crAssphage abundance and genomic selective pressure correlate with altered bacterial abundance in the fecal microbiota of South African mother-infant dyads

Bryan P. Brown[a,b]#, Jerome Wendoh[c], Denis Chopera[d], Enock Havyarimana[c], Shameem Jaumdally[c], Donald D. Nyangahu[a,b], Clive M. Gray[c], Darren P. Martin[e], Arvind Varsani[f,g]#, and Heather B. Jaspan[a,b,c].

[a]Seattle Children's Research Institute, Seattle, Washington, USA

[b]Schools of Medicine and Public Health, University of WA, Seattle, WA, USA

[c]Department of Pathology, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

[d]Africa Health Research Institute, University of KwaZulu-Natal, Durban, South Africa

[e]Structural Biology Research Unit, Department of Integrative Biomedical Sciences, University of Cape Town, Observatory, Cape Town, South Africa

[f]The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life sciences, Arizona State University, Tempe, Arizona, USA

[g]Department of Integrative Biomedical Sciences, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town, South Africa

Running Head: Correlations between crAssphage and bacterial abundance

#Address correspondence to Arvind Varsani, Arvind.varsani@asu.edu or Bryan P Brown, bryan.brown@seattlechildrens.org.

24 Abstract word count: 241

25 Main text word count: 4,667

26

27 Keywords: bacteriophage, crAssphage, South Africa, transmission, microbiota, selection

28

Abstract

crAssphages are a class of bacteriophages that are highly abundant in the human gastrointestinal tract. Accordingly, crAssphage genomes have been identified in most human fecal viral metagenome studies. However, we currently have an incomplete understanding of factors impacting the transmission frequencies of these phages between mothers and infants, and the evolutionary pressures associated with such transmissions. Here, we use metagenome sequencing of stool-associated virus-like particles to identify the prevalence of crAssphage across ten South African mother-infant dyads that are discordant for HIV infection. We report the identification of a complete 97kb crAssphage genome, parts of which are detected at variable levels across each mother-infant dyad. We observed average nucleotide sequence identities of >99% for crAssphages from related mother-infant pairs but ~97% identities between crAssphages from unrelated mothers and infants: a finding strongly suggestive of vertical mother to infant transmission. We further analyzed patterns of nucleotide diversity across the crAssphage sequences described here, identifying particularly elevated positive selection in RNA polymerase and phage tail protein encoding genes, which we validated against a crAssphage genome from previous studies. Using 16S rRNA gene sequencing, we found that the relative abundances of *Bacteroides thetaiotaomicron* and *Parabacteroides merdae* (Order: Bacteroidales) were differentially correlated with crAssphage abundance. Together, our results reveal that crAssphages may be vertically transmitted from mothers to their infants and that hotspots of selection within crAssphage RNA polymerase and phage tail protein encoding genes are potentially mediated by interactions between crAssphages and their bacterial partners.

Importance

crAssphages are an ubiquitous member of the human gut microbiome and modulate interactions with key bacterial associates within the order Bacteroidales. However, the role of this interaction in the genomic evolution of crAssphage remains unclear. Across a longitudinally sampled cohort of ten South African mother-infant dyads, we use metagenome sequencing of the fecal virome and 16S rRNA gene sequencing of the fecal bacterial microbiota to elucidate the ecological and evolutionary dynamics of these interactions. Here, we demonstrate elevated levels of crAssphage average nucleotide identity between related mother-infant dyads as compared to unrelated individuals, suggesting vertical transmission. We report strong positive selection in crAssphage RNA polymerase and phage tail protein genes. Finally, we demonstrate that crAssphage abundance is linearly correlated ($P < 0.014$) with the abundance of two bacterial taxa, *Bacteroides thetaiotaomicron* and *Parabacteroides merdae.* These results suggest that phage-bacterial interactions may help shape ecological and evolutionary dynamics in the gut.

Introduction

The human gut virome is dominated by bacteriophages (1-3). In the last decade, a bacteriophage species' ~97kb circular DNA genome, commonly referred to as crAssphage (from cross assembly of 12 fecal metagenomes), was identified and found to be the most abundant virus in the human gut (2). crAssphage-like contigs have been *de novo* assembled from fecal samples from various regions of the world and across

health and disease states, emphasizing its high prevalence among human enteric microbiota. Among all the currently published complete and partial crAssphage genomes, irrespective of whether they have been isolated from environmental samples, humans or primates (4-6), there exists >90% sequence similarity between homologous genome regions (2, 7-9). Though the evolutionary relationships of crAssphages to other phage families remains unresolved (some evidence points toward a distant relationship to phages in the family *Podoviridae* (10)), comparative genomic analyses of human-associated crAssphage taxa have partitioned them into four candidate subfamilies composed of ten putative genera.

Despite the abundance of crAssphage-like contigs detected in human fecal microbiota, little is known about the lifestyle or selective pressures acting on crAssphage genomes. Based on analysis of CRISPR spacers and other genomic features, it was speculated that crAssphages prey upon members of *Bacteroides* (2). Intuitively, this fits well with the Bacteroidetes-dominated community profile of human gut microbiota. Shkoporov et al. (11) recently confirmed that members of the crAssphage group stably infect *Bacteroides intestinalis* and are able to maintain long-term persistence *in vivo,* though the mechanisms underlying this relationship are unknown. Furthermore, after long-term (23 days) interaction experiments between a crAssphage isolate (crAss001) and *B. intestinalis*, approximately half of isolated *B. intestinalis* colonies demonstrated complete resistance to the phage, indicating rapid evolution of bacterial interaction factors (11). However, it remains unclear how this relationship impacts selective pressures acting across the crAssphage genome.

Our knowledge of bacterial-phage coevolution in the human gut is largely

unexplored (12, 13). Antagonistic interactions between bacteria and phages are crucial determinants of genomic evolution for both partners, and several studies have described these trends across diverse systems and environments (14-17). In the gut environment, Minot et al. (12) described rapid nucleotide substitution rates (>10^{-5} per nucleotide per day) in the lytic phage family *Microviridae*. With the establishment of a bacterial host for crAssphage taxa, evolutionary studies of selection in these host cells are likely to reveal some of the genomic consequences of bacterial-phage relationships in the gut.

Here, we report the identification of a complete genome sequence of a crAssphage variant (M186D4) from a South African adult stool sample. To our knowledge, this is the first complete genome obtained from a single individual. Metagenomic sequencing of the viral microbiota of ten mother-infant dyads, that were differentially infected/exposed to HIV, yielded similar levels of crAssphage genomic coverage between mothers and related infants at one week postpartum, potentially indicating comparable crAssphage abundance between related mothers and infants and, therefore, possible vertical transmission. Across the genome of isolate M186D4, we identify selective “hot spots” of nucleotide variation including phage tail protein genes and, to a lesser extent, the RNA polymerase genes. These results were validated across the first published crAssphage genome, illustrating selective pressures acting on human associated crAssphages that potentially arise as a consequence of antagonistic interactions with bacterial consortia in the gastrointestinal tract.

Materials and Methods

121 *Sample Collection and virus-like particle isolation*

122 Stools were collected from 10 infants and their mothers from 4 days to 15 weeks post
123 vaginal delivery (Table 2) at a periurban clinic in Cape Town, South Africa. Stools were
124 transported on ice and stored at -80°C until nucleic acid extraction. Samples were
125 defrosted and approximately 0.5g of fecal sample was homogenized in 20ml SM buffer
126 as previously described (18) and centrifuged at 10,000 x g for 10 min. The resulting
127 supernatant was filtered sequentially through a 0.45µm and 0.2µm syringe filter. Filtered
128 supernatants were incubated with lysozyme and Turbo DNAse (Thermo Fisher
129 Scientific, USA) at 37°C for 1 hour to degrade nucleic acids not enclosed in virus-like
130 particles.

