AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

CrAssphage May Be Viable Markers of Contamination in Pristine and Contaminated River Water

Nyasha Mafumo,[a] Oliver K. I. Bezuidt,[a] Wouter le Roux,[b] Thulani P. Makhalanyane[a]

[a]DSI/NRF SARChI in Marine Microbiomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa
[b]Water-Related Microbiology Laboratory, Water Centre, Pretoria, South Africa

ABSTRACT Viruses are the most biologically abundant entities and may be ideal indicators of fecal pollutants in water. Anthropogenic activities have triggered drastic ecosystem changes in rivers, leading to substantial shifts in chemical and biological attributes. Here, we evaluate the viability of using the presence of crAssphage as indicators of fecal contamination in South African rivers. Shotgun analysis revealed diverse crAssphage viruses in these rivers, which are impacted by chemical and biological pollution. Overall, the diversity and relative abundances of these viruses was higher in contaminated sites compared to pristine locations. In contrast to fecal coliform counts, crAssphage sequences were detected in pristine rivers, supporting the assertion that the afore mentioned marker may be a more accurate indicator of fecal contamination. Our data demonstrate the presence of diverse putative hosts which includes members of the phyla *Bacteroidota, Pseudomonadota, Verrucomicrobiota*, and *Bacillota*. Phylogenetic analysis revealed novel subfamilies, suggesting that rivers potentially harbor distinct and uncharacterized clades of crAssphage. These data provide the first insights regarding the diversity, distribution, and functional roles of crAssphage in rivers. Taken together, the results support the potential application of crAssphage as viable markers for water quality monitoring.

IMPORTANCE Rivers support substantial populations and provide important ecosystem services. Despite the application of fecal coliform tests and other markers, we lack rapid and reproducible approaches for determining fecal contamination in rivers. Waterborne viral outbreaks have been reported even after fecal indicator bacteria (FIB) were suggested to be absent or below regulated levels of coliforms. This indicates a need to develop and apply improved indicators of pollutants in aquatic ecosystems. Here, we evaluate the viability of crAssphage as indicators of fecal contamination in two South African rivers. We assess the abundance, distribution, and diversity of these viruses in sites that had been predicted pristine or contaminated by FIB analysis. We show that crAssphage are ideal and sensitive markers for fecal contamination and describe novel clades of crAss-like phages. Known crAss-like subfamilies were unrepresented in our data, suggesting that the diversity of these viruses may reflect geographic locality and dependence.

KEYWORDS bacteriophages, bacteria, crAssphage, metagenome assembled genomes, phylogeny, faecal pollution, viruses

The exposure to anthropogenic pollutants has resulted in the drastic decline in the quantity and quality of potable water. Aquatic pollutants include an assortment of toxic chemicals and pathogenic microorganisms (1, 2). In developing countries, threats to drinking water are exacerbated by the inadequate access to wastewater treatment facilities (3–5). As a result, drinking water sources may be exposed to human fecal contamination, which can result in low-quality drinking water, with increased potential of

Editor Hans C. Bernstein, UiT - The Arctic University of Norway

Address correspondence to Thulani P. Makhalanyane, thulani.makhalanyane@up.ac.za.

Received 22 December 2022
Accepted 10 January 2023
Published 6 February 2023

spreading waterborne diseases (6–9). Active surveillance to monitor and detect potential pathogens is vital for protecting public health and ensuring potable water (10–12). Currently, for microbial contaminants, these efforts have relied on the use of fecal indicator bacteria (FIB) such as *Escherichia coli* (5, 13, 14). However, the concentrations of FIB do not always correlate with the presence of some biological pollutants present in aquatic environments (15–20). This has been demonstrated by several studies, which have reported on waterborne viral outbreaks that occurred after analyses based on FIB were suggested to be absent or below regulated levels of coliforms (21, 22). This discrepancy demonstrates the failure of current methods and suggests the need to further develop new monitoring strategies, which account for all biological risks associated with contaminated water. This indicates a need for a universal marker, which is highly sensitive, for integration into modern microbial contamination surveillance protocols (23).

Viruses are the most biologically abundant entities on Earth and have been proposed to be better indicators of fecal pollutants (24–27). Several metagenomic studies have shown that crAssphage are the most abundant viruses in the human gut (28–31). The current data suggest that crAssphage cluster into five discrete groups designated as alpha-gamma, beta, delta, epsilon, and zeta (28), and span 10 distinct and phylogenetically diverse genera (32, 33). Based on the analysis of CRISPR spacer regions, and functional characterization, it appears that crAssphage mainly infect bacteria from the phylum Bacteroidota (28, 31, 32, 34, 35). Due to this abundance, these double-stranded (ds) DNA viruses may be ideal microbial source tracking (MST) markers of human fecal contamination (23, 36–38). Stachler and Bibby showed the potential utility of using crAssphage-based markers for tracking human fecal waste. The authors showed that crAssphage was highly host specific and highly abundant in sewage and biosols in Europe and the United States. The study also showed that crAssphage were detected in sewage samples from Asia and Africa, albeit at lower abundances. The results suggest that this phage is prevalent globally and support its application as a MST marker. Following these observations, several qPCR assays have been developed (37, 39). These assays have been successfully applied to quantify crAssphage in feces, wastewater, and surface waters in several regions, including parts of Europe (39–41), Asia (42, 43), North America (23, 44), South America, (45) and Australia (46). Evidence from these studies suggests that crAssphage may be highly specific for detecting human feces and sewage, with little or no cross-reactivity with animal feces, and hence, ideal MST markers. However, several studies have also demonstrated that crAssphage may not occur exclusively in the human gut, but may be present in the guts of animals and feces, albeit at lower concentrations (36, 47, 48). Nevertheless, there is an urgent need to investigate the suitability of using crAssphage as a biomarker of fecal contamination in underrepresented and understudied geographic locations such as Africa, to assess the feasibility of using the virus as a universal marker (48–50).

To reduce this knowledge deficit, we used shotgun metagenomic analysis to explore the diversity of crAss-like phages in two South African rivers (Fig. 1A). Based on previous studies (36, 37, 40, 49), we predict that crAssphage in contaminated sites will be more abundant and diverse. We selected three conserved capsid and genome-packaging proteins (terminase large subunit, portal proteins, and major capsid proteins) as markers for detecting crAss-like phages in samples collected from both pristine and contaminated sites. We characterized and classified crAss-like sequences from these environments into subfamilies and identified distinct clades of diverse crAss-like phages.

RESULTS

The diversity of bacterial genomes in river systems. In total, 297 bacterial metagenome assembled genomes (MAGs) were reconstructed from our six metagenomes. These MAGs constituted 115 high and 149 medium quality bins, which were dominated by *Pseudomonadota*, *Actinobacteriota*, and *Bacteroidota* phyla, respectively (Table S1). A phylogenetic reconstruction of *Bacteroidota* showed that MAGs from this study clustered separately from NCBI complete reference genomes (Fig. 1B). Furthermore, the data suggest a clear

