Interrogation of the integrated mobile genetic elements in gut-associated
Bacteroidaceae with a consensus prediction approach

Running title: Integrated MGEs in gut-associated *Bacteroidaceae*

Danielle E. Campbell^{a,b,c}, Joseph R. Leigh^b, Ted Kim^{b,c}, Whitney E. England^{b,c,d}, Rachel
J. Whitaker^{b,c}, Patrick H. Degnan^{c,##}

^a Department of Medicine, Division of Infectious Diseases, Edison Family Center for
Genome Sciences & Systems Biology, Washington University School of Medicine, Saint
Louis, Missouri, USA

^b Carl R. Woese Institute for Genomic Biology, University of Illinois, Urbana, Illinois, USA

^c Department of Microbiology, University of Illinois, Urbana, Illinois, USA

^d Department of Pharmaceutical Sciences, University of California, Irvine, Irvine, CA,
USA

^e Department of Microbiology and Plant Pathology, University of California, Riverside,
California, USA

[#]Address correspondence to Patrick H. Degnan, patrick.degnan@ucr.edu

ABSTRACT

Exploration of mobile genetic element (MGE) diversity and relatedness is vital to understanding microbial communities, especially the gut microbiome, where the mobilization of antibiotic resistance and pathogenicity genes has important clinical consequences. Current MGE prediction tools are biased toward elements similar to previously-identified MGEs, especially tailed phages of proteobacterial hosts. Further, there is a need for methods to examine relatedness and gene sharing among MGEs. We present VIC SIN, a consensus approach for MGE prediction and clustering of predictions to provide classification. Testing of VIC SIN on datasets of *Pseudomonas aeruginosa* and *Bacteroides fragilis* genomes suggests VIC SIN is the optimal approach to predict integrated MGEs from poorly-explored host taxa, because of its increased sensitivity and accuracy. We applied VIC SIN to a dataset of gut-associated *Bacteroidaceae* genomes, identifying 816 integrated MGEs falling into 95 clusters, most of which are novel. VIC SIN's fast and simple network-building scheme revealed a high degree of gene sharing within and between related MGE clusters. Shared gene functions across MGEs include core mobilization functions and accessory gene content, such as type VI secretion systems and antibiotic resistance genes. The MGEs identified here encode a large portion of unknown gene content, emphasizing the fact that the full diversity of MGEs and the factors they encode remain very poorly understood. Together, this work motivates more exploration of the gut mobilome, which is likely one of the most potent drivers of microbial evolution in the human microbiome.

IMPORTANCE

Mobile genetic elements (MGEs), including phages and integrative and conjugative elements (ICEs), drive the diversity and function of microbial communities through horizontal gene transfer. Current tools to predict MGEs in genomic sequence data are highly focused on phages, and are biased against the discovery of novel MGEs. We present VIC SIN, a consensus approach to MGE prediction that is able to find a diversity of MGEs, particularly in poorly-understood bacterial taxa. By applying VIC SIN to a large database of diverse *Bacteroidaceae* genomes, we have been able to get a distinct view of the gut mobilome, extending beyond the phageome. These novel MGEs belong to related groups, sharing a significant amount of functional gene content within and between groups, supporting a mosaic model of evolution for ICEs. Understanding how phages evolve in *Bacteroidaceae* hosts, however, remains elusive and highlights the need for more experimental research.

KEYWORDS

Mobile genetic element, integrative and conjugative element, bacteriophage, recombination, *Bacteroidaceae*, *Bacteroides*

INTRODUCTION

Horizontal gene transfer (HGT), the sharing of DNA between co-existing organisms, is a primary mechanism shaping microbial communities, and a driver of genomic plasticity and adaptation (1–5). Across microbial communities, it is postulated that the vast majority of HGT is mediated by mobile genetic elements (MGEs) (4–7). MGEs encode their own mechanisms that allow for their transfer between and their

maintenance within cellular hosts, thereby circumventing two of the main barriers to HGT: cell entry and persistence (4). HGT occurs at especially high rates in the human gut microbiome (8, 9), which has been attributed to both the high density of microbes (10, 11) and the activity of some MGEs to host-associated conditions, such as inflammation (12, 13). MGE activity has important consequences for all microbial hosts and communities, including in the human gut. Within the gut microbiome, MGEs disseminate antibiotic resistance genes, virulence factors, and other advantageous traits, both in symbiotic and pathogenic contexts (3, 14–17). Characterizing MGE diversity is crucial for understanding how genetic diversity, genes of interest, and relevant traits are maintained within and disseminated between microbes.

The process of characterizing MGE diversity presents two major challenges: (i) accurate and sensitive detection of MGEs, and (ii) MGE classification. An increasingly large number of tools have been developed to search genomic sequence data for putative MGEs using distinct computational approaches (18). Each of these individual approaches to MGE prediction has advantages and limitations. Similarity-based search tools are biased towards the identification of MGEs similar to previously characterized MGEs, and therefore perform poorly at identifying novel MGEs (18, 19). *De novo* prediction tools use generic models of MGE sequence characteristics but exhibit variable accuracy and sensitivity depending on the microbial host (18, 19). Genome alignment tools identify all non-core genomic elements, which include MGEs but can also include other accessory genomic features, such as recombination hotspots (20), cell surface modification loci (20, 21), and CRISPR-Cas systems (22). Machine learning and deep learning-based tools may require large dataset inputs or customized training

94 data, which may not be available for most microbial systems (18, 23, 24). Furthermore,
 95 MGE identification is most thoroughly tested on clinically-relevant and well-studied host
 96 microbes, especially those in the phylum *Proteobacteria* (19, 23–27), and is largely
 97 tailored to identifying integrated viruses (*i.e.*, prophages), neglecting the numerous other
 98 types of MGEs (18, 28). Although testing these tools on lesser-understood microbial
 99 hosts and MGEs is intrinsically difficult, the current lack of representation across host
 100 and MGE diversity has resulted in biases in MGE prediction (27).

101 The second major challenge to understanding MGE diversity is understanding
 102 how they are related, and their further classification into related groups. Although many
 103 types of MGEs have been identified, we will focus on viruses and integrative and
 104 conjugative elements (ICEs), which appear to be the predominant means of HGT in the
 105 gut microbiome (29). Traditionally, viral classification has been accomplished by the
 106 International Committee for the Taxonomy of Viruses (ICTV), which relies on a
 107 combination of viral characteristics to classify new viruses, including viral particle
 108 morphology, genome composition, and homology (30). There is no universally accepted
 109 method for the classification of most other MGEs. One method for ICE classification is
 110 based on integrase gene homology and synteny of ICE backbone features (31),
 111 however the mosaic nature of ICEs suggests classification based heavily on a single
 112 gene may not be accurate (32). The mosaicism within ICEs and across all MGE types is
 113 a result of recombination events between otherwise unrelated MGEs coexisting within a
 114 single host cell (32, 33). While the continuous nature of MGE diversity is perhaps the
 115 biggest challenge to their classification, it also calls for a generalizable classification

scheme that does not currently exist, in order to better understand their diversity and relatedness.

The genus *Bacteroides* represents one of the most common and abundant microbes associated with the human gut (34), having coevolved with hominid hosts for millions of years (35). *Bacteroides* genomes encode diverse suites of accessory genomic elements including polysaccharide utilization loci (36, 37), capsular polysaccharide synthesis loci (38), CRISPR-Cas systems (39, 40), and a diversity of integrated MGEs (41, 42). Although *Bacteroides* species engage in beneficial interactions with hosts (e.g., production of short chain fatty acids, immune system development), they can also be opportunistic pathogens outside and inside the gut lumen. *Bacteroides* species commonly carry ICEs encoding antibiotic resistance genes (ARGs) that can be transferred among symbiotic and pathogenic microbes in the gut, thus acting as important antibiotic resistance reservoirs (15, 43, 44). Further, enterotoxigenic *B. fragilis* strains secrete hemolysins and toxins associated with inflammation and colon cancer, some of which are carried by ICEs (45–48). Given their clinical relevance, genomic diversity, and large amount of sequence data, the *Bacteroides* presents an ideal case study for examining MGE diversity.

