**Full title:** Biogeographic study of human gut-associated crAssphage suggests impacts from industrialization and recent expansion.

**Short title:** Biogeographic analysis of global human crAssphage strains.

**Authors and affiliations:**

Tanvi P. Honap[1,2,†], Krithivasan Sankaranarayanan[1,3,†], Stephanie L. Schnorr[1,#a,#b], Andrew T. Ozga[1,2, #c], Christina Warinner[1,2,#d], and Cecil M. Lewis Jr. [1,2,*]

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

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

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

[#a] Current address: Department of Anthropology, University of Nevada, Las Vegas, 4505 S. Maryland Pkwy., Las Vegas, Nevada 89154, USA

[#b] Current address: Konrad Lorenz Institute for Evolution and Cognition Research, Martinstrasse 12, 3400 Klosterneuburg, Austria

[#c] Current address: Halmos College of Natural Sciences and Oceanography, Nova Southeastern University, 3301 College Ave, Fort Lauderdale, Florida 33314, USA

23 [#d] Current address: Department of Anthropology, Harvard University, 11 Divinity Ave,
24 Cambridge, Massachusetts 02138, USA
25
26 * Corresponding author:
27 E-mail: cmlewis@ou.edu (C. M. L.)
28
29 [†] These authors contributed equally to this work.
30

# 31 Abstract

32 CrAssphage (cross-assembly phage) is a bacteriophage that was first discovered in
33 human gut metagenomic data. CrAssphage belongs to a diverse family of crAss-like
34 bacteriophages thought to infect gut commensal bacteria belonging to *Bacteroides* species.
35 However, not much is known about the biogeography of crAssphage and whether certain strains
36 are associated with specific human populations. In this study, we screened publicly available
37 human gut metagenomic data from 3,341 samples for the presence of crAssphage *sensu stricto*
38 (NC_024711.1). We found that crAssphage prevalence is low in traditional, hunter-gatherer
39 populations, such as the Hadza from Tanzania and Matses from Peru, as compared to
40 industrialized, urban populations. Statistical comparisons showed no association of crAssphage
41 prevalence with variables such as age, sex, body mass index, and health status of individuals.
42 Phylogenetic analyses show that crAssphage strains reconstructed from the same individual over
43 multiple time-points, cluster together. CrAssphage strains from individuals from the same study
44 population do not always cluster together. Some evidence of clustering is seen at the level of
45 broadly defined geographic regions, however, the relative positions of these clusters within the

crAssphage phylogeny are not well-supported. We hypothesize that this lack of strong biogeographic structuring is suggestive of a recent expansion event within crAssphage. Using a Bayesian dating approach, we estimate this expansion has occurred within the past 200 years. Overall, we determine that crAssphage presence is associated with an industrialized lifestyle. The absence of strong biogeographic structuring within global crAssphage strains is likely due to a recent population expansion within this bacteriophage.

# Introduction

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 gut virome is composed of double-stranded DNA bacteriophages [4], which are argued to be the primary regulators of bacterial biomass [5]. Bacteriophages have known impacts on the human gut; for example, they are involved in the transmission of bacterial antibiotic resistance genes, and facilitate bacterial carbohydrate utilization [4]. The majority of the bacteriophages found in the human gut belong to the families *Siphoviridae*, *Podoviridae*, and *Myoviridae* within the order Caudovirales [6]; however, our knowledge of the true bacteriophage diversity is limited [7]. 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 ribosomal RNA gene in archaea and bacteria), preventing characterization using current amplicon-based sequencing strategies.

