predictions and other reference phages
Predictions, Neisseria phages, other reference phages

Fig. 2. vConTACT clustering of PhiSpy predictions, *Neisseria* phages, and phages that infect other bacterial taxa. vConTACT v.2.0-generated network of dereplicated PhiSpy predictions and reference viruses visualized with Cytoscape using an edge-weighted spring-embedded algorithm. Nodes represent predicted prophages (color corresponding to the *Neisseria* species in which prophages were identified), *Neisseria* phages (dark grey outlined in the color corresponding to the bacterial host species), or other reference viruses (dark grey without outline). Edges represent the vConTACT-generated similarity score between each pair of viruses (only similarity scores ≥1 are included in the network). Highly similar viruses are positioned close together. Only reference viruses that are connected to ≥1 predicted prophage are included in the network. Information about vConTACT subclusters is included in Data Set S1, Tab 3, and similarity scores (edge weights) in Tab 4.

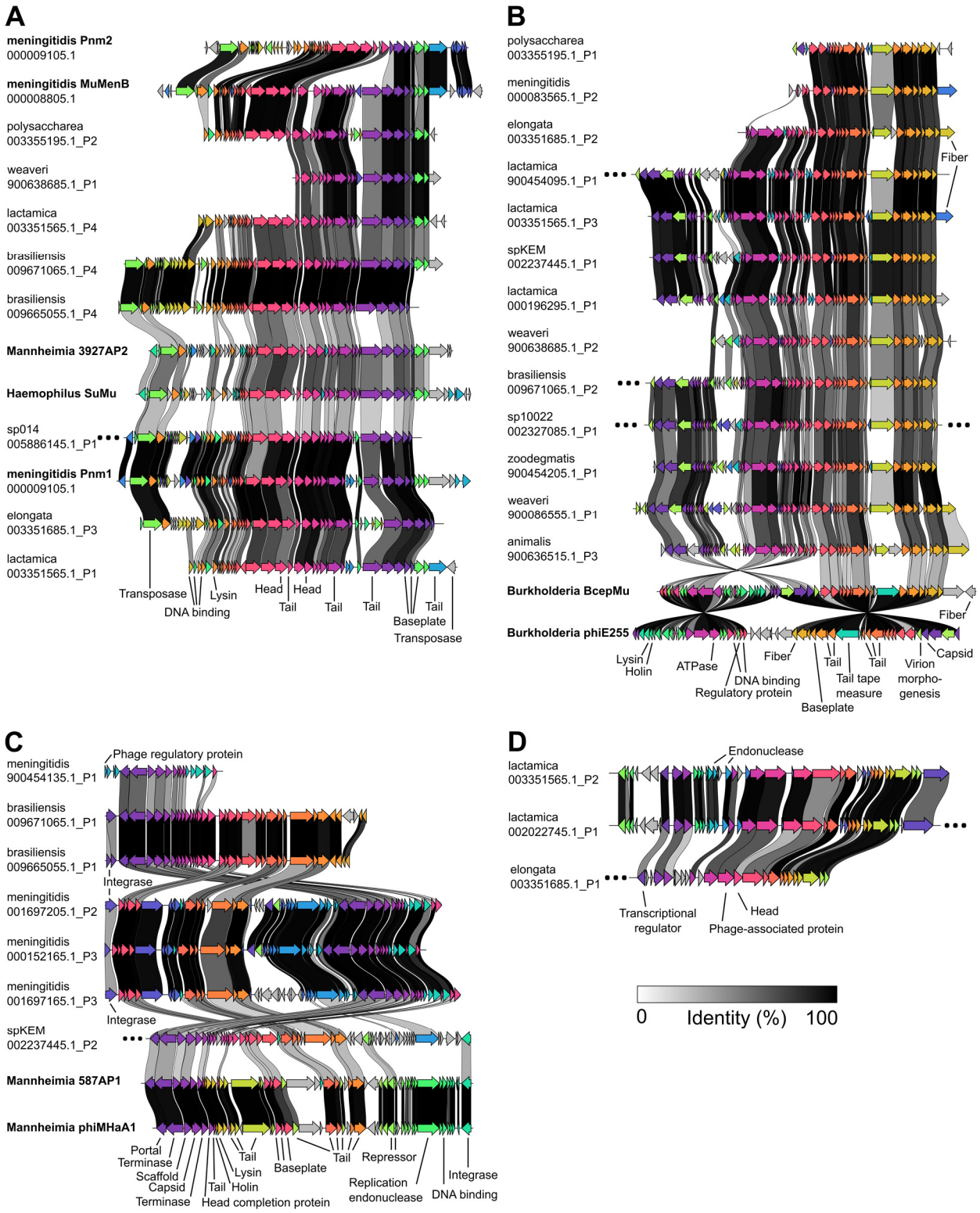

A
meningitidis Pnm2
000009105.1
meningitidis MuMenB
000008805.1
polysaccharea
003355195.1_P2
weaveri
900638685.1_P1
lactamica
003351565.1_P4
brasiliensis
009671065.1_P4
brasiliensis
009665055.1_P4
Mannheimia 3927AP2
Haemophilus SuMu
sp014
005886145.1_P1
meningitidis Pnm1
000009105.1
elongata
003351685.1_P3
lactamica
003351565.1_P1
Transposase
DNA binding
Lysin
Head
Tail
Head
Tail
Tail
Baseplate
Tail
Transposase
B
polysaccharea
003355195.1_P1
meningitidis
000083565.1_P2
elongata
003351685.1_P2
lactamica
900454095.1_P1
lactamica
003351565.1_P3
spKEM
002237445.1_P1
lactamica
000196295.1_P1
weaveri
900638685.1_P2
brasiliensis
009671065.1_P2
sp10022
002327085.1_P1
zoodegmatis
900454205.1_P1
weaveri
900086555.1_P1
animalis
900636515.1_P3
Burkholderia BcepMu
Burkholderia phiE255
Fiber
Fiber
Lysin
Holin
ATPase
DNA binding
Regulatory protein
Fiber
Baseplate
Tail
Tail
Tail
Tail tape
measure
Virion
morpho-
genesis
Capsid
C
meningitidis
900454135.1_P1
Phage regulatory protein
brasiliensis
009671065.1_P1
brasiliensis
009665055.1_P1
Integrase
meningitidis
001697205.1_P2
meningitidis
000152165.1_P3
meningitidis
001697165.1_P3
Integrase
spKEM
002237445.1_P2
Mannheimia 587AP1
Mannheimia phiMHaA1
Portal
Terminase
Scaffold
Capsid
Terminase
Tail
Tail
Lysin
Holin
Head completion protein
Baseplate
Tail
Repressor
Replication
endonuclease
Integrase
DNA binding
D
lactamica
003351565.1_P2
Endonuclease
lactamica
002022745.1_P1
elongata
003351685.1_P1
Transcriptional
regulator
Head
Phage-associated protein
0
Identity (%)
100

Fig. 3. Shared genes between predictions and reference phages. (A-D) Clinker-generated visualizations showing genes that are shared between members of the following vConTACT subclusters: 181_0 (A), 239_0 (B), 7_0 (C), and 479_0 (D). The location of each subcluster within the vConTACT-generated network is shown in Fig. 2. Reference phages are indicated in bold. Arrows with the same color indicate genes that are similar between phages; connections between arrows indicate amino acid sequence identity as described in the key. Grey arrows indicate genes that are not shared between phages; ellipses indicate that >2 unshared genes are present at either end of a phage genome. Shared genes are annotated with predicted functions of encoded proteins (except for hypothetical proteins). In Panel A, several late phage genes are present/absent between *Burkholderia* phages and *Neisseria* predictions (their order in phiE255 from left to right: lysin, holin, tail tape measure, and fiber proteins). Panel C also includes two reference viruses that are not part of the subcluster: *Mannheimia* phages 587AP1 and phiMHaA1. Information about vConTACT subclusters is included in Data Set S1, Tab 3.

Highly similar predicted prophages are found in distantly related *Neisseria* species

Next, we explored whether different *Neisseria* species may be infected by highly similar phages by examining whether any vConTACT clusters include PhiSpy predictions found in genomes of different *Neisseria* species (**Fig. 4A**).

We found that out of the 12 subclusters that include PhiSpy predictions, 9 contain predictions identified in different *Neisseria* species (**Fig. 4A**). Strikingly, subclusters 181_0 and 239_0 (**Fig. 3A and B**) include predictions found in 7 and 10 different *Neisseria* species, respectively (**Fig. 4A**).

Afterwards, we investigated whether the bacterial species (in which these predictions were identified) are closely or distantly related to each other. For every subcluster that includes ≥5 predictions, we calculated the pairwise average nucleotide identity (ANI) between each bacterial genome in which the predicted prophages were identified.

Every subcluster we examined includes multiple phages found in the same species, as shown by ANI >95% (**Fig. 4B**). Also, 4 clusters include predictions found in closely related species (ANI 90-95%).

Finally, 5 clusters include predictions found in more distantly related species (ANI <90%, **Fig. 4B**). Strikingly, host bacterial genomes of predictions in 3 of these clusters (181_0, 239_0, 7_0, **Fig. 3A-C**) have ANI ~80% (**Fig. 4B**). Therefore, these results suggest that even distantly related *Neisseria* species may be infected by closely related phages.