Here we describe a new tool: Viral, Integrative, & Conjugative Sequence Identification & Networking (VIC SIN). VIC SIN uses a consensus approach to combine similarity-based searching, *de novo* prediction, and genome alignment for integrated MGE prediction. VIC SIN then classifies MGEs using nucleotide homology to network and cluster them into related groups. We test VIC SIN using two distinct benchmark datasets: *Pseudomonas aeruginosa*, a very well-understood microbial host of diverse

MGEs from the phylum *Proteobacteria*, and *Bacteroides fragilis*, a lesser-studied organism from the divergent phylum *Bacteroidetes*. Benchmark analyses suggest that although VICSIN performs comparably to its constituent tools for *P. aeruginosa* genomes, VICSIN is more accurate and sensitive for the *B. fragilis* benchmark dataset. Thus, we suggest VICSIN is a superior tool for MGE discovery in poorly characterized hosts. We applied VICSIN to a large dataset of *Bacteroides* genomes, revealing a highly diverse set of related MGEs that has gone unexplored.

RESULTS

VICSIN predicts and networks integrated MGEs

We developed VICSIN as a tool for (i) comprehensive prediction of MGEs integrated in microbial genomes and (ii) clustering of MGEs as a method of classification (Fig. 1). VICSIN uses a consensus approach to MGE prediction, utilizing the pre-existing MGE prediction tools VirSorter (19), PhiSpy (25), and AGEnt (26). Further, VICSIN integrates optional nucleotide BLAST steps with user-input databases of CRISPR spacers and/or known MGEs. Consensus boundaries of overlapping predictions are determined for each genome. These consensus regions are then compared across the set of analyzed genomes using nucleotide BLAST and boundaries are extended if they are sufficiently similar to an independent consensus prediction (“Re-BLAST”), producing the final VICSIN predictions. Finally, VICSIN predictions are networked on the basis of percent length aligned (PLA) using reciprocal nucleotide BLAST. MGE pairs with PLAs greater than 20% are passed to MCL for clustering (Fig. 1).

VICSIN is designed to take groups of ≥ 5 related bacterial genomes as input. To test VICSIN, two benchmark datasets were developed for the divergent bacterial species *P. aeruginosa* and *B. fragilis* (Fig. 2). Each benchmark dataset is comprised of 6 closely related genomes (Fig. 2). In order to prevent drops in prediction precision, VICSIN requires input genomes be grouped below the species level, using a cutoff of $\geq 70\%$ PLA by Spine, approximately corresponding to $>98\%$ average nucleotide identity (ANI) in both benchmark datasets (Fig. S1). For both datasets, “known MGEs” integrated in each genome were identified based on literature reports or manual annotation (Table S1), identifying 40 known MGEs in the *P. aeruginosa* genomes, and 38 known MGEs in the *B. fragilis* genomes. Though VICSIN can use known MGEs for prediction by BLAST, this option was not used for benchmarking purposes.

Using the known MGEs in these two benchmark datasets, we compared the accuracy and sensitivity of VICSIN to its individual constituent tools (Fig. 3). We used four main metrics to evaluate the effectiveness of the tools: precision (*i.e.*, the number of true positive predictions, divided by the total number of predictions), recall (*i.e.*, the number of true positive predictions, divided by the total number of known MGEs), coverage (*i.e.*, the percent length of a known MGE covered by a true positive prediction), overprediction (*i.e.*, the excess length of a true positive prediction that does not cover a known MGE), and total false positive length (*i.e.*, the sum of the lengths of all false positive predictions in a genome). True positive predictions were defined to be any prediction overlapping a known MGE in a benchmark genome.

Analysis of the *B. fragilis* dataset found that VICSIN is more accurate and sensitive than its constituent tools VirSorter, PhiSpy, and AGEnt for all six genomes

(Fig. 3A). For some of the *P. aeruginosa* genomes (*i.e.* strains DK2, PAO1, and RP73), VirSorter has greater precision and recall than VICSIN (Fig 3B). Inspection of these predictions revealed that the dropoff in VICSIN's accuracy was primarily driven by PhiSpy and AGEnt (Fig. 3C), an expected result given the low precision of PhiSpy (Fig. 3A, 3B), and that AGEnt detects all variable components of microbial genomes (26). However, 28 of 60 (47%) false positive predictions were supported by CRISPR BLAST matches, suggesting there may remain a number of unrecognized MGEs in the benchmark genomes. Overall, VICSIN predictions exhibit high coverage of known MGEs (Fig. 3D) and low overprediction (Fig. 3E), especially in the *B. fragilis* dataset. Together, this analysis suggests VICSIN is at least as accurate and sensitive as its constituent tools for *P. aeruginosa* host genomes, and exceeds the accuracy and precision of those constituent tools for *B. fragilis*. This may be a result of *P. aeruginosa* genomes being key elements in the testing and training of VirSorter, PhiSpy, and AGEnt [19, 25, 26]. We suggest VICSIN is the optimal approach to MGE prediction in poorly-characterized host genomes, especially those outside the phylum *Proteobacteria*.

Classification of predicted MGEs is a critical step for subsequent analyses of evolutionary history, mode of transfer, and host range. The existing tools ClustAGE (50) and vConTACT 2.0 (51) use different methods to classify MGEs, each with advantages and limitations. ClustAGE divides input genomic elements into subelements (*e.g.* gene blocks) and bins these elements using nucleotide similarity. These subelement bins allow ClustAGE to examine the mosaic sharing of nucleotide blocks, offering a fine-scale view of relatedness, although it does not attempt cluster the MGEs as a whole. vConTACT, which was designed for viral sequence data, first clusters the

predicted proteins encoded by input genetic elements, and subsequently uses these shared protein profiles to generate viral clusters. Like ClustAGE, VIC SIN's clustering does not require open reading frame annotation, translation, or protein clustering. However, instead of identifying shared subelements, VIC SIN generates MGE clusters based on the extent of shared nucleotide sequence among the input MGEs. To evaluate the effect these methods have, the known MGEs in the *P. aeruginosa* benchmark dataset (Table S1) were analyzed in parallel with VIC SIN, ClustAGE (50), and vConTACT 2.0 (51). From the 40 known MGEs used as input, both VIC SIN and vConTACT produced the fewest clusters (Table 1), and nearly identical networks (Fig. S2A, S2B). This is expected, since both methods were designed for the similar purposes of classifying MGEs or viruses, respectively, and employ large lengths of sequence similarity. The fundamentally different approach of ClustAGE results in a large number of subelement bins, illustrating the high degree of connectivity among the known MGEs over short sequence lengths; however, it does little to provide a means of classifying the MGEs more broadly (Fig. S2C).

Importantly, VIC SIN offers several practical advantages while achieving similar results to vConTACT. Prior to running vConTACT the user must also predict protein encoded on their MGEs, and then generate protein clusters using programs in the vConTACT toolbox. This process, however, is not designed to handle multiple MGEs and requires additional work to reformat and compile protein data inputs. VIC SIN eliminates these additional steps by operating directly on the nucleotide sequences. Further, by using nucleotide data, VIC SIN avoids the possibility of protein prediction inaccuracies, as MGEs can be challenging to properly annotate (52). vConTACT and

VICSIN are both built on BLAST and MCL, making them equally scalable. Because of its practical advantages and its reduced susceptibility to error and compute time during protein calling and data format conversion, we conclude that VICSIN is a superior tool for MGE clustering.