68 Recent advances in metagenomic sequencing and development of data mining tools now
69 allow for large-scale studies of human gut bacteriophage diversity using shotgun-sequencing
70 data from human fecal samples [7]. From such data, one particular DNA bacteriophage,
71 crAssphage, appears to be remarkably abundant in the human gut [8]. CrAssphage (cross-
72 assembly phage) was named after the crAss (cross-assembly) software originally used to
73 discover the bacteriophage [8]. CrAssphage is prevalent in available gut metagenomic data from
74 the U.S. [9], comprising up to 90% of viral particle-derived metagenomes and up to 22% of
75 reads in a total fecal community from the U.S. National Institutes of Health's Human
76 Microbiome Project cohort [10]. Moreover, crAssphage is up to six times more abundant than all
77 other known bacteriophages that could be reconstructed from these publicly available
78 metagenomes [8]. Bacteria belonging to phylum Bacteroidetes, which are commonly found gut
79 commensals, were predicted to be the host of this bacteriophage [8].
80 A recent study has shown that crAssphage, which is now referred to as crAssphage *sensu*
81 *stricto*, is actually a member of a diverse group of crAss-like bacteriophages [11]. The first
82 isolated member of this group, ΦCrAss001, shows podovirus-like morphology and infects the
83 bacterium, *Bacteroides intestinalis* [12]. In addition to the human gut, crAss-like phages are
84 found in nonhuman primate and termite guts, terrestrial and groundwater sources, and oceanic
85 environments [11, 13-15]. Across metagenomic studies, crAssphage *sensu stricto* is nearly
86 exclusively associated with the human gut microbiome [8, 11] and has been proposed as a
87 potential biomarker for fecal contamination [16-20].
88 Recent research suggests that crAssphage may be present in infants as early as one-month
89 after birth [21], which is partly explained by the fact that *Bacteroides* species are among the most
90 abundant members of the gut microbiomes of newborns [22]. It has also been suggested that

91 while most adults contain a single dominant crAssphage strain [21], a minority can carry
92 multiple (hundreds) crAssphage strains [20]. However, the means by which crAssphage is
93 initially acquired [13, 21], strain variation within an individual [20], overall geographic
94 distribution [20, 23], and prevalence in other environmental reservoirs has not been fully
95 elucidated. Although crAssphage has been detected globally using PCR-based assays [13, 20], a
96 systematic analysis of the strain biogeography of crAssphage is necessary, considering the
97 diverse range of currently reported crAssphage strains and the existence of other crAss-like
98 phages in the human gut [11] that may confound crAssphage detection using PCR-based assays.
99 Here, we report strain-level prevalence and diversity of crAssphage *sensu stricto*
100 observed across 3,341 gut metagenomic samples originating from globally distributed human
101 populations (Table 1). In this study, we consider genetically non-identical crAssphage sequences
102 as different crAssphage strains [24]. Further, we evaluate crAssphage prevalence, abundance,
103 and strain diversity as a function of biogeography, human dietary lifestyle and variables
104 including age, sex, body mass index (BMI), and health status.
105

106 **Table 1. Publicly available gut metagenomic datasets used in this study**

| Dataset | Description | Samples | Individuals | Accession | Reference |
|---|---|---|---|---|---|
| BCK | Urban mother-infant pairs from Sweden | 400 | 200 | ERP005989 | [22] |
| CNA | Urban individuals of Cheyenne, Arapaho, and non-native ancestry from Oklahoma, USA | 61 | 61 | PRJNA299502 | [25] |
| HAD | Hadza hunter-gatherers from Tanzania | 67 | 67 | SRP056480, SRP110665 | [26], [27] |
| HMP | Urban individuals from USA | 204 | 123 | phs000228.v3.p1 | [10] |
| ITA | Urban individuals from Italy | 11 | 11 | SRP056480 | [26] |
| ISR | Urban individuals from Israel | 950 | 851 | PRJEB11532 | [28] |
| KRL | Urban individuals from Sweden | 145 | 145 | ERP002469 | [29] |
| LIU | Traditional pastoralists from Mongolia | 110 | 110 | SRP080787 | [30] |
| MAT | Matses hunter-gatherers from Peru | 25 | 25 | PRJNA268964 | [31] |
| MHC | Urban individuals from China | 363 | 363 | SRA045646, SRA050230 | [32] |
| MHE | Urban individuals from Sweden and Denmark | 756 | 606 | ERP003612, ERP004605, ERP002061 | [33], [34], [35] |
| XIE | Urban twin-pairs from the UK | 249 | 249 | ERP010708 | [36] |

107

108 # Results and discussion

109 ## CrAssphage screening

110 Publicly available gut metagenome data from 3,341 samples were screened for the
111 presence of crAssphage *sensu stricto* using a two-pronged approach: first, a direct reference-
112 based mapping to crAssphage (NC_024711.1) using Bowtie2 [37], and second, a *de novo*
113 assembly using MEGAHIT [38] followed by a protein-protein BLAST search (BLASTP) [39]
114 against the crAssphage proteins (see Methods). Mapping and assembly statistics are provided in
115 S1 Table.