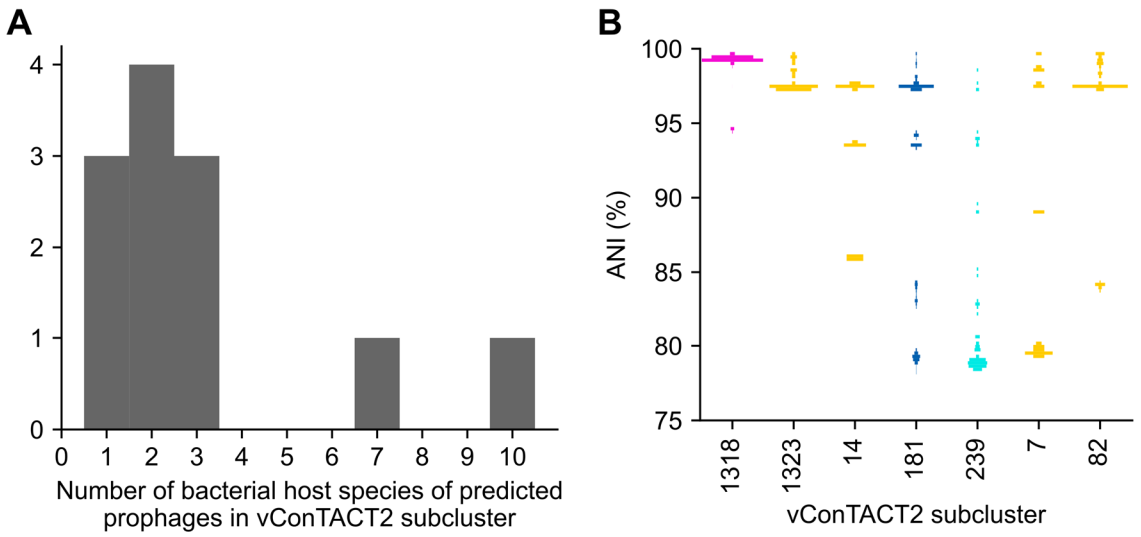


Fig. 4. Several clusters include PhiSpy predictions identified in different host bacterial species. (A) The number of different host bacterial species of PhiSpy predictions within each vConTACT v2.0 subcluster. (B) The percent average nucleotide identity (ANI) between the bacterial genomes in which PhiSpy predictions were identified. Only vConTACT subclusters that include 5 or more predicted prophages are shown ("_0" was omitted from the end of subcluster names). The distribution of ANI values is represented as a histogram where the width of bars at a given ANI corresponds to the proportion of genome pairs with that ANI. The color of each subcluster indicates the types of predictions that belong to that subcluster (orange: predictions only; pink: predictions and *Neisseria* phages; light blue: predictions and other reference phages; dark blue: predictions, *Neisseria* phages, and other reference phages). Information about vConTACT subclusters is included in Data Set S1, Tab 3.

Identification of CRISPR arrays and matches within *Neisseria* genomes

Here, we surveyed a larger set of 2619 *Neisseria* genomes (see Methods, **Table S1D**) for the presence of CRISPR arrays (see **Data Set S1**).

Consistent with previous findings (40–42), we identified Type II-C CRISPR arrays in 45% of *N. meningitidis* genomes (862/1894), and no CRISPR arrays in *N. gonorrhoeae* genomes (0/630). Also, we identified CRISPR arrays in ≥1 genome of every commensal species included in this study. Repeat sequences in arrays of commensal species are associated with 7 different CRISPR subtypes (I-A, I-C, I-F, II-C, III-A, III-B, III-D). In total, we found 3676 unique CRISPR spacers.

Next, we used BLASTn to search for matches between CRISPR spacers and sequences within *Neisseria* genomes. We only kept matches that had 100% identity over the entire length of the spacer (i.e., 0 mismatches), and we looked for both intra- and inter-species matches. We found that 22% of spacers (800/3676) target sequences within the smaller set of high-quality *Neisseria* genomes. Out of these targeting spacers, 67% (539/800) match known or predicted prophages.

Previously, Zhang et al. identified 5 self-targeting spacers in 6 *N. meningitidis* genomes (41). Here, we found that 52% of CRISPR-positive *N. meningitidis* genomes (23/44) encode self-targeting spacers.

Examining the locations of CRISPR matches within *Neisseria* genomes

We next examined the genomic locations of spacer matches. In addition to providing defense against MGEs, the Type II-C CRISPR system of *N. meningitidis* has been proposed to play a role in limiting natural transformation (41, 42). If CRISPR systems limit transformation, we would expect to see targeting evenly distributed along the length of the bacterial chromosome with no obvious enrichment of targeting in any location. If, however, prophages are targeted by CRISPR immunity, we would expect that matches would be enriched within prophages.

Fig. 5 shows the genomic locations of spacer matches within 2 genomes that encode CRISPR arrays (*N. meningitidis,* sp. 10022) and 2 that do not (*N. gonorrhoeae*, *N. weaveri*). In *N. gonorrhoeae* and *N. meningitidis*, we observe a low level of targeting across the length of the genome (**Fig. 5A**), consistent with a role for CRISPR systems in restricting transformation. There are also regions of high targeting: in *N. gonorrhoeae*, peaks correspond to the location of several known prophages, whereas peaks in the *N. meningitidis* genome may correspond to as of yet unidentified MGEs (**Fig. 5A**).

Matches within sp. 10022 and *N. weaveri* genomes mostly correspond to predicted prophages (**Fig. 5B**). Overall, many fewer spacers appear to target sp. 10022 and *N. weaveri* genomes, which is due at least in part to the few available genomes of commensal species, leading to a small pool of targeting spacers. Thus, our ability to make comparisons of targeting between *N. meningitidis* and commensals is limited.

Finally, we examined which species encode the targeting spacers (color of each circle, **Fig. 5**). Previously, *N. meningitidis* spacers were reported to match protospacers in *N. gonorrhoeae* genomes (41); here, we observe that *N. meningitidis* is largely responsible for the low-level targeting of the *N. gonorrhoeae* and *N. meningitidis* genomes in **Fig. 5A**. In contrast, prophages in these 4 genomes are matched by spacers from *N. meningitidis* or other species (**Fig. 5**). In subsequent analyses, we quantify CRISPR targeting of each prophage and bacterial genome and further investigate interspecies targeting.

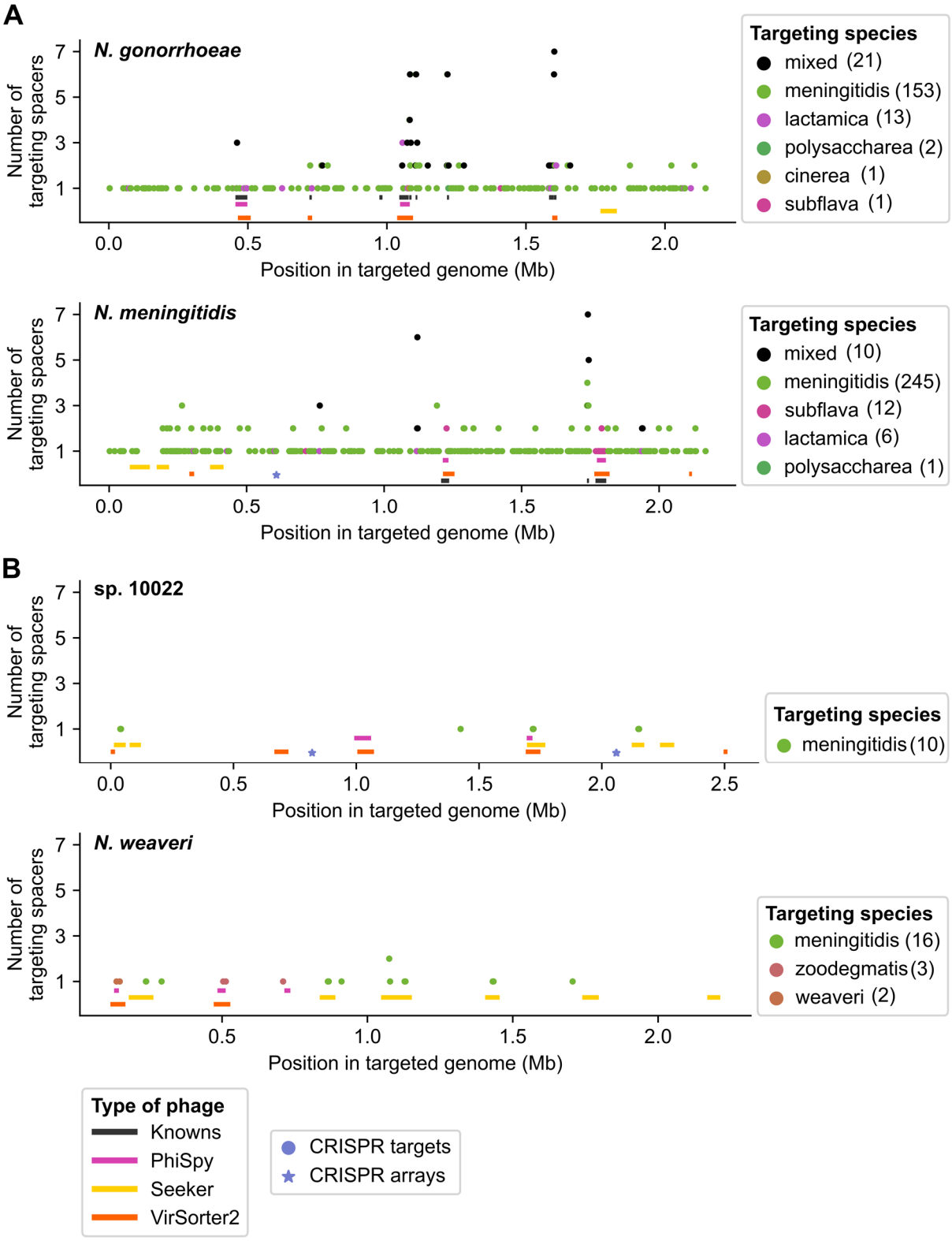

A
N. gonorrhoeae
Number of targeting spacers
Position in targeted genome (Mb)
Targeting species
mixed (21)
meningitidis (153)
lactamica (13)
polysaccharea (2)
cinerea (1)
subflava (1)
N. meningitidis
Targeting species
mixed (10)
meningitidis (245)
subflava (12)
lactamica (6)
polysaccharea (1)
B
sp. 10022
Targeting species
meningitidis (10)
N. weaveri
Targeting species
meningitidis (16)
zoodegmatis (3)
weaveri (2)
Type of phage
Knowns
PhiSpy
Seeker
VirSorter2
CRISPR targets
CRISPR arrays

Fig. 5. Locations of matches between CRISPR spacers and *Neisseria* genomes. (A and B) Genomic locations of matches between *Neisseria* CRISPR spacers and bacterial genomes. Each plot shows a genome of a different species: *N. gonorrhoeae* (GCA_000006845.1) and *N. meningitidis* (GCA_000009105.1) (A), and sp. 10022 (GCA_002327085.1) and *N. weaveri* (GCA_900638685.1) (B). Each data point (circle) represents the number of spacers that match each position (1-kb bin) in the bacterial genome. The color of the circle corresponds to the species encoding the spacer (as indicated in the key, with mixed indicating >1 unique species); the number in parentheses indicates the number of spacers encoded by each species that match the genome (at 100% identity over the entire spacer length). CRISPR arrays are denoted by stars, and rectangles represent prophages (as indicated in the key). Information about CRISPR targeting of each genome is provided in Data Set S1, Tab 10; locations of CRISPR arrays are in Tab 7; and locations of known and predicted prophages are in Tables S1B and S2A, respectively.