131

132 *Viral nucleic acid extraction and sequencing*

133 Viral nucleic acid was extracted from 200µl of the filtrate using the High Pure viral
134 nucleic acid kit (Roche Diagnostics, USA) following the standard protocol. Circular viral
135 DNA was amplified using rolling circle amplification (RCA) with Illustra TempliPhi 100
136 amplification kit (GE Healthcare, USA). The RCA products were used to prepare a 350
137 bp insert DNA library for each sample following the manufacturer's standard protocol.
138 Shotgun sequencing was performed on an Illumina HiSeq 2500 platform using 150bp
139 PE chemistry by Novogene (Hong Kong).

140

141 *Bacterial genomic DNA (gDNA) extraction and sequencing*

142 Bacterial gDNA was extracted from the same stool samples as described above. Stools
143 were incubated with 6µl mutanolysin (25kU/ml), 50µl lysozyme (450kU/ml), and 3µl

lysostaphin (4kU), and incubated at 37°C for 60 minutes with shaking. gDNA was then extracted using the MoBio Powersoil kit following the manufacturer's instructions. Libraries of the V6 region of 16S rDNA were prepared and generated as described previously (19). Individual libraries were purified using the QIAquick 96 PCR purification kit, quantitated with the Quant-iT dsDNA Broad Range assay, and pooled in equimolar quantities. Pooled libraries were visualized on an agarose gel, excised, and purified using the QIAquick Gel Extraction kit. Library QC and sequencing was performed by the Canadian Centre for Applied Genomics on an Illumina HiSeq 2000 using a 100bp PE approach.

Viral metagenome assembly and annotation

Raw sequencing reads were trimmed using Trimmomatic v0.36 (20) and then *de novo* assembled using spades v 3.12 (21). All *de novo* assembled contigs of >500 nucleotides (nt) were analyzed using a BLASTx search (22) against a local viral RefSeq protein database compiled from GenBank. An approximate ~97kb circular (identified by terminal redundancy) contig was identified that had similarities to the first published crAssphage genome (2). Open reading frames (ORFs) were identified using Glimmer (23, 24) and annotated using an in-house reference protein database compiled from NCBI's GenBank resource.

Bacterial 16S sequence processing

Forward and reverse indices and marker gene primers were removed using cutadapt (25). Trimmed reads were then quality filtered, dereplicated into amplicon sequence

variants (ASVs), and taxonomically classified using the dada2 package (26). Taxonomic classification of ASVs was performed using the Ribosomal Database Project's Naïve Bayesian classifier (27) against training set 16. Resulting sequence tables were imported into the Phyloseq (28) framework for further processing and analysis. ASV and sample filtering parameters were estimated and applied using custom R functions available at https://github.com/itsmisterbrown/microfiltR. A full vignette detailing our filtering strategy is available at the same location.

Compositional transformations and bacterial 16S sequence analysis

Filtered datasets were transformed into centered log ratio (CLR) coordinates and subset to taxa within Bacteroidales. Pearson correlations were performed between the CLR transformed abundance of each ASV and the genomic coverage of crAssphage for that sample. Taxa with absolute Pearson's $r > 0.4$ were included in downstream analyses.

To reduce compositional effects and infer relationships between crAssphage abundance and bacterial abundance, we generated isometric log ratio (ILR) balances of bacterial taxa that were identified as significantly different between samples with high and low crAssphage abundance. Wilcoxon rank sum tests were performed on the CLR abundance of each taxon to determine significance. The ILR transform is a compositional data analytical method that reduces compositional effects by translating relative abundance datasets into orthonormal coordinates suitable for standard statistical analyses (29). Balances enable highly interpretable results of ILR coordinates by using selected sets of taxa in the log ratios. ILR balances were generated using the following equation:

$$b_i = \sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}} log \frac{g_p(b_i^+)}{g_p(b_i^-)}$$

where n_i^+ and n_i^- indicate the set of taxa involved the numerator and denominator of that balance (b_i), respectively, and $n_i^{\pm} = \sum_{\theta_{ij=\pm 1}} p_j$ provides the weight integration (30). The first part of the listed balance equation, $\sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}}$, acts as a scaling factor that normalizes each balance to unit length, regardless of weighting scheme. The weighted geometric mean (30) of the taxa associated with a given balance is represented as $g_p(b_i^{\pm})$ and can be formalized as:

$$g_p(y) = exp\left(\frac{\sum_{i=1}^{D} p_i log y_i}{\sum_{i=1}^{D} p_i}\right)$$

Here, p_i represents the weight assigned to a taxon$_i$, which is either uniform or reflective of its abundance across the dataset. When uniform weights are applied, the geometric mean and isometric log ratio equations default to the original form (29, 31).

Variant analysis and statistics

Trimmed paired end reads were mapped to the closed M186D4 genome using the BWA with default parameters (32). Samtools mpileup utility (33) was used to calculated the per-nucleotide read coverage and variant frequency. Variants were detected and filtered using VarScan (34). Variants were only considered authentic if there were, at minimum, five high quality reads supporting the variant, which was the lowest threshold for significance in our dataset utilizing an α value of 0.05. P values were calculated using a Fisher's Exact test on read counts supporting the reference and variant alleles (34).

Functional effects of identified variants were predicted using SnpEff (35) against a custom annotation database generated as part of this study and annotated as nonsynonymous (N_d) or synonymous (S_d) substitutions. The total number of nonsynonymous (N) and synonymous (S) sites for each protein coding sequence was calculated using SnpGenie (36). The numbers of nonsynonymous (d_N) and synonymous (d_S) substitutions per site were estimated using the Jukes-Cantor formula:

$$dN = \frac{3}{4}\ln\left(1 - \frac{4p_N}{3}\right) \qquad dS = \frac{3}{4}\ln\left(1 - \frac{4p_S}{3}\right)$$

where p_N and p_S indicate the proportions of nonsynonymous and synonymous substitutions, respectively, and can be estimated by $p_N = \frac{N_d}{N}$ and $p_S = \frac{S_d}{S}$.

Genome wide average nucleotide identity (ANI) calculations were performed as described previously (37). Briefly, a sliding window algorithm was used to assess nucleotide identity in 1000bp fragments. Alignments were required to span 200bp and have a minimum identity of 70%. For incomplete genomes generated in this study, trimmed reads were recruited to the genome of isolate M186D4 using Bowtie2 (38) with "very sensitive local" parameters (not requiring end-to-end read alignment). Reads that aligned to genome M186D4 were then used for genome assembly with the spades assembler (39). Assembled contigs were used for ANI calculations for samples without closed genomes. P values for ANI comparisons were calculated from Wilcoxon rank sum tests due to variations in sample sizes between groups. Downstream statistical analysis and visualization was conducted using the base R framework and the ggplot2 package (40).

Validation of results against a previously published crAssphage genome

The vast majority of publicly available crAssphage genomes have been assembled from pooled samples of many individuals (2, 3, 8, 11), rendering those datasets mostly unsuitable for variant analysis. However, the individual datasets from which the original crAssphage genome sequence was cross assembled (3) were primarily composed of related mother-infant pairs with low interpersonal viral diversity, thus representing the most attractive option for cross validation. Raw nucleotide sequences were download from the NCBI Sequence Read Archive from Study SRP002523. The genome assembly and raw reads from Run SRR073438 were selected for variant analysis. Run SRR073438 was pooled DNA isolated from virus-like particles from a mother and twin pair, as well as an additional maternal sample. Genes and annotations were transferred from the associated published genome (GenBank accession #BK010471) using Prokka (41). Quality filtered reads from each run were aligned to the genome from which they were assembled using BWA (32). For run SRR073438, an α value of 0.05 was used to set read thresholds for variant significance because the vast majority of reads were isolated from a mother and related twins and the broader interpersonal diversity across the dataset was low (3).