116 Using the reference-based mapping approach, we identified 614 samples wherein a high-
117 quality crAssphage strain could be reconstructed (>70% of the genome at >10-fold mean

118 coverage). However, this approach is suitable only for samples containing a single dominant
119 crAssphage strain. 60 samples showed more than 100 heterozygous sites (which corresponds to
120 0.1% of the total genome), with the maximum being 1,118 heterozygous sites in a single sample.
121 In order to determine whether the samples comprised multiple crAssphage strains, we performed
122 a *de novo* assembly of each metagenome. We used BLASTP to query the predicted open reading
123 frames (ORFs) for each sample against the crAssphage reference (NC_024711.1) proteins. The
124 BLASTP hits were then filtered to include only those which showed 95% query coverage and
125 95% identity to crAssphage proteins. These criteria were chosen to avoid false positive hits from
126 crAss-like phages, since the average pairwise identity between members of the crAss-like phage
127 family ranges from 20-40% [11, 40]. Using this approach, we identified 963 samples wherein at
128 least one crAssphage protein was recovered (S2 Table). We were not able to recover all 90
129 proteins from a single sample, with the maximum number of crAssphage proteins recovered
130 from a single sample being 67. We hypothesize that this is due to the limitations of our *de novo*
131 metagenomic assembly as well as the stringency of our BLASTP search parameters, wherein our
132 approach is likely biased towards avoiding false positive hits. However, our approach is not
133 biased towards recovery of specific crAssphage proteins, since all 90 crAssphage reference
134 proteins were recovered across all samples. Since we were only able to recover a maximum of 67
135 proteins for a single sample, we decided to use a 50% coverage threshold to call a sample
136 "crAss-positive", i.e. at least 33 crAssphage proteins had to be recovered from the sample. Using
137 this criterion, 719 samples were found to be crAss-positive. Further, we determined that 497
138 samples had at least one crAssphage reference protein matched by multiple sample ORFs,
139 suggesting presence of multiple strains.
140

141 ## crAssphage prevalence

142 To assess the prevalence of crAssphage among individuals from each study population as
143 well as potential associations with health, individuals were divided into categories based on
144 health status (Table 2). For some datasets (BCK, HMP, ISR, and MHE), multiple samples were
145 available from the same individual; in this case, an individual was considered crAss-positive if at
146 least one sample from that individual was positive. None of the HAD and ITA samples, acquired
147 from Hadza individuals from Tanzania (N=67) and urban Italians (N=11), respectively, were
148 crAss-positive. For the MAT samples acquired from Matses individuals from Peru (N=25), only
149 two were crAss-positive. The Hadza peoples from Tanzania as well as the Matses from the
150 Amazon jungle of Peru have a subsistence practice primarily based on hunting and gathering [31,
151 41]. Among healthy, urban individuals leading an industrialized lifestyle, crAssphage prevalence
152 ranged from 14.0 % among Chinese individuals from the MetaHIT cohort (MHC; N=185) to
153 35.7 % among U.S. residents from the HMP cohort (N=123). CrAssphage prevalence showed no
154 significant associations with human health status (Table 2).
155