Application of VICSIN to characterize the Bacteroidaceae integrated mobilome

A subset of genomes was selected from a database of 341 *Bacteroides* genomes accessible on NCBI or sequenced in-house (Table S2). As was performed for the benchmark datasets, genomes were compared by pairwise Spine-PLA (Fig. S3). From the similarity matrix, two groups of interest were chosen for analysis with VICSIN: (1) 61 *B. fragilis* genomes, (2) and 27 *P. vulgatus* or *P. dorei* genomes. *B. fragilis* is commonly associated with extraintestinal infections and perhaps the best-studied *Bacteroides* species (53). As a result, many known MGEs have been characterized for it, including ICEs (54–60) and phages (61–65). *P. vulgatus* and *P. dorei* constitute a clade divergent from that of *B. fragilis*, and together have fewer known MGEs (66, 67). This scheme was chosen in order to characterize MGEs both within host groups and also to sample across the diverse *Bacteroidaceae* family. The *B. fragilis* and *P. vulgatus* groups were arbitrarily split into subgroups of 7-10 genomes, in order to keep AGEnt sensitivity high and compute times low. Because VICSIN prediction does not scale, each subgroup was run independently for prediction, and those predictions were later aggregated for VICSIN MGE cluster analysis.

VICSIN identified a total of 816 MGEs across the 88 host genomes (Table S3). This constitutes 42.3 Mb of predicted mobile sequence out of 468 Mb of total genome length (9.0%). This is greater than the mobile content of the initial *P. aeruginosa* (4.3%)

and *B. fragilis* (5.0%) benchmark datasets. This may be the result of false positive predictions, reflecting other accessory genomic elements. Therefore, we recommend manual examination of predictions, as for all MGE prediction tools, especially from large datasets (≥ 50 genomes). However, it should be noted that 9% of a genome consisting of MGEs is well within the range what has been observed for other organisms (68, 69). Across the 816 VICSIN predictions, 96% were supported by AGenT, 50% by VirSorter, 38% by CRISPR match, and 31% by PhiSpy (Fig. S4).

All 816 MGE predictions were pooled with a set of known *Bacteroidaceae* MGEs including phages and ICEs (Table S4). These predicted and known *Bacteroidaceae* MGEs were clustered with VICSIN, generating 95 MGE clusters that represented 770 individual MGE predictions and 41.1 Mb of predicted mobile sequence ($\sim 97\%$ of predicted MGEs) (Fig. 4). Only 46 predicted MGEs were determined to be singletons, having $\leq 20\%$ PLA with the other MGEs. Each MGE was assigned an MGE type (*i.e.* phage, ICE, mobile, or unknown) on the basis of Pfam annotations (Table S3, S5).

Genetic recombination in VICSIN-predicted phage clusters

It was observed that three clusters with known phages are isolated in the network having low/no PLA above 20% with other MGEs (Fig. 4, S4). This includes the crAssphages (Cluster 21), Hankyphage and its putative relatives (Cluster 29), and two siphophages (Cluster 77), but not *Salyersviridae* temperate phages (Cluster 23). Given the low connectivity of some phage clusters in the network analysis, it suggests that the rate of recombination between MGEs may vary by MGE type or cluster. Notably, this pattern is not determined by phage lifestyle (*i.e.* obligately lytic versus temperate). Hankyphage is a Mu-like phage capable of being stably integrated, but we did not

identify sufficient shared DNA sequences with other MGEs. In contrast, a recently-described group of temperate phages from the proposed family *Salyersviridae* does exhibit some gene sharing with other MGE clusters. Further research will be needed to determine the mechanisms affecting recombination rates between specific MGE types. While further sampling of MGEs may find more evidence of gene sharing between phages and other MGEs in *Bacteroidaceae* hosts, these results suggest it is less common, and certainly much rarer than the rate of gene sharing among ICE-like MGEs (Fig. 4).

Encoded protein functions of VICSIN-predicted MGEs

We hypothesized that the gene content uniting an MGE cluster would be comprised of core genes encoding the essential functions for MGE transmission and persistence, and the gene content shared between MGE clusters would be comprised of accessory genes that benefit the *Bacteroidaceae* host. To examine gene sharing at a functional level, predicted MGE genes were annotated using Pfams (Fig. 5).

This analysis supports our hypothesis that sharing within MGE clusters would be defined by core transmission genes. Many of the most frequently shared genes between ICEs within a cluster (ICE-ICE within) are those that encode the Type IV secretion system (T4SS). Likewise, some the most frequently shared between phages within a cluster (phage-phage within) encode structural components of the phage capsid, or the machinery for packaging the phage genome. Gene sharing between ICEs and phages within a cluster (ICE-phage within) is largely driven by functions not readily assignable to an MGE type, or to accessory genes like those associated with capsular polysaccharide synthesis (CPS) loci or polysaccharide utilization loci (PUL). This is

reflective of the ambiguous nature of those MGE clusters apparently containing ICEs and phages, which further investigation may find are comprised of novel MGE types or chimeric MGEs.

The Pfam analysis, however, only partially supports our hypothesis that gene sharing between clusters would be driven by accessory genes. Functional gene sharing between MGE clusters is dominated by the same core genes which drive gene sharing within those clusters. This sharing of core genes between MGE clusters does support the hypothesis that recombination is a powerful contributor to MGE evolution, leading to their modular and mosaic nature. While some accessory genes, such as Type VI secretion system (T6SS) genes, are shared between clusters, their frequencies are dwarfed by that of core genes. This is likely due to several factors, including difficulty in annotating potential accessory genes using Pfams and the potential for a high diversity of accessory genes compared to core functional genes.

Extensive synteny within VICIN MGE clusters

The sharing of genetic content was also examined through representative clusters. Clusters of interest were chosen on the basis of (i) having a known MGE member (Fig. S5), and (ii) having members from more than one species (Fig. S6). We postulated that clusters with MGEs from more than one host species are least likely to contain false positive predictions, by virtue of being recently mobile. Two clusters each representing ICEs and phages were selected; no attempt was made to examine clusters composed of unknown MGE types, which are difficult to annotate. Cluster 1 (CTn86-like elements) and Cluster 3 (CTnDOT-like elements) were chosen as representative of

ICEs (Fig. 6A, 6B), and both show extensive sharing of core functional genes. Shared genes within these clusters include the *tra* operon, which encodes the T4SS, and can include the *mob* operon, encoding the relaxosome. Further, accessory genes in ICEs (*i.e.* bile salt hydrolase, xylanase, antibiotic resistances) are variable within their respective clusters. Representative phage clusters were Cluster 23 (*Salyersviridae*-like elements) and Cluster 29 (Hankyphage-like elements) (Fig. 6C, 6D). Both phage clusters exhibit a greater degree of similarity within the cluster than do the representative ICE-like clusters, covering the majority of these phage genomes. This is likely because more of a phage genome can be considered core; phages encode a large number of genes involved in capsid structure and assembly, genome packaging, host lysis, DNA replication, and regulation.

Detection of clinically-relevant traits in VICSIN-predicted MGEs

Finally, functional gene sharing was further interrogated by searching for specific clinically relevant functions among the VICSIN predictions. First, we examined antibiotic resistance genes (ARGs) (Fig. 7A). Twenty of the 816 VICSIN predicted MGEs were determined to encode a total 24 resistance genes. All were characterized resistance genes common among the Bacteroidetes: *cfxA3*, *ermF*, *ermG*, *tetX*, *tetQ* and *linA*, each conferring cefoxitin, clindamycin, erythromycin, tetracycline, tetracycline, and lincomycin resistance, respectively.

