Title

Biogeographic study of human gut associated crAssphage suggests impacts from industrialization and recent expansion.

Author list

Tanvi P. Honap[a,b†], Krithivasan Sankaranarayanan[a,c†], Stephanie L. Schnorr[a,d], Andrew T. Ozga[a,b, e], Christina Warinner[a,b], Cecil M. Lewis Jr.,[a,b]*

[a]Laboratories of Molecular Anthropology and Microbiome Research, University of Oklahoma, 101 David L. Boren Blvd., Norman, Oklahoma 73019, USA,

[b]Department of Anthropology, University of Oklahoma, 455 W. Lindsey St., Norman, Oklahoma 73019, USA,

[c]Department of Microbiology and Plant Biology, University of Oklahoma, 730 Van Vleet Oval, Norman, Oklahoma 73019, USA,

[d]Konrad Lorenz Institute for Evolution and Cognition Research, Martinstrasse 12, 3400 Klosterneuburg, Austria,

[e]Center for Evolution and Medicine, Arizona State University, 427 E. Tyler Mall, Tempe, Arizona 85287, USA

Email addresses

tanvi.honap@ou.edu, ksankar@ou.edu, stephanie.schnorr@kli.ac.at, aozga@asu.edu, christina.warinner@ou.edu, cmlewis@ou.edu

***Corresponding Author**

Cecil M. Lewis Jr.

Department of Anthropology

University of Oklahoma

101 David L. Boren Blvd.

Norman, OK 73019

cmlewis@ou.edu

[†]These authors contributed equally to this work.

Abstract

Background: CrAssphage (cross-assembly phage) is a bacteriophage that was first discovered in human gut metagenomic data. CrAssphage belongs to a diverse family of crAss-like bacteriophages thought to infect gut commensal bacteria belonging to *Bacteroides* species. However, not much is known about the biogeography of crAssphage and whether certain strains are associated with specific human populations.

Results: Gut metagenomic data for a total of 3,263 samples from ten studies were *de novo* assembled. 755 samples contained open reading frames (ORFs) with matches to >=50% of crAssphage reference proteins and were considered positive for crAssphage. For these studies, we find that crAssphage prevalence is low in traditional, hunter-gatherer populations as compared to industrialized, urban populations. Statistical comparisons showed no association of crAssphage prevalence with host variables such

as age, sex, body mass index, and health status. Furthermore, we show that crAssphage sequences from the same individual inferred over multiple timepoints are identical suggesting persistence of crAssphage within individuals. A subset of 366 samples showing presence of single crAssphage viral genomes was used for phylogenetic analyses; however, these viral sequence types showed a lack of biogeographic structuring. A clustering approach identified nine and 33 crAssphage sequence clusters at the 97% and 98% identity thresholds, respectively. At the 98% identity threshold, two major clusters were observed – the first is restricted to Europe, the Middle East, and the Americas, and the second is predominantly present in Asia. We hypothesize that this structuring is suggestive of a recent expansion event within a major crAssphage clade.

Conclusion: We determined that crAssphage presence is associated with an industrialized lifestyle/diet and is not associated with host variables including health status. CrAssphage exhibits limited phylogeographic structure, and its major clade likely underwent a recent global expansion.

Keywords

crAssphage; human virome; metagenomics; bacteriophage

Background

The virome harbors the most abundant and diverse set of genes on earth, with deep impacts on host biology, including the very architecture of host genomes (1). Often described as the “dark matter” of biology (2), the virome has an underexplored role in

human health (3). Much of the human virome is composed of DNA bacteriophages (4), which are argued to be the primary regulators of bacterial biomass (5). Among their known impacts within the human gut, bacteriophages are involved in the transmission of bacterial antibiotic resistance genes, and they facilitate bacterial carbohydrate utilization (4). Yet, very little is known about bacteriophage diversity in humans (6). While the human virome is a component of the human microbiome, it is not as amendable to high throughput taxonomic characterization as the archaeal and bacterial subsets. Specifically, viruses lack “universal” taxonomically diagnostic marker genes (for example, the 16S rRNA gene in archaea and bacteria), preventing characterization using current amplicon-based sequencing strategies.

Recent advances in shotgun metagenome sequencing and development of data mining tools now allow for large scale studies of DNA bacteriophages within the human microbiome. From such data, one particular DNA bacteriophage, crAssphage, appears to be remarkably abundant in the human gut (7). CrAssphage (cross-assembly phage) was named after the crAss (cross-assembly) software originally used to discover the bacteriophage (7). CrAssphage is prevalent in available metagenomic data from the U.S. (8), comprising up to 90% of viral particle-derived metagenomes and up to 22% of reads in a total fecal community from the U.S. National Institutes of Health’s Human Microbiome Project cohort (9). Moreover, crAssphage is up to six times more abundant than all other known phages that could be reconstructed from these publicly available metagenomes (7). CrAssphage is a member of what has been subsequently described as a diverse family of crAss-like bacteriophages (10).