156 **Table 2. Association of health status with prevalence of crAssphage**

| Dataset | Total number of individuals | Health Status | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Healthy | IGT | T2D | CD | IBD | UC | Not specified |
| **BCK** | **200** | 39 (200) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| **CNA** | **61** | 14 (37) | 0 (0) | 11 (18) | 0 (0) | 0 (0) | 0 (0) | 2 (6) |
| **HAD** | **67** | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (67) |
| **HMP** | **123** | 47 (123) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| **ISR** | **851** | 185 (851) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| **ITA** | **11** | 0 (11) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| **KRL** | **145** | 15 (43) | 12 (49) | 20 (53) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| **LIU** | **110** | 11 (110) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| **MAT** | **25** | 2 (25) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| **MHC** | **363** | 37 (185) | 0 (0) | 23 (176) | 0 (0) | 0 (0) | 0 (0) | 1 (2) |
| **MHE** | **606** | 87 (350) | 0 (0) | 0 (0) | 2 (9) | 7 (25) | 12 (48) | 53 (174) |
| **XIE** | **249** | 64 (211) | 0 (0) | 5 (10) | 0 (0) | 0 (0) | 0 (0) | 12 (28) |

157 Values indicate number of crAss-positive individuals (total number of individuals in the
158 category). Health status categories refer to: IGT – Impaired Glucose Tolerance, T2D – Type 2
159 Diabetes, CD – Crohn's Disease, IBD – Inflammatory Bowel Disorder, UC – Ulcerative Colitis.
160

161 Our findings suggest that crAssphage prevalence may be associated with an
162 industrialized lifestyle. A similar finding was reported in a manuscript that was published while
163 our manuscript was under review [15]. This latter study found that gut metagenomic data for
164 rural Malawi and Amazonas individuals from Venezuela [42] as well as the ~5300-year old
165 Tyrolean Iceman [43] show very low numbers of crAssphage sequences. The association of
166 crAssphage with industrialized practices can also be partially explained by its putative bacterial
167 host, *Bacteroides* species [8]. CrAss-like phages can infect *Bacteroides* [44], although it is not
168 known if they are limited to this host genus [45]. Gut microbiome studies have shown that
169 *Bacteroides* tend to be more abundant in industrialized human populations as compared to

170 traditional peoples [46], and the Hadza and Matses individuals are consistent with this pattern
171 [31, 41]. However, the two crAss-positive Matses individuals do not show higher relative
172 abundance of *Bacteroides* as compared to crAss-negative Matses individuals (S1 Fig).

173 For healthy individuals, we also analyzed metadata, such as age, sex, and BMI, to study
174 potential associations with prevalence of crAssphage. The HAD, ISR, ITA, LIU, and MAT
175 datasets were excluded from this analysis due to unavailability of metadata or insufficient
176 number of crAss-positive individuals. There were no statistically significant differences in the
177 prevalence of crAssphage among individuals based on age, sex, or BMI (Table 3).

178

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

| Dataset | CNA | HMP | KRL | MHC | MHE | XIE |
|---|---|---|---|---|---|---|
| **Age** | | | | | | |
| **18 - 40** | 9 (23) | 47 (123) | 0 (0) | 16 (74) | 11 (35) | 1 (4) |
| **41 - 65** | 4 (12) | 0 (0) | 0 (0) | 19 (103) | 47 (215) | 44 (134) |
| **> 65** | 1 (2) | 0 (0) | 15 (43) | 2 (7) | 2 (19) | 19 (73) |
| **Sex** | | | | | | |
| **Male** | 5 (18) | 30 (65) | 0 (0) | 19 (95) | 31 (128) | 0 (0) |
| **Female** | 9 (19) | 17 (58) | 15 (43) | 18 (90) | 29 (142) | 64 (211) |
| **BMI** | | | | | | |
| **< 18.5** | 0 (0) | 0 (0) | 1 (1) | 2 (17) | 0 (3) | 0 (4) |
| **18.5-24.9** | 2 (13) | 29 (68) | 6 (18) | 17 (104) | 29 (130) | 35 (108) |
| **>= 25.0** | 12 (24) | 18 (55) | 8 (24) | 18 (64) | 58 (207) | 29 (98) |

180 Values denote number of crAss-positive individuals (total number of individuals in the category)

181

## 182 Phylogenetic analyses

183 To study the phylogenetic relationships between the crAssphage strains, we selected ten
184 genes which are present in all members of the crAss-like phage family [11]. These genes include
185 those encoding the five putative capsid proteins, a single-stranded DNA-binding protein, a DNA-
186 G family primase, a PD-(D/E)XK family nuclease, and two hypothetical proteins. We then