We hypothesized that antibiotic resistance genes would be associated with particular MGE clusters due to strong selection for MGEs to maintain their ARGs over evolutionary time. Yet, given the large size of the MGEs predicted (average length = 51.8 Kb), the short length of the ARGs observed (1,974 bp for the longest ARG, *tetQ*),

and the 20% PLA cutoff for clustering, we did not expect the sharing of ARGs to drive MGE clustering. However, our initial hypothesis was true only for *linA*, which was detected exclusively in two MGEs in Cluster 2 (Table 2). Further sampling may reveal that *linA* is not restricted to a single MGE cluster. We determined that all of the other ARGs detected were distributed in multiple MGE clusters. The tetracycline resistance gene *tetQ* is common and highly conserved in Cluster 3, and although it is also present in two other clusters, is found in 8 out of 39 predictions (21%) within the group. This suggests that although *tetQ* is maintained over evolutionary time by some members of a cluster, stochastic loss events may cause its absence in other MGEs of the cluster. Two MGE predictions encode two copies of the same ARG (*i.e.* *tetX* in *P. vulgatus* VPI-4496.2 prediction 23 and *ermG* in *B. fragilis* OF05-13AC prediction 17), although these ARGs are not identical to one another. These may be unique alleles of the same ARG and confer resistance to the same antibiotics, or they might confer unique resistance phenotypes for which a resistance gene is not yet known. Three MGEs encode both *tetX* and *ermF* (*i.e.* *P. vulgatus* VPI-4496.2 prediction 23, *B. fragilis* CL07T12C05 prediction 72, *P. vulgatus* VPI-BV8526 prediction 108), although these shared ARGs do not correlate with their associated clusters. In two of these three MGE predictions, *ermF* and *tetX* occur in a conserved gene cassette. Together, the distribution of *tetX* and *ermF* supports the previous theory that ARGs are readily transferred individually and in cassettes between co-infecting MGEs via recombination (15, 43).

Bacteroidaceae MGEs are also known to carry T6SSs, which confer a competitive advantage to their hosts in the gut by antagonizing neighboring bacteria. To search for T6SSs in VICSIN predictions, three genes were used as queries: *hcp*, *tssB*,

and *vgrG*. These genes are universal among the three T6SS genetic architectures known to occur in *Bacteroidaceae* genomes. Of the 816 VIC SIN predictions, 71 were found to encode predicted T6SS. T6SSs primarily occurred in three clusters: Cluster 0, Cluster 5 (CTnBST-like ICEs), and Cluster 6 (*P. vulgatus*-associated ICEs). Evidence of sharing of these T6SS genes is apparent between clusters (*i.e.* between Clusters 0 and 9, and between Clusters 6 and 52), suggesting they may also move between MGEs by recombination.

DISCUSSION

Here we describe VIC SIN, a tool for the prediction and networking of integrated MGEs. VIC SIN uses a consensus approach to MGE prediction that results in greater sensitivity than its constituent tools. It is especially powerful for poorly-studied host genomes such as those of the family *Bacteroidaceae*. Our data suggests that this may be a result of these tools being developed and tested using extant model organisms like those of *P. aeruginosa* and *E. coli* (18). VIC SIN then uses nucleotide similarity to generate networks and clusters of the MGE predictions as a basis for classifying the MGEs and visualizing their relatedness. These networks provide insight into the frequency and routes of HGT and recombination occurring among MGEs. VIC SIN analyses provide an import framework for understanding the variable fraction of microbial genomes.

We applied VIC SIN to 88 *Bacteroidaceae* genomes from two groups within the family: (i) *B. fragilis* and (ii) *P. vulgatus* and *P. dorei*. *Bacteroidaceae* genomes are highly plastic, with large portions of these genomes predicted to be mobilized by HGT

(66, 71–73). Although the mechanisms causing the entirety of these events are largely unknown, one of the principle drivers of HGT is likely uncharacterized integrated MGEs. Presently, between the NCBI and ICEberg databases only have 27 annotated *Bacteroidaceae*-specific phage or ICE elements. Our results found that 60% (57/95) of VICIN-predicted MGE clusters lack a known MGE. Further, we show evidence of recombination within and between clusters of predicted MGEs, and hypothesize at least some of these MGEs are capable of transfer across a broad host range. It has been hypothesized that a combination of these factors is an especially potent driver of HGT across the *Bacteroidaceae* family (71, 73).

We found very limited gene sharing between known phages and integrated MGE predictions in the current *Bacteroidaceae* dataset. This suggests some barrier to recombination exists between some MGEs. We do not believe the barrier is caused by short infection times for these phages; Hankyphages are Mu-like viruses capable of stable integration into host chromosomes (74), and cells infected with CrAssphages have been observed to enter a “carrier state” where lysis and progeny release can be delayed for long periods of time (75). Part of this could be caused by structural constraints of phages versus ICEs. Whereas phages have limited space within the capsid to store their genome, ICEs which transfer directly from cell to cell via the T4SS have fewer constraints on genome size. Known ICEs can be hundreds of kilobases in length (e.g. PAIst in *Streptomyces turgidiscabies* >674 Kb) (76). Further, the current data reinforces the view that ICEs are likely the primary drivers of HGT in these genomes, though this may be due to a lack of knowledge about lysogenic phages infecting *Bacteroidaceae* hosts. Phages are known to drive HGT in other bacterial

414 systems (77), so it stands to reason they should also play a role in *Bacteroides* and
415 *Phocaeicola*.

416 Despite the limited gene sharing between phages and other MGE clusters, gene
417 sharing between phages of the same cluster was found to be very high, covering the
418 majority of the genomes examined. This further suggests phage genomes are under
419 strong selection to only encode the content necessary for their transfer, to the exclusion
420 of a diversity of accessory genes, including ARGs and T6SSs.

421

422 METHODS

423 **Data availability.** VIC SIN, genome_grouper.py, and test datasets are available
424 at github.com/IGBIllinois/VICSIN.

425 **Benchmark datasets.** Two datasets were created and used for testing VIC SIN,
426 one of *Pseudomonas aeruginosa* genomes and another of *Bacteroides fragilis*
427 genomes. Genomes in benchmark datasets were annotated based on published MGEs
428 integrated in those genomes, or through manual annotation (Table S1).

429 The *P. aeruginosa* dataset consists of six ingroup genomes: *P. aeruginosa* DK2
430 (NC_018080), *P. aeruginosa* UCBPP-PA14 (NC_008463), *P. aeruginosa* PAO1
431 (NC_002516), *P. aeruginosa* LESB58 (NC_011770), *P. aeruginosa* PA38182
432 (HG530068), *P. aeruginosa* RP73 (NC_021577); and three outgroup genomes: *P.*
433 *denitrificans* ATCC 13867 (NC_020829), *P. stutzeri* CGMCC 1.1803 (NC_015740), and
434 *P. putida* KT2440 (NC_002947).

435 The *B. fragilis* dataset consists of six ingroup genomes: *B. fragilis* CL03T12C07
436 (NZ_AGXL000000000), *B. fragilis* NCTC 9343 (NC_003228), *B. sp.* 3_2_5

437 (NZ_ACIB000000000), *B. fragilis* 638R (NC_016776), *B. fragilis* CL05T00C42
 438 (NZ_AGXO000000000), *B. fragilis* YCH46 (NC_006347); and three outgroup genomes:
 439 *B. fragilis* 3_1_12 (NZ_ABZX000000000), *B. nordii* CL02T12C05 (NZ_AGXS000000000),
 440 *B. salyersiae* CL02T12C01 (NZ_AGXV000000000).