Across metagenomic studies, crAssphage is nearly exclusively associated with the gut microbiome (7, 10), and has been proposed as a potential biomarker for fecal contamination (11-13). While crAssphage is characteristic of the adult human gut, information regarding its initial acquisition (13), geographic distribution (14), and prevalence in other environmental reservoirs is limited. Although crAssphage has been detected globally using PCR-based assays (15), there remains little systematic analysis of the strain biogeography of crAssphage, which is particularly important considering the diverse range of currently reported strains and the existence of closely related crAss-like phages in the human gut (10) that may confound crAssphage detection using PCR-based assays.

Here we report crAssphage prevalence and diversity observed across 3,263 gut microbiome samples originating from ten globally distributed human populations (Table 1), including a crAssphage strain characterized from the gut microbiome of a traditional hunter-gatherer population. Further, we evaluate crAssphage prevalence, abundance, and strain diversity as a function of biogeography, dietary lifestyle, and host variables including age, sex, body mass index (BMI), and health status.

Table 1. Previously published metagenomics data used in this study

Studies/Datasets	**Geographic location**	**Number of samples**	**Number of unique individuals**	**References**
Bäckhed	Sweden	400	200	(16)
HMP	USA	204	123	(9)
Karlsson	Sweden	145	145	(17)
Liu	Mongolia	110	110	(18)

MetaHIT-China	China	363	363	(19)
MetaHIT-Europe	Denmark, Spain	756	609	(20), (21), (22)
Obregon-Tito	Peru	25	25	(23)
Sankaranarayanan	USA	61	61	(24)
Xie	UK	249	249	(25)
Zeevi	Israel	950	851	(26)
TOTAL		**3,263**	**2,736**	

Results

CrAssphage prevalence. Shotgun gut metagenome data from 3,263 samples were de novo assembled using MEGAHIT (27). Assembly statistics including total assembled length, number of contigs, and N50, are provided in Additional File 1: Table S1. A BLASTP search identified 755 samples containing open reading frames (ORFs) with matches to >=50% of crAssphage proteins (Additional File 1: Table S2). For individuals with multiple samples (Bäckhed, HMP, Metahit-Europe, and Zeevi datasets), crAssphage was scored as present if at least one sample contained matches to >=50% of crAssphage reference proteins. To assess the prevalence of crAssphage among individuals from each study population as well as potential associations with health, individuals were divided into categories based on health status (Table 2).

Among healthy individuals, crAssphage prevalence ranged from 8% among Peruvian hunter-gatherers from the Obregon-Tito et al. study (23) to 39% among urban-industrialized U.S. residents from the HMP cohort (9). For each study, no statistically significant associations were found between the prevalence of crAssphage and health status (Table 2).

Table 2. Association of health status with prevalence of crAssphage

Study/Dataset	Health Status	Number of individuals positive for crAssphage (Total number of individuals in category)
HMP	Healthy	47 (123)
Karlsson	Healthy Impaired Glucose Tolerance Type-2 Diabetes	15 (43) 13 (49) 22 (53)
Liu	Healthy	15 (110)
MetaHIT-China	Healthy Type-2 Diabetes N/A	37 (185) 24 (176) 1 (2)
MetaHIT-Europe	Healthy Crohn's Disease Inflammatory Bowel Disease Ulcerative Colitis N/A	95 (350) 2 (9) 7 (25) 13 (48) 58 (177)
Obregon-Tito	Healthy	2 (25)
Sankaranarayanan	Healthy Type-2 Diabetes N/A	14 (37) 11 (18) 2 (6)
Xie	Healthy Diabetes N/A	69 (211) 5 (10) 12 (28)
Zeevi	Healthy	190 (851)

N/A denotes health status information was unavailable.

Metadata such as age, sex, and BMI (Table S1) were analyzed to study potential associations with prevalence of crAssphage. No metadata was available for the Liu et al. (18) and Zeevi et al. (26) studies, and hence, they were excluded from this analysis. The Obregon-Tito et al. study (23) was also excluded due to low prevalence of crAssphage. Prevalence of crAssphage across the studies, grouped by age, sex, and BMI are summarized in Table 3. There were no statistically significant differences in the prevalence of crAssphage among individuals on the basis of age and sex. In the Sankaranarayanan et al. study (24), individuals in the overweight category, i.e. BMI >= 25, had an increased prevalence of crAssphage (chi-square: 7.3589, p-value: 0.00667,

α: 0.05). However, this result was likely influenced by the small sample size, as evidenced by a lack of association in larger cohorts such as MetaHIT and HMP.