Comparing CRISPR targeting of each predicted prophage to backbone targeting

Above, we observed CRISPR targeting along the entire length of the chromosome in *N. gonorrhoeae* and *N. meningitidis*. To distinguish whether predicted prophages are preferentially targeted, it is necessary to compare the level of targeting of predicted prophages to the background level across the rest of the genome. Therefore, we quantified the density of CRISPR targeting of every predicted prophage and the rest of the bacterial genome in which it was identified (i.e., the backbone).

We defined prophage targeting density as the number of CRISPR matches within the prophage divided by the prophage length. The backbone targeting density was obtained by dividing backbone targeting (the number of CRISPR matches within a bacterial genome excluding targets in all known or predicted prophages and CRISPR arrays) by the length of the backbone (length of the entire bacterial genome minus the combined lengths of the prophages identified in that genome).

First, we compared the targeting density of each prophage to the targeting density of the backbone genome (**Fig. 6A**). We then compared ratios of prophage/backbone targeting between different *Neisseria* species (**Fig. 6B**). Although many *N. gonorrhoeae* and *N. meningitidis* prophages have high targeting densities (**Fig. 6A**), the high degree of backbone targeting of these genomes results in mostly low targeting ratios (**Fig. 6B**). For example, even though all 13 known *Neisseria* prophages are matched by spacers, only 5 of them (MDAΦ and NgoΦ6-9) are significantly more highly targeted compared to the backbone (**Data Set S1**).

In multiple commensal species, the ratio of prophage/backbone targeting is very high (**Fig. 6B**), in many cases due to little or no backbone targeting (**Fig. 6A**). Low levels of backbone targeting could be due to several, non-mutually exclusive reasons: 1) the small number of commensal spacers sampled in this study resulting in lower apparent targeting; 2) infrequent encounters between certain species (e.g., *N. weaveri* is an opportunistic pathogen rather than a resident of the human mucosa) (63); or 3) species-specific barriers to transformation, including differences in DNA uptake sequences (64) and in whether CRISPR systems target chromosomal DNA.

Overall, 20% of dereplicated prophages predicted in this study (259/1306) have a significantly higher targeting density compared to the backbone (**Table S2C, Data Set S1**). Furthermore, the majority of significantly targeted predictions (74%; 191/259) do not cluster with known *Neisseria* phages, plasmids, or the GGI (**Table S2C**); these predictions belong to 30 different vConTACT subclusters (**Data Set S1**), including 3 of the PhiSpy subclusters highlighted above (239_0, 7_0, 479_0, **Fig. 3**). Therefore, these 191 predictions represent likely candidates for novel *Neisseria* phages.

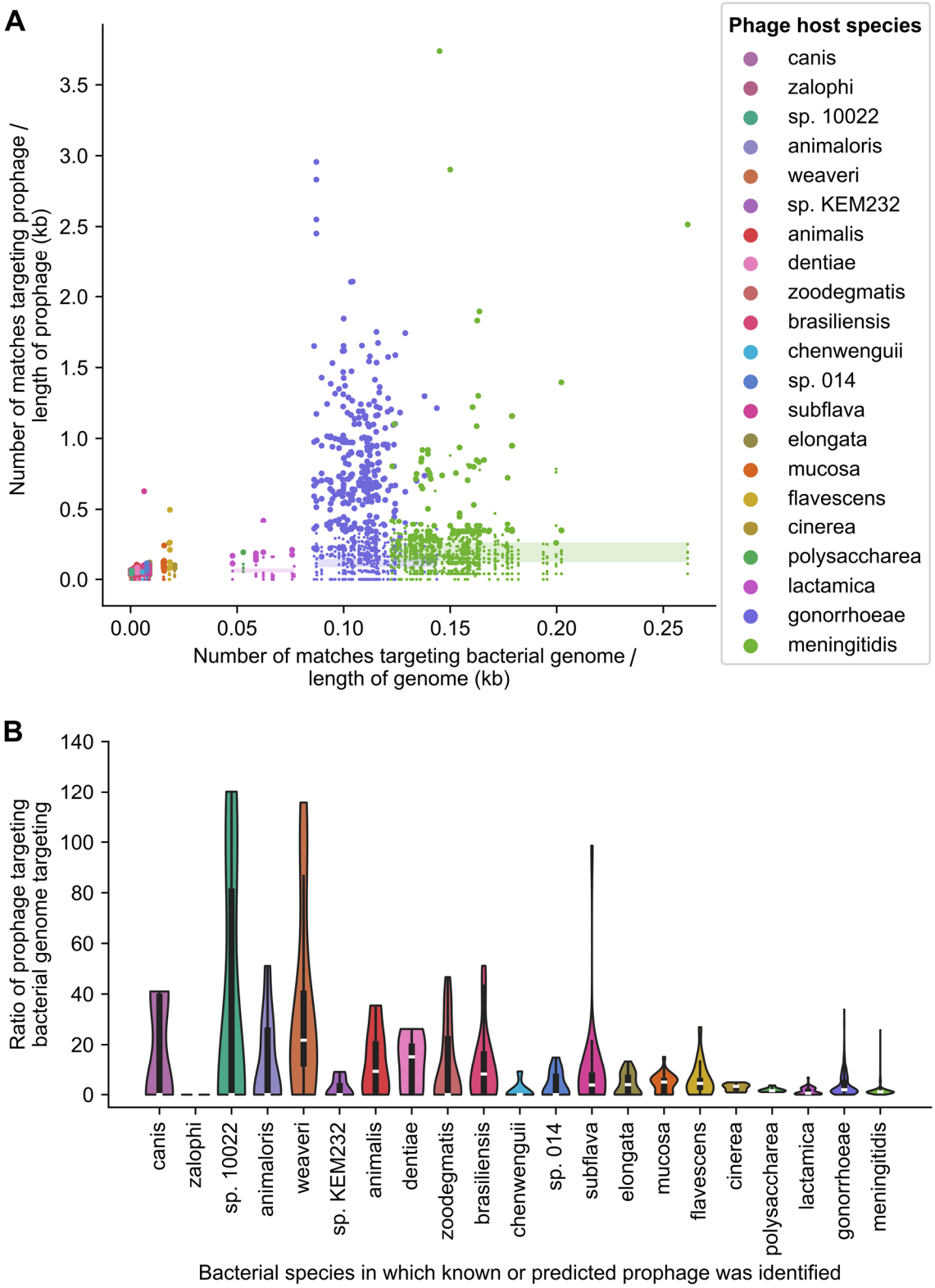

A
Number of matches targeting prophage / length of prophage (kb)
Number of matches targeting bacterial genome / length of genome (kb)
Phage host species
canis
zalophi
sp. 10022
animaloris
weaveri
sp. KEM232
animalis
dentiae
zoodegmatis
brasiliensis
chenwenguii
sp. 014
subflava
elongata
mucosa
flavescens
cinerea
polysaccharea
lactamica
gonorrhoeae
meningitidis
B
Ratio of prophage targeting bacterial genome targeting
Bacterial species in which known or predicted prophage was identified

Fig. 6. Comparing CRISPR targeting density of prophages to targeting of bacterial genome backbones. (A) Matches were identified between CRISPR spacers and bacterial genomes (at 100% identity over the entire spacer length). For each prophage, the density of matches within the prophage was compared to the backbone genome in which the prophage was identified (targeting densities are provided in Data Set S1, Tab 10). Y coordinates represent the number of matches within each prophage divided by the length of the prophage. X coordinates represent the number of matches within the bacterial genome in which the prophage was identified (excluding spacers targeting all prophages that were predicted in that genome) divided by the length of the bacterial genome (minus the length of all predicted prophages within that genome). Each data point represents a known *Neisseria* prophage or a predicted prophage. The color of the circle corresponds to the species in which the prophage was identified. Larger circles indicate prophages that have significantly higher targeting densities compared to the backbone (as described in the Methods and listed in Data Set S1, Tab 11). To highlight which prophages are targeted more highly than the backbone sequence, the range of targeting densities for genomes of each species is shown as a translucent box; the color of the box corresponds to the species. (B) Distribution of prophage/backbone CRISPR targeting ratios grouped according to the species in which the prophage was identified. Only dereplicated predicted prophages are included. CRISPR targeting ratios are provided in Data Set S1, Tab 10.

Interspecies CRISPR targeting is widespread among *Neisseria* species

We observed a high degree of interspecies targeting in our dataset. Out of 539 spacers that target known or predicted prophages, 53% (288/539) are involved in interspecies targeting of prophages. Furthermore, 35% (186/539) *only* target prophages found in other species.

We further examined interspecies CRISPR targeting using a network representation of targeting relationships between *Neisseria* species (**Fig. 7**). Specifically, the network includes targeting of backbone sequences, known *Neisseria* prophages, and predicted prophages that are more highly targeted compared to the backbone and that do not cluster with *Neisseria* plasmids.

The network is highly interconnected: all 21 *Neisseria* species included in this study are connected to ≥1 other species in the network, and 16 species are connected to ≥2 others (**Fig. 7**). Interestingly, *N. meningitidis* spacers match prophage and backbone sequences of 17 and 20 different species, respectively. Additionally, *N. gonorrhoeae* and *N. meningitidis* are each targeted by 7 *Neisseria* species.

Moreover, there are differences in the type of sequences targeted (edge color, **Fig. 7**). *N. meningitidis* spacers seem to predominantly target the backbone genome sequences of *N. meningitidis* and several other species (many pink arrows from *N. meningitidis*). In contrast, *N. subflava* and *N. lactamica* spacers primarily target prophages of other species (green arrows).

While these results suggest that interspecies targeting of *Neisseria* sequences is widespread, an alternative explanation is that spacers were exchanged between species. However, out of 3676 total spacers, only 2 identical spacers were present in genomes from different species (**Data Set S1**). Taken together, the findings above indicate that interspecies CRISPR targeting is common between *Neisseria* species.

Finally, we investigated whether *Neisseria* prophages may be targeted by other bacterial taxa using the tool CRISPRopenDB and its database of 11 million spacers (65). Four predictions identified in *N. animalis* are matched by a single spacer from *Eikenella corrodens* (another member of the Neisseriaceae), while a *N. elongata* prediction is matched by one *Aggregatibacter aphrophilus* spacer (**Data Set S1**).