441 ***Viral, Integrative, & Conjugative Sequence Identification & Networking***
 442 **(VICSIN)**. VICSIN takes genomes in nucleotide FASTA format or GenBank format.
 443 General Feature Format (GFF) files defining ORFs are optional inputs. In the absence
 444 of GFF or GenBank files, VICSIN uses ORFs called by VirSorter or Prodigal in later
 445 steps. Multiple genomes are input simultaneously and should represent a group of
 446 closely-related organisms. VICSIN performs an initial alignment of all genomes with
 447 Spine using default parameters to define a core genome and calculate Spine-PLA. If the
 448 Spine-PLA is less than 70%, a warning is passed to the user.

449 Prediction of MGEs by VICSIN is accomplished by three default tools (PhiSpy,
 450 VirSorter, AGEnt) and two optional BLAST-based searches (Known BLAST and
 451 CRISPR BLAST). VICSIN runs PhiSpy on each individual genome in a genome group
 452 with a window size (w) of 40 and threshold (n) of 20. VICSIN implements VirSorter with
 453 default parameters and the provided “Viromes” search database. Spine and AGEnt are
 454 run in sequence using default parameters except using a minimum core length of 500
 455 bp. Known BLAST is done by default nucleotide BLAST using a user-input FASTA file of
 456 relevant known MGEs. CRISPR BLAST is based on CLdb (78). VICSIN finds
 457 overlapping predictions and merges them when they are within 5 bp of each other or a
 458 contig boundary. VICSIN breaks PhiSpy predictions at overlapping prediction
 459 boundaries to limit overprediction by PhiSpy. Finally, VICSIN uses Re-BLAST to

compare all predictions, and will extend prediction boundaries to match those of other predictions made using multiple methods.

After prediction, VICsIN filters out predictions shorter than 12 kb, a cutoff chosen because this was the length of the shortest known MGE in the benchmark datasets; neither VICsIN or its constituent tools are optimized for the prediction of short MGEs, like insertion sequences. Next, VICsIN performs reciprocal nucleotide BLAST to compare all predictions to each other, and calculates total PLA for every prediction pair, which serves as the basis for networking. PLAs greater than or equal to 20% are passed to MCL with an inflation parameter of 2, generating MGE clusters. Predictions smaller than 12 kb are clustered separately.

Other networking methods. vConTACT was run on CyVerse using default parameters except where noted. ClustAGE was run with default parameters and visualized with <http://vfsm spineagent.fsm.northwestern.edu/results/cpQeC6nUglST/results.html>. All networks were visualized with Cytoscape.

Bacteroidaceae experimental dataset. From the NCBI web interface, 300 *Bacteroidaceae* genomes were downloaded (Table S2). An additional 18 genomes were added to the dataset that were sequenced as part of this work (Table S7). Known MGEs integrated in the networking steps were downloaded from NCBI or ICEberg databases (31).

MGE cluster typing. Open reading frames from MGE prediction were annotated with the Pfam 33.1 database (79) and hmmscan using trusted cutoffs (--cut_tc). A list of Pfam protein families associated with ICEs and general mobile sequences was

compiled (Table S5). Phage-associated Pfams were defined as any with the words “terminase,” “portal,” “capsid,” “tail,” “base plate,” “spike,” “neck,” “head,” or “phage protein” in their description. The number of phage, ICE, and mobile genes were counted. MGEs with more phage genes than ICE genes were assigned to the phage type. MGEs with more ICE genes than phage genes were assigned to the ICE type. MGEs with no ICE or phage genes, or an equal number of phage and ICE genes, were assigned to the “mobile” type. Any MGE with no genes of interest were assigned to the “unknown” type.

Gene annotation. Using the Pfam predictions from above additional Pfam IDs were assigned to functional categories based on gene annotations from *B. thetaiotaomicron* VPI-5482 and *B. fragilis* NCTC 9343 (Table S5). This included genes involved in capsule biosynthesis (CPS), polysaccharide utilization loci (PULs) and type VI secretion systems (T6SS). In addition, antibiotic resistance gene nucleotide sequences from the Comprehensive Antibiotic Resistance Database v1 ($n = 2,163$; <https://card.mcmaster.ca/>) (80) and an additional 17 genes missing from the database (e.g., *tetA*, *ermA*, *tetC*) were grouped into 90% identity clusters with USEARCH v5.2.236 (81). A BLAST database was then generated with the protein sequences for the 734 representative (seed) antibiotic resistance genes for these clusters. ORFs from the VICIN predictions were then queried using an E-value cut off of 10^{-10} , and the results were filtered for matches $\geq 50\%$ aligned and $\geq 50\%$ AA identity.

A BLAST database was generated with the protein sequences for 46 representative Hcp, TssB, and VgrG from *Bacteroidaceae* genomes (Table S6). ORFs

from the VICIN predictions were then queried using an E-value cut off of 10^{-10} , and the results were filtered for matches $\geq 50\%$ aligned and $\geq 50\%$ AA identity.

Genome sequencing. Cells were pelleted from 5 mL overnight culture in TYG (82) by centrifugation at $4,000 \times g$ for 5 min at 4°C , resuspended in 0.5 mL TE buffer (10 mM Tris, 1 mM EDTA), and lysed by adding sodium dodecyl sulfate (SDS) and proteinase K (GoldBio, Olivette, MO) to final concentrations of 0.07% and 300 $\mu\text{g/mL}$, respectively, and incubating for 2 hr at 55°C . Cellular material was removed by washing twice in an equal volume of buffered phenol, phenol-chloroform-isoamyl alcohol (VWR, Radnor, PA), and DNA precipitated with 100% ethanol in the presence of 0.3M sodium acetate at -20°C overnight. DNA pellets were washed with 70% ethanol, dried, and resuspended in TE buffer.

DNA libraries were constructed with the Nextera XT Library Preparation Kit and Index Kit (Illumina, San Diego, CA). DNA libraries were pooled and sequenced on both the Illumina HiSeq5000 and HiSeq2500 and fastq files were generated from demultiplexed reads with bcl2fastq Conversion Software (Illumina, San Diego, CA). Reads were trimmed and assembled using the A5ud pipeline (83). Sequencing methods and assembly data are summarized in Table S7.

ACKNOWLEDGEMENTS

We thank Ted Pacyga for his invaluable help with the genome_grouper.py program, Nadja Shoemaker and Abigail Salyers for access to an impressive collection of *Bacteroidaceae* isolates, and Alvaro Hernandez and Chris Wright for DNA sequencing.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. D.E.C. was supported by the University of Illinois at Urbana-Champaign Department of Microbiology and the NIH grant T32 DK077653-29. W.E.E. was supported by a James R. Beck Graduate Research Fellowship and NIH training grant T32 AI078876. R.J.W. was supported by the Allen Foundation with an Allen Distinguished Investigator Award.

REFERENCES

1. Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. 8. Nat Rev Genet 16:472–482.
2. Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. 6784. Nature 405:299–304.
3. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev 33:376–393.
4. Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3:711–721.
5. Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55:709–742.
6. Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. J Mol Biol 222:851–856.