Table 3. Association of age, sex, and BMI with prevalence of crAssphage

Study/ Dataset	MetaHIT-China	MetaHIT-Europe	HMP	Karlsson	Sankara-narayanan	Xie
Age (in years)						
< 18	1 (2)	0 (1)	0 (0)	0 (0)	0 (0)	0 (0)
18-40	19 (104)	21 (69)	47 (123)	0 (0)	11 (25)	1 (4)
41-65	34 (216)	70 (262)	0 (0)	0 (0)	11 (26)	61 (168)
> 65	7 (39)	2 (19)	0 (0)	50 (145)	3 (4)	24 (77)
Sex						
Male	36 (206)	42 (158)	30 (65)	0 (0)	10 (25)	0 (0)
Female	25 (155)	46 (194)	17 (58)	50 (145)	15 (30)	86 (249)
BMI						
< 18.5	3 (29)	0 (8)	0 (0)	1 (1)	0 (0)	1 (4)
18.5-24.9	29 (193)	44 (176)	29 (68)	13 (48)	2 (14)	41 (128)
>= 25.0	29 (139)	72 (247)	18 (55)	36 (96)	23 (41)	40 (116)

Values denote number of individuals positive for crAssphage (Total number of individuals in the category)

CrAssphage diversity. Genes encoding for capsid proteins have been previously used as molecular markers in viral phylogenetics (10, 14, 28-30). A recent study identified five putative capsid architecture genes in the crAssphage genome (reference ORFs 75-79) (10). To determine crAssphage biogeographic diversity, we focused on a multi-gene phylogenetic approach utilizing these five genes. Samples positive for crAssphage were filtered to remove those with: 1) absence of ORFs matching any one of the five genes, and 2) presence of more than one ORF matching any of the five genes. This limited our analyses to samples containing a single reconstructed viral sequence type. A total of

366 samples (out of 755) matched this criterion. Further, depth of coverage data was used to verify that recovered ORFs represent the same viral genome.

A Neighbor-Joining (NJ) tree constructed from a concatenated amino-acid alignment across all five genes (Additional File 2: Figure S1) supports the phylogenetic structure of crAss-like phages reported by (10). Further, all 366 crAssphage viral sequence types reconstructed in this study fall within the crAssphage clade. Within the crAssphage clade, the sequence types show a Hennigian comb-like phylogenetic structure (31) and relationships between them could not be further resolved due to low bootstrap support for the external branches.

Since biogeographic structuring was not observed at the level of individual viral sequence types, we used a clustering approach to group the 366 crAssphage sequences prior to phylogenetic analysis. The ORFs from the reference crAssphage strain (NC024711.1) was also included in this analysis, resulting in a total of 367 crAssphage sequences. Clustering analysis resulted in the identification of nine and 33 clusters, at identity thresholds of 97% and 98%, respectively (Figure 1 and Additional File 1: Table S3, Table S4). At the 97% identity threshold, the 367 crAssphage sequences group into nine clusters, with ~94% of sequences being present in one of two major globally distributed clusters – Clusters B and E (Figure 1A). Among the nine clusters, Clusters B and E are most closely related to each other. 44% of clusters were singleton, and typically contain sequences from the Zeevi et al. study. At the 98% identity threshold, a total of 33 clusters were identified, with Clusters 2 and 4 being the

major clusters. While Cluster 2 is globally distributed, Cluster 4 is restricted to populations in Europe, the Middle East, and the Americas, and is absent in Asia. On the other hand, one of the minor clusters, Cluster 1, is predominantly Asian, with a low prevalence in Europe. At this threshold, a comb-like structure similar to that seen in the NJ tree comprising the unclustered sequences was observed (Figure 1B), suggesting that further resolution of phylogenetic relationships between clusters is unlikely.

To assess whether the lack of biogeographic structure at the 98% identity threshold could be explained by a recent expansion event in crAssphage, we used a Bayesian Evolutionary Analysis Sampling Trees (BEAST) approach. A subset of 338 crAssphage sequences which make up the clusters showing the comb-like structure at the 98% identity threshold (as seen in Figure 1B) was used (Table S4). The centroid strain from the most closely related cluster outside this group, Cluster 22, was also included. The nucleotide alignment of the sequences for the five capsid genes was used as input. Since the crAssphage strains included in this analysis were mostly contemporaneous, a fixed mean substitution rate of 7.2×10^{-5} substitutions per site per year was used, based on the estimations for the double-stranded DNA human Adenovirus subtype-B (32). Our Bayesian Skyline plot (BSP) (Additional File 3: Figure S2) shows that crAssphage strains likely underwent an expansion event within the past 200 years.

CrAssphage acquisition and persistence. Acquisition and persistence of crAssphage sequence types in the human gut was evaluated using data from studies focusing on 1) mother-infant pairs, 2) longitudinal sampling of individuals, and 3) twin-pairs. We

assessed the potential to identify vertical transmission of crAssphage using the Bäckhed et al. study (16), which comprises samples from healthy Swedish mothers (N=100) and their infants (N=100) at birth, 4-months, and 12-months of age. CrAssphage prevalence among the healthy mothers was 25%. At birth, none of the infants showed presence of crAssphage. Three and 16 individuals were positive for crAssphage at the 4-month and 12-month stages, respectively. Thus, the prevalence of crAssphage among 4-month old infants was significantly less as compared to that in the mothers (3% and 25%, respectively) (chi-square=20.0997, p-value=0.00001, α=0.05). However, there was no significant difference between the prevalence of crAssphage among 12-month old infants and mothers (16% and 25%, respectively), suggesting that crAssphage presence is established by this age. All three infants who were positive for crAssphage at the 4-month stage showed continued presence of crAssphage at the 12-month stage as well. According to their placement in the NJ tree (Additional File 4: Figure S3), crAssphage sequence types from the same individual at 4- and 12-month stage were identical. crAssphage sequences from mother and infant from the same family could not be included in the tree, because they did not pass our criteria for inclusion in the phylogenetic analyses.