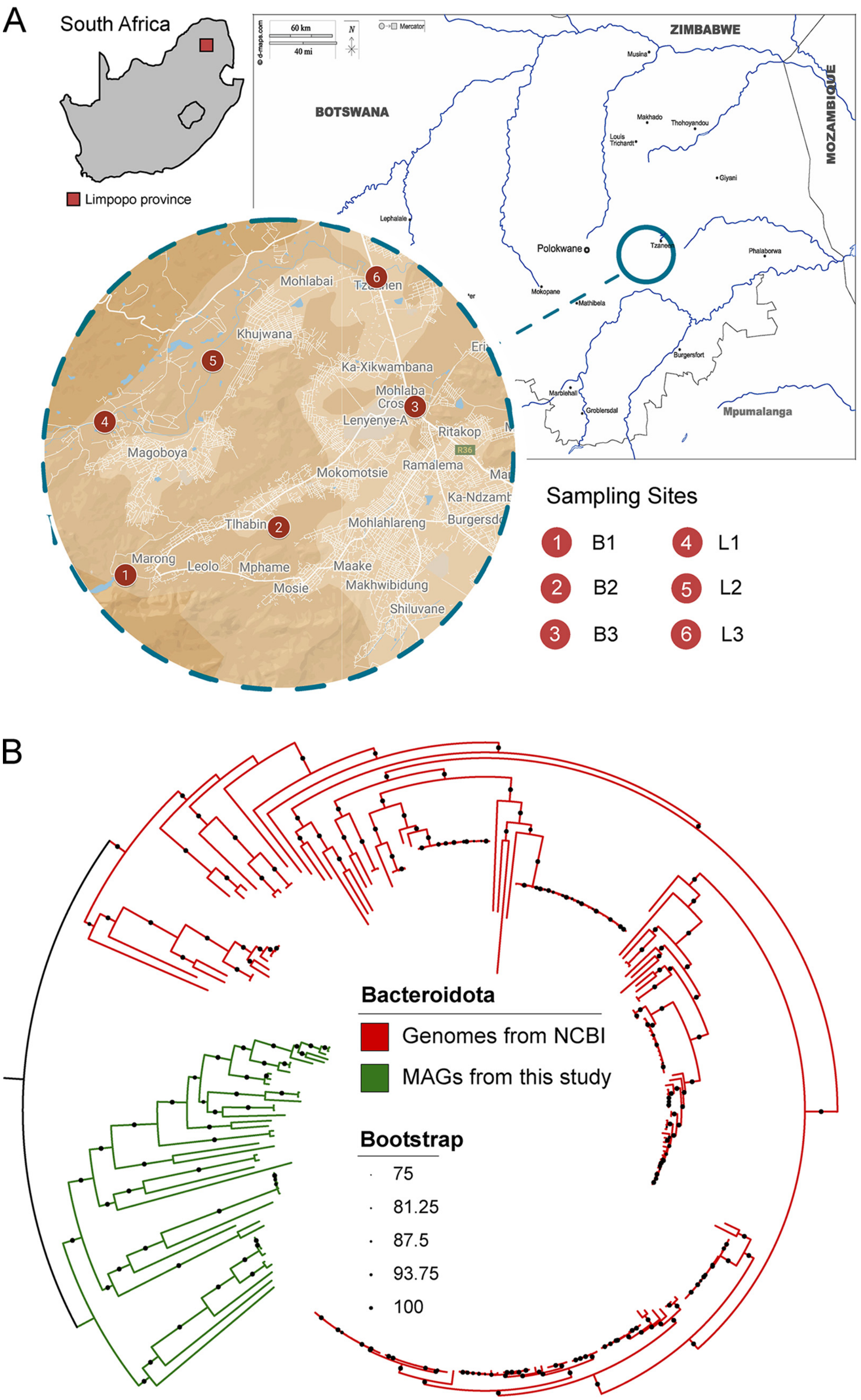

FIG 1 Map of sampling sites and a phylogenetic tree of Bacteroidota. (A) The six sampling sites in the Limpopo, Province of South Africa. The letters designate specific sampling locations as with L and B for Letsitele and Thabina

(Continued on next page)

genetic discontinuity between our *Bacteroidota* and those from the NCBI, based on ANI scores ≤88% (Table S2). Similarity searches using 298 CRISPR spacers, detected in the high and medium quality MAGs, predicted 16 potential bacterial hosts. These hosts were associated with crAss-like viruses recovered from our samples. Among these 16 putative hosts, 14 belonged to *Bacteroidota*, and remaining hosts were affiliated with *Verrucomicrobiota* and *Pseudomonadota*. Further analysis, using iPHoP, resulted in the prediction of six *Bacteroidota* and two *Bacillota* putative hosts (Table S3). This expanded the known pool of putative hosts to three, and added increased the repertoire of phyla infected by our crAss-like viruses.

CrAssphage relative abundances in contaminated and pristine water. As a proxy for crAssphage relative abundance, transcripts per million (TPM) values (Table S4) were generated to evaluate the distribution and diversity of crAss like phages in the different sampling sites (Fig. 2). The relative abundances suggest that some crAss-like phages may be site specific. For instance, several crAss-like phage genomes (crAss 13577, crAss 22689, crAss 50418) were found in high abundances in samples from Letsitele but were absent in Thabina. Furthermore, we also identified specific crAssphage genomes (crAss 12948, crAss 16291, crAss 3238), which were completely absent upstream of settlements (sites B1 and L1). However, these genomes were identified in downstream sites, which were located within the vicinity of human settlements and are subject to human fecal contamination due to the use of pit latrines (sites B2, B3, L2, and L3) (Table S4). Overall, as predicted, we observed higher relative abundances and diversity of crAss-like phages in the contaminated sites (B2, B3, L2) compared to pristine sites (Fig. 2).

Diversity of uncharacterized crAssphage in rivers. HMM profile searches, using major capsids, terminase large subunits (TerL), and portal proteins, resulted in the prediction of 384 crAss-like contigs across all metagenomes. Of the overall predicted crAss-like contigs, only 57 harbored ≥2 hallmark genes, representing a total of 50 vOTUs. Of these, we found six high-quality viral contigs, five medium-quality, and 46 low-quality viral contigs due to the highly fragmented nature of these sequences (Table S5). Moreover, the reconstruction of phylogeny, using dereplicated TerL protein sequences (51), showed the diversity of crAss-like phages associated with our data (Fig. 3). From this analysis, we observed that around 64% (48 of 75) of our crAss-like TerL sequences clustered separately from the previously proposed groups, whereas the remaining (34) sequences clustered with known Epsilon and Delta subfamilies.

Of the total crAss-like phage predictions, 20 contigs had all three hallmark genes. The size of the contigs ranged from 12 to 119.3 kbp. Of these, 17 contigs were classified as putative Epsilon (16) and Delta (1) subfamilies, respectively, based phylogenetic placements using TerL proteins. The Epsilon crAss-like phages were the largest contigs (Data set S1). We selected five high-quality and near-complete contigs, with size ≥100 kbp, and compared these to three human gut associated crAss-like genomes (Fig. 4). The analyses suggest that crAss-like contigs from our data were highly similar in comparison to reference genomes. However, the order of genes related to the RNAP subunit and to the capsid gene module, was conserved across all genomes (Fig. 4).

DISCUSSION

CrAssphage are potentially ideal MST markers for detecting human fecal contamination due to their abundance, specificity, and sensitivity (36, 37, 52–54). However, few studies have investigated the viability of using crAssphage as potential markers in environmental water at both global and local scales (36, 48, 50). In this study, we conducted *in silico* analysis to deter-

FIG 1 Legend (Continued)
(known locally as the Bathabina) River, respectively. Specific sites include the following: L1, upstream of the settlements and irrigation; L2, midstream; L3, downstream of the settlements; B1, upstream of the settlements; B2, midstream; B3, downstream of the settlements. (B) Bacteroidota maximum-likelihood phylogenetic tree. The metagenome assembled genomes (MAGs) obtained from this study are indicated in green. Those retrieved from the NCBI database are shown in red. The tree illustrates separate clustering between the MAGs obtained in this study and those from the NCBI. The genomes from this study form a separate and distinct cluster. Bootstrap values were calculated to support the robustness of the different clades and are indicated by the insert. The Limpopo map was sourced from dmaps (https://d-maps.com/pays.php?num_pay=1641&lang=en), and the inset was sourced from Google Maps.