- 549 7. Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of
550 change and exchange. *J Mol Evol* 44:383–397.
- 551 8. Loy A, Pfann C, Steinberger M, Hanson B, Herp S, Brugiroux S, Neto JCG,
552 Boekschoten MV, Schwab C, Urich T, Ramer-Tait AE, Rattei T, Stecher B, Berry D.
553 2017. Lifestyle and horizontal gene transfer-mediated evolution of *Mucispirillum*
554 *schaedleri*, a core member of the murine gut microbiota. *mSystems* 2:e00171-16.
- 555 9. Frazão N, Sousa A, Lässig M, Gordo I. 2019. Horizontal gene transfer overrides
556 mutation in *Escherichia coli* colonizing the mammalian gut. *PNAS* 116:17906–
557 17915.
- 558 10. Sousa A, Frazão N, Ramiro RS, Gordo I. 2017. Evolution of commensal bacteria in
559 the intestinal tract of mice. *Curr Opin in Microbiol* 38:114–121.
- 560 11. Kelly BG, Vespermann A, Bolton DJ. 2009. Gene transfer events and their
561 occurrence in selected environments. *Food Chem Toxicol* 47:978–983.
- 562 12. Stecher B, Denzler R, Maier L, Bernet F, Sanders MJ, Pickard DJ, Barthel M,
563 Westendorf AM, Krogfelt KA, Walker AW, Ackermann M, Dobrindt U, Thomson NR,
564 Hardt W-D. 2012. Gut inflammation can boost horizontal gene transfer between
565 pathogenic and commensal *Enterobacteriaceae*. *PNAS* 109:1269–1274.
- 566 13. Diard M, Bakkeren E, Cornuault JK, Moor K, Hausmann A, Sellin ME, Loverdo C,
567 Aertsen A, Ackermann M, Paepe MD, Slack E, Hardt W-D. 2017. Inflammation
568 boosts bacteriophage transfer between *Salmonella* spp. *Science* 355:1211–1215.

- 569 14. Balcazar JL. 2014. Bacteriophages as vehicles for antibiotic resistance genes in
570 the environment. PLoS Pathog 10:e1004219.
- 571 15. Salyers A, Shoemaker NB. 2006. Reservoirs of antibiotic resistance genes. Animal
572 Biotechnol 17:137–146.
- 573 16. Fortier L-C, Sekulovic O. 2013. Importance of prophages to evolution and virulence
574 of bacterial pathogens. Virulence 4:354–365.
- 575 17. Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in
576 pathogenic and environmental microorganisms. 5. Nat Rev Microbiol 2:414–424.
- 577 18. Khot V, Strous M, Hawley AK. 2020. Computational approaches in viral ecology.
578 Comput Struct Biotechnol 18:1605–1612.
- 579 19. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from
580 microbial genomic data. PeerJ 3:e985.
- 581 20. Kung VL, Ozer EA, Hauser AR. 2010. The accessory genome of *Pseudomonas*
582 *aeruginosa*. Microbiol Mol Biol Rev 74:621–641.
- 583 21. Patrick S, Blakely GW, Houston S, Moore J, Abratt VR, Bertalan M, Cerdeño-
584 Tárraga AM, Quail MA, Corton N, Corton C, Bignell A, Barron A, Clark L, Bentley
585 SD, Parkhill J. 2010. Twenty-eight divergent polysaccharide loci specifying within-
586 and amongst-strain capsule diversity in three strains of *Bacteroides fragilis*.
587 Microbiol 156:3255–3269.

- 588 22. Belkum A van, Soriaga LB, LaFave MC, Akella S, Veyrieras J-B, Barbu EM,
589 Shortridge D, Blanc B, Hannum G, Zambardi G, Miller K, Enright MC, Mugnier N,
590 Bami D, Schicklin S, Felderman M, Schwartz AS, Richardson TH, Peterson TC,
591 Hubby B, Cady KC. 2015. Phylogenetic distribution of CRISPR-Cas systems in
592 antibiotic-resistant *Pseudomonas aeruginosa*. mBio 6.
- 593 23. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. 2017. VirFinder: a novel k-mer
594 based tool for identifying viral sequences from assembled metagenomic data.
595 Microbiome 5:69.
- 596 24. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Poplin R, Sun F.
597 2020. Identifying viruses from metagenomic data using deep learning. Quant Biol
598 8:64–77.
- 599 25. Akhter S, Aziz RK, Edwards RA. 2012. PhiSpy: a novel algorithm for finding
600 prophages in bacterial genomes that combines similarity- and composition-based
601 strategies. Nucleic Acids Res 40:e126.
- 602 26. Ozer EA, Allen JP, Hauser AR. 2014. Characterization of the core and accessory
603 genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt.
604 BMC Genom 15:737.
- 605 27. Abdelkareem AO, Khalil MI, Elaraby M, Abbas H, Elbehery AHA. 2018. VirNet:
606 Deep attention model for viral reads identification, p. 623–626. In 2018 13th
607 International Conference on Computer Engineering and Systems (ICCES).

- 608 28. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS. 2020. A
609 bioinformatic analysis of integrative mobile genetic elements highlights their role in
610 bacterial adaptation. *Cell Host & Microbe* 27:140-153.e9.
- 611 29. Yaffe E, Relman DA. 2020. Tracking microbial evolution in the human gut using Hi-
612 C reveals extensive horizontal gene transfer, persistence and adaptation. 2. *Nat*
613 *Microbiol* 5:343–353.
- 614 30. Adriaenssens EM, Brister JR. 2017. How to name and classify your phage: an
615 informal guide. *Viruses* 9.
- 616 31. Bi D, Xu Z, Harrison EM, Tai C, Wei Y, He X, Jia S, Deng Z, Rajakumar K, Ou H-Y.
617 2012. ICEberg: a web-based resource for integrative and conjugative elements
618 found in Bacteria. *Nucleic Acids Res* 40:D621–D626.
- 619 32. Osborn AM, Böltner D. 2002. When phage, plasmids, and transposons collide:
620 genomic islands, and conjugative- and mobilizable-transposons as a mosaic
621 continuum. *Plasmid* 48:202–212.
- 622 33. Wozniak RAF, Waldor MK. 2010. Integrative and conjugative elements: mosaic
623 mobile genetic elements enabling dynamic lateral gene flow. 8. *Nat Rev Microbiol*
624 8:552–563.
- 625 34. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes
626 GR, Tap J, Bruls T, Batto J-M, Bertalan M, Borruel N, Casellas F, Fernandez L,
627 Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc
628 M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J,

629 Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner
630 F, Pedersen O, de Vos WM, Brunak S, Doré J, Weissenbach J, Ehrlich SD, Bork P.
631 2011. Enterotypes of the human gut microbiome. *Nature* 473:174–180.

632 35. Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV, Muller MN,
633 Pusey AE, Peeters M, Hahn BH, Ochman H. 2016. Cospeciation of gut microbiota
634 with hominids. *Science* 353:380–382.

635 36. Kaoutari AE, Armougom F, Gordon JI, Raoult D, Henrissat B. 2013. The
636 abundance and variety of carbohydrate-active enzymes in the human gut
637 microbiota. *Nat Rev Microbiol* 11:497–504.

638 37. Martens EC, Chiang HC, Gordon JI. 2008. Mucosal glycan foraging enhances
639 fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host*
640 *Microbe* 4:447–457.

641 38. Krinos CM, Coyne MJ, Weinacht KG, Tzianabos AO, Kasper DL, Comstock LE.
642 2001. Extensive surface diversity of a commensal microorganism by multiple DNA
643 inversions. *Nature* 414:555–558.

644 39. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, Amitai G, Sorek R.
645 2018. Systematic discovery of antiphage defense systems in the microbial
646 pangenome. *Science* 359.

647 40. Tajkarimi M, Wexler HM. 2017. CRISPR-Cas systems in *Bacteroides fragilis*, an
648 important pathobiont in the human gut microbiome. *Front Microbiol* 8.