To overcome the detection limits we experienced for data from infant samples, we additionally looked for signatures of crAssphage sequence type continuity that could lend evidence as to its mode of acquisition and transmission. To this end, we evaluated samples from different time-points for individuals from the MetaHIT-Europe cohort, finding that crAssphage sequences from the same individual were most closely related

to each other (Figure S1). From the Xie et al. study (25) comprising samples from 124 adult twin-pairs, we found 14 twin-pairs where both individuals were positive for crAssphage, and of these, only two twin-pairs (P7 and P92) that could be included in the phylogenetic analysis (Additional File 5: Figure S4). While the crAssphage sequences from both P7 individuals were identical, those from P92 individuals were not closely related to each other.

Discussion

The geographic distribution of crAssphage is global (15), but as observed here, the prevalence of crAssphage is lower within samples from more traditional populations. Only two Matses individuals, out of a total of 25, are positive for crAssphage, the lowest of all populations studied. The Matses populations live in the Amazon jungle of Peru and have a subsistence practice primarily based on hunting and gathering (23). Similarly, crAssphage prevalence is lower, on average, among healthy Mongolian individuals (Table 2); the majority of these individuals practice a traditional pastoral lifestyle (18). The overall picture from the data presented here is that crAssphage prevalence is associated with an industrialized lifestyle/diet, but with no associations to health, sex, or body-size variables.

The association of crAssphage with industrialized practices can be partially explained by the putative bacterial host, *Bacteroides* (7). Evidently, crAss-like phages can infect *Bacteroides* (33), although it is not known if they are limited to this host genus. Gut microbiome studies show that *Bacteroides* tend to be more abundant in industriallized

populations when compared to traditional peoples (34), and the Matses samples are consistent with this pattern (23). However, our two Matses individuals bearing crAssphage do not show higher relative abundance of *Bacteroides* as compared to those Matses individuals without crAssphage (Additional File 6: Figure S5).

Beyond the higher prevalence in industrialized populations, biogeographic structuring among the crAssphage sequence types was minimally observed. Typically, sequences belonging to individuals from the same study/cohort did not cluster together in phylogenetic analyses (Figure S1). The exception was the MetaHIT-China cohort, wherein the majority of the crAssphage sequences from these individuals cluster together. However, these strains could be described as a sub-clade within a broader comb-like clade characteristic of recent expansion, which includes strains from all geographic regions studied. Moreover, some of the common crAssphage sequences in the MetaHIT-China cohort are also common globally (Additional File 1: Table S3). This suggests a recent global expansion of some closely related crAssphage strains, with additional subsequent expansion event(s) within China, minimally. Interestingly, the crAssphage sequence recovered from one of the two crAssphage-positive Matses individuals belongs to a cluster otherwise comprising strains from Sweden and the U.S. (9, 16).

Our centroid-based clustering approach (at 97% and 98% identity thresholds) provides a relative picture for timing the expansion event. The comb-like aspect of the tree is retained at 98% identity threshold and is lost at 97% identity threshold (Figure 1). Most

clusters with sufficient information were globally distributed. In our trees, the more evolutionarily distant sequences tend to be observed in Europe, but this is likely attributed to sampling biases as samples from European populations are much more represented (1,864 out of 3,263 samples) and these divergent strains are rare even among samples from European populations. Overall, there is little support for biogeographic structure. The exception is the phylogeny based on 98% identity threshold (Figure 1B) in which one of the two major clusters, Cluster 4, was restricted to Europe, the Middle East, and the Americas, but absent in Asia, and another more minor cluster, Cluster 1, was predominately present in Asia.

We estimated the timing of the expansion event to be approximately within the past 200 years. However, we stress the uncertainty of this estimate given the limitations of our Bayesian dating approach – firstly, due to the lack of appropriate external calibration points associated with the samples and a substitution rate for crAssphage, we used an estimated substitution rate for a double stranded DNA virus (human Adenovirus subtype-B), and secondly, our input alignment comprises only the five genes that encode the capsid proteins rather than complete crAssphage genomes. Further investigation is necessary to provide a better estimate using a substitution rate specific to crAssphage. Given the recent successful propagation of crAssphage in culture (33), the determination of this substitution rate has become a possibility.

Evidence for biogeographic structure is minimal in these analyses suggesting crAssphage has substantial horizontal transfer in humans. Children acquiring

crAssphage by 4 months of age have the same strain by 12 months of age (Figure S2), and adults sampled at different time points tend to carry the same strain (Figure S1). Overall, these patterns suggest crAssphage presence over time is stable.