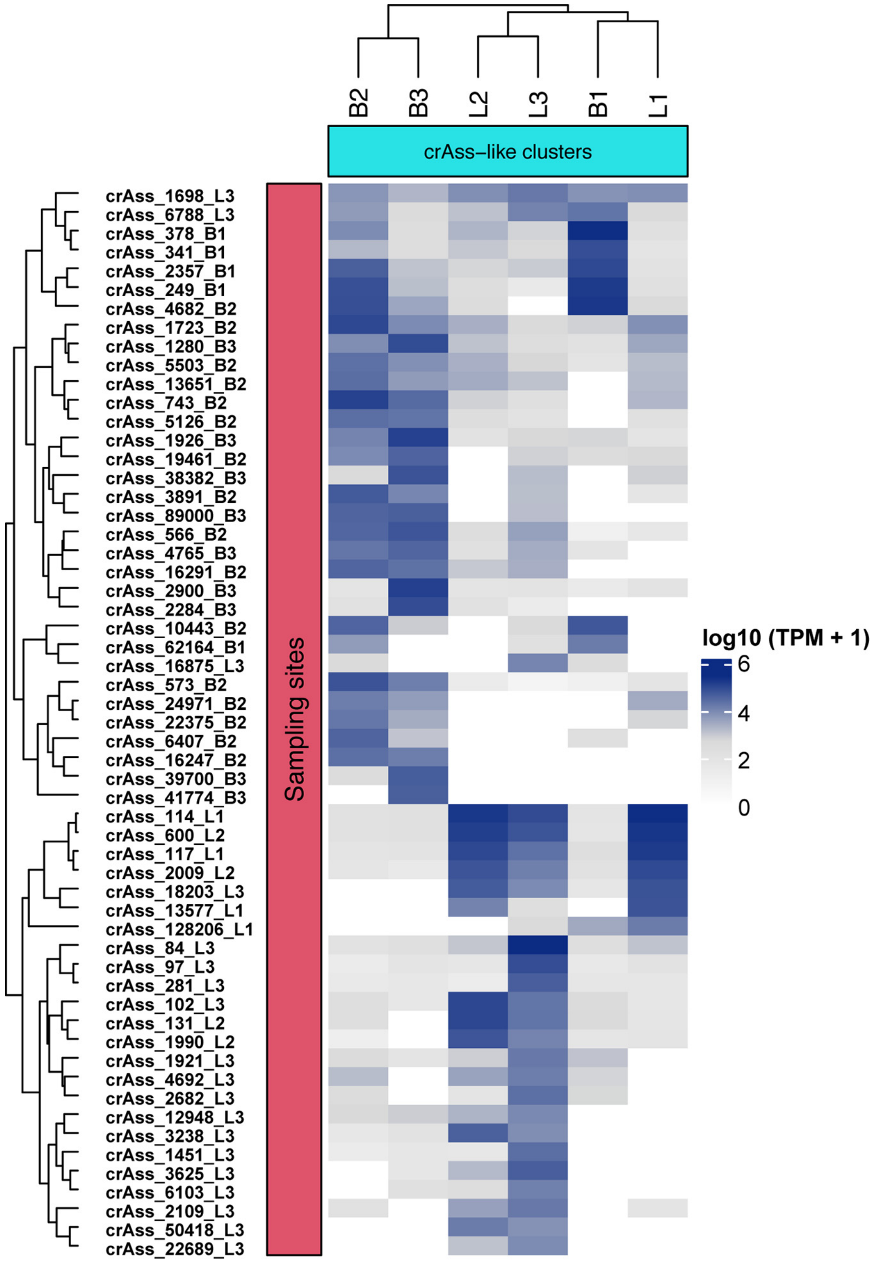

FIG 2 The relative abundances of crAssphage viruses. The relative abundance was calculated by read mapping sequences from the different sampling locations. The six sites (L1, L2, L3, B1, B2, B3) are shown in the *x* axis. The labels B and L correspond to Thabina and Letsitele river sampling sites, respectively. The cluster analysis was based on Euclidean distances. The putative crAssphage sequences, obtained from this study, are shown on the right *y* axis. The dark blue color indicates high crAssphage abundances, in a specific site, while white shows the absence of crAssphage per location. The heatmap shows generally higher diversity and distribution of crAssphage observed in contaminated sites compared with pristine sampling locations.

mine the feasibility of crAssphage as indicators of water contamination in two South African rivers. The sampling sites in the rivers were designated as pristine or contaminated based on fecal quality assessments (Table S6). Our analysis suggests that consistent with our prediction, crAssphage appear to be more abundant in contaminated sites (B2, B3, and L2) compared to pristine sites. We also reveal highly diverse and uncharacterized crAssphage in these rivers, expanding the genomic repertoire of known crAss-like clades.

By expanding the known crAss-like clade, our data shed new light regarding the established potential hosts of these viruses. Previous studies suggest that crAssphage primarily

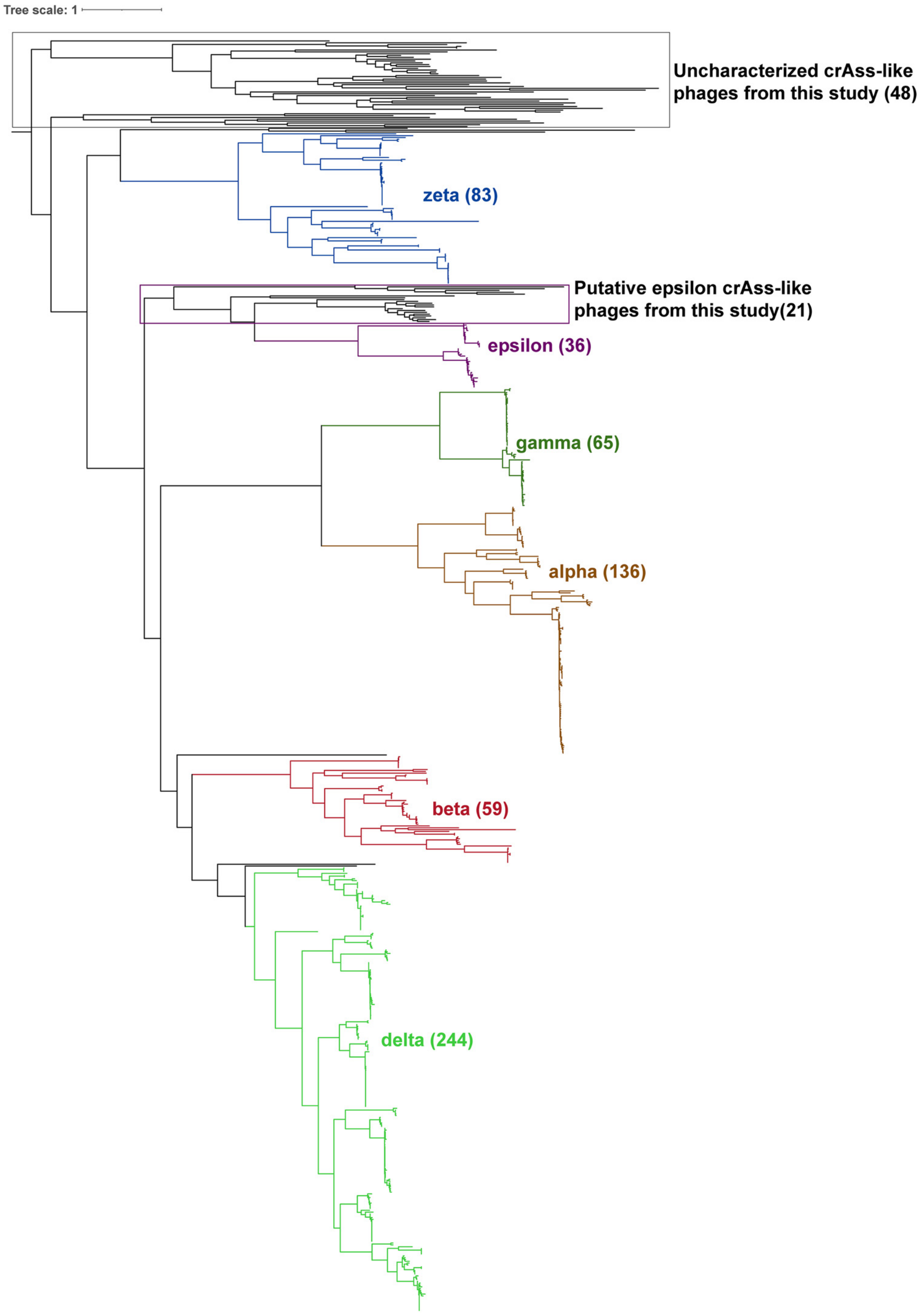
Tree scale: 1
Uncharacterized crAss-like phages from this study (48)
zeta (83)
Putative epsilon crAss-like phages from this study(21)
epsilon (36)
gamma (65)
alpha (136)
beta (59)
delta (244)