Data availability

The annotated genome sequence for crAssphage M186D4 is available under GenBank accession number MK238400. Partial crAssphage genome sequences are available under NCBI BioProject PRJNA526942. The data, functions, and R script required to reproduce the 16S analyses used in this study are available at https://github.com/itsmisterbrown/crAssphage_M186D4_analyses

Results

Metagenome sequencing and assembly

From a 24 year old, HIV-infected female's stool sample, a closed circular genome of 97,757 kb (Figure 1) was identified as having high similarity (96.6% ANI; Figure 2A) to the first identified crAssphage (GenBank accession #BK010471) (2). The participant was one week postpartum, had a CD4 count of 265 cells/mm^3, and had initiated antiretroviral therapy during pregnancy. The genome was fully closed with 451-fold mean coverage from 140,984 150bp PE reads (Table 1, Figure 3B). We identified and annotated 82 ORFs and protein coding sequences. Of the 10 candidate genera proposed by Guerin et al (9), crAssphage M186D4 belonged in candidate genus I, with a genome-wide GC content of 29%.

Average nucleotide identity and persistence across mother-infant dyads

Read mapping to the crAssphage M186D4 genome yielded complete or partially complete crAssphage genome hits in all mother-infant dyads, regardless of HIV infection or exposure status (Table 2). The percent coverage of the genome varied across samples, ranging from 2.2% to 100%. Mother-infant dyads typically yielded consistent degrees of genome coverage, though this varied across time points (Table 2). When analyzing patterns of nucleotide diversity between related maternal and infant samples, high levels of ANI between related individuals were evident (~99%; Figure 2D), though only two mother-infant dyads had enough genomic coverage for robust assessment. Average nucleotide identities between unrelated individuals within our

277 cohort were similar (~97%) to levels between isolate M186D4 and previously described
278 crAssphage variants from Mexico (Figure 2B) (8), Malawi (Figure 2C) (42), and the
279 United States (3) (~96%). Longitudinal sampling and analysis of infant fecal microbiota
280 suggest that crAssphage infection persists through, at least, the first 15 weeks of life
281 (Table 2).
282
283 *Variant analysis*
284 We detected 40 significant single nucleotide polymorphisms (SNP) and 1
285 insertion/deletion in the genome of crAssphage M186D4. After filtering, 40 SNPs were
286 retained, of which 38 fell within coding regions (Table 1), leading to a variation rate of 1
287 variant per 2,443 bases. The nucleotide mutation profile was composed of 33 transitions
288 and 7 transversion, for a transition to transversion ratio of 4.71. Of the 38 variants that
289 occurred within coding regions, 78.9% (30) resulted in nonsynonymous substitutions
290 (N_d), 18.4% (7) resulted in synonymous (S_d) mutations, and only 2.6% (1) resulted in a
291 nonsense mutation. The only insertion fell within a coding sequence for a tail tubular
292 protein P22 (Table 1).
293 There was a nonuniform distribution of SNPs across the genome of M186D4.
294 Our results show an accumulation of mutations in two distinct genomic regions (Figure
295 3). These mutations accumulated in RNA polymerase genes that lie within genomic
296 regions from ~30,000-50,000bps and phage tail protein in genomic regions from
297 ~58,000-76,000bps. Averaged across the genome, the d_N/d_S ratio was 0.54. When
298 considering only RNA polymerase genes, the d_N/d_S ratio was 1.89. In phage tail
299 proteins, we report a further elevated d_N/d_S ratio of 2.66. The gene with the high SNP

count was a putative phage tail-collar protein (DUF3751), which had a total of seven SNPs. The variant with the highest frequency across the dataset was a T -> C transition occurring within a putative Bacteroidetes-associated carbohydrate-binding (BACON) domain containing protein (Figure S1). When removing genes encoding RNA polymerase subunits and phage tail proteins from consideration, the genome-wide d_N/d_S ratio fell to 0.34.

To ensure that these results were consistent across crAssphage taxa, we performed variant annotation and analysis on an additional crAssphage genome assembled from a cohort in the United States (3). As observed for crAssphage M186D4, we detected a nonuniform distribution of SNPs across the genome of crAssphage SRR073438, despite even sequencing depth (Figure S2). Consistent with observations from M186D4, SNPs persistently occurred in genes annotated as RNA polymerase and tail proteins, often resulting in nonsynonymous substitutions (Table S2). For crAssphage SRR073438, the genes with the highest SNP counts were a putative tail fiber protein and putative tail protein UGP073, both with three nonsynonymous substitutions (N_d) and zero synonymous substitutions (S_d, Table S2). Because all synonymous substitutions (S_d) occurring at significant levels were distributed across protein coding genes other than RNA polymerase and tail protein genes, we were not able to calculate d_S values for those genes. However, average d_N values were similar to those observed in tail protein genes in crAssphage M186D4 (SRR073438: 0.0017, M186D4: 0.0029). Similarly, for RNA polymerase genes, d_N values were comparable between both genomes (SRR073438: 0.0002, M186D4: 0.0008).

323 *Bacterial-crAssphage interactions*

324 We report differential shifts in bacterial abundance between samples with high and low
325 abundance (genomic coverage) of crAssphage. We chose to use genomic coverage,
326 rather than read count, to mitigate compositional effects and amplification bias
327 associated with rolling circle amplification. Though genomic coverage will not eliminate
328 bias, it remains relative to the crAssphage genome size, which appears to be
329 comparatively static across locations and studies (9). This is in contrast to relative
330 abundance and read counts, which are compositional in nature (43), and which are
331 further skewed by amplification biases (44). Eleven members of the order Bacteroidales
332 were found to have moderate or strong correlations (absolute Pearson's $r > 0.4$) with
333 crAssphage abundance (Figure 4a). Of these, strains of *Bacteroides thetaiotaomicron*
334 and *Parabacteroides merdae* displayed significantly different CLR abundances between
335 samples with high and low crAssphage abundance. *B. thetaiotaomicron* was
336 significantly elevated in samples with high crAssphage abundance and *P. merdae*
337 abundance was significantly reduced in samples with high crAssphage abundance. To
338 infer interactions between the abundance of these taxa and crAssphage abundance
339 beyond binary comparisons, we generated isometric log ratio balances between these
340 two taxa. We report that crAssphage abundance (fold coverage) is a significant
341 predictor ($P < 0.014$) of the abundance of *B. thetaiotaomicron* and *P. merdae*,
342 independent of HIV infection/exposure status (Figure 4b).

343

344 **Discussion**

345 As an ecological niche, the gut environment imposes a dynamic range of selective

pressures on the genomes of gut microbiota. We assembled and analyzed a circularized full genome of a human crAssphage with ~97% sequence similarity to that of the first described crAssphage genome [2]. We identified differential selective pressures along the genome, with hotspots of selection targeting groups of genes. We present evidence of substantial genetic heterogeneity and dynamic selective pressure across the genome of South African crAssphage M186D4. Many nucleotide polymorphisms occurred in protein coding regions and yielded a high ratio of substitution rates at nonsynonymous to synonymous sites (d_N/d_S) differentially across the genome, specifically implying strong positive selection at RNA polymerase genes and phage tail protein genes, and purifying selection genome wide. We detected complete or partially complete crAssphage genomes in all mother-infant dyads, regardless of HIV infection or exposure status. Longitudinal sampling and analysis of infant fecal microbiota suggest that crAssphage infection persists through, at least, the first 15 weeks of life. We analyzed patterns of nucleotide diversity between related maternal and infant samples, reporting high levels of average nucleotide identity between related individuals and levels of nucleotide similarity between unrelated individuals ~2-3% lower and comparable to levels from previously published crAssphage variants from elsewhere in the world. Further, we detected differential abundance of *B. thetaiotaomicron* and *P. merdae* between samples with high and low crAssphage abundance.