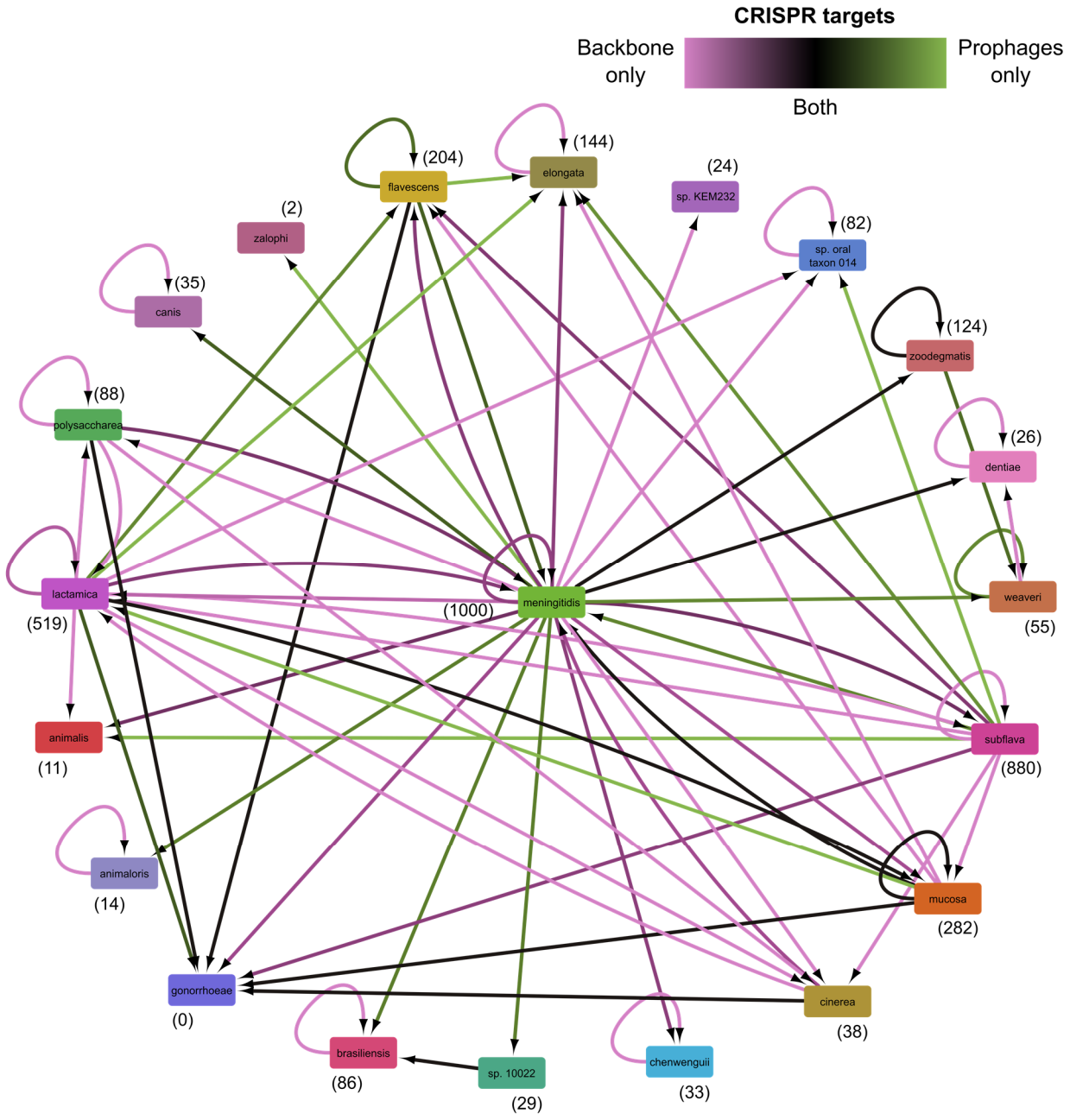

CRISPR targets
Backbone only
Prophages only
Both
flavescens
(204)
elongata
(144)
sp. KEM232
(24)
zalophi
(2)
sp. oral taxon 014
(82)
canis
(35)
zoodegmatis
(124)
polysaccharea
(88)
dentiae
(26)
lactamica
(519)
meningitidis
(1000)
weaveri
(55)
animalis
(11)
subflava
(880)
animaloris
(14)
mucosa
(282)
gonorrhoeae
(0)
cinerea
(38)
brasiliensis
(86)
sp. 10022
(29)
chenwenguii
(33)

Fig. 7. Interspecies CRISPR targeting of *Neisseria* prophages and bacterial genome backbones. Network representing intra- and inter-species CRISPR targeting of prophages and bacterial backbone sequences visualized using Cytoscape. Each node represents a bacterial species, and the adjacent number in parentheses indicates the total number of spacers encoded by that species. Nodes are connected by an edge if CRISPR spacers encoded by one species target another species. The direction of CRISPR targeting is indicated using an arrow that points to the species being targeted. Edge color indicates the relative number of spacers targeting prophages compared to the targeting of the backbone (i.e., bacterial genome excluding CRISPR arrays or sequence contained within any known or predicted prophages — not only prophages that are significantly targeted). Included in this analysis are backbone sequences, known *Neisseria* prophages, and predicted prophages that are targeted significantly more than the backbone and do not cluster with *Neisseria* plasmids. Information about CRISPR spacers and targeting is provided in Data Set S1, Tabs 9-10.

Many known and predicted *Neisseria* prophages have additional inferred host species

Elucidating the host range of phages is critical for understanding how they influence their microbial hosts, including their role in mobilizing DNA (12). CRISPR targeting data are frequently used to predict the bacterial hosts of phages (35–39); therefore, we took advantage of the extensive interspecies CRISPR targeting observed above (**Fig. 7**) to infer the hosts of known *Neisseria* prophages and predictions from all 3 tools.

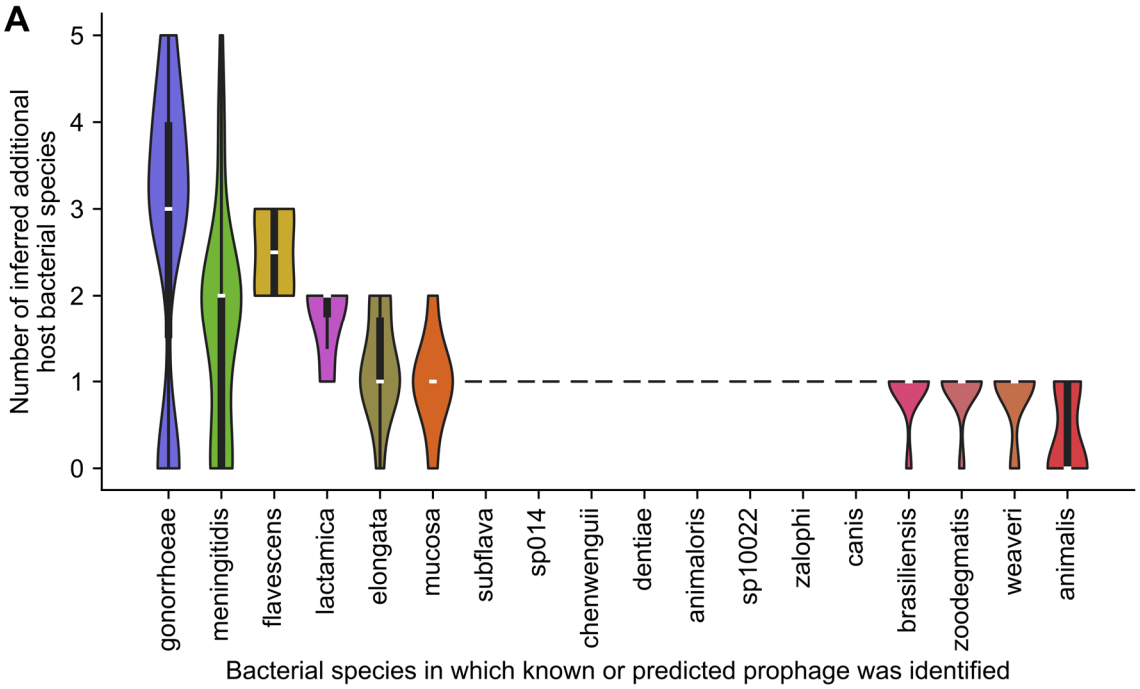
To increase the likelihood of examining true prophages (instead of chromosomal sequences), we only included dereplicated predictions that have significantly higher targeting densities compared to the backbone (i.e., significantly targeted predictions) identified above (**Fig. 6**).

We found that 98% of significantly targeted predicted prophages are likely shared with ≥1 other species (254/259) (**Fig. 8A, Table S2D**). Additionally, 100% of significantly targeted *N. gonorrhoeae* predictions (145/145) and 93% of significantly targeted *N. meningitidis* predictions (26/28) have ≥2 inferred additional host species. Furthermore, 50% (72/145) of predicted *N. gonorrhoeae* prophages have ≥4 additional host species (**Fig. 8A**).

Interspecies targeting data suggest that prophages are likely shared between a variety of *Neisseria* species (**Fig. 8B**). *N. gonorrhoeae* prophages (indicated by the many purple dots in multiple rows, **Fig. 8B**) are shared primarily with 3 closely related species, *N. meningitidis*, *N. polysaccharea*, and *N. lactamica* (median ANI between each one and *N. gonorrhoeae*: 93-95%, **Fig. 1**), and also with *N. cinerea*, which is less closely related to *N. gonorrhoeae* (median ANI: 90%).

Furthermore, *N. meningitidis* is predicted to share phages with a wide range of species, especially predictions that do not cluster with known *Neisseria* MGEs (many differently colored dots in the row corresponding to *N. meningitidis*, **Fig. 8B**). Interestingly, 7 of these species (*N. animaloris, N. canis*, *N. dentiae*, sp. 10022, *N. weaveri*, *N. zalophi*, *N. zoodegmatis*, **Fig. 8B**) are distantly related to *N. meningitidis* (median ANI: <80%, **Fig. 1**).

Finally, we compared the number of inferred additional host species of PhiSpy predictions between vConTACT subclusters. Significantly targeted members of 1318_0 (which includes the known prophages NgoΦ1 and NgoΦ2) have 3-5 additional host species, while members of other subclusters have 0-3 (**Fig. S6**). Together, our findings suggest that diverse *Neisseria* species may be infected by the same phages.