187 identified a subset of 232 samples wherein only one sample ORF matched these crAssphage
188 reference proteins, suggesting the presence of a single strain of crAssphage in the samples.
189 Furthermore, we confirmed that these ORFs had similar depths of coverage to verify that they
190 represent the same viral genome. We also included the reference crAssphage genome
191 (NC_024711.1), resulting in a total of 233 taxa. A nucleotide alignment comprising these ten
192 genes contained 12,642 sites. A Maximum Likelihood (ML) tree of the 233 strains based on this
193 multi-gene alignment is given in S2 Fig.

194 According to the ML tree, crAssphage strains from the same individual cluster together.
195 While strains from individuals belonging to the same population/study dataset do cluster
196 together, many such crAssphage clusters are found in each population. Unfortunately, the
197 relative positions of these clusters could not be well-established due to insufficient bootstrap
198 support. Interestingly, we found that the strain from a crAss-positive Matses individual clusters
199 with crAssphage strains from European individuals from the MetaHIT cohort, suggesting that it
200 belongs to a widespread crAssphage clade.

201 To assess whether the lack of biogeographic structure in the multi-gene phylogeny was
202 due to lack of sufficient information for phylogenetic analysis, we identified a subset of 118
203 samples fulfilling the following criteria: 1) according to the *de novo* assembly results, none of
204 crAssphage proteins were matched by more than one sample ORF, and 2) according to the
205 reference-based mapping, more than 70% of the crAssphage genome was covered at least 10-
206 fold. Furthermore, we determined that these strains showed a range of 2 - 280 heterozygous sites
207 (i.e. a maximum of 0.3% of sites across the genome were heterozygous). Taken altogether, we
208 considered these samples to contain only one dominant crAssphage strain. The reference
209 crAssphage genome was also included. A multi-genome alignment comprising all 97,065

210 nucleotide sites was built and used to generate an ML tree (Fig 1). This phylogeny supports the
211 findings of our multi-gene analysis. CrAssphage strains from the same individual cluster
212 together, with high bootstrap support. Strains from individuals from the same study dataset do
213 not all necessarily cluster together; many such phylogenetically dissimilar strains are found in
214 individuals within a study population. At the level of broadly defined geographic regions, some
215 biogeographic structure is observed. However, even at phylogenomic resolution, due to low
216 bootstrap values, the relative positions of these clusters within the crAssphage phylogeny cannot
217 be well-defined.