The genomic selective pressures that we detected were primarily focused on two regions undergoing relatively elevated positive selection. These regions, from 30-50kb and 58-73kb, are comprised of genes from various clusters of orthologous groups

(COGs), but the variation was localized into genes annotated as RNA polymerase subunits and phage tail proteins, respectively. Phage tail proteins are essential in mediating bacterial cell attachment and genome delivery. The putative protein with the highest SNP count and most variants (seven) is the phage tail fiber protein (DUF3751). This protein, while functionally uncharacterized, has been shown to mediate antagonistic interactions with bacterial taxa (45) and is a conserved element in phage-derived bacterial tailocin complexes (45). Tailocins are bacterial protein complexes co-opted from bacteriophage that are morphologically and functionally similar to phage tail proteins and are critical to eukaryotic and bacterial cell binding (46, 47). Furthermore, T4 tail adhesion protein, which is homologous to DUF3751, mediates adsorption of T4-like bacteriophages to *Escherichia coli* and contains a lipopolysaccharide (LPS) binding site. T4 tail protein gp12 has been shown to modulate host inflammatory responses to LPS *in vivo* (48).

Though it is unclear how this domain physically interacts with bacterial cell receptors, positive selection along this gene and other tail proteins may represent a Red Queen scenario where the phage tail protein is adapting to constantly evolving target cell surface proteins. Adaptive evolution of phage tail proteins has been well documented in marine bacteriophage communities (49) and has been shown to facilitate an expanded host range. Additionally, long-term (23 day) phage/*B. intestinalis* co-cultivation experiments have shown that founder strains of crAssphage have severely limited ability to infect bacterial strains passaged for the length of the experiment (11), indicating rapid, likely antagonistic, coevolution between crAssphage and bacterial host strains. This reduction in infection rate likely reflects nonsynonymous

changes in bacterial cell surface receptors that are mediated by interaction with crAssphage tail proteins.

We detected genomic regions of crAssphages in all mother-infant pairs and samples sequenced, regardless of HIV infection/exposure status or time point sampled (Table 2). Interestingly, the percent of genome coverage from read alignment to isolate M186D4 was relatively consistent within a mother-infant pair. Though this observation is not a truly quantitative assay of viral abundance, consistent coverage of the crAssphage genome between related mother and infant samples at one week postpartum is suggestive of comparable abundance of crAssphage taxa, potentially from vertical transmission. Importantly, this differs from findings in infants from Missouri, USA, where crAssphage-like sequences were only identified at 24 months and not in earlier samples (50). We identify crAssphage sequences in feces as early as 4 days of age. We also note that genomic coverage of crAssphage in the same infant varied across time, with differential coverage during later sample points.

At the nucleotide level, our results support vertical transmission of crAssphage from mothers to infants. We report elevated average nucleotide identities (~99%) shared between genomes from related mothers and infants, as compared to unrelated pairs (~97%). This trend was consistent between maternal and infant samples collected at 4 days postpartum, but also held true for infant samples collected at week 4. Whether this is due to continued viral transmission via breastfeeding or other sources remains unclear. Average nucleotide identities between unrelated individuals within our cohort were similar to, though slightly higher than, levels between the genomes of isolate M186D4 and the three previously described global strains with complete genome

sequences. Overall, our data are suggestive of maternal transmission of crAssphage to infants, though further investigation is needed to accurately quantitate viral load and patterns of nucleotide diversity between mothers and infants, as well as over longitudinally collected samples starting at birth to assay viral dynamics.

With regard to the gut bacterial cohort, we report altered bacterial dynamics between samples with high and low crAssphage abundance. Previous studies have demonstrated that members of the Bacteroidales are preyed upon by crAssphage taxa (2, 11), though the dynamics of this relationship are unclear. In this cohort, we report significantly altered abundance of *B. thetaiotaomicron* and *P. merdae* in samples with high versus low crAssphage abundance, independent of HIV infection/exposure status. Using compositional data analytical approaches (isometric log ratio balances), we report that crAssphage abundance is a significant predictor of the log ratio abundance of these taxa (Figure 4b). *B. thetaiotaomicron* has been previously hypothesized as a potential host of crAssphage taxa (2, 9), though our data are the first to demonstrate this positive correlation. *P. merdae* belongs to the recently reassigned genus, *Parabacteroides,* which is sister to *Bacteroides* and a persistent member of the gastrointestinal tract (51, 52). The opposite shifts in abundance between these two taxa may represent unique relationships with crAssphage taxa (e.g. non-disruptive proliferation in *B. thetaiotaomicron* as reported for *B. intestinalis* (11), or comparably elevated lytic activity in *P. merdae*) or shared ecological niches in the gastrointestinal tract. In scenarios where crAssphage and *B. thetaiotaomicron* are elevated, *P. merdae* may decrease in abundance due to lack of resource access or other antagonistic bacterial interactions, though further studies dissecting this relationship are needed.

438 In summary, our data argue for vertical transmission of crAssphage across
439 mother-infant dyads, potentially as a component of the inherited microbiome. Our
440 results suggest genome-wide purifying selection in crAssphage, with episodic shifts of
441 strong positive selection within RNA polymerase and phage tail protein genes. Elevated
442 selective pressure of phage tail proteins may be due to antagonistic coevolution
443 between crAssphage and bacterial targets, though further work is needed to
444 demonstrate this conclusively. Additionally, our data suggest that crAssphage
445 abundance may direct gut bacterial dynamics, providing an ecological basis for the
446 associated genomic consequences. Collectively, our results argue that crAssphage
447 alters bacterial consortia and may act as a driver of ecological and evolutionary
448 dynamics in the gut.
449

450 **Acknowledgements**

451 This study was funded in part by the University of Washington Center for AIDS
452 Research, an NIH funded program under award number AI027757, supported by the
453 following NIH Institutes and Centers (NIAID, NCI, NIMH, NIDA, NICHD, NHLBI, NIA,
454 NIGMS, NIDDK). The InFANT study cohort was supported in part by the Canadian
455 Institutes of Health Research HIV Vaccine Initiative grant (#01044-000), NIH
456 R01AI131302 and AI120714-01A1. We would like to thank the InFANT study team for
457 collecting samples. We also thank all participants in the study for providing samples. DC
458 is supported by a Wellcome Trust DELTAS Africa grant to the SANTHE programme
459 (grant #107752/Z/15/Z).
460

Contributions

HBJ, DC, AV and DM designed the study. JW and DDN prepared bacterial 16S rDNA libraries. DC, EH, SJ prepared the samples for sequencing. BPB and AV analyzed the data. HBJ, BPB, and AV wrote the manuscript. All authors reviewed and edited the manuscript.

References

1. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD. 2011. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21:1616-25.
2. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GG, 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.
3. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466:334-8.
4. Garcia-Aljaro C, Balleste E, Muniesa M, Jofre J. 2017. Determination of crAssphage in water samples and applicability for tracking human faecal pollution. Microb Biotechnol 10:1775-1780.
5. Cinek O, Mazankova K, Kramna L, Odeh R, Alassaf A, Ibekwe MU, Ahmadov G, Mekki H, Abdullah MA, Elmahi BME, Hyoty H, Rainetova P. 2018. Quantitative

CrAssphage real-time PCR assay derived from data of multiple geographically distant populations. J Med Virol 90:767-771.