A
Number of inferred additional host bacterial species
0
1
2
3
4
5
gonorrhoeae
meningitidis
flavescens
lactamica
elongata
mucosa
subflava
sp014
chenwenguii
dentiae
animaloris
sp10022
zalophi
canis
brasiliensis
zoodegmatis
weaveri
animalis
Bacterial species in which known or predicted prophage was identified

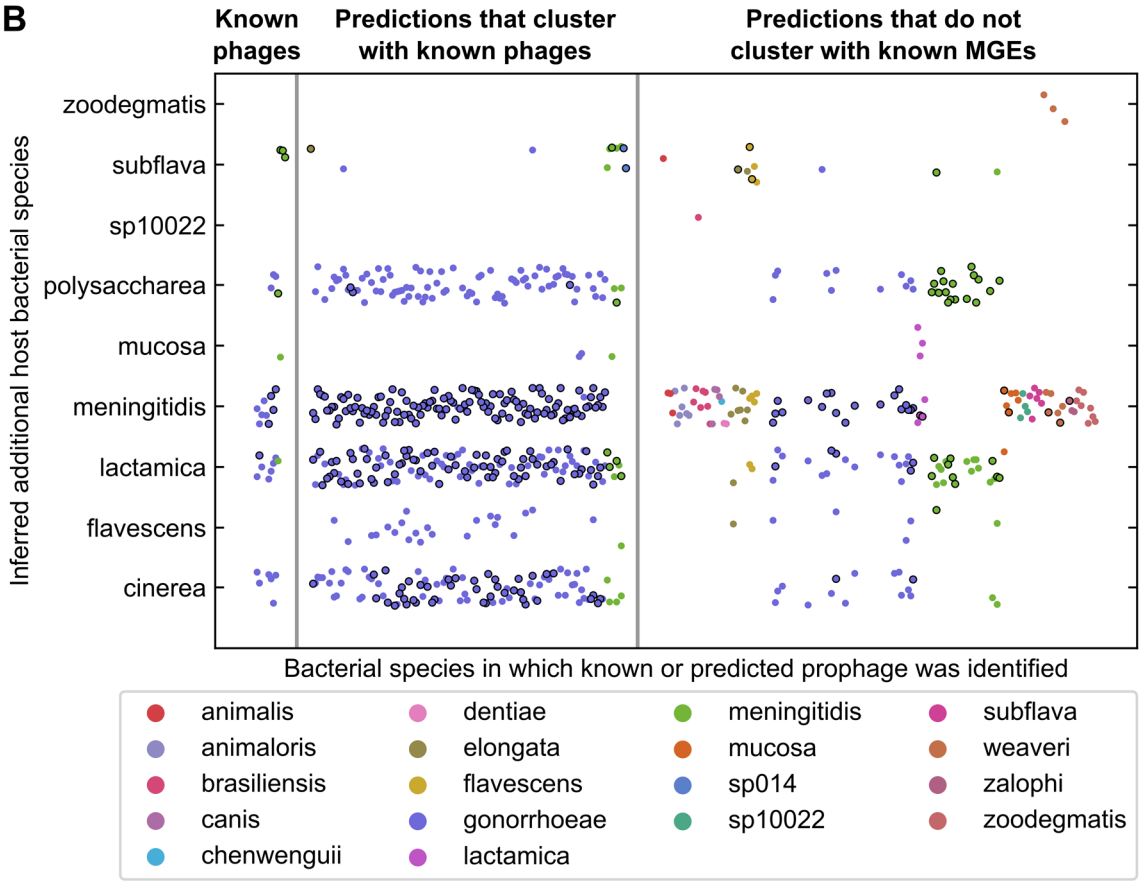

B
Known phages
Predictions that cluster with known phages
Predictions that do not cluster with known MGEs
Inferred additional host bacterial species
zoodegmatis
subflava
sp10022
polysaccharea
mucosa
meningitidis
lactamica
flavescens
cinerea
Bacterial species in which known or predicted prophage was identified
animalis
animaloris
brasiliensis
canis
chenwenguii
dentiae
elongata
flavescens
gonorrhoeae
lactamica
meningitidis
mucosa
sp014
sp10022
subflava
weaveri
zalophi
zoodegmatis

Fig. 8. Inference of additional host species of *Neisseria* phages using interspecies CRISPR targeting. (A) Distribution of the number of inferred additional host species of known *Neisseria* phages and dereplicated phages predicted by PhiSpy, VirSorter2, and Seeker (only those that have a significantly higher CRISPR targeting density compared to the bacterial genome backbone, Data Set S1, Tab 11). Interspecies matches between CRISPR spacers and prophages were used to infer the phage hosts. The data are grouped according to the species in which the prediction was identified. (B) Inferred sharing of prophages between the indicated *Neisseria* species. Each circle represents a known or predicted prophage, and the color of the circle corresponds to the species in which the prophage was identified (as indicated in the key). Circles with a black outline indicate that >1 spacer encoded by that species matches that prophage, while circles without an outline represent a single spacer match. Only prophages that have ≥1 inferred additional host species are shown. Predicted prophages are divided into 2 categories: those that cluster with known *Neisseria* prophages and those that do not cluster with known *Neisseria* MGEs (phages, plasmids, or the Gonococcal Genetic Island). The species names along the Y axis are the additional bacterial host species inferred using interspecies targeting data.

DISCUSSION

In this study, we sought to broaden the diversity of phages known to infect *Neisseria* species. We used 3 different virus prediction tools to scan 248 genomes of commensal and pathogenic *Neisseria* species for prophages. Clustering approaches revealed that many of these predictions are dissimilar from known *Neisseria* MGEs (prophages, plasmids, or the GGI) and phages described to infect other taxa. Therefore, we may have uncovered novel *Neisseria* prophage diversity.

We also identified prophages in several commensal *Neisseria* species that are highly similar to the *N. meningitidis* prophages Pnm1-2 and MuMenB, as well as Mu-like phages that infect *Haemophilus parasuis* and *Mannheimia haemolytica*, two Gammaproteobacteria. Although Pnm2 and MuMenB are defective, they may retain the ability to contribute genes to other co-infecting phages (55). Interestingly, several predicted *Neisseria* prophages are highly similar to Mu-like phages that infect other Betaproteobacteria: *Burkholderia cepacia* and *B. thailandensis*.

Commensal *Neisseria* species frequently colonize the upper respiratory tracts of humans and animals. Thus, these species may encounter *Haemophilus*, *Mannheimia*, and *Burkholderia* species within these niches and could be exposed to the same or similar phages. Additionally, we found highly similar predicted prophages in different *Neisseria* species, including distantly related species.

CRISPR systems provide immunity against MGEs and can be used as a historical record of encounters between microbes and MGEs. Previously, Zhang et al. identified CRISPR spacers that match the filamentous phage MDAΦ (41), and we found that all 13 known *Neisseria* prophages that we examined are matched by spacers. In addition to defending against MGEs, CRISPR systems of *N. meningitidis* and other microbes may also play a role in restricting the exchange of chromosomal DNA between species (41, 42, 66, 67). Our observations that backbone sequences of many *Neisseria* species are targeted by *N. meningitidis* spacers at a low level are consistent with this model.

To identify predictions that are likely to be MGEs, we compared targeting of each predicted prophage to targeting of the backbone. We found that 20% of dereplicated predicted prophages (259/1306) have a significantly higher targeting density compared to the backbone, and 74% of

these (191/259) do not cluster with known *Neisseria* MGEs. Therefore, these 191 predictions are candidates that warrant further study.

Furthermore, we found evidence of widespread interspecies targeting of predicted prophages by *Neisseria* spacers and used those data to infer additional host bacterial species of each prediction. Specifically, we focused on predictions that are significantly more highly targeted than the backbone, which are more likely to be actual prophages. Building upon previous findings (33, 34), our results suggest that multiple known and predicted phages may be able to infect multiple species of *Neisseria*, including species that are distantly related.

This study has several important limitations. Commensal species are underrepresented among the available *Neisseria* genome assemblies, and thus, also in this study. This underrepresentation limits our ability to compare patterns of CRISPR targeting between species. Additionally, our inference of additional host species is limited by whether genomes encode CRISPR spacers (e.g., *N. gonorrhoeae* genomes do not encode CRISPR arrays). Moreover, we only used 3 virus prediction tools and among them only PhiSpy was specifically designed to predict prophages (52).

Prophages are known to influence the fitness and virulence of many bacterial species, including *N. meningitidis* (15–20). Furthermore, considerable evidence suggests that accessory genes are shared extensively between *Neisseria* species and that commensal species are a reservoir of AR and virulence genes (43, 68–72). Therefore, it is critical to understand the diversity and host range of phages, which have the potential to mobilize genes among *Neisseria* species and alter their evolutionary trajectories. Further research on *N. gonorrhoeae* phages is also crucial for developing phage therapy approaches (73).

By combining clustering and CRISPR targeting analyses, we have identified candidate, novel *Neisseria* phages and inferred that several may infect multiple species within this bacterial genus. We hope that our findings may inform future studies seeking to elucidate the impact of viruses on *Neisseria* biology. Finally, we believe that our work may have implications for understanding the interactions occurring among the diverse *Neisseria* species that colonize the oropharynx and the phages that infect them.

METHODS

Generation of bacterial genome datasets. A set of high-quality bacterial genome assemblies was selected for prophage prediction. Specifically, *Neisseria* genome assemblies with N50 ≥250 kb and contigs ≤10 were downloaded from GenBank (44, 45). This set of 248 assemblies is referred to as the 'smaller *Neisseria* genome dataset' and is described in **Table S1A**. Additionally, for the identification of CRISPR arrays, a second dataset of bacterial genome assemblies was compiled as follows: *Neisseria* genome assemblies with N50 ≥15 kb were downloaded from GenBank (44, 45) and limited to the species represented in the smaller genome dataset. This second set of 2619 assemblies is referred to as the 'larger *Neisseria* genome dataset' and is described in **Table S1D**. All assemblies described above were downloaded on 3/30/2020.

Construction of bacterial phylogenetic tree and heatmap of average nucleotide identity. PubMLST was used to concatenate sequences of the 53 genes encoding ribosomal proteins of each bacterial genome within the smaller *Neisseria* genome dataset (74, 75). Then, the concatenated protein sequences were used to create a maximum-likelihood ribosomal multilocus sequence typing (rMLST) tree using RAxML (with GTRCAT model and 100 bootstrap replicates) (76) and visualized using iTOL v5 (77). The tree was rooted using midpoint rooting.

The pairwise average nucleotide identity (ANI) between bacterial genomes in the smaller dataset was calculated using FastANI (78) and visualized as a heatmap using the R package pheatmap (79, 80).