- 649 41. Nguyen M, Vedantam G. 2011. Mobile genetic elements in the genus *Bacteroides*,
650 and their mechanism(s) of dissemination. *Mobile Genetic Elements* 1:187–196.
- 651 42. Koonin EV, Yutin N. 2020. The crAss-like phage group: How metagenomics
652 reshaped the human virome. *Trends Microbiol* 28:349–359.
- 653 43. Whittle G, Shoemaker NB, Salyers AA. 2002. The role of *Bacteroides* conjugative
654 transposons in the dissemination of antibiotic resistance genes. *Cell Mol Life Sci*
655 59:2044–2054.
- 656 44. Salyers AA, Shoemaker NB. 1996. Resistance gene transfer in anaerobes: new
657 insights, new problems. *Clin Infect Dis* 23:S36–S43.
- 658 45. Zakharzhevskaya NB, Vanyushkina AA, Altukhov IA, Shavarda AL, Butenko IO,
659 Rakitina DV, Nikitina AS, Manolov AI, Egorova AN, Kulikov EE, Vishnyakov IE,
660 Fisunov GY, Govorun VM. 2017. Outer membrane vesicles secreted by pathogenic
661 and nonpathogenic *Bacteroides fragilis* represent different metabolic activities. 1.
662 *Sci Rep* 7:5008.
- 663 46. Ko SH, Rho DJ, Jeon JI, Kim Y-J, Woo HA, Lee YK, Kim JM. 2016. *Bacteroides*
664 *fragilis* enterotoxin upregulates heme oxygenase-1 in intestinal epithelial cells via a
665 mitogen-activated protein kinase- and NF-κB-dependent pathway, leading to
666 modulation of apoptosis. *Infect Immun* 84:2541–2554.
- 667 47. Pierce JV, Bernstein HD. 2016. Genomic diversity of enterotoxigenic strains of
668 *Bacteroides fragilis*. *PLOS ONE* 11:e0158171.

48. S3ki J, Eitel Z, Terhes G, Nagy E, Urb3n E. 2013. Occurrence and analysis of rare cfiA–bft doubly positive *Bacteroides fragilis* strains. *Anaerobe* 23:70–73.
49. Chen H, Brinkac LM, Mishra P, Li N, Lympelopoulou DS, Dickerson TL, Gordon-Bradley N, Williams HN, Badger JH. 2015. Draft genome sequences for the obligate bacterial predators *Bacteriovorax* spp. of four phylogenetic clusters. *Stand in Genomic Sci* 10:11.
50. Ozer EA. 2018. ClustAGE: a tool for clustering and distribution analysis of bacterial accessory genomic elements. *BMC Bioinformatics* 19:150.
51. Bolduc B, Jang HB, Doulier 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.
52. McNair K, Aziz RK, Pusch GD, Overbeek R, Dutilh BE, Edwards R. 2018. Phage Genome Annotation Using the RAST Pipeline, p. 231–238. *In* Clokie, MRJ, Kropinski, AM, Lavigne, R (eds.), *Bacteriophages: Methods and Protocols*, Volume 3. Springer, New York, NY.
53. Wexler AG, Goodman AL. 2017. An insider’s perspective: *Bacteroides* as a window into the microbiome. 5. *Nat Microbiol* 2:1–11.
54. Husain F, Veeranagouda Y, Boente R, Tang K, Mulato G, Wexler HM. 2014. The Ellis Island Effect: A novel mobile element in a multi-drug resistant *Bacteroides fragilis* clinical isolate includes a mosaic of resistance genes from Gram-positive bacteria. *Mob Genet Elements* 4:e29801.

55. Husain F, Tang K, Veeranagouda Y, Boente R, Patrick S, Blakely G, Wexler HM. 2017. Novel large-scale chromosomal transfer in *Bacteroides fragilis* contributes to its pan-genome and rapid environmental adaptation. Microb Genom 3.
56. Franco AA. 2004. The *Bacteroides fragilis* pathogenicity island is contained in a putative novel conjugative transposon. J Bacteriol 186:6077–6092.
57. Whittle G, Hamburger N, Shoemaker NB, Salyers AA. 2006. A *Bacteroides* conjugative transposon, CTnERL, can transfer a portion of itself by conjugation without excising from the chromosome. J Bacteriol 188:1169–1174.
58. Kuwahara T, Yamashita A, Hirakawa H, Nakayama H, Toh H, Okada N, Kuhara S, Hattori M, Hayashi T, Ohnishi Y. 2004. Genomic analysis of *Bacteroides fragilis* reveals extensive DNA inversions regulating cell surface adaptation. PNAS 101:14919–14924.
59. Peed L, Parker AC, Smith CJ. 2010. Genetic and functional analyses of the mob operon on conjugative transposon CTn341 from *Bacteroides* spp. J Bacteriol 192:4643–4650.
60. Hecht DW, Kos IM, Knopf SE, Vedantam G. 2007. Characterization of BctA, a mating apparatus protein required for transfer of the *Bacteroides fragilis* conjugal element BTF-37. Res Microbiol 158:600–607.
61. Jofre J, Blanch AR, Lucena F, Muniesa M. 2014. Bacteriophages infecting *Bacteroides* as a marker for microbial source tracking. Water Res 55:1–11.

- 710 62. Payan A, Ebdon J, Taylor H, Gantzer C, Ottoson J, Papageorgiou GT, Blanch AR,
711 Lucena F, Jofre J, Muniesa M. 2005. Method for isolation of *Bacteroides*
712 bacteriophage host strains suitable for tracking sources of fecal pollution in water.
713 Appl Environ Microbiol 71:5659–5662.
- 714 63. Tartera C, Araujo R, Michel T, Jofre J. 1992. Culture and decontamination methods
715 affecting enumeration of phages infecting *Bacteroides fragilis* in sewage. Appl
716 Environ Microbiol 58:2670–2673.
- 717 64. Tartera C, Jofre J. 1987. Bacteriophages active against *Bacteroides fragilis* in
718 sewage-polluted waters. Appl Environ Microbiol 53:1632–1637.
- 719 65. Puig A, Queralt N, Jofre J, Araujo R. 1999. Diversity of *Bacteroides fragilis* strains
720 in their capacity To recover phages from human and animal wastes and from
721 fecally polluted wastewater. Appl Environ Microbiol 65:1772–1776.
- 722 66. Lange A, Beier S, Steimle A, Autenrieth IB, Huson DH, Frick J-S. 2016. Extensive
723 mobilome-driven genome diversification in mouse gut-associated *Bacteroides*
724 *vulgatus* mpk. Genome Biol Evol 8:1197–1207.
- 725 67. Campbell DE, Ly LK, Ridlon JM, Hsiao A, Whitaker RJ, Degnan PH. 2020. Infection
726 with Bacteroides phage BV01 alters the host transcriptome and bile acid
727 metabolism in a common human gut microbe. Cell Reports 32.
- 728 68. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson
729 NR, Roberts AP, Cerdeño-Tárraga AM, Wang H, Holden MTG, Wright A, Churcher
730 C, Quail MA, Baker S, Bason N, Brooks K, Chillingworth T, Cronin A, Davis P,

731 Dowd L, Fraser A, Feltwell T, Hance Z, Holroyd S, Jagels K, Moule S, Mungall K,
732 Price C, Rabinowitsch E, Sharp S, Simmonds M, Stevens K, Unwin L, Whithead
733 S, Dupuy B, Dougan G, Barrell B, Parkhill J. 2006. The multidrug-resistant human
734 pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat Genet*
735 38:779–786.

736 69. Newton ILG, Bordenstein SR. 2011. Correlations between bacterial ecology and
737 mobile DNA. *Curr Microbiol* 62:198–208.

738 70. Darling AE, Mau B, Perna NT. 2010. progressiveMauve: Multiple genome
739 alignment with gene gain, loss and rearrangement. *PLOS ONE* 5:e11147.

740 71. Shoemaker NB, Vlamakis H, Hayes K, Salyers AA. 2001. Evidence for extensive
741 resistance gene transfer among *Bacteroides* spp. and other genera in the human
742 colon. *Appl Environ Microbiol* 67:561–568.

743 72. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC, Henrissat B,
744 Coutinho PM, Minx P, Latreille P, Cordum H, Brunt AV, Kim K, Fulton RS, Fulton
745 LA, Clifton SW, Wilson RK, Knight RD, Gordon JI. 2007. Evolution of symbiotic
746 bacteria in the distal human intestine. *PLOS Biology* 5:e156.