FIG 3 Phylogenetic tree of crAssphage sequences. The tree was constructed using TerL protein sequences from crAssphage. CrAssphage, obtained from this study, are colored black. CrAssphage sequences obtained from a recent study (28) representing Zeta, Epsilon, Gamma, Alpha, Beta, and Delta clusters are shown in blue, green, purple, brown, red, and green, respectively. CrAssphage obtained in this study clustered with the Zeta, Epsilon, and Delta subfamilies. In general, crAssphage retrieved from this study clustered separately from those obtained from the Yutin et al. (28) study. This suggests that they may potentially represent novel, and as yet, uncharacterized crAssphage sequences.

infect Bacteriodata (28, 31, 32, 34, 35). Host prediction using two independent approaches suggests the presence of three other potential host phyla, *Bacillota*, *Verrucomicrobia*, and *Pseudomonadota*. These results corroborate a recent *in silico* prediction by N. Yutin et al. (28), which proposed *Pseudomonadota* and *Bacillota* as other putative hosts of crAssphage. These findings provide additional evidence confirming that crAssphage may have a wider

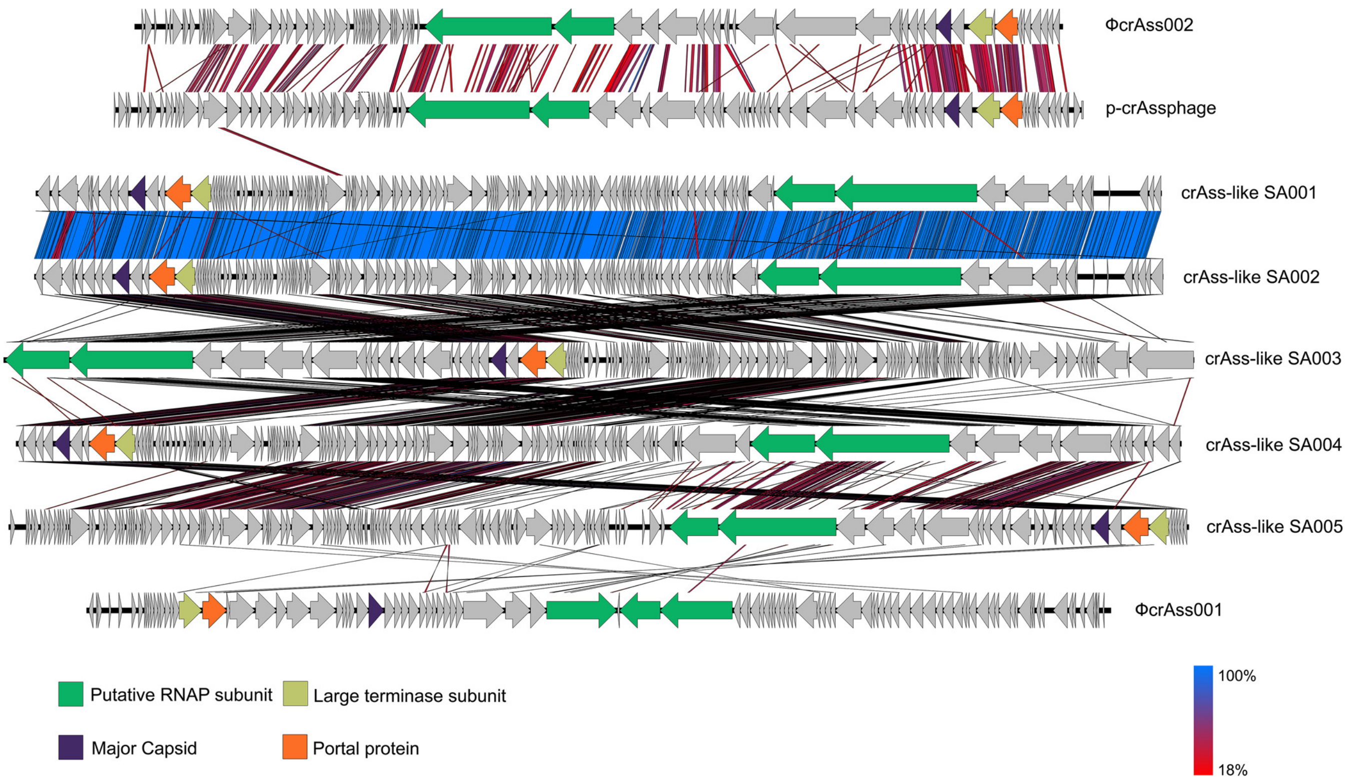

FIG 4 Genomic structure comparison of near complete Epsilon crAssphage. The crAssphage sequences from this study were compared with the prototypical crAssphage (p-crAssphage), and two other gut associated crAssphage. Regions of amino acid sequence homology are shown. Blue indicates 100% homology and red indicates the lowest similarities. The RNAP subunit and the capsid gene module were found in all the genomes analyzed.

range of hosts than initially predicted. In addition to the expanded host phyla, phylogenetic analysis suggests that the *Bacteriodata* from our study may be more diverse than previously described genomes. This is perhaps unsurprising as the majority of current *Bacteroidota* genomes were isolated from human guts of western populations (55–57). Our MAGs were retrieved from African rivers, which are under different evolutionary selective pressures (e.g., temperature, physicochemical variables) compared to gut bacteria (58). It has been established that differences in geographical origin, ethnicity, and urbanization substantially shape the diversity of microbiota (59–61). It is possible that these variables may contribute to the disparities observed in this study at the species level (60). Our findings suggest that crAssphage host diversity may be substantially under characterized. It is possible that, depending on the environmental niche, crAssphage may associate with other hosts. For instance, Verrucomicrobia, which are known to be more abundant in soils (62, 63), may be more ideal hosts for crAssphage in these niches.

The relative abundance estimates suggest that crAssphage dominate impacted river sites. This result validates our proposal that these sequences may be viable markers of contamination and is consistent with previous studies based on quantitative PCR (37–40, 42–44). A previous study by Stachler et al. used sequence data to demonstrate the potential of crAssphage based biomarkers. To the best of our knowledge, this study provides the first evidence, based on genome resolved metagenomics, showing the applicability of crAssphage sequences as markers of contamination in South African rivers. While FIB were useful in providing a general indication of pristine and contaminated sites, our analysis suggests that crAssphage may be more sensitive markers. In addition to their sensitivity (38, 43, 54, 64), previous studies suggest that crAssphage are more resistant to environmental stress compared to FIB (65). The ability to withstand environmental stress may favor the use of these indicators as ideal markers of contamination.