Prediction of prophages within bacterial genomes. PhiSpy, Virsorter2, and Seeker were used to predict prophages within the smaller set of *Neisseria* genomes (47, 49, 50). PhiSpy was run in strict mode without HMM searches after training using a custom training set. To generate the custom set, we combined the PhiSpy default reference genomes with the *N. gonorrhoeae* genome FA 1090 (GCA_000006845.1) annotated with proteins from the dsDNA tailed phages (NgoΦ1 – 5) and filamentous phages (NgoΦ6 – 9) (26, 28). We did not add any *N. meningitidis* genomes to the training set because *N. meningitidis* MC58 (GCA_000008805.1) and *N. meningitidis* Z2491 (GCA_000009105.1) were already included in the reference dataset. VirSorter2 and Seeker were run using default settings.

Dereplication of predicted prophages. An all-by-all BLASTn was performed separately with prophages predicted by each tool (81). Predicted prophages were dereplicated at 95% length aligned using a custom script (blast_average_link_hier_clust_output_clusters.py; https://github.com/Alan-Collins/Neisseria-prophage-paper). Information about dereplicated predictions and the predictions used as their representatives is found in **Table S2B**. Known *Neisseria* plasmids were dereplicated using the same method.

Hierarchical clustering of predicted prophages with *Neisseria* MGEs based on percent length aligned nucleotide sequence. First, an all-by-all BLASTn was performed on dereplicated predicted prophages, dereplicated known *Neisseria* plasmids, and the Gonococcal Genetic Island (81). Next, a distance matrix was created based on the percent length of aligned (PLA) sequence between pairs of MGEs (distance = 1 - PLA). The Python package SciPy v1.6.1 was then used to perform average/UPGMA linkage clustering on the distance matrix (82, 83). Next, a custom script (identify_blast_clusters.py; https://github.com/Alan-Collins/Neisseria-prophage-paper) was used to extract cluster memberships and extract the tree in Newick format, which was visualized using iTOL v5 (77).

vConTACT v.2.0 clustering of phages based on shared genes. Prodigal was used to predict the protein-coding genes of known *Neisseria* phages and dereplicated predicted prophages (84). Afterwards, Prodigal-generated protein sequences were clustered with reference viral genomes using vConTACT v.2.0 (60, 62). In addition to viruses from the RefSeq database (45, 85), we used 12,892 virus sequences provided by the Millard lab as reference viral sequences (http://millardlab.org/bioinformatics/lab-scripts/supplementing-and-colouring-vcontact2-clusters/). Protein sequence files and mapping files generated on 5/30/2020 were downloaded on 5/20/2021. Clustering with reference sequences was performed separately on prophages predicted by each virus prediction tool. The following vConTACT settings were used: --rel-mode 'Diamond', --db 'ProkaryoticViralRefSeq94-Merged' --pcs-mode MCL --vcs-mode ClusterONE. Networks were visualized in Cytoscape v3.8.2 using the edge-weighted spring-embedded layout algorithm, which positions highly similar viruses close together (86). Duplicate edges were removed from the network, and reference viruses are only shown if they are connected by an edge to a known or predicted *Neisseria* prophage.

Analyses of vConTACT viral clusters. We compared the similarity of predicted prophages to reference viruses between the three tools as follows. First, we examined each prediction's connections to reference viruses and identified the edge with the highest similarity score. Then, we compared the distributions of similarity scores between each tool using the Mann-Whitney U test implemented in the Python package SciPy (82, 83). To compare gene clusters between the members of each cluster, Prokka was used to predict and annotate ORFs within each phage

(using Pfam, TIGRFAM, and HAMAP databases) (87–90), and Clinker was used to generate comparisons of annotated predicted proteins (91).

Identification of CRISPR arrays in bacterial genomes. MinCED and CRISPRCasDB were used to identify CRISPR repeats in the smaller set of *Neisseria* genomes (92, 93). Using the CRISPR repeats identified in high-quality genomes, we next used a custom script (reps2spacers.py; https://github.com/Alan-Collins/Neisseria-prophage-paper) to run BLASTn (using -task blastn-short) and process results to identify spacers in the larger set of *Neisseria* genomes (81). CRISPRCasTyper was used to predict the CRISPR subtype associated with each of the identified repeats (94). To investigate sharing of identical spacers between genomes of different species, an all-by-all BLASTn (using -task blastn-short) was performed on all spacers that were identified in the larger *Neisseria* genome dataset (81).

Prediction of CRISPR targeting of prophages and bacterial genomes. BLASTn (using -task blastn-short) was used to identify matches between *Neisseria* CRISPR spacers (identified in the larger *Neisseria* genome dataset) and either prophages or high-quality bacterial genomes (81). Matches were filtered as follows: the spacer had to match the target with 100% identity over the entire length of the spacer. Additionally, matches between spacers and CRISPR arrays found within bacterial genomes or prophages were removed. CRISPRopenDB was used to predict CRISPR targeting of predicted prophages by other bacterial taxa using the default setting of 2 mismatches (65).

Comparison of CRISPR targeting of prophages vs. bacterial genome backbones and statistical testing. To compare the CRISPR targeting of predicted prophages and the genomes in which they were found (i.e., the backbone), we used the following method. As described above, matches between CRISPR spacers and targets in prophages or bacterial genomes were identified using BLASTn (81), and only hits with 100% identity over the full length of the spacer were kept. CRISPR spacers were excluded if they matched a CRISPR array found in either a predicted prophage or bacterial genome. Next, we quantified CRISPR targeting per kb; importantly, this was done differently for prophages and backbones as follows. For prophages, the targeting density is the number of CRISPR matches divided by the prophage length in kb. For backbones, targeting is the number of matches within the entire bacterial genome minus any matches to locations that are known/predicted to be part of an MGE; the length is calculated by subtracting the length of all MGEs identified in the genome from the length of the entire bacterial genome. Then, the backbone targeting density was calculated as the number of CRISPR targets (not in a known/predicted MGE) divided by the length of the genome minus the lengths of all known/predicted MGEs. Afterwards, we performed statistical testing to test whether there is a difference between the targeting density of the prophage and the backbone for each prophage; to do this, each kb of prophage or bacterial genome was treated as a separate datapoint. Specifically, we performed a Mann-Whitney U test to compare each of the CRISPR targeting counts for the separate kb bins between each prophage and the backbone using the Python package SciPy (82, 83). P-values were adjusted with the Holm-Sidak correction using the Python package Statsmodels (82, 95).

Inferring host bacterial species of known and predicted prophages. Results from the interspecies CRISPR targeting analysis described above were used to infer the additional host species of known *Neisseria* phages and dereplicated predictions made by PhiSpy, VirSorter2, and Seeker. Any species that was found to target a predicted prophage with ≥1 spacer was inferred to be a host of that predicted prophage. Only dereplicated predicted prophages that were found to have a significantly higher targeting density compared to the rest of the genome in which they were identified (as described above) were included in this analysis.

Data availability. All of the genome sequences used in this study were downloaded from GenBank. Accession numbers of the 2619 bacterial genome assemblies used in this study are included in **Table S1D**. Accession numbers of phage genomes and plasmids are found in **Table S1B** and **S1C**, respectively. Custom scripts created for analysis of data in this study are available at https://github.com/Alan-Collins/Neisseria-prophage-paper. Key data sets generated in this study and resulting analyses are provided in the Supplemental Material.

ACKNOWLEDGMENTS

We thank Julie Pryde, Awais Vaid, and other members of the Champaign-Urbana Public Health District for their collaboration, which inspired us to conduct this study.

This work was supported by an Allen Distinguished Investigator award to R.J.W (ADI12345) and the Carl R. Woese Institute for Genomic Biology Postdoctoral Fellowship to G.O.

REFERENCES

1. Liu G, Tang CM, Exley RMY 2015. Non-pathogenic *Neisseria*: members of an abundant, multi-habitat, diverse genus. Microbiology 161:1297–1312.

2. Unemo M, Seifert HS, Hook EW, Hawkes S, Ndowa F, Dillon J-AR. 2019. Gonorrhoea. Nat Rev Dis Primers 5:1–23.

3. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, Monnet DL, Pulcini C, Kahlmeter G, Kluytmans J, Carmeli Y, Ouellette M, Outterson K, Patel J, Cavaleri M, Cox EM, Houchens CR, Grayson ML, Hansen P, Singh N, Theuretzbacher U, Magrini N, Aboderin AO, Al-Abri SS, Awang Jalil N, Benzonana N, Bhattacharya S, Brink AJ, Burkert FR, Cars O, Cornaglia G, Dyar OJ, Friedrich AW, Gales AC, Gandra S, Giske CG, Goff DA, Goossens H, Gottlieb T, Guzman Blanco M, Hryniewicz W, Kattula D, Jinks T, Kanj SS, Kerr L, Kieny M-P, Kim YS, Kozlov RS, Labarca J, Laxminarayan R, Leder K, Leibovici L, Levy-Hara G, Littman J, Malhotra-Kumar S, Manchanda V, Moja L, Ndoye B, Pan A, Paterson DL, Paul M, Qiu H, Ramon-Pardo P, Rodríguez-Baño J, Sanguinetti M, Sengupta S, Sharland M, Si-Mehand M, Silver LL, Song W, Steinbakk M, Thomsen J, Thwaites GE, van der Meer JW, Van Kinh N, Vega S, Villegas MV, Wechsler-Fördös A, Wertheim HFL, Wesangula E, Woodford N, Yilmaz FO, Zorzet A. 2018. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. The Lancet Infectious Diseases 18:318–327.

4. Centers for Disease Control and Prevention (U.S.). 2019. Antibiotic resistance threats in the United States, 2019. Centers for Disease Control and Prevention (U.S.).

5. Linz B, Schenker M, Zhu P, Achtman M. 2000. Frequent interspecific genetic exchange between commensal neisseriae and *Neisseria meningitidis*. Molecular Microbiology 36:1049–1058.

6. Maiden MC. 2008. Population genomics: diversity and virulence in the *Neisseria*. Curr Opin Microbiol 11:467–471.

7. Corander J, Connor TR, O'Dwyer CA, Kroll JS, Hanage WP. 2012. Population structure in the *Neisseria*, and the biological significance of fuzzy species. Journal of The Royal Society Interface 9:1208–1215.

8. Arnold BJ, Huang I-T, Hanage WP. 2021. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol 1–13.

9. Brüssow H, Canchaya C, Hardt W-D. 2004. Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. Microbiol Mol Biol Rev 68:560–602.