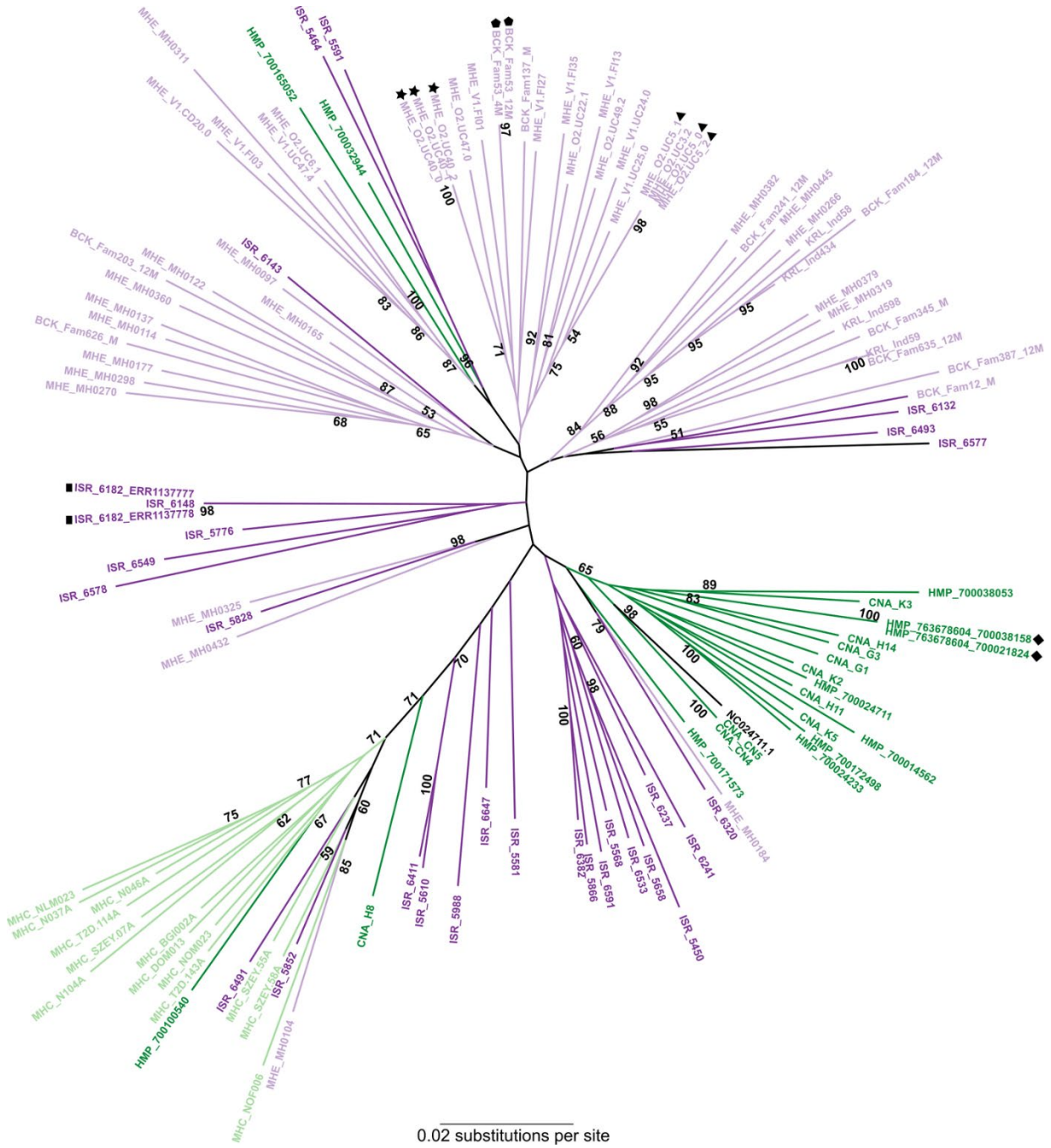



**Fig 1. Phylogenomic analysis of crAssphage strains.** The Maximum Likelihood tree was based on a multi-genome alignment comprising 97,065 sites. The tree was built using the Generalized Time-Reversible model with gamma-distributed rate variation and proportion of invariant sites. Bootstrap support was estimated from 100 replicates; only values greater than 50% are shown. Strains are color-coded according to the geographic location: dark green – The Americas, light

224 green – Asia, light violet – Europe, and dark purple – Middle East (Israel). Symbols are used to
225 denote strains from the same individual.
226

## 227 CrAssphage population expansion

228 We hypothesized that the lack of strong biogeographic clustering could be explained by a
229 recent expansion event in crAssphage. To asses this, we used a Bayesian Evolutionary Analysis
230 Sampling Trees (BEAST) approach [47]. The analysis was performed using the multi-gene
231 alignment described previously, and a fixed mean substitution rate, since these crAssphage
232 strains are mostly contemporaneous. Because a substitution rate for crAssphage or other
233 *Podoviridae* bacteriophages has not been estimated as of yet, we used a rate of $1.9 \times 10^{-4}$
234 substitutions per site per year as estimated for bacteriophages in the family *Siphoviridae* [48], the
235 members of which also infect bacteria and archaea. Our Bayesian Skyline plot (BSP) (S3 Fig)
236 shows that crAssphage strains underwent an expansion event likely within the past 200 years.
237 However, we stress the uncertainty of this estimate given the limitations of our Bayesian dating
238 approach – namely, the lack of appropriate external calibration points and a substitution rate
239 specific to crAssphage. Given the recent successful propagation of crAssphage in culture [44],
240 the determination of this substitution rate has become a possibility. Further investigation using a
241 crAssphage-specific substitution rate will provide a more robust estimate of the timing of the
242 population expansion.
243