6. Holmfeldt K, Solonenko N, Shah M, Corrier K, Riemann L, Verberkmoes NC, Sullivan MB. 2013. Twelve previously unknown phage genera are ubiquitous in global oceans. Proc Natl Acad Sci U S A 110:12798-803.

7. McCann A, Ryan FJ, Stockdale SR, Dalmasso M, Blake T, Ryan CA, Stanton C, Mills S, Ross PR, Hill C. 2018. Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. PeerJ 6:e4694.

8. Cervantes-Echeverria M, Equihua-Medina E, Cornejo-Granados F, Hernandez-Reyna A, Sanchez F, Lopez-Contreras BE, Canizales-Quinteros S, Ochoa-Leyva A. 2018. Whole-genome of Mexican-crAssphage isolated from the human gut microbiome. BMC Res Notes 11:902.

9. 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.

10. 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. Nature microbiology 3:38.

11. 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. Nature communications 9:4781.

12. Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. 2013. Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A 110:12450-5.

13. Scanlan PD. 2017. Bacteria-Bacteriophage Coevolution in the Human Gut: Implications for Microbial Diversity and Functionality. Trends Microbiol 25:614-623.

14. Brockhurst MA, Morgan AD, Fenton A, Buckling A. 2007. Experimental coevolution with bacteria and phage: the pseudomonas fluorescens—Φ2 model system. Infection, Genetics and Evolution 7:547-552.

15. Gomez P, Buckling A. 2011. Bacteria-phage antagonistic coevolution in soil. Science 332:106-9.

16. Martiny JB, Riemann L, Marston MF, Middelboe M. 2014. Antagonistic coevolution of marine planktonic viruses and their hosts. Annual review of marine science 6:393-414.

17. Schwartz DA, Lindell D. 2017. Genetic hurdles limit the arms race between Prochlorococcus and the T7-like podoviruses infecting them. The ISME journal 11:1836.

18. Kraberger S, Waits K, Ivan J, Newkirk E, VandeWoude S, Varsani A. 2018. Identification of circular single-stranded DNA viruses in faecal samples of Canada lynx (Lynx canadensis), moose (Alces alces) and snowshoe hare (Lepus americanus) inhabiting the Colorado San Juan Mountains. Infect Genet Evol 64:1-8.

19. Arthur JC, Perez-Chanona E, Muhlbauer M, Tomkovich S, Uronis JM, Fan TJ, Campbell BJ, Abujamel T, Dogan B, Rogers AB, Rhodes JM, Stintzi A, Simpson

529 KW, Hansen JJ, Keku TO, Fodor AA, Jobin C. 2012. Intestinal inflammation
530 targets cancer-inducing activity of the microbiota. Science 338:120-3.
531 20. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for
532 Illumina sequence data. Bioinformatics 30:2114-20.
533 21. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new
534 versatile metagenomic assembler. Genome Res 27:824-834.
535 22. Altschul SF, Lipman DJ. 1990. Protein database searches for multiple
536 alignments. Proc Natl Acad Sci U S A 87:5509-13.
537 23. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S,
538 Cooper A, Markowitz S, Duran C. 2012. Geneious Basic: an integrated and
539 extendable desktop software platform for the organization and analysis of
540 sequence data. Bioinformatics 28:1647-1649.
541 24. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial
542 gene identification with GLIMMER. Nucleic Acids Res 27:4636-41.
543 25. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput
544 sequencing reads. EMBnet journal 17:pp. 10-12.
545 26. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. 2016.
546 DADA2: High-resolution sample inference from Illumina amplicon data. Nat
547 Methods 13:581-3.
548 27. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for
549 rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied
550 and environmental microbiology 73:5261-5267.

28. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217.

29. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C. 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology 35:279-300.

30. Egozcue JJ, Pawlowsky-Glahn V. 2016. Changing the reference measure in the simplex and its weighting effects. Austrian Journal of Statistics 45:25-44.

31. Egozcue JJ, Pawlowsky-Glahn V. 2005. Groups of parts and their balances in compositional data analysis. Mathematical Geology 37:795-828.

32. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-60.

33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-9.

34. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25:2283-2285.

35. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6:80-92.

573 36. Nelson CW, Moncla LH, Hughes AL. 2015. SNPGenie: estimating evolutionary
574 parameters to detect natural selection using pooled next-generation sequencing
575 data. Bioinformatics 31:3709-11.
576 37. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje
577 JM. 2007. DNA–DNA hybridization values and their relationship to whole-
578 genome sequence similarities. International journal of systematic and
579 evolutionary microbiology 57:81-91.
580 38. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat
581 Methods 9:357-9.
582 39. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
583 Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler
584 G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly
585 algorithm and its applications to single-cell sequencing. J Comput Biol 19:455-77.
586 40. Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer.
587 41. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics
588 30:2068-9.
589 42. Reyes A, Blanton LV, Cao S, Zhao G, Manary M, Trehan I, Smith MI, Wang D,
590 Virgin HW, Rohwer F, Gordon JI. 2015. Gut DNA viromes of Malawian twins
591 discordant for severe acute malnutrition. Proc Natl Acad Sci U S A 112:11941-6.
592 43. Li H. 2015. Microbiome, metagenomics, and high-dimensional compositional
593 data analysis. Annual Review of Statistics and Its Application 2:73-94.

44. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A. 2019. Minimum information about an uncultivated virus genome (MIUVIG). Nature biotechnology 37:29.

45. Ghequire MG, Dillen Y, Lambrichts I, Proost P, Wattiez R, De Mot R. 2015. Different Ancestries of R Tailocins in Rhizospheric Pseudomonas Isolates. Genome Biol Evol 7:2810-28.

46. Hockett KL, Renner T, Baltrus DA. 2015. Independent co-option of a tailed bacteriophage into a killing complex in Pseudomonas. MBio 6:e00452-15.

47. Ghequire MGK, De Mot R. 2015. The Tailocin Tale: Peeling off Phage Tails. Trends Microbiol 23:587-590.

48. Miernikiewicz P, Klopot A, Soluch R, Szkuta P, Keska W, Hodyra-Stefaniak K, Konopka A, Nowak M, Lecion D, Kazmierczak Z, Majewska J, Harhala M, Gorski A, Dabrowska K. 2016. T4 Phage Tail Adhesin Gp12 Counteracts LPS-Induced Inflammation In Vivo. Front Microbiol 7:1112.

49. Angly F, Youle M, Nosrat B, Srinagesh S, Rodriguez-Brito B, McNairnie P, Deyanat-Yazdi G, Breitbart M, Rohwer F. 2009. Genomic analysis of multiple Roseophage SIO1 strains. Environ Microbiol 11:2863-73.

50. Lim ES, Zhou Y, Zhao G, Bauer IK, Droit L, Ndao IM, Warner BB, Tarr PI, Wang D, Holtz LR. 2015. Early life dynamics of the human gut virome and bacterial microbiome in infants. Nat Med 21:1228-34.

51. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC, Henrissat B, Coutinho PM, Minx P, Latreille P, Cordum H, Van Brunt A, Kim K, Fulton RS,

Fulton LA, Clifton SW, Wilson RK, Knight RD, Gordon JI. 2007. Evolution of symbiotic bacteria in the distal human intestine. PLoS Biol 5:e156.

52. Asnicar F, Manara S, Zolfo M, Truong DT, Scholz M, Armanini F, Ferretti P, Gorfer V, Pedrotti A, Tett A, Segata N. 2017. Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. mSystems 2.