An additional benefit of using crAssphage, as molecular markers of contamination, is their diversity and global distribution (28, 32, 66, 67). Based on evaluating the diversity and phylogeny of crAss-like viruses, the 48 TerL protein sequences from our data were highly distinct relative to known clades. This distinct clustering suggests that crAss-like phages from our data set may represent highly diverse subfamilies, reflects geographic locality and dependence (48), and further supports our assertion regarding evolutionary niche selection of these bacteriophages. Previous work on wastewater has provided strong evidence of geographic dependence of crAssphage (36, 48). These studies hypothesized that the abundances and diversity of crAssphage would be higher in European and U.S. samples, compared to those in Africa and Asia (36, 48, 66). Similar to findings on host microbiota, the variation in the diversity of crAssphage may be explained by the differences in urbanization and diet (59, 68, 69). These differences have been shown to drive changes in both host and phage genomes (66, 69). A small proportion of the remaining TerL sequences (27) identified in this study were classified with known Epsilon, Delta, and Zeta subfamilies. Epsilon sequences were recently described by N. Yutin et al. (28) and appear to be dominated by sequences from gut microbiota. The Delta subfamilies constitute the largest group of gut crAss virome and are known to be distantly related to the Epsilon subfamilies (28). The identification of sequences from these subfamilies hints at the presence of human fecal contamination in the river samples. However, relative to the other sequences, these subfamilies were underrepresented in the sampling area. The majority of these were affiliated with uncharacterized crAss-like sequences, reported in this study, representing a potentially novel clade. This finding further supports the effects of selective pressures and crAssphage niche specificity.

To further elucidate the diversity of crAssphage, comparison of the genomic structure of five near-complete sequences retrieved from this study was done. By comparing these sequences with ΦcrAss001 (35), ΦcrAss002 (70), and prototypical crAssphage (p-crAssphage) (31), the analysis revealed high variability in similarity and gene order conservation among the phages. From the five epsilon crAssphage, two were nearly identical, and three were highly homologous. This finding supports the assertion that sequences from this study were both novel and distinct from previously characterized phages. The observed differences suggest that, although the five crAss-like phages were classified within the epsilon subfamily, these may in fact represent four new genera (32, 33) and calls for the revision and possible expansion of the current taxonomy. Together these results support our proposed view regarding the potentially diverse assortment of uncharacterized crAss-like sequences in various environments.

Conclusion and future prospects. Our findings increase the repertoire of known crAss-like phages. The expanded taxonomy, including potentially novel clades, establishes a baseline for the identification of as yet unknown environmental crAssphage. These novel clades reported in this study were linked to new host phyla (*Bacillota*, *Verrucomicrobia*, and *Pseudomonadota*), confirming previous reports which demonstrated that crAssphage are not exclusively associated with *Bacteroidota*. In addition, our analysis supports the use of crAssphage viruses as biomarkers of fecal contamination. Using crAssphage as microbial source tracking markers may assist in mitigating the spread of waterborne diseases due to their robustness and sensitivity. However, future studies in environmental crAssphage viruses are required to validate these observations across a variety of river systems. These validations are required to establish the diversity of these sequences, which may result in the application of crAssphage as quantitative markers of fecal contamination.

MATERIALS AND METHODS

Study area, sample collection, and processing. The samples were collected in the Limpopo Province of South Africa, in the Bathlabile tribal area, following consent from the Bathlabile Traditional Council. The area relies on water sourced from the Thabina and Letsitele rivers, which are subject to chemical and biological pollutants. The chemical pollutants in the Thabina river are primarily due to the application of organic fertilizers to nearby citrus farms, while biological pollutants in both rivers are in the form of fecal matter from humans (due to the use of pit latrines) and free roaming animals.

Six water samples were collected from pristine and polluted river waters. The selected locations were distributed at three sites along the Thabina and Lestsitele rivers (Fig. 1A). The B1 and L1 sites being from a pristine source, upstream of settlements, and agricultural practices. The B2 and L2 sites were in the middle of settlements, whereas B3 and L3 sites were downstream of the settlements (Fig. 1A). Five liters of water were collected from each location for metagenomic analysis and stored on ice. An additional liter of water from each sampling location was collected for standard water quality analysis at the Council of Scientific and Industrial Research (Table S6). The pH, turbidity, coliform, and *E. coli* counts were determined for each sample (Table S6). Each sample was filtered through a 0.2 μm polycarbonate (PES) filter membranes (Merck, RSA). DNA extractions from these filters followed, using the Qiagen Power Soil DNA Isolation Kits (Qiagen, USA) according to the manufacturer's instructions. The quality of the resultant DNA was evaluated using gel electrophoresis (1% agarose) and its concentration was determined with Qubit dsDNA assay kit in Qubit 4 Fluorometer (Thermo Fisher Scientific, USA). High-quality DNA from each sample was sent for shotgun sequencing using an Illumina MiSeq (2 × 150 bp) at Admera Health (NJ, USA).

Metagenomic analysis. To elucidate potential crAssphage hosts, we generated MAGs from shotgun data. Raw paired-end metagenomic reads (2 × 150 bp) were quality trimmed, assembled, and binned using the ATLAS pipeline version 2.4.4 with the qc, assembly, and binning parameters (51). Briefly, using default versions of the tools below, the reads were quality filtered using BBTools version 37.99 (https://jgi.doe.gov/data-and-tools/bbtools/), assembled using metaSPAdes version 3.13.1 (71), and MAGs were generated using both metabat2 (72) and maxbin2 (73). The resultant bins were combined, refined and dereplicated using DAS_Tool version 1.1.2 (74) and dREP version 3.0.0 (75). The dereplicated MAGs were assessed for quality and completeness using CheckM version 1.1.5 (76). Using previously defined standards, MAGs with completeness $>$90% and contamination $<$5% were classified as high-quality (Table S7) and those with completeness of $\geq$50% and contamination of $<$10% were classified as medium-quality (77) (Tables S8 to 10).

Taxonomic annotation of bacterial MAGs and host detection. The taxonomy of all medium- and high-quality MAGs were inferred using the Genome Taxonomy Database Toolkit version 1.6.0 (78). Of these, we reconstructed a maximum likelihood phylogenetic tree using MAGs classified as Bacteroidota as these are primary hosts for crAssphage using GTOtree version 1.6.11 (79). We used single copy gene sets (80) specific for the *Bacteroidota* phylum. This tree comprised of 47 MAGs, generated from this study, and included 450 reference *Bacteroidota* complete genomes acquired from the NCBI RefSeq release 205 database (81). The phylogenetic tree was visualized and annotated using iTol version 6.6 (82). Following this, we conducted pairwise comparisons between the Bacteroidota MAGs from this study and the 450 reference Bacteroidota. For these comparisons, we established the criteria of shared average nucleotide identity (ANI) using FastANI version 1.32 (83).

Hidden Markov models-based detection of crAssphage. For the detection of putative crAssphage, we acquired 81,246 crAss-related protein sequences from both the NCBI RefSeq (81) and a recent study by N. Yutin et al. (28). These sequences were clustered into a nonredundant set of 20,039 protein sequences using CD-HIT version 4.8.1 (84) with parameters: -c 0.9 -n 5 -aS 0.8. The representative sequences from these clusters were further compared for shared similarities using BLASTp (80) with e-value 1e-05. The blast results were clustered using Markov Clustering algorithm (MCL) version 14.137 (85) with 1.5 inflation. Clusters associated with major capsids, terminase large subunits (TerL), and portal proteins were individually aligned using MAFFT version 7.487 (86) with the -auto parameter, and the resultant alignments were converted to hidden Markov models (HMM) profiles using hmmbuild version 3.3.0 (87). To search for putative crAss-like viruses in our data, contigs were predicted for open reading frames (ORFs), using Prodigal v2.6.3 (88), with the -a and -p meta parameters. The profiles were subsequently searched against all protein sequences predicted in our contigs using the -T 50 parameter in hmmscan version 3.3.0 (87). Contigs predicted to be in possession of $\geq$2 of these crAss-like hallmark genes were clustered based on 95% sequence identity with over 80% of the shortest contig resulting in 50 viral OTUs and further assessed for quality using CheckV version 0.9.0 (Table S5).