## 244 CrAssphage acquisition and persistence

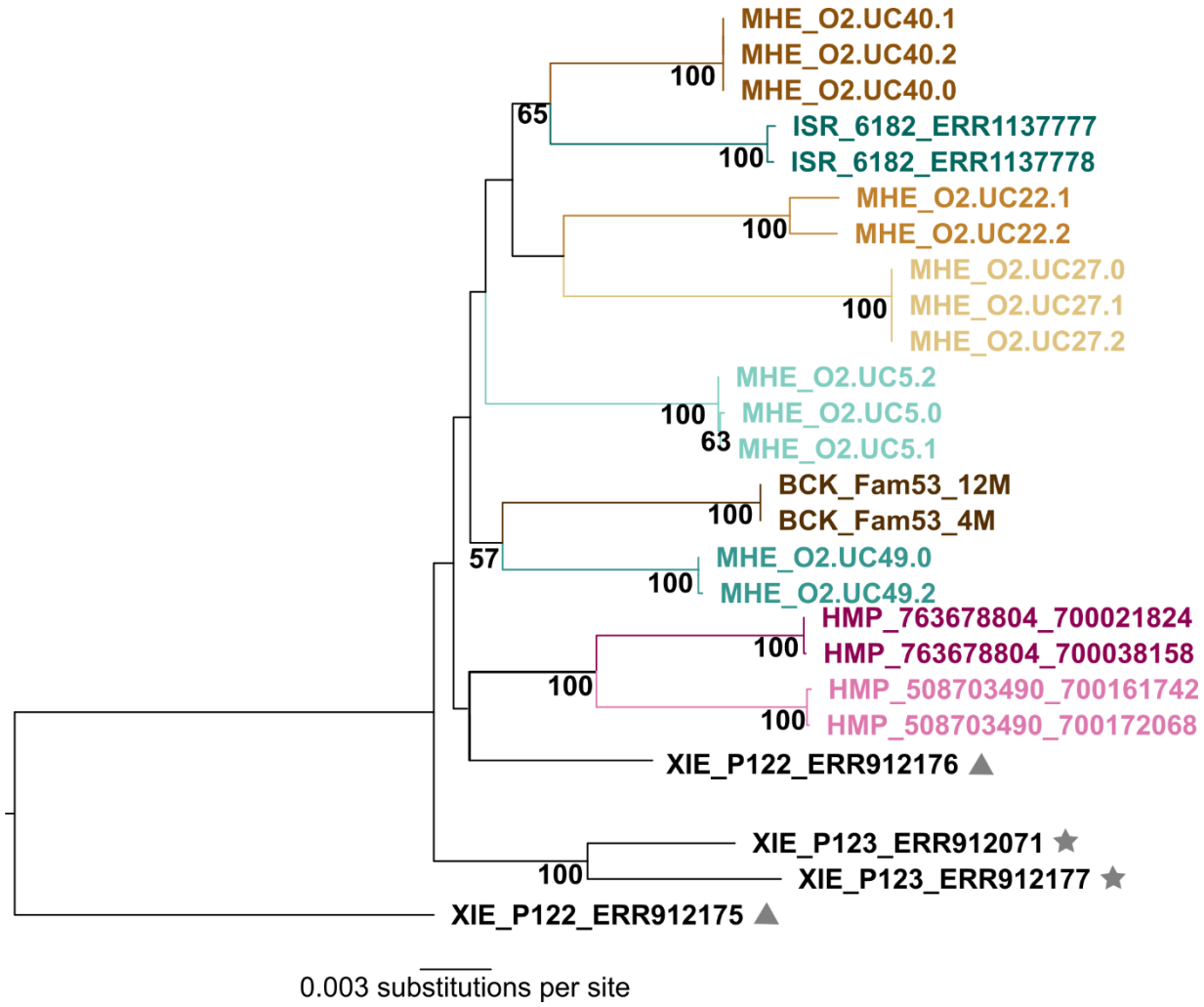
245 Acquisition and persistence of crAssphage strains in the human gut was evaluated using
246 data from studies focusing on 1) mother-infant pairs, 2) twin-pairs, and 3) longitudinal sampling