Conclusion

CrAssphage is associated with the gut microbiota of industrialized human populations and is less prevalent in non-industrialized or more traditional populations. CrAssphage exhibits limited phylogeographic structure, and its major clade likely underwent a recent global expansion. The estimated timing of the expansion is approximately within the past 200 years; however, the mechanism behind this expansion remains uncertain.

Methods

Data acquisition and processing. Gut metagenomic data for a total of 3,263 samples from ten studies/cohorts were acquired (Table 1). Shotgun datasets were processed using AdapterRemoval v2 (35) to remove reads with ambiguous bases ('N"), trim at low-quality bases (Q<30), and merge overlapping read-pairs. Processed reads longer than 30 base pairs (bp) were retained for downstream analysis. These analysis-ready reads were screened for the presence of crAssphage.

De novo metagenomic assembly. Analysis-ready reads were assembled into contigs using MEGAHIT (27). Depth of coverage was calculated by mapping analysis-ready reads to assembled contigs using Bowtie2 (36), followed by processing of resulting alignment files using SAMTools (37) and custom R scripts. Open reading frame (ORF)

prediction was carried out using Prodigal (38). A custom BLAST database was created using the collection of 91 proteins previously predicted from the crAssphage reference genome (NC024711.1) (10). The predicted ORFs (amino acid) from each shotgun metagenome were queried against this custom database using BLASTP (39), and matches were identified using the following criteria: 1) query coverage (length of alignment / query length) ≥ 70%, 2) percent identity ≥ 60%, and 3) E value < 1e-5. A sample was considered positive for crAssphage if matches were recovered for ≥ 50% of the reference crAssphage proteins.

Association testing – crAssphage prevalence and metadata. Individuals were grouped on the basis of 1) health status, 2) age: < 18 years, 18 – 40 years, 41 – 65 years, and > 65 years, 3) sex: male and female, and 4) BMI: underweight (BMI < 18.5), normal (BMI 18.5 – 24.99), and overweight (BMI ≥ 25), according to the World Health Organization recommendations. Associations between the presence of crAssphage and health status, age, sex, and BMI categories were assessed using the Chi square test.

Taxonomic profiles. For the Matses individuals from the Obregon-Tito et al. study (23), analysis-ready reads for each individual were mapped to the Greengenes database of 16S rRNA gene sequences (40) using Bowtie2 (36). Unmapped reads were removed using the --no-unal option. The resulting SAM files were converted to BAM files, sorted, and duplicates were removed using SAMTools v1.3 (37). The sequences of the reads mapping to the Greengenes database were obtained from the BAM files. These sequences were clustered into Operational Taxonomic Units (OTUs) using vsearch

(41), employing the cluster_fast algorithm and comparing to the GreenGenes database. Other parameters used included a minimum sequence length of 70 bp and 97% similarity for clustering. The resulting OTU table was rarefied to a depth of 10,000 sequences per sample and singleton OTUs were removed using QIIME (42). Taxonomic summaries were generated at the phylum- and genus-levels. Relative abundances of phylum Bacteroidetes in individuals were determined and plotted using R (43).

Phylogenetic analysis. For each assembled metagenome, the number of unique ORFs matching each crAssphage reference protein was calculated from the BLAST results and used to infer the number of concurrent phage strains carried in the sample. Samples identified as containing one unique ORF for each crAssphage reference protein were selected for phylogenetic analysis. Depth of coverage information was used to verify that identified ORFs were representative of the same genome. A previously published study demonstrated the use of putative capsid proteins (genes 75, 76, 77, 78, and 79) to document diversity among crAss-like phages (10). Phylogenetic analysis in this study were focused on these genes and included sample- and cluster-based trees to evaluate phylogeography, and Bayesian approaches to estimate strain divergence.

Amino acid sequences for the ORFs mapping to the putative capsid genes were extracted from each sample selected for phylogenetic analysis. Additionally, amino acid sequences for these genes were extracted from the reference crAssphage strain and 32

crAss-like phages (10), and included in the phylogenetic analysis. Multiple sequence alignments were constructed for each gene separately using MUSCLE v3.8.31 (44) with default parameters. The alignments were concatenated to produce a multi-gene alignment comprising 4,543 sites, of which 1,102 sites were retained following curation (removal of sites with gaps and missing data). The curated alignment was used to construct an Neighbor-Joining tree in MEGA7 (45) using the Jones-Taylor-Thornton substitution model with uniform rates among sites. Branch support was estimated from bootstrap analysis using 500 replicates.

Additionally, a concatenated nucleotide alignment was generated for the five putative capsid genes across all samples selected for phylogenetic analysis. This alignment comprised 8,633 sites. Centroid based clustering was performed on this alignment using UCLUST (46) at sequence similarity thresholds of 97% and 98%, respectively. For each cluster analysis, centroid sequences for individual clusters were aligned using MUSCLE, and the resulting alignment curated to remove sites with >5% missing data. The curated alignments were used to construct Neighbor-Joining trees in MEGA7 using the Tamura-Nei substitution model with uniform rates among sites, and bootstrap analysis with 500 replicates to estimate branch support.