CrAssphage relative abundance and host detection. As a proxy for determining the distribution and relative abundances of crAss-like phages across sampling sites, the CoverM version 0.6.1 (https://github.com/wwood/CoverM) tool was used to calculate TPM values. We used 57 crAss-like contigs with $\geq$2 hallmark genes and parent metagenomic reads. The resultant relative abundance estimates were then visualized using ggplot2 (89) in R v3.6.0 (90). Furthermore, these crAss-like contigs were also processed for host detection. MAGs generated from this study were searched for CRISPR spacers using MINCED version 0.4.2 (https://github.com/ctSkennerton/minced/tree/master). The identified spacers were probed for shared sequence similarity against the 57 crAss-like contigs, using BLASTn with the following parameters: -e-value 0.01 -word_size 8 -dust no -perc_identity 90. The crAss-like contigs were analyzed for host detection, using the iPHoP (integrated Phage Host Predictions) version 1.1.0 (https://bitbucket.org/srouxjgi/iphop). The taxonomies of our MAGs were first reclassified with GTDB-Tk version 2.1.0 using the de_novo_wf parameter as required by the pipeline and were further integrated into the database of hosts. Following this, the iPHoP pipeline was run using the default parameters.

Genome comparisons and subfamily classification of crAss-like phages. To determine the crAss-like subfamilies in our data set, we retrieved 96 complete TerL protein sequences from all our metagenomes. Protein sequences were clustered based on 95% sequence identity, over 80% of the shortest contig, resulting in the representation of 75 viral OTUs. These were supplemented, and subsequently aligned with 623 other TerL sequences acquired from N. Yutin et al. (28) using MAFFT version 7.487 (86) with the -auto parameter. The resultant alignment file was used to reconstruct a maximum likelihood tree with 1,000 bootstraps using IQ-TREE version 1.6.6 (91). The tree was then midpoint-rooted, visualized, and annotated using iToL (82). This was followed by comparisons of the five high-quality near complete ($\geq$100 kbp) epsilon crAss-like genomes against each other, as well as ΦcrAss001, (35), ΦcrAss002 (70) (two pure culture isolates), and a prototypical crAssphage (*in silico* derived) (31) to assess relatedness using EasyFig version 2.2.2 (92) using tblastx with parameters: e-value cut-off 0.001 and length filter 30.

Data availability. The metagenomic data have been deposited at NCBI under BioProject ID PRJNA894350. Metagenomic assembled genomes are available from https://doi.org/10.6084/m9.figshare.21640316. The bash script and HMM profiles used to search for crAss-like viruses in our data can be accessed from https://github.com/SAmicrobiomes/crAssZA.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TABLE S1, XLSX file, 0.02 MB.

TABLE S2, XLSX file, 0.01 MB.

TABLE S3, XLSX file, 0.01 MB.

TABLE S4, XLSX file, 0.01 MB.

TABLE S5, XLSX file, 0.01 MB.

TABLE S6, XLSX file, 0.01 MB.

TABLE S7, XLSX file, 0.01 MB.

TABLE S8, XLSX file, 0.01 MB.

TABLE S9, XLSX file, 0.01 MB.

TABLE S10, XLSX file, 0.01 MB.

ACKNOWLEDGMENTS

We thank the Bathlabile Traditional Council for granting us permission to access and sample within their jurisdiction. We wish to acknowledge the South African Department of Science and Innovation, the Technology Innovation Agency (TIA) and the University of Pretoria for funding. T.P.M. wishes to acknowledge the National Research Foundation for support (UID 129225, 136491). N.M. wishes to acknowledge the Technology Innovation Agency for providing a scholarship. We acknowledge the Centre for High Performance Computing (Cape Town, South Africa) and the University of Pretoria's Centre for Bioinformatics and Computation Biology for providing computational resources. This manuscript is part of ongoing research supported by the DSI/NRF SARChI in Marine Microbiomics.

REFERENCES

1. Ouyang Y. 2005. Evaluation of river water quality monitoring stations by principal component analysis. Water Res 39:2621–2635. https://doi.org/10.1016/j.watres.2005.04.024.
2. Magana-Arachchi DN, Wanigatunge RP. 2020. Ubiquitous waterborne pathogens, p 15–42. *In* Prasad MN, Grobelak A (ed), Waterborne pathogens detection and determination. Butterworth-Heinemann, Oxford. https://doi.org/10.1016/B978-0-12-818783-8.00002-5.
3. Massoud MA, Tarhini A, Nasr JA. 2009. Decentralized approaches to wastewater treatment and management: applicability in developing countries. J Environ Manage 90:652–659. https://doi.org/10.1016/j.jenvman.2008.07.001.
4. Edokpayi JN, Odiyo JO, Durowoju OS. 2017. Impact of wastewater on surface water quality in developing countries: a case study of South Africa, p 401–416. *In* Tutu H (ed), Water quality. TechOpen, London.
5. Bartram J, Fewtrell L. 2001. Water quality: guidelines, standards and health: assessment of risk and risk management for water-related infectious disease. IWA.
6. Ashbolt NJ. 2015. Microbial contamination of drinking water and human health from community water systems. Curr Environ Health Rep 2:95–106. https://doi.org/10.1007/s40572-014-0037-5.
7. Pal M, Ayele Y, Hadush M, Panigrahi S, Jadhav V. 2018. Public health hazards due to unsafe drinking water. Air Water Borne Dis 7:2.
8. Leclerc H, Schwartzbrod L, Dei-Cas E. 2002. Microbial agents associated with waterborne diseases. Crit Rev Microbiol 28:371–409. https://doi.org/10.1080/1040-840291046768.
9. Ashbolt NJ. 2004. Microbial contamination of drinking water and disease outcomes in developing regions. Toxicology 198:229–238. https://doi.org/10.1016/j.tox.2004.01.030.
10. Stewart JR, Boehm AB, Dubinsky EA, Fong T-T, Goodwin KD, Griffith JF, Noble RT, Shanks OC, Vijayavel K, Weisberg SB. 2013. Recommendations following a multi-laboratory comparison of microbial source tracking methods. Water Res 47:6829–6838. https://doi.org/10.1016/j.watres.2013.04.063.
11. McLellan SL, Eren AM. 2014. Discovering new indicators of fecal pollution. Trends Microbiol 22:697–706. https://doi.org/10.1016/j.tim.2014.08.002.
12. Herrig I, Seis W, Fischer H, Regnery J, Manz W, Reifferscheid G, Böer S. 2019. Prediction of fecal indicator organism concentrations in rivers: the shifting role of environmental factors under varying flow conditions. Environ Sci Eur 31:59. https://doi.org/10.1186/s12302-019-0250-9.
13. Edberg SC, Rice EW, Karlin RJ, Allen MJ. 2000. Escherichia coli: the best biological drinking water indicator for public health protection. J Applied Microbiology 88:106S–116S. https://doi.org/10.1111/j.1365-2672.2000.tb05338.x.
14. Goh SG, Saeidi N, Gu X, Vergara GGR, Liang L, Fang H, Kitajima M, Kushmaro A, Gin KYH. 2019. Occurrence of microbial indicators, pathogenic bacteria and viruses in tropical surface waters subject to contrasting land use. Water Res 150:200–215. https://doi.org/10.1016/j.watres.2018.11.058.
15. Jiang S, Noble R, Chu W. 2001. Human adenoviruses and coliphages in urban runoff-impacted coastal waters of Southern California. Appl Environ Microbiol 67:179–184. https://doi.org/10.1128/AEM.67.1.179-184.2001.
16. Noble RT, Fuhrman JA. 2001. Enteroviruses detected by reverse transcriptase polymerase chain reaction from the coastal waters of Santa Monica Bay, California: low correlation to bacterial indicator levels, p 175–184. *In* Porter JW (ed), The ecology and etiology of newly emerging marine diseases. Springer, Dordrecht, DL.
17. Rosario K, Symonds EM, Sinigalliano C, Stewart J, Breitbart M. 2009. Pepper mild mottle virus as an indicator of fecal pollution. Appl Environ Microbiol 75:7261–7267. https://doi.org/10.1128/AEM.00410-09.
18. McKee AM, Cruz MA. 2021. Microbial and viral indicators of pathogens and human health risks from recreational exposure to waters impaired by fecal contamination. J Sustainable Water Built Environ 7:e03121001. https://doi.org/10.1061/JSWBAY.0000936.
19. Vivier J, Ehlers M, Grabow W. 2004. Detection of enteroviruses in treated drinking water. Water Res 38:2699–2705. https://doi.org/10.1016/S0043-1354(01)00433-X.