747 73. Coyne MJ, Zitomersky NL, McGuire AM, Earl AM, Comstock LE. 2014. Evidence of
748 extensive DNA transfer between *Bacteroidales* species within the human gut. *mBio*
749 5:e01305-14.

74. Benler S, Cobián-Güemes AG, McNair K, Hung S-H, Levi K, Edwards R, Rohwer F. 2018. A diversity-generating retroelement encoded by a globally ubiquitous *Bacteroides* phage. *Microbiome* 6:191.
75. Keller R, Traub N. 1974. The characterization of *Bacteroides fragilis* bacteriophage recovered from animal sera: Observations on the nature of *Bacteroides* phage carrier cultures. *J Gen Virol*, 24:179–189.
76. Huguet-Tapia JC, Bignell DRD, Loria R. 2014. Characterization of the integration and modular excision of the integrative conjugative element PAISt in *Streptomyces turgidiscabies* Car8. *PLOS ONE* 9:e99345.
77. Dixit PD, Pang TY, Studier FW, Maslov S. 2015. Recombinant transfer in the basic genome of *Escherichia coli*. *PNAS* 112:9070–9075.
78. Bautista MA, Black JA, Youngblut ND, Whitaker RJ. 2017. Differentiation and structure in *Sulfolobus islandicus* rod-shaped virus populations. *5. Viruses* 9:120.
79. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR. 2004. The Pfam protein families database. *Nucleic Acids Res* 32:D138–D141.
80. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJV, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. 2013. The

770 comprehensive antibiotic resistance database. Antimicrob Agents Chemother
771 57:3348–3357.

772 81. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST.
773 Bioinformatics 26:2460–2461.

774 82. Holdeman LV, Cato EP, Moore WEC. 1977. Anaerobic laboratory manual, 4th
775 edition. Anaerobe Laboratory, Virginia Polytechnic Institute and State University,
776 Blacksburg.

777 83. Riley AB, Kim D, Hansen AK. 2017. Genome sequence of “Candidatus Carsonella
778 ruddii” strain BC, a nutritional endosymbiont of *Bactericera cockerelli*. Genome
779 Announc 5:e00236-17.

780

781 TABLES

782 **Table 1. Clustering of *P. aeruginosa* benchmark MGEs.**

	VICSIN	vConTACT 2.0	ClustAGE ^a
No. of clusters/bins	6	6	83
Largest cluster/bin	6	6	6
No. singletons	23	24	67

783 ^aClustAGE does not generate clusters; bins were counted instead.

784 FIGURES

Figure 1. Overview of the VIC SIN process. VIC SIN takes groups of closely related microbial genomes as the only required input; CRISPR spacer and known MGE databases are optional inputs. VIC SIN runs each of its constituent prediction tools: Spine and AGEnt, PhiSpy, VirSorter, and nucleotide BLAST (optional). Consensus predictions are determined and their boundaries may be extended in the Re-BLAST step. MGE predictions are split into two groups by their length. Reciprocal nucleotide BLAST is used to calculate percent length aligned (PLA), which is the basis for MCL clustering. Network and cluster outputs are formatted for viewing with external network visualization software. Dashed outline boxes indicate processes; solid outline boxes indicate input or intermediate result data; black filled boxes indicate primary result outputs.

Figure 2. Benchmark datasets for testing VICsIN. Benchmark datasets were constructed from genomes of (A) *P. aeruginosa* and (B) *B. fragilis*. Phylogenies for each dataset were estimated by maximum likelihood from concatenated alignment of 13 core genes. Support values above tree branches represent bootstraps from 100 replications. Adjacent heatmaps depict average nucleotide identity (ANI) (49) and PLA calculated by Spine (Spine-PLA) for each genome pair except for self-pairs. Each dataset includes one intermediate genome (*P. denitrificans* ATCC 13867 and *B. fragilis* 3_1_12) and two outgroup genomes (*P. stutzeri* GCMC 1.1803, *P. putida* KT2440, *B. salyersiae* CL02T12C01, and *B. nordii* CL02T12C05), which were used to test the effects of genome relatedness on VICsIN accuracy and sensitivity (Fig. S1). Genome sizes are indicated for each genome in grey.

810

Figure 3. MGE prediction with VIC SIN is more sensitive and accurate than with its constituent tools. MGE prediction precision and recall were calculated for each of the six genomes from the (A) *B. fragilis* and (B) *P. aeruginosa* benchmark datasets. Precision was calculated as the proportion of true positive predictions to total predictions made. Recall was calculated as the proportion of true positive predictions to known MGEs. (C) The constituent tools (VirSorter, AGEnt, PhiSpy, CRISPR BLAST) comprising true positive (n = 56) and false positive (n = 60) MGE predictions were summed for both *B. fragilis* and *P. aeruginosa* datasets. Except for VirSorter predictions, VIC SIN requires consensus between two methods to accept a prediction. (D) For each in-group known MGE with a corresponding prediction, percent coverage was calculated as the ratio of the correctly predicted basepairs to the actual total basepairs of the known MGE. The black bars represent averages and standard deviations are displayed by error bars. (E) For each in-group known MGE with a corresponding prediction, the length of overprediction was calculated and each represented as a colored circle. Averages are displayed as black rectangles and standard deviations are displayed by error bars.

Figure 4. Network of clustered MGEs predicted from *Bacteroidaceae* hosts. VICSIN predicted MGEs from *B. fragilis* and *P. vulgatus* were clustered using VICSIN with a set of known *Bacteroidaceae* MGEs (Table S4). Each node nodes represents an MGE. Node size indicates if an MGE is a prediction (small) or known (large); node colors correspond to the MGE type, either based on known type or on Pfam assignment. Edges are drawn between nodes when $PLA \geq 20\%$. Clusters 1, 3, 21, 23, 29, 31, and 77 are indicated.

Figure 5. Analysis of functional gene sharing within and between MGE clusters using Pfams. Genes which were shared between MGE predictions were annotated using the Pfam database and assigned to a category on the basis of the MGE types involved, and whether that gene is shared within or between clusters. Mobile and unknown MGE categories are not shown. The number of sharing events for each Pfam was normalized to the total number of Pfams counted in each category (Pfam Domain

Count). 105 Pfams are shown, representing the 25 most frequently shared Pfams in each category. Pfams were assigned to functional groups based on the genetic elements they are known to be associated with; ambiguous Pfams were not assigned to a group. Functional groups highlighted: phage-associated (pink), capsule synthesis (blue), ICE-associated (cyan), carbohydrate utilization (green), and T6SS (orange). Rows were ordered and the dendrogram were generated using heatmaps.2 in the gplots package in R (<https://www.rdocumentation.org/packages/gplots/versions/3.0.4/topics/heatmap.2>).

Figure 6. Alignment of three example VIC SIN MGE clusters. Nucleotide alignments were made for select nodes from (A) Cluster 1, (B) Cluster 3, and (C) Cluster 23, and (D) Cluster 29. Alignments each show four VIC SIN predictions and one known MGE: CTn9343, CTnDOT, Bacteroides phage BV01, and Bacteroides phage p00. MGE genome maps at top represent the known MGE. Alignments made in Mauve (70). Colors represent locally collinear blocks and curves represent sequence conservation.

Figure 7. Sharing of ARGs and T6SS genes among *Bacteroidaceae* integrated MGEs. (A) ARGs from the CARD database were identified in VICsIN-predicted MGEs. (B) T6SS predictions were made for MGEs encoding *hcp*, *tssB*, and *vgrG*, which are conserved in all T6SS genetic architectures known to occur in *Bacteroidaceae* genomes (Table S6).