For estimating the timing of a potential population expansion among crAssphage sequences, a subset of 338 sequences belonging to the crAssphage clusters showing a comb-like structure at the 98% clustering level (see Figure 1B), was used. Additionally, the centroid strain from Cluster 22 was included as an outgroup. The list of strains

included in the BEAST analysis is given in Table S4. The nucleotide alignment of the five capsid genes was used as input for BEAST v1.8.4 (47). Since substitution rates for crAssphage or crAss-like phages remain undetermined, we used a substitution rate of 7.2×10^{-5} substitutions per site per year as estimated for the human Adenovirus subtype-B (32), which is a double-stranded DNA virus. We used a General Time Reversible (GTR) model of nucleotide substitution, an uncorrelated lognormal clock model with uniform rate across branches, and a Bayesian Coalescent Skyline plot tree prior. One Markov Chain Monte Carlo (MCMC) run was carried out with 10,000,000 iterations, sampling every 1,000 steps. The first 1,000,000 iterations were discarded as burn-in. Tracer (48) was used to visualize the results of the MCMC run and generate a Bayesian Skyline Plot (BSP).

Author Contributions

Conceptualization: C.M.L. and K.S.

Methodology: K.S., C.M.L, and T.P.H

Investigation: K.S., T.P.H, S.L.S, and A.T.O.

Formal analysis: K.S. and T.P.H.

Funding acquisition and resources: C.M.L., K.S., and C.W.

Writing - Original draft: T.P.H., C.M.L., and K.S.

Writing - Review and editing: All authors

Funding

This study was supported by a grant from the National Institutes of Health (NIH R01 GM089886). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Virgin HW. The virome in mammalian physiology and disease. Cell. 2014;157(1):142-50.
2. Rohwer F, Youle M. Consider something viral in your research. Nat Rev Micro. 2011;9(5):308-9.
3. De Paepe M, Leclerc M, Tinsley CR, Petit MA. Bacteriophages: an underestimated role in human and animal health? Frontiers in cellular and infection microbiology. 2014;4:39.
4. Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI. Going viral: next-generation sequencing applied to phage populations in the human gut. Nature reviews Microbiology. 2012;10(9):607-17.
5. Williams SC. The other microbiome. Proc Natl Acad Sci U S A. 2013;110(8):2682-4.
6. Rampelli S, Turroni S, Schnorr SL, Soverini M, Quercia S, Barone M, et al. Characterization of the human DNA gut virome across populations with different subsistence strategies and geographical origin. Environmental Microbiology. 2017;19(11):4728-35.

7. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun. 2014;5.

8. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature. 2010;466(7304):334-8.

9. NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, et al. The NIH Human Microbiome Project. Genome Res. 2009;19(12):2317-23.

10. Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA, et al. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat Microbiol. 2018;3(1):38-46.

11. Ahmed W, Lobos A, Senkbeil J, Peraud J, Gallard J, Harwood VJ. Evaluation of the novel crAssphage marker for sewage pollution tracking in storm drain outfalls in Tampa, Florida. Water Res. 2017;131:142-50.

12. Garcia-Aljaro C, Balleste E, Muniesa M, Jofre J. Determination of crAssphage in water samples and applicability for tracking human faecal pollution. Microb Biotechnol. 2017;10(6):1775-80.

13. Stachler E, Kelty C, Sivaganesan M, Li X, Bibby K, Shanks OC. Quantitative CrAssphage PCR Assays for Human Fecal Pollution Measurement. Environ Sci Technol. 2017;51(16):9146-54.

14. Guerin E, Shkoporov A, Stockdale S, Clooney AG, Ryan FJ, Draper LA, et al. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. bioRxiv. 2018.

15. Cinek O, Mazankova K, Kramna L, Odeh R, Alassaf A, Ibekwe MU, et al. Quantitative CrAssphage real-time PCR assay derived from data of multiple geographically distant populations. J Med Virol. 2018;90(4):767-71.

16. Backhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe. 2015;17(5):690-703.

17. Karlsson FH, Tremaroli V, Nookaew I, Bergstrom G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. 2013;498(7452):99-103.

18. Liu W, Zhang J, Wu C, Cai S, Huang W, Chen J, et al. Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. Sci Rep. 2016;6:34826.

19. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490(7418):55-60.

20. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013;500(7464):541-6.

21. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014;32(8):834-41.

22. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32(8):822-8.

23. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, et al. Subsistence strategies in traditional societies distinguish gut microbiomes. Nat Commun. 2015;6:6505.

24. Sankaranarayanan K, Ozga AT, Warinner C, Tito RY, Obregon-Tito AJ, Xu J, et al. Gut Microbiome Diversity among Cheyenne and Arapaho Individuals from Western Oklahoma. Curr Biol. 2015;25(24):3161-9.