10. Canchaya C, Fournous G, Brüssow H. 2004. The impact of prophages on bacterial chromosomes. Molecular Microbiology 53:9–18.

11. Fortier L-C, Sekulovic O. 2013. Importance of prophages to evolution and virulence of bacterial pathogens. Virulence 4:354–365.

12. Koskella B, Brockhurst MA. 2014. Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. FEMS Microbiol Rev 38:916–931.

13. Cehovin A, Lewis SB. 2017. Mobile genetic elements in *Neisseria gonorrhoeae*: movement for change. Pathogens and Disease 75.

14. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. 2021. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. Nat Rev Microbiol 19:347–359.

15. Waldor MK, Mekalanos JJ. 1996. Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. Science.

16. Karaolis DKR, Somara S, Maneval DR, Johnson JA, Kaper JB. 1999. A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. Nature 399:375–379.

17. Bille E, Ure R, Gray SJ, Kaczmarski EB, McCarthy ND, Nassif X, Maiden MCJ, Tinsley CR. 2008. Association of a Bacteriophage with Meningococcal Disease in Young Adults. PLOS ONE 3:e3885.

18. Bille E, Meyer J, Jamet A, Euphrasie D, Barnier J-P, Brissac T, Larsen A, Pelissier P, Nassif X. 2017. A virulence-associated filamentous bacteriophage of *Neisseria meningitidis* increases host-cell colonisation. PLOS Pathogens 13:e1006495.

19. Mai-Prochnow A, Hui JGK, Kjelleberg S, Rakonjac J, McDougald D, Rice SA. 2015. 'Big things in small packages: the genetics of filamentous phage and effects on fitness of their host.' FEMS Microbiology Reviews 39:465–487.

20. Taylor VL, Fitzpatrick AD, Islam Z, Maxwell KL. 2019. Chapter One - The Diverse Impacts of Phage Morons on Bacterial Fitness and Virulence, p. 1–31. *In* Kielian, M, Mettenleiter, TC, Roossinck, MJ (eds.), Advances in Virus Research. Academic Press.

21. Morgan GJ, Hatfull GF, Casjens S, Hendrix RW. 2002. Bacteriophage Mu genome sequence: analysis and comparison with Mu-like prophages in *Haemophilus*, *Neisseria* and *Deinococcus*. Edited by M. Gottesman. Journal of Molecular Biology 317:337–359.

22. Clokie MR, Millard AD, Letarov AV, Heaphy S. 2011. Phages in nature. Bacteriophage 1:31–45.

23. Grose JH, Casjens SR. 2014. Understanding the enormous diversity of bacteriophages: The tailed phages that infect the bacterial family Enterobacteriaceae. Virology 468–470:421–443.

24. Kawai M, Uchiyama I, Kobayashi I. 2005. Genome Comparison In Silico in *Neisseria* Suggests Integration of Filamentous Bacteriophages by their Own Transposase. DNA Research 12:389–401.

25. Bille E, Zahar J-R, Perrin A, Morelle S, Kriz P, Jolley KA, Maiden MCJ, Dervin C, Nassif X, Tinsley CR. 2005. A chromosomally integrated bacteriophage in invasive meningococci. J Exp Med 201:1905–1913.

26. Piekarowicz A, Majchrzak M, Kłyz A, Adamczyk-Popławska M. 2006. Analysis of the filamentous bacteriophage genomes integrated into *Neisseria gonorrhoeae* FA1090 chromosome. Pol J Microbiol 55:251–260.

27. Meyer J, Brissac T, Frapy E, Omer H, Euphrasie D, Bonavita A, Nassif X, Bille E 2016. Characterization of MDAΦ, a temperate filamentous bacteriophage of *Neisseria meningitidis*. Microbiology 162:268–282.

28. Piekarowicz A, Kłyż A, Majchrzak M, Adamczyk-Popławska M, Maugel TK, Stein DC. 2007. Characterization of the dsDNA prophage sequences in the genome of *Neisseria gonorrhoeae* and visualization of productive bacteriophage. BMC Microbiol 7:66.

29. Klee SR, Nassif X, Kusecek B, Merker P, Beretti J-L, Achtman M, Tinsley CR. 2000. Molecular and Biological Analysis of Eight Genetic Islands That Distinguish *Neisseria meningitidis* from the Closely Related Pathogen *Neisseria gonorrhoeae*. Infect Immun 68:2082–2095.

30. Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T, Davies RM, Davis P, Devlin K, Feltwell T, Hamlin N, Holroyd S, Jagels K, Leather S, Moule S, Mungall K, Quail MA, Rajandream M-A, Rutherford KM, Simmonds M, Skelton J, Whitehead S, Spratt BG, Barrell BG. 2000. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature 404:502–506.

31. Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ, Nelson WC, Gwinn ML, DeBoy R, Peterson JD, Hickey EK, Haft DH, Salzberg SL, White O, Fleischmann RD, Dougherty BA, Mason T, Ciecko A, Parksey DS, Blair E, Cittone H, Clark EB, Cotton MD, Utterback TR, Khouri H, Qin H, Vamathevan J, Gill J, Scarlato V, Masignani V, Pizza M, Grandi G, Sun L, Smith HO, Fraser CM, Moxon ER, Rappuoli R, Venter JC. 2000. Complete Genome Sequence of *Neisseria meningitidis* Serogroup B Strain MC58. Science.

32. Masignani V, Giuliani MM, Tettelin H, Comanducci M, Rappuoli R, Scarlato V. 2001. Mu-Like Prophage in Serogroup B *Neisseria meningitidis* Coding for Surface-Exposed Antigens. Infect Immun 69:2580–2588.

33. Piekarowicz A, Kłyż A, Majchrzak M, Szczêsna E, Piechucki M, Kwiatek A, Maugel TK, Stein DC. 2014. *Neisseria gonorrhoeae* Filamentous Phage NgoΦ6 Is Capable of Infecting a Variety of Gram-Negative Bacteria. Journal of Virology 88:1002–1010.

34. Al Suwayyid BA, Rankine-Wilson L, Speers DJ, Wise MJ, Coombs GW, Kahler CM. 2020. Meningococcal Disease-Associated Prophage-Like Elements Are Present in *Neisseria gonorrhoeae* and Some Commensal *Neisseria* Species. Genome Biol Evol 12:3938–3950.

35. Khot V, Strous M, Hawley AK. 2020. Computational approaches in viral ecology. Computational and Structural Biotechnology Journal 18:1605–1612.

36. Kothari A, Roux S, Zhang H, Prieto A, Soneja D, Chandonia J-M, Spencer S, Wu X, Altenburg S, Fields MW, Deutschbauer AM, Arkin AP, Alm EJ, Chakraborty R, Mukhopadhyay A. Ecogenomics of Groundwater Phages Suggests Niche Differentiation Linked to Specific Environmental Tolerance. mSystems 6:e00537-21.

37. Andersson AF, Banfield JF. 2008. Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. Science 320:1047–1050.

38. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. 2016. Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol Rev 40:258–272.

39. Benler S, Yutin N, Antipov D, Rayko M, Shmakov S, Gussow AB, Pevzner P, Koonin EV. 2021. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. Microbiome 9:78.

40. Tesson F, Herve A, Touchon M, d'Humieres C, Cury J, Bernheim A. 2021. Systematic and quantitative view of the antiviral arsenal of prokaryotes.

41. Zhang Y, Heidrich N, Ampattu BJ, Gunderson CW, Seifert HS, Schoen C, Vogel J, Sontheimer EJ. 2013. Processing-Independent CRISPR RNAs Limit Natural Transformation in *Neisseria meningitidis*. Molecular Cell 50:488–503.

42. Zhang Y. 2017. The CRISPR-Cas9 system in *Neisseria* spp. Pathog Dis 75:ftx036.

43. Calder A, Menkiti CJ, Çağdaş A, Lisboa Santos J, Streich R, Wong A, Avini AH, Bojang E, Yogamanoharan K, Sivanesan N, Ali B, Ashrafi M, Issa A, Kaur T, Latif A, Mohamed HAS, Maqsood A, Tamang L, Swager E, Stringer AJ, Snyder LASY 2020. Virulence genes and previously unexplored gene clusters in four commensal *Neisseria* spp. isolated from the human throat expand the neisserial gene repertoire. Microbial Genomics 6:e000423.

44. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. Nucleic Acids Research 41:D36–D42.

45. NCBI Resource Coordinators. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 46:D8–D13.

46. Mustapha MM, Lemos APS, Griffith MP, Evans DR, Marx R, Coltro ESF, Siebra CA, Timm L, Ribeiro H, Monteiro A, Pasculle AW, Marsh JW, Van Tyne D, Harrison LH, Sacchi CT. 2020. Two Cases of Newly Characterized *Neisseria* Species, Brazil. Emerg Infect Dis 26:366–369.

47. Akhter S, Aziz RK, Edwards RA. 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Research 40:e126–e126.

48. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ 3:e985.

49. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, Pratama AA, Gazitúa MC, Vik D, Sullivan MB, Roux S. 2021. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome 9:37.

50. Auslander N, Gussow AB, Benler S, Wolf YI, Koonin EV. 2020. Seeker: alignment-free identification of bacteriophage genomes by deep learning. Nucleic Acids Research 48:e121–e121.

51. Glickman C, Hendrix J, Strong M. 2021. Simulation study and comparative evaluation of viral contiguous sequence identification tools. BMC Bioinformatics 22:329.

52. Roach MJ, McNair K, Giles SK, Inglis L, Pargin E, Roux S, Decewicz P, Edwards RA. 2021. Philympics 2021: Prophage Predictions Perplex Programs.

53. Ho SFS, Millard AD, Schaik W van. 2021. Comprehensive benchmarking of tools to identify phages in metagenomic shotgun sequencing data.

54. Hendrix RW, Hatfull GF, Ford ME, Smith MCM, Neil Burns R. 2002. Chapter 12 - Evolutionary Relationships Among Diverse Bacteriophages and Prophages: All The World's a Phage, p. 133–VI. *In* Syvanen, M, Kado, CI (eds.), Horizontal Gene Transfer (Second Edition). Academic Press, London.

55. Casjens S. 2003. Prophages and bacterial genomics: what have we learned so far? Molecular Microbiology 49:277–300.

56. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR, Hendrix RW, Hatfull GF. 2003. Origins of Highly Mosaic Mycobacteriophage Genomes. Cell 113:171–182.