Figure legends

Figure 1. The 97,757bp genome of crAssphage M186D4. CDS are located on the outermost circle and are colored blue. BLAST alignments of isolate M186D4 to three additional crAssphage genomes are colored red green and blue and their GenBank accession number is listed in the legend. GC content is displayed in black. GC skew is displayed in purple and green.

Figure 2. Average nucleotide identity of crAssphage genomes is higher between related mother-infant dyads that between unrelated individuals. A. Histogram of the average nucleotide identity of 1,000bp fragments of crAssphage genomes from Missouri (2, 3) and this study. **B**. Histogram of the average nucleotide identity of 1,000bp fragments of crAssphage genomes from Malawi (42) and this study. **C**. Histogram of the average nucleotide identity of 1,000bp fragments of crAssphage genomes from Mexico (8) and this study. **D**. Boxplots of average nucleotide identity between related mother-infant dyads sequenced in this study (Within), unrelated individuals sequenced in this study (Between), and genomes from previous studies

(Global). Blue lines indicate mean average nucleotide identity. P values: NS > 0.05, * < 0.05, ** < 0.01.

Figure 3. Selective pressures are variable across the genome of crAssphage M186D4. A. The distribution of variants per kilobase across the genome. The region encoding RNA polymerase genes is colored teal. The region encoding phage tail proteins is colored red. dN/dS ratios are listed in each respective region. B. The fold coverage of quality filtered reads across the genome of crAssphage M186D4.

Figure 4. crAssphage abundance (coverage) predicts abundance of select Bacteroidales taxa. A. Centered log ratio (CLR) abundance of Bacteroidales taxa correlated with crAssphage abundance. CLR abundance of *P. merdae* is significantly decreased in samples with greater than 50% crAssphage coverage (red) as compared to samples with less than 50% crAssphage coverage (blue). CLR abundance of *B. thetaiotaomicron* is significantly increased in samples with greater than 50% crAssphage coverage. **B**. crAssphage coverage is a significant linear predictor of the isometric log ratio balance of *B. thetaiotaomicron* to *P. merdae,* independent of HIV infection/exposure status. HIV infected/exposed (green), HIV uninfected/unexposed (yellow), P values $\leq$ 0.05 are indicated with an asterisk (*).

Table 1. Functional annotation of variants in crAssphage M186D4. N_d, Nonsynonymous mutations; S_d, synonymous mutations; N, nonsynonymous sites; S, synonymous sites

662 **Table 2. crAssphage genomic coverage across all dyads sequenced in this study.**

663 Infected/exposed dyads are separated from uninfected/unexposed by the dashed line.

664

665 Figure S1. Variant frequency across the genome of crAssphage isolate M186D4. Each

666 significant variant is indicated by a point. Lines are added for visual aid only.

667

668 Figure S2. The fold coverage of quality filtered reads across the genome of crAssphage

669 SRR073438.

670

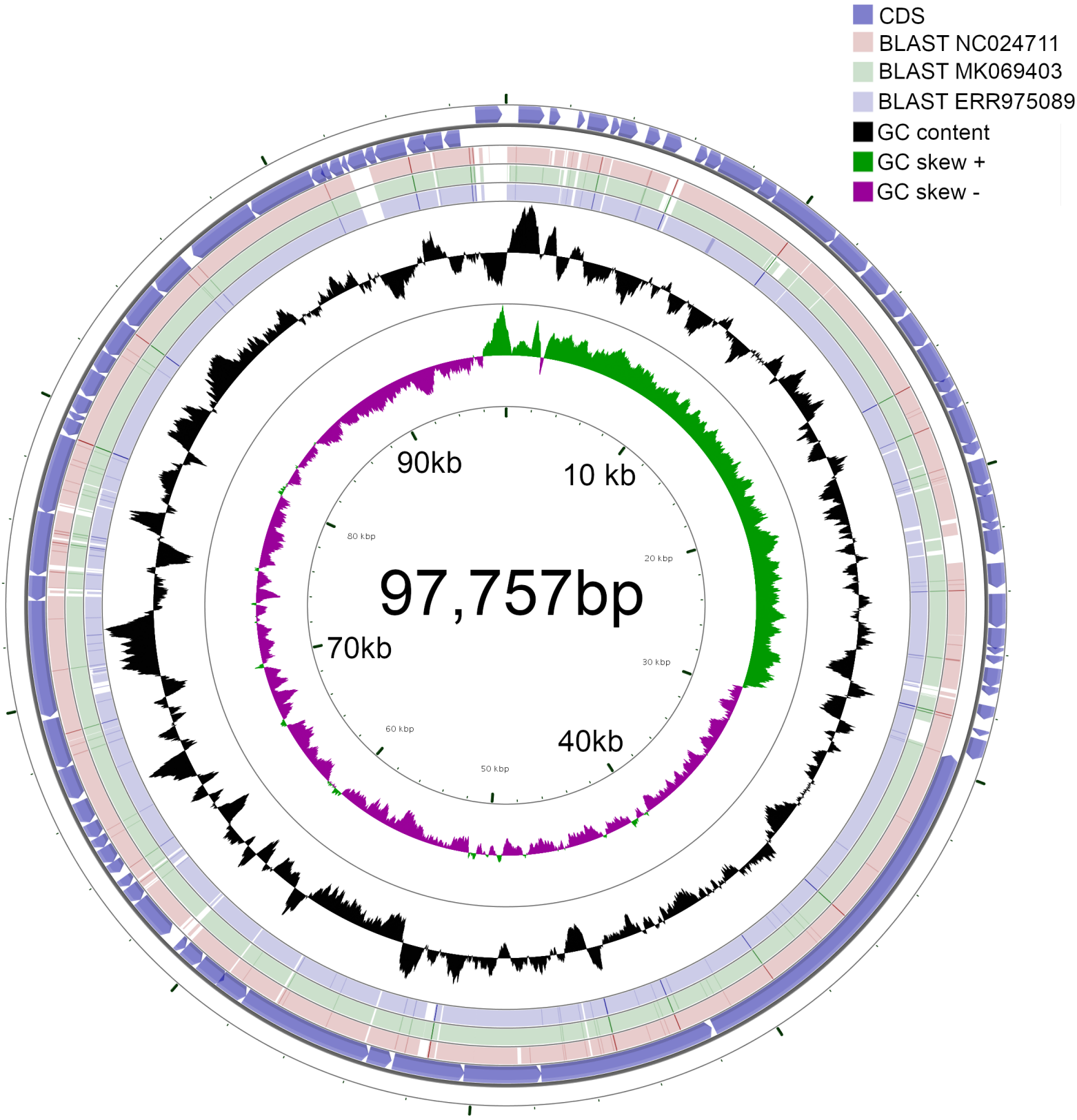
671 Table S1. Single nucleotide variants across the genome of crAssphage M186D4.