25. Xie H, Guo R, Zhong H, Feng Q, Lan Z, Qin B, et al. Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. Cell Syst. 2016;3(6):572-84 e3.

26. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized Nutrition by Prediction of Glycemic Responses. Cell. 2015;163(5):1079-94.

27. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674-6.

28. Dolja VV, Koonin EV. Phylogeny of capsid proteins of small icosahedral RNA plant viruses. Journal of General Virology. 1991;72:1481-6.

29. Rowe JM, Fabre MF, Gobena D, Wilson WH, Wilhelm SW. Application of the major capsid protein as a marker of the phylogenetic diversity of Emiliania huxleyi viruses. FEMS Microbiol Ecol. 2011;76(2):373-80.

30. Sinclair RM, Ravantti JJ, Bamford DH. Nucleic and Amino Acid Sequences Support Structure-Based Viral Classification. J Virol. 2017;91(8).

31. Yang J, Li J, Dong L, Grünewald S. Analysis on the reconstruction accuracy of the Fitch method for inferring ancestral states. BMC Bioinformatics. 2011;12:18-.

32. Firth C, Kitchen A, Shapiro B, Suchard MA, Holmes EC, Rambaut A. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. Mol Biol Evol. 2010;27(9):2038-51.

33. Shkoporov A, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, et al. ΦCrAss001, a member of the most abundant bacteriophage family in the human gut, infects Bacteroides. bioRxiv. 2018.

34. Gupta VK, Paul S, Dutta C. Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. Front Microbiol. 2017;8:1162.

35. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Research Notes. 2016;9.

36. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9:357.

37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

38. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11(1):119.

39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403-10.

40. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069-72.

41. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.

42. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature Methods. 2010;7:335.

43. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.

44. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32(5):1792-7.

45. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular Biology and Evolution. 2016;33(7):1870-4.

46. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460-1.

47. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution. 2018;4(1):vey016-vey.

48. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Systematic Biology. 2018:syy032-syy.

Figures

Figure 1. Clustering of crAssphage strains based on percentage identity thresholds. Panels A and B show phylogenetic relationships between crAssphage clusters at the 97% and 98% identity thresholds, respectively. Both panels show Neighbor-Joining trees constructed using MEGA7. The trees were based on the Tamura Nei model (26) using a partial deletion cut-off set at 95%. The scale denotes substitutions per site. Bootstrap support based on 500 replicates is given on the branches; only values greater than 50% are shown. The geographic prevalence is shown next to each cluster: red triangles = Europe, blue circles = Asia, green stars = The Americas, and purple squares = The Middle East. Studies/datasets included in each geographic category are - Europe: Bäckhed et al. (Sweden), Karlsson et al. (Sweden), MetaHIT-Europe (Denmark, Spain), Xie et al. (U.K.); Asia: Liu et al. (Mongolia), MetaHIT-China (China); The Americas: HMP (U.S.A), Obregon-Tito et al. (Peru), Sankaranarayanan et al. (U.S.A); The Middle East: Zeevi et al. (Israel).

Additional Files

Additional File 1 – Supplementary Tables S1 – S4

This Microsoft Excel-compatible spreadsheet contains Supplementary Tables S1 – S4 in separate tabs. Table S1: A list of all 3,263 samples included in this study, associated meta-data, and results of the crAssphage screening using de novo metagenomic

assembly. Table S2: A list of 755 samples found to be positive for crAssphage and the numbers of unique ORFs mapping to all crAssphage proteins (1 - 91) for each sample. Table S3: Summary of results of clustering crAssphage strains at the 97- and 98-percentage identity thresholds. Table S4: A list of 367 crAssphage strains included in phylogenetic analyses along with corresponding cluster number at both 97- and 98-percentage identity thresholds.

Additional File 2 - Supplementary Figure S1

Neighbor-Joining tree of 367 crAssphage sequences and 32 crAss-like phages. The tree is based on amino acid sequences for the five genes encoding the capsid proteins. Sites with gaps or missing data were completely deleted, resulting in a final alignment of 1,102 sites. The tree was constructed on the Jones-Taylor-Thornton model of amino acid substitution with uniform rates among sites. Bootstrap support was estimated from 500 replicates. Bootstrap values for outer branches having support greater than 50% are shown. The bootstrap support for the crAssphage clade is 100%. Bootstrap support for branches within this clade is not shown in this figure for clarity. These are shown in the newick file provided as Additional File 7. Taxa are color-coded according to study/dataset.

Additional File 3 – Supplementary Figure S2

Bayesian Skyline Plot of 339 crAssphage strains. The X-axis denotes time in years before present (YBP) and Y-axis denotes estimated effective population size.