57. Rohwer F, Edwards R. 2002. The Phage Proteomic Tree: a Genome-Based Taxonomy for Phage. Journal of Bacteriology 184:4529–4535.

58. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. 2008. Reticulate Representation of Evolutionary and Functional Relationships between Phage Genomes. Molecular Biology and Evolution 25:762–777.

59. Roux S, Hallam SJ, Woyke T, Sullivan MB. 2015. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. eLife 4:e08490.

60. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol 37:632–639.

61. Turner D, Kropinski AM, Adriaenssens EM. 2021. A Roadmap for Genome-Based Phage Taxonomy. 3. Viruses 13:506.

62. Bolduc B, Jang HB, Doulcier G, You Z-Q, Roux S, Sullivan MB. 2017. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. PeerJ 5:e3243.

63. Andersen BM, Steigerwalt AG, O'Connor SP, Hollis DG, Weyant RS, Weaver RE, Brenner DJ. 1993. *Neisseria weaveri* sp. nov., formerly CDC group M-5, a gram-negative bacterium associated with dog bite wounds. Journal of Clinical Microbiology 31:2456–2466.

64. Frye SA, Nilsen M, Tønjum T, Ambur OH. 2013. Dialects of the DNA Uptake Sequence in *Neisseriaceae*. PLOS Genetics 9:e1003458.

65. Dion MB, Plante P-L, Zufferey E, Shah SA, Corbeil J, Moineau S. 2021. Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. Nucleic Acids Research 49:3127–3138.

66. Brodt A, Lurie-Weinberger MN, Gophna U. 2011. CRISPR loci reveal networks of gene exchange in archaea. Biology Direct 6:65.

67. Turgeman-Grott I, Joseph S, Marton S, Eizenshtein K, Naor A, Soucy SM, Stachler A-E, Shalev Y, Zarkor M, Reshef L, Altman-Price N, Marchfelder A, Gophna U. 2019. Pervasive acquisition of CRISPR memory driven by inter-species mating of archaea can limit gene transfer and influence speciation. Nat Microbiol 4:177–186.

68. Bennett JS, Bentley SD, Vernikos GS, Quail MA, Cherevach I, White B, Parkhill J, Maiden MC. 2010. Independent evolution of the core and accessory gene sets in the genus *Neisseria*: insights gained from the genome of *Neisseria lactamica* isolate 020-06. BMC Genomics 11:652.

69. Marri PR, Paniscus M, Weyand NJ, Rendón MA, Calton CM, Hernández DR, Higashi DL, Sodergren E, Weinstock GM, Rounsley SD, So M. 2010. Genome Sequencing Reveals Widespread Virulence Gene Exchange among Human *Neisseria* Species. PLOS ONE 5:e11835.

70. Fiore MA, Raisman JC, Wong NH, Hudson AO, Wadsworth CB. 2020. Exploration of the *Neisseria* Resistome Reveals Resistance Mechanisms in Commensals That May Be Acquired by *N. gonorrhoeae* through Horizontal Gene Transfer. Antibiotics (Basel) 9:656.

71. Spratt BG, Zhang QY, Jones DM, Hutchison A, Brannigan JA, Dowson CG. 1989. Recruitment of a penicillin-binding protein gene from *Neisseria flavescens* during the emergence of penicillin resistance in *Neisseria meningitidis*. PNAS 86:8988–8992.

72. Wadsworth CB, Arnold BJ, Sater MRA, Grad YH. 2018. Azithromycin Resistance through Interspecific Acquisition of an Epistasis-Dependent Efflux Pump Component and Transcriptional Regulator in *Neisseria gonorrhoeae*. mBio 9.

73. Cater K, Międzybrodzki R, Morozova V, Letkiewicz S, Łusiak-Szelachowska M, Rękas J, Weber-Dąbrowska B, Górski A. 2021. Potential for Phages in the Treatment of Bacterial Sexually Transmitted Infections. 9. Antibiotics 10:1030.

74. Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res 3:124.

75. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MCJY 2012. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology 158:1005–1015.

76. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

77. Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research 49:W293–W296.

78. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114.

79. R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

80. Kolde, Raivo. 2019. pheatmap: Pretty Heatmaps. R package version 1.0.12:https://CRAN.R-project.org/package=pheatmap.

81. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403–410.

82. van Rossum, G. 2001. Python/C API Reference Manual.

83. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272.

84. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

85. Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 33:D501–D504.

86. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 13:2498–2504.

87. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

88. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: The protein families database in 2021. Nucleic Acids Research 49:D412–D419.

89. Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res 31:371–373.

90. Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, de Castro E, Baratin D, Cuche BA, Bougueleret L, Poux S, Redaschi N, Xenarios I, Bridge A. 2015. HAMAP in 2015: updates to the protein family classification and annotation system. Nucleic Acids Research 43:D1064–D1070.

91. Gilchrist CLM, Chooi Y-H. 2021. clinker & clustermap.js: automatic generation of gene cluster comparison figures. Bioinformatics 37:2473–2475.

92. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. 2007. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics 8:209.

93. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, Rocha EPC, Vergnaud G, Gautheret D, Pourcel C. 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Research 46:W246–W251.

94. Russel J, Pinilla-Redondo R, Mayo-Muñoz D, Shah SA, Sørensen SJ. 2020. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. The CRISPR Journal 3:462–469.

95. Seabold S, Perktold J. 2010. Statsmodels: Econometric and Statistical Modeling with Python, p. 92–96. *In* . Proceedings of the 9th Python in Science Conference, Austin, Texas.

SUPPLEMENTAL MATERIAL

Fig. S1. Comparison of known *Neisseria* prophages to predictions made by three bioinformatic tools. (A and B) Visualizations of locations of known and predicted prophages made using SnapGene software (from Insightful Science; available at snapgene.com) and modified. PhiSpy, VirSorter2, or Seeker were used to predict prophages in two bacterial genomes: *N. gonorrhoeae* FA10 (A) and *N. meningitidis* Z2491 (B). Known dsDNA prophages are shown in dark grey, filamentous prophages in light grey, and predictions are shown in different colors as indicated. Known *Neisseria* prophages are described in Table S1B, and information about each prediction is provided in Table S2A.

Fig. S2. The distribution of predicted prophage lengths for each virus prediction tool. (A to C) Length in kb is shown for dereplicated PhiSpy (A), VirSorter2 (B), and (C) Seeker predictions. Information about each prediction is provided in Table S2A and dereplication of predictions in Table S2B.

Fig. S3. Few predictions cluster with known *Neisseria* plasmids or genetic island based on percent-length aligned nucleotide sequences. (A and B) Dendrogram representing average-linkage hierarchical clustering of known *Neisseria* plasmids (A) and the Gonococcal Genetic Island (GGI) (B) with dereplicated predictions from all tools based on percent length aligned by BLASTn. Clustering was performed using SciPy and the resulting dendrograms were visualized with iTOL. Distinct clusters are highlighted by alternating light and dark grey clade shading. Color strips indicate the *Neisseria* species in which the MGE was identified and whether the MGE was previously known or predicted in this study. Beneath the color strips is an identifying cluster number for each cluster. Only predictions that cluster with known *Neisseria* plasmids or the GGI are included in the dendrograms. Information about plasmids and the GGI are provided in Table S1C. Hierarchical clustering memberships of each indicated cluster are presented in Data Set S1, Tab 2.

Fig. S4. A greater number of PhiSpy and VirSorter2 predictions are connected to reference viruses compared to Seeker predictions. (A and B) vConTACT v.2.0-generated networks of dereplicated VirSorter2 (A) or Seeker (B) predictions with reference viruses visualized with Cytoscape using an edge-weighted spring-embedded algorithm. Nodes represent reference viruses (dark grey), predicted prophages (color corresponding to the *Neisseria* species in which they were identified), or known *Neisseria* phages (dark grey outlined in the color corresponding to the bacterial host species). Edges show the vConTACT v.2.0-generated similarity score between each pair of viruses (only similarity scores ≥1 are included in the network). Highly similar viruses are positioned close together. Only reference viruses that connect to ≥1 predicted prophage are included in the network.

Fig. S5. PhiSpy and VirSorter2 predictions are more similar to reference viruses compared to Seeker predictions. vConTACT v.2.0-generated similarity scores between each predicted prophage and its most similar reference virus in the network. Similarity scores are shown for each virus pair and grouped by the tool that predicted the prophage (PhiSpy, VirSorter2, or Seeker). Distributions of similarity scores were compared between each tool using the Mann-Whitney U test. Similarity scores (edge weights) for each network are provided in Data Set S1, Tabs 4-6.

Fig. S6. Inferred additional host species of each PhiSpy prediction grouped by vConTACT cluster. Interspecies CRISPR targeting data were used to infer additional host species for each predicted prophage. The distribution of the number of inferred additional host species is shown for PhiSpy-predicted prophages (only dereplicated predictions that have a significantly higher CRISPR targeting density compared to the bacterial genome backbone;

listed in Data Set S1, Tab 11). The data are grouped according to the vConTACT v.2.0 subcluster.

Table S1. (A) Smaller set of *Neisseria* genome assemblies used in this study to predict prophages. (B) Prophages previously identified in *Neisseria* genomes. (C) *Neisseria* plasmids and the Gonococcal Genetic Island. (D) Larger set of *Neisseria* genome assemblies used in this study to identify CRISPR arrays.

Table S2. (A) Prophages predicted in this study. (B) Dereplication of predicted prophages at 95% length aligned. (C) Comparison of clustering and CRISPR targeting of predicted prophages between tools. (D) Inferred additional *Neisseria* host species of significantly targeted prophages.

Data Set S1. Data included in each tab: (1) Bacterial genome maximum-likelihood ribosomal MLST tree file. (2) Predictions that cluster with *Neisseria* plasmids and the Gonococcal Genetic Island by hierarchical clustering. (3) vConTACT subcluster information and memberships. (4-6) Edge weights of vConTACT networks. (7) CRISPR arrays identified in each *Neisseria* genome. (8) *Neisseria* CRISPR repeat sequences and associated subtypes. (9) *Neisseria* CRISPR spacer sequences. (10) CRISPR targeting densities of known and predicted prophages compared to genome backbone targeting density. (11) Dereplicated prophages with CRISPR targeting densities that are significantly higher than the bacterial genome. (12) vConTACT subclusters that include significantly targeted predicted prophages. (13) Identical spacers identified in genomes of different *Neisseria* species. (14) Matches between spacers encoded by other bacterial taxa and *Neisseria* predicted prophages.