20. Fong T-T, Lipp EK. 2005. Enteric viruses of humans and animals in aquatic environments: health risks, detection, and potential water quality assessment tools. Microbiol Mol Biol Rev 69:357–371. https://doi.org/10.1128/MMBR.69.2.357-371.2005.
21. Hauri AM, Schimmelpfennig M, Walter-Domes M, Letz A, Diedrich S, Lopez-Pila J, Schreier E. 2005. An outbreak of viral meningitis associated with a public swimming pond. Epidemiol Infect 133:291–298. https://doi.org/10.1017/s0950268804003437.
22. Karmakar S, Rathore AS, Kadri SM, Dutt S, Khare S, Lal S. 2008. Post-earthquake outbreak of rotavirus gastroenteritis in Kashmir (India): an epidemiological analysis. Public Health 122:981–989. https://doi.org/10.1016/j.puhe.2008.01.006.
23. Ahmed W, Lobos A, Senkbeil J, Peraud J, Gallard J, Harwood VJ. 2018. Evaluation of the novel crAssphage marker for sewage pollution tracking in storm drain outfalls in Tampa, Florida. Water Res 131:142–150. https://doi.org/10.1016/j.watres.2017.12.011.
24. Hamza IA, Jurzik L, Überla K, Wilhelm M. 2011. Evaluation of pepper mild mottle virus, human picobirnavirus and Torque teno virus as indicators of fecal contamination in river water. Water Res 45:1358–1368. https://doi.org/10.1016/j.watres.2010.10.021.
25. Farkas K, Walker DI, Adriaenssens EM, McDonald JE, Hillary LS, Malham SK, Jones DL. 2020. Viral indicators for tracking domestic wastewater contamination in the aquatic environment. Water Res 181:115926. https://doi.org/10.1016/j.watres.2020.115926.
26. Lin J, Ganesh A. 2013. Water quality indicators: bacteria, coliphages, enteric viruses. Int J Environ Health Res 23:484–506. https://doi.org/10.1080/09603123.2013.769201.
27. Skraber S, Gassilloud B, Gantzer C. 2004. Comparison of coliforms and coliphages as tools for assessment of viral contamination in river water. Appl Environ Microbiol 70:3644–3649. https://doi.org/10.1128/AEM.70.6.3644-3649.2004.
28. Yutin N, Benler S, Shmakov SA, Wolf YI, Tolstoy I, Rayko M, Antipov D, Pevzner PA, Koonin EV. 2021. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative crAss-like phages with unique genomic features. Nat Commun 12:1044. https://doi.org/10.1038/s41467-021-21350-w.
29. Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, McDonnell SA, Khokhlova EV, Draper LA, Forde A, Guerin E, Velayudhan V, Ross RP, Hill C. 2019. The human gut virome is highly diverse, stable, and individual specific. Cell Host Microbe 26:527–541.e5. https://doi.org/10.1016/j.chom.2019.09.009.
30. Manrique P, Bolduc B, Walk ST, van der Oost J, de Vos WM, Young MJ. 2016. Healthy human gut phageome. Proc Natl Acad Sci U S A 113:10400–10405. https://doi.org/10.1073/pnas.1601060113.
31. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, Felts B, Dinsdale EA, Mokili JL, Edwards RA. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun 5:4498. https://doi.org/10.1038/ncomms5498.
32. Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA, Koonin EV. 2018. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat Microbiol 3:38–46. https://doi.org/10.1038/s41564-017-0053-y.
33. Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ, Sutton TDS, Draper LA, Gonzalez-Tortuero E, Ross RP, Hill C. 2018. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. Cell Host Microbe 24:653–664.e6. https://doi.org/10.1016/j.chom.2018.10.002.
34. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. 2017. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res 45:39–53. https://doi.org/10.1093/nar/gkw1002.
35. Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, Hill C. 2018. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. Nat Commun 9. https://doi.org/10.1038/s41467-018-07225-7.
36. Stachler E, Bibby K. 2014. Metagenomic evaluation of the highly abundant human gut bacteriophage crAssphage for source tracking of human fecal pollution. Environ Sci Technol Lett 1:405–409. https://doi.org/10.1021/ez500266s.
37. Stachler E, Kelty C, Sivaganesan M, Li X, Bibby K, Shanks OC. 2017. Quantitative CrAssphage PCR assays for human fecal pollution measurement. Environ Sci Technol 51:9146–9154. https://doi.org/10.1021/acs.est.7b02703.
38. Sala-Comorera L, Reynolds LJ, Martin NA, Pascual-Benito M, Stephens JH, Nolan TM, Gitto A, O'Hare GMP, O'Sullivan JJ, García-Aljaro C, Meijer WG. 2021. CrAssphage as a human molecular marker to evaluate temporal and spatial variability in faecal contamination of urban marine bathing waters. Sci Total Environ 789:147828. https://doi.org/10.1016/j.scitotenv.2021.147828.
39. García-Aljaro C, Ballesté E, Muniesa M, Jofre J. 2017. Determination of crAssphage in water samples and applicability for tracking human faecal pollution. Microb Biotechnol 10:1775–1780. https://doi.org/10.1111/1751-7915.12841.
40. Farkas K, Adriaenssens EM, Walker DI, McDonald JE, Malham SK, Jones DL. 2019. Critical evaluation of crAssphage as a molecular marker for human-derived wastewater contamination in the aquatic environment. Food Environ Virol 11:113–119. https://doi.org/10.1007/s12560-019-09369-1.
41. Crank K, Li X, North D, Ferraro GB, Iaconelli M, Mancini P, La Rosa G, Bibby K. 2020. CrAssphage abundance and correlation with molecular viral markers in Italian wastewater. Water Res 184:116161. https://doi.org/10.1016/j.watres.2020.116161.
42. Malla B, Ghaju Shrestha R, Tandukar S, Sherchand JB, Haramoto E. 2019. Performance evaluation of human-specific viral markers and application of pepper mild mottle virus and crAssphage to environmental water samples as fecal pollution markers in the Kathmandu Valley, Nepal. Food Environ Virol 11:274–287. https://doi.org/10.1007/s12560-019-09389-x.
43. Kongprajug A, Mongkolsuk S, Sirikanchana K. 2019. CrAssphage as a potential human sewage marker for microbial source tracking in Southeast Asia. Environ Sci Technol Lett 6:159–164. https://doi.org/10.1021/acs.estlett.9b00041.
44. Stachler E, Akyon B, de Carvalho NA, Ference C, Bibby K. 2018. Correlation of crAssphage qPCR markers with culturable and molecular indicators of human fecal pollution in an impacted urban watershed. Environ Sci Technol 52:7505–7512. https://doi.org/10.1021/acs.est.8b00638.
45. Jennings WC, Gálvez-Arango E, Prieto AL, Boehm AB. 2020. CrAssphage for fecal source tracking in Chile: covariation with norovirus, HF183, and bacterial indicators. Water Res X 9:100071. https://doi.org/10.1016/j.wroa.2020.100071.
46. Ahmed W, Payyappat S, Cassidy M, Besley C, Power K. 2018. Novel crAssphage marker genes ascertain sewage pollution in a recreational lake receiving urban stormwater runoff. Water Res 145:769–778. https://doi.org/10.1016/j.watres.2018.08.049.
47. Li Y, Gordon E, Shean RC, Idle A, Deng X, Greninger AL, Delwart E. 2021. CrAssphage and its bacterial host in cat feces. Sci Rep 11:815. https://doi.org/10.1038/s41598-020-80076-9.
48. Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K, Dinsdale EA, Cinek O, Aziz RK, McNair K, Barr JJ, Bibby K, Brouns SJJ, Cazares A, de Jonge PA, Desnues C, Díaz Muñoz SL, Fineran PC, Kurilshikov A, Lavigne R, Mazankova K, McCarthy DT, Nobrega FL, Reyes Muñoz A, Tapia G, Trefault N, Tyakht AV, Vinuesa P, Wagemans J, Zhernakova A, Aarestrup FM, Ahmadov G, Alassaf A, Anton J, Asangba A, Billings EK, Cantu VA, Carlton JM, Cazares D, Cho G-S, Condeff T, Cortés P, Cranfield M, Cuevas DA, De la Iglesia R, Decewicz P, Doane MP, Dominy NJ, Dziewit L, Elwasila BM, Eren AM, et al. 2019. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. Nat Microbiol 4:1727–1736. https://doi.org/10.1038/s41564-019-0494-6.
49. Ward LM, Ghaju SR, Tandukar S, Sherchand JB, Haramoto E, Sherchan SP. 2020. Evaluation of crAssphage marker for tracking fecal contamination in river water in Nepal. Water Air Soil Pollut 231:282. https://doi.org/10.1007/s11270-020-04648-1.
50. Sabar MA, Honda R, Haramoto E. 2022. CrAssphage as an indicator of human-fecal contamination in water environment and virus reduction in wastewater treatment. Water Res 221:118827. https://doi.org/10.1016/j.watres.2022.118827.
51. Kieser S, Brown J, Zdobnov EM, Trajkovski M, McCue LA. 2020. ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. BMC Bioinformatics 21:257. https://doi.org/10.1186/s12859-020-03585-4.
52. Chen H, Liu C, Li Y, Teng Y. 2021. Integrating metagenomic and Bayesian analyses to evaluate the performance and confidence of crAssphage as an indicator for tracking human sewage contamination in China. Environ Sci Technol 55:4992–5000. https://doi.org/10.1021/acs.est.1c00071.
53. Ahmed W, Gyawali P, Feng S, McLellan SL. 2019. Host specificity and sensitivity of established and novel sewage-associated marker genes in human and nonhuman fecal samples. Appl Environ Microbiol 85:e00641-19. https://doi.org/10.1128/AEM.00641-19.