Additional File 4 – Supplementary Figure S3

Neighbor-Joining tree of 27 crAssphage strains from individuals from the Bäckhed dataset. The tree is based on the nucleotide sequences for the five capsid genes. Sites with less than 95% site coverage were excluded, resulting in a final alignment of 8,610 sites. The tree was constructed using the Tamura-Nei model of nucleotide substitution with uniform rates among sites. Bootstrap support was estimated from 500 replicates. Bootstrap values for outer branches having support greater than 50% are shown. Taxa names are denoted as sample number followed by the family number and identifier (M: mothers, 4M: infant at 4-months, 12M: infant at 12-months).

Additional File 5 – Supplementary Figure S4

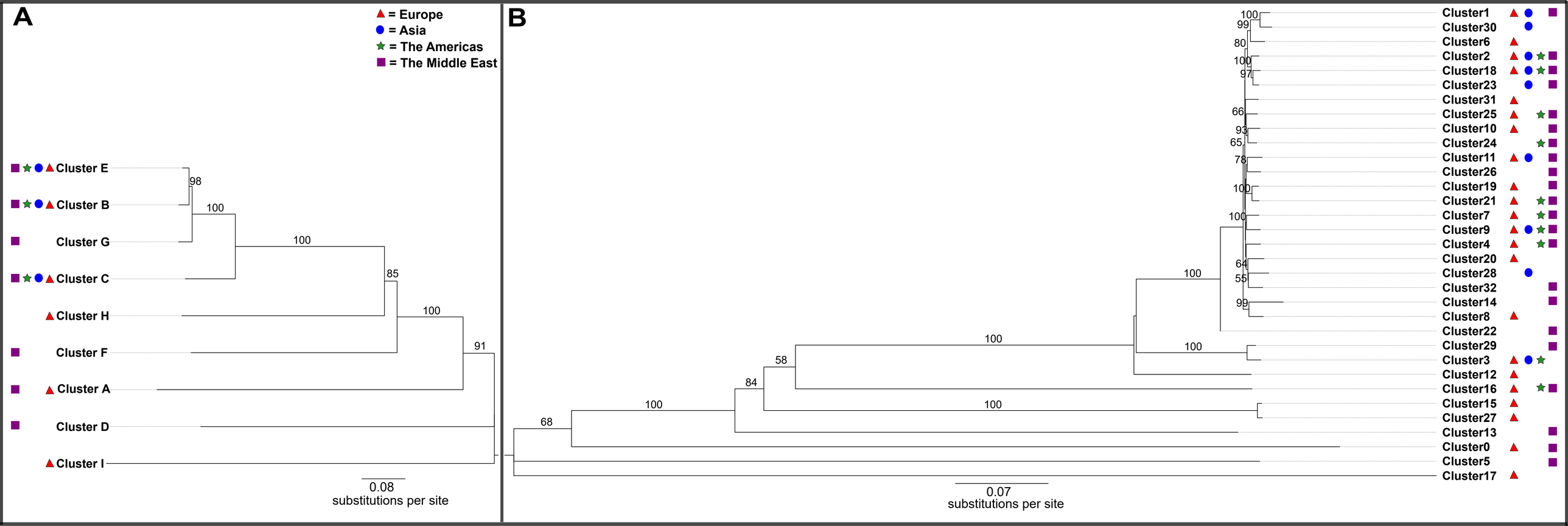
Neighbor-Joining tree of 36 crAssphage strains from individuals from the Xie dataset. The tree is based on the nucleotide sequences for the five capsid genes. Sites with less than 95% site coverage were excluded, resulting in a final alignment of 8,220 sites. The tree was constructed using the Tamura-Nei model of nucleotide substitution with uniform rates among sites. Bootstrap support was estimated from 500 replicates. Bootstrap values for outer branches having support greater than 50% are shown. Taxa names are denoted as sample number followed by the twin-pair number. Samples from pairs of twins included in this analysis are highlighted in red.

Additional File 6 – Supplementary Figure S5

Relative abundance of genus *Bacteroides* among Matses individuals. Boxplot denoting percentage relative abundance of *Bacteroides* sp. among Matses individuals. Values corresponding to individuals positive for crAssphage are denoted in red.

Additional File 7

This is a newick file containing the Neighbor-Joining tree described in Figure S1. This file can be viewed using any tree-viewing program such as Figtree.


A
▲ = Europe
● = Asia
★ = The Americas
■ = The Middle East
Cluster E
Cluster B
Cluster G
Cluster C
Cluster H
Cluster F
Cluster A
Cluster D
Cluster I
98
100
100
85
100
91
0.08
substitutions per site
B
Cluster1
Cluster30
Cluster6
Cluster2
Cluster18
Cluster23
Cluster31
Cluster25
Cluster10
Cluster24
Cluster11
Cluster26
Cluster19
Cluster21
Cluster7
Cluster9
Cluster4
Cluster20
Cluster28
Cluster32
Cluster14
Cluster8
Cluster22
Cluster29
Cluster3
Cluster12
Cluster16
Cluster15
Cluster27
Cluster13
Cluster0
Cluster5
Cluster17
100
99
80
100
97
66
93
65
78
100
100
64
55
99
100
100
100
58
84
100
100
68
0.07
substitutions per site