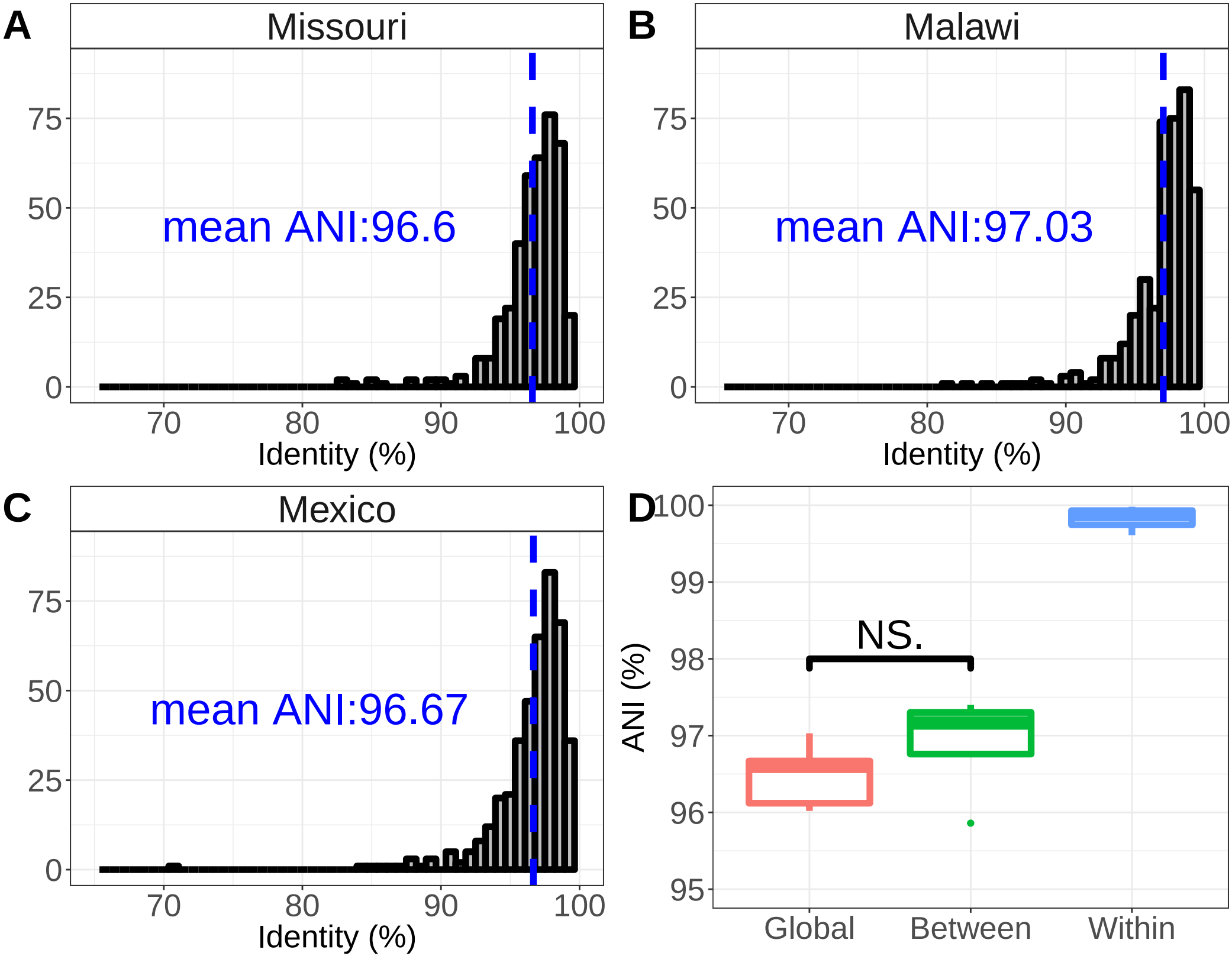
672

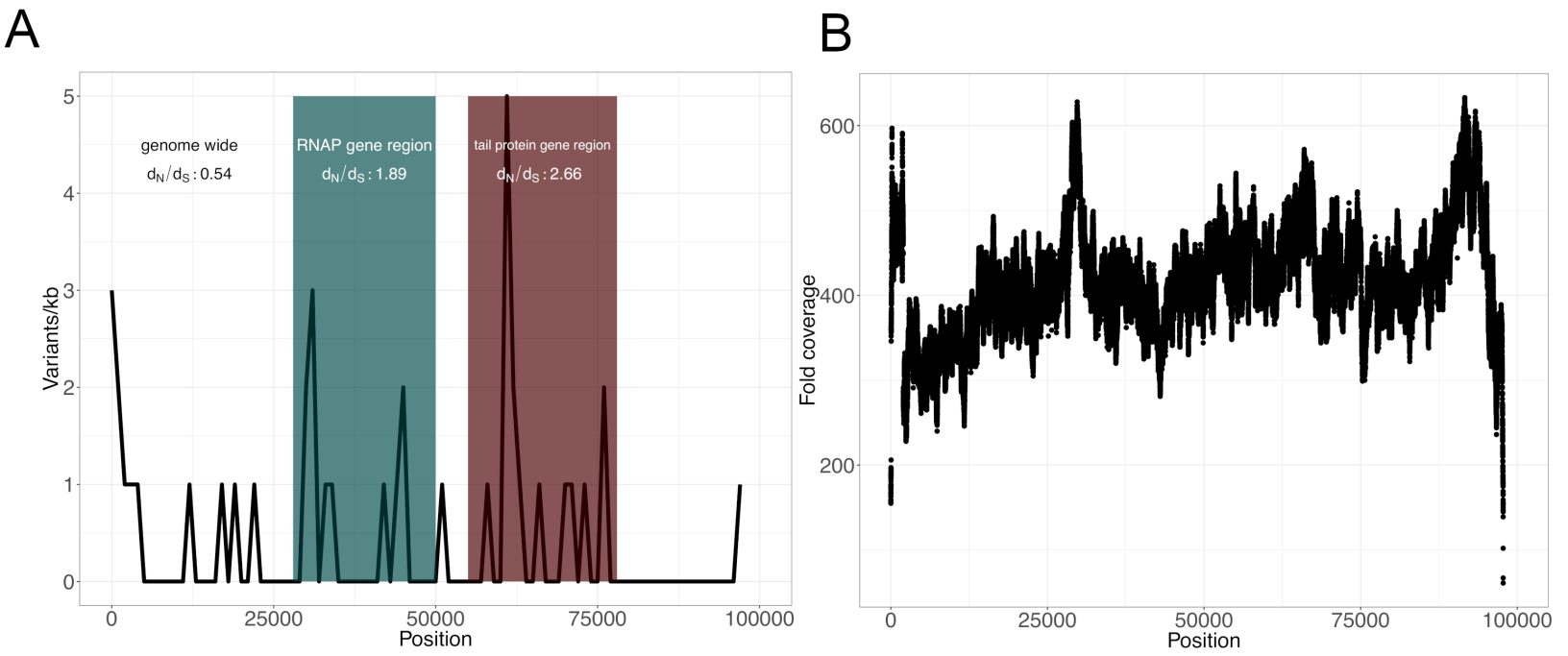
673 Table S2. Functional annotation of variants in crAssphage SRR073438

674

675




A
Missouri
mean ANI:96.6
75
50
25
0
70
80
90
100
Identity (%)
B
Malawi
mean ANI:97.03
75
50
25
0
70
80
90
100
Identity (%)
C
Mexico
mean ANI:96.67
75
50
25
0
70
80
90
100
Identity (%)
D
ANI (%)
100
99
98
97
96
95
NS.
Global
Between
Within