54. Sangkaew W, Kongprajug A, Chyerochana N, Ahmed W, Rattanakul S, Denpetkul T, Mongkolsuk S, Sirikanchana K. 2021. Performance of viral and bacterial genetic markers for sewage pollution tracking in tropical Thailand. Water Res 190:116706. https://doi.org/10.1016/j.watres.2020.116706.
55. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, MetaHIT Consortium., et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65. https://doi.org/10.1038/nature08821.
56. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Paul Brooks J, Buck GA, Buhay CJ, Busam DA, Campbell JL, et al. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214.
57. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD. 2021. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 39:105–114. https://doi.org/10.1038/s41587-020-0603-3.
58. Cooper VS, Honsa E, Rowe H, Deitrick C, Iverson AR, Whittall JJ, Neville SL, McDevitt CA, Kietzman C, Rosch JW. 2020. Experimental evolution in vivo to identify selective pressures during pneumococcal colonization. mSystems 5:e00352-20. https://doi.org/10.1128/mSystems.00352-20.
59. Zuo T, Sun Y, Wan Y, Yeoh YK, Zhang F, Cheung CP, Chen N, Luo J, Wang W, Sung JJY, Chan PKS, Wang K, Chan FKL, Miao Y, Ng SC. 2020. Human-gut-DNA virome variations across geography, ethnicity, and urbanization. Cell Host Microbe 28:741–751.e4. https://doi.org/10.1016/j.chom.2020.08.005.
60. Gupta VK, Paul S, Dutta C. 2017. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. Front Microbiol 8:1162. https://doi.org/10.3389/fmicb.2017.01162.
61. Dwiyanto J, Hussain MH, Reidpath D, Ong KS, Qasim A, Lee SWH, Lee SM, Foo SC, Chong CW, Rahman S. 2021. Ethnicity influences the gut microbiota of individuals sharing a geographical location: a cross-sectional study from a middle-income country. Sci Rep 11:2618. https://doi.org/10.1038/s41598-021-82311-3.
62. Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA, Knight R, Fierer N. 2011. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil Biol Biochem 43:1450–1455. https://doi.org/10.1016/j.soilbio.2011.03.012.
63. Janssen PH. 2006. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. Appl Environ Microbiol 72:1719–1728. https://doi.org/10.1128/AEM.72.3.1719-1728.2006.
64. Nam SJ, Hu WS, Koo OK. 2022. Evaluation of crAssphage as a human-specific microbial source-tracking marker in the Republic of Korea. Environ Monit Assess 194. https://doi.org/10.1007/s10661-022-09918-5.
65. Wu Z, Greaves J, Arp L, Stone D, Bibby K. 2020. Comparative fate of CrAssphage with culturable and molecular fecal pollution indicators during activated sludge wastewater treatment. Environ Int 136:105452. https://doi.org/10.1016/j.envint.2019.105452.
66. Honap TP, Sankaranarayanan K, Schnorr SL, Ozga AT, Warinner C, Lewis CM, Jr. 2020. Biogeographic study of human gut-associated crAssphage suggests impacts from industrialization and recent expansion. PLoS One 15:e0226930. https://doi.org/10.1371/journal.pone.0226930.
67. Koonin EV, Yutin N. 2020. The crAss-like phage group: how metagenomics reshaped the human virome. Trends Microbiol 28:349–359. https://doi.org/10.1016/j.tim.2020.01.010.
68. Holtz LR. 2020. Putting the virome on the map: the influence of host geography and ethnicity on the gut virome. Cell Host Microbe 28:636–637. https://doi.org/10.1016/j.chom.2020.10.007.
69. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. 2021. Massive expansion of human gut bacteriophage diversity. Cell 184:1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029.
70. Guerin E, Shkoporov AN, Stockdale SR, Comas JC, Khokhlova EV, Clooney AG, Daly KM, Draper LA, Stephens N, Scholz D, Ross RP, Hill C. 2021. Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. Microbiome 9:89. https://doi.org/10.1186/s40168-021-01036-7.
71. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834. https://doi.org/10.1101/gr.213959.116.
72. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7:e7359. https://doi.org/10.7717/peerj.7359.
73. Wu YW, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32:605–607. https://doi.org/10.1093/bioinformatics/btv638.
74. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol 3:836–843. https://doi.org/10.1038/s41564-018-0171-1.
75. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 11:2864–2868. https://doi.org/10.1038/ismej.2017.126.
76. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.
77. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu W-T, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Kyrpides NC, Schriml L, Garrity GM, Hugenholtz P, Sutton G, The Genome Standards Consortium., et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731. https://doi.org/10.1038/nbt.3893.
78. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36:1925–1927. https://doi.org/10.1093/bioinformatics/btz848.
79. Lee MD. 2019. GToTree: a user-friendly workflow for phylogenomics. Bioinformatics 35:4162–4164. https://doi.org/10.1093/bioinformatics/btz188.
80. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
81. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189.
82. Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293–W296. https://doi.org/10.1093/nar/gkab301.
83. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9.
84. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England) 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565.
85. Van Dongen SM. 2000. Graph clustering by flow simulation. PhD Thesis, University of Utrecht, The Netherlands.
86. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.

87. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR. 2015. HMMER web server: 2015 update. Nucleic Acids Res 43:W30–W38. https://doi.org/10.1093/nar/gkv397.

88. 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. https://doi.org/10.1186/1471-2105-11-119.

89. Wickham H. 2016. ggplot2-elegant graphics for data analysis. Springer International Publishing. Cham, Switzerland.

90. R Core Team. 2013. R: a language and environment for statistical computing. R Core Team. Vienna, Austria.

91. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300.

92. Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer. Bioinformatics 27:1009–1010. https://doi.org/10.1093/bioinformatics/btr039.