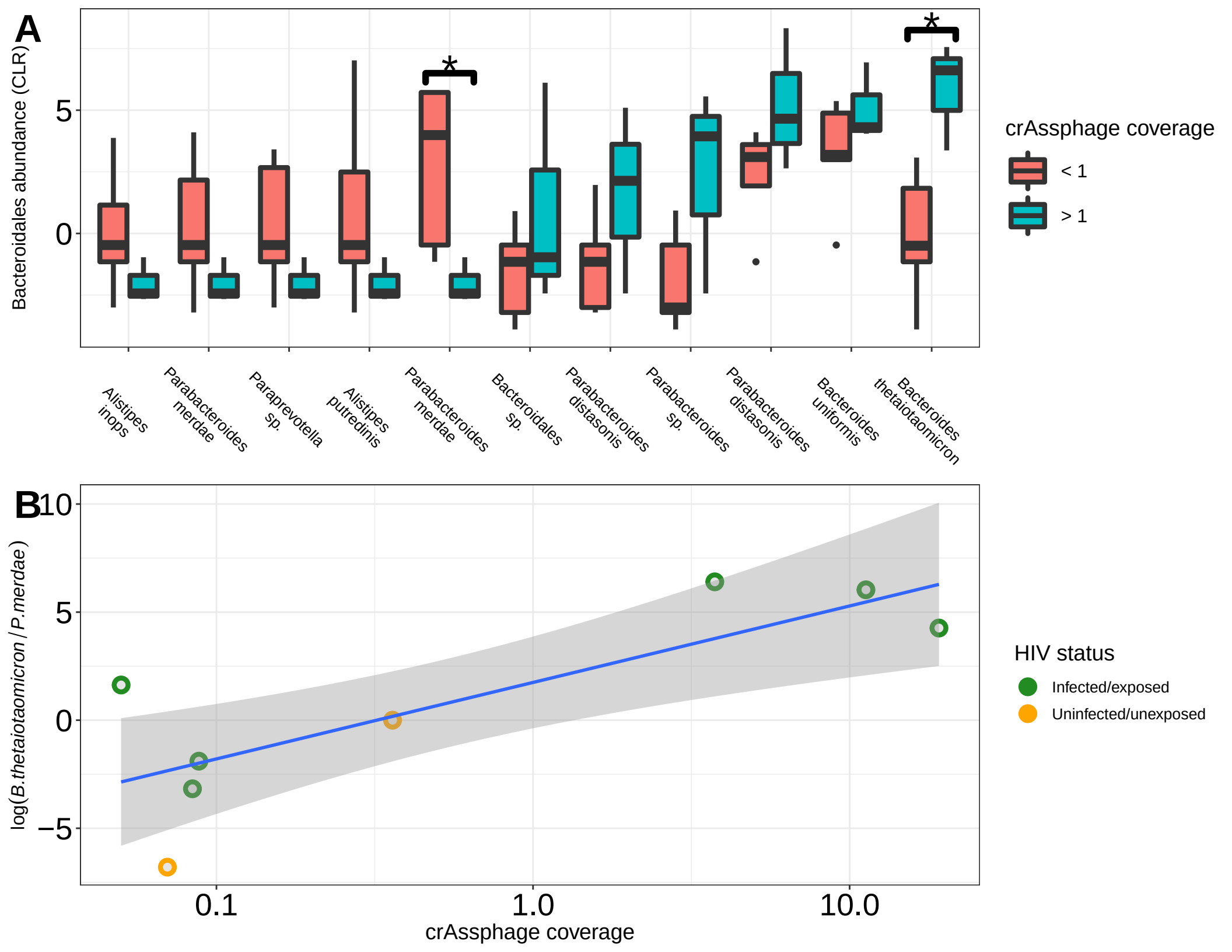

A
Bacteroidales abundance (CLR)
5
0
*
*
Alistipes inops
Parabacteroides merdae
Paraprevotella sp.
Alistipes putredinis
Parabacteroides merdae
Bacteroidales sp.
Parabacteroides distasonis
Parabacteroides sp.
Parabacteroides distasonis
Bacteroides uniformis
Bacteroides thetaiotaomicron
crAssphage coverage
< 1
> 1
B
log(B.thetaiotaomicron / P.merdae)
10
5
0
−5
0.1
1.0
10.0
crAssphage coverage
HIV status
Infected/exposed
Uninfected/unexposed

Table 1. Functional annotation of variants in crAssphage M186D4.

Gene	Start	Stop	Length	Direction	N_d	Nonsense mutations	S_d	Insertion	N	S
putative ssb single stranded DNA-binding protein	388	1215	827	forward	3	0	1	0	631.5	193.5
hypothetical protein	1381	1725	345	forward	2	0	3	0	267.8333	74.16667
putative SWI2/SNF2 ATPase 252C non-canonical Walker A motif	19772	20590	819	forward	1	0	0	0	645	171
putative deoxynucleoside monophosphate kinase	20936	21811	876	forward	1	0	0	0	684	189
coil containing protein	22045	22827	783	forward	0	0	1	0	619.6667	160.3333
putative RNAP catalytic subunit fused to unknownRNAP subunit	29489	41764	12276	reverse	5	0	2	0	8090.833	2295.167
putative RNAP associated protein fused to zincinprotease	41832	47720	5889	reverse	4	0	0	0	3900.833	1085.167
putative tail tubular protein P22 gp4	58024	59040	1016	reverse	0	0	1	1	683.6667	174.3333
putative phage tail fiber protein (DUF3751)	61309	62769	1461	reverse	7	0	0	0	984.3	260.67
putative tail protein	65848	66618	770	reverse	2	0	0	0	498.3333	137.6667
putative Bacon (Bacteroidetes-Associated Carbohydrate-binding) domain containing protein	69585	73469	3885	reverse	2	0	0	0	2600.5	780.5
putative tail protein	73473	74294	821	reverse	1	0	0	0	528.6667	155.3333
putative tail fiber protein	74320	76413	2093	reverse	2	0	0	0	1367.833	402.1667
putative plasmid replication initiation protein RepL	96778	97632	855	forward	0	1	0	0	678.1667	173.8333
Total					30	1	8	1		

N_d, Nonsynonymous mutations; S_d, synonymous mutations; N, nonsynonymous sites; S, synonymous sites

Table 2. crAssphage genomic coverage across all dyads sequenced in this study.

Dyad	Individual	HIV status	Time point	Coverage (fold)	Coverage (%)	Bases covered	Plus reads	Minus reads	Read GC (%)
186	Mother 1	Infected	Week 1	451.6449	100	97757	140984	141121	29.39
	Infant 1	Exposed	Week 1	0.0858	7.3376	7173	27	29	28.62
	Infant 1	Exposed	Week 4	3.7521	73.5814	71931	1026	1032	29.75
	Infant 1	Exposed	Week 15	0.221	15.5365	15188	72	72	29.39
197	Infant 2	Exposed	Week 1	0.0427	4.0181	3928	15	13	27.38
	Infant 2	Exposed	Week 15	0.0737	6.2799	6139	24	24	29.1
519	Mother 5	Infected	Week 1	0.0537	4.5787	4476	17	18	26.97
	Infant 5	Exposed	Week 1	0.0614	5.3868	5266	21	19	28.94
521	Mother 6	Infected	Week 1	19.1483	99.1295	96906	4371	4438	29.44
	Infant 6	Exposed	Week 1	11.2559	81.3947	79569	2546	2553	29.17
526	Mother 7	Infected	Week 1	0.0875	6.5049	6359	33	24	29.81
	Infant 7	Exposed	Week 1	0.0844	6.9939	6837	26	29	30.73
389	Infant 3	Unexposed	Week 1	0.0446	3.8166	3731	15	14	28.87
	Infant 3	Unexposed	Week 4	0.0875	7.5677	7398	29	28	28.93
	Infant 3	Unexposed	Week 15	4.2816	75.08	73396	1053	1059	29.82
395	Infant 4	Unexposed	Week 4	1.6697	49.3018	48196	536	545	32.54
701	Mother 8	Uninfected	Week 1	0.0322	2.8254	2762	11	10	28.76
	Infant 8	Unexposed	Week 15	0.0246	2.2986	2247	9	7	28.63
703	Mother 9	Uninfected	Week 1	0.0275	2.541	2484	9	9	31.26
	Infant 9	Unexposed	Week 1	0.0389	3.3542	3279	13	13	28.54
	Infant 9	Unexposed	Week 4	0.0821	7.0205	6863	24	29	28.7
720	Mother 10	Uninfected	Week 1	0.0706	5.9065	5774	23	23	29.87
	Infant 10	Unexposed	Week 1	0.3564	27.2963	26684	54	54	29.68