247 of individuals. We assessed the potential to identify vertical transmission of crAssphage using
248 the BCK dataset [22], which comprises samples from healthy Swedish mothers (N=100) and
249 their infants (N=100) at birth, 4-months, and 12-months of age. CrAssphage prevalence among
250 the healthy mothers was 23%. None of the samples from newborns were considered crAss-
251 positive. Most of the samples from newborn infants showed no reads mapping to crAssphage and
252 no crAssphage proteins recovered from the *de novo* metagenomic assembly approach (S2 Table),
253 with the exception of the infants from Family 549 and Family 263, who showed presence of
254 three and five crAssphage proteins, respectively, at birth. The mothers from both families were
255 also crAss-positive. The infant from Family 263 was determined to be crAss-positive at the four-
256 and 12-month stages, whereas the infant from Family 549 was never crAss-positive, despite
257 showing a steady increase in the number of crAssphage proteins recovered. At the four-month
258 stage, a total of three infants were crAss-positive, all of whom remained crAss-positive at the 12-
259 month stage as well. A total of 16 infants were crAss-positive at the 12-month stage, suggesting
260 that by the end of the first year of life, crAssphage prevalence among infants was similar to that
261 in mothers (chi-square=1.5608; p-value=0.2115, $\alpha$=0.05). Interestingly, in nine out of 16 mother-
262 infant pairs, the 12-month old infants were crAss-positive whereas mothers were completely
263 crAss-negative (no crAssphage proteins were recovered from the samples from the mothers).
264 This supports the hypothesis that crAssphage may be acquired by means other than vertical
265 transmission, as recently reported by [21].

266 To study concordance of crAssphage prevalence among twin-pairs, we screened the XIE
267 dataset comprising samples from twin-pairs from the UK (N=124 pairs). We found 15 cases
268 wherein both twins were crAss-positive and 57 cases wherein both twins were crAss-negative.
269 There were 51 twin-pairs with discordant crAss-positive status. CrAssphage strains from only

270 two crAss-positive twin-pairs could be included in our phylogenetic analyses: in one case, the

271 strains from twin-pair P98 clustered together, whereas those from twin-pair P122 did not (Fig 2).




272

273 **Fig 2. Relationships of crAssphage strains recovered from the same individual and twin-**

274 **pairs.** The Maximum Likelihood tree was based on the multi-gene alignment comprising 12,642

275 nucleotide sites. Sites with missing data were eliminated. The tree was built using the

276 Generalized Time-Reversible model with gamma-distributed rate variation and proportion of

277 invariant sites. Bootstrap support values estimated from 100 replicates are given; only values

278 greater than 50% are shown. CrAssphage strains recovered from multiple samples from the same

individual are color-coded accordingly (MHE, ISR, HMP, and BCK datasets). Symbols are used to denote crAssphage strains recovered from the same twin-pair (XIE dataset).

Additionally, we looked for signatures of crAssphage strain continuity among individuals using the BCK, HMP, and MHE datasets. As seen in Fig 2, crAssphage strains recovered at different time-points from the same individual cluster together, as also reported by [20]. In certain individuals (for example, MHE_O2.UC40), crAssphage strains from different time-points are a 100% identical, suggesting persistence of a single crAssphage strain. In others (for example, MHE_O2.UC22), the strains are closely related but not identical, suggesting that at different time-points, different crAssphage strains might be dominant in an individual.

# Conclusions

The geographic distribution of crAssphage is global [13, 20], but as observed here, the prevalence of crAssphage is lower within samples from more traditional, hunter-gatherer populations such as the Hadza from Tanzania and Matses from Peru. 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, age, sex, or body-size variables. CrAssphage strains from the same individual tend to cluster together phylogenetically. Overall, crAssphage shows limited biogeographic clustering as seen in cases of a recent population expansion event. We estimate that this expansion occurred 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 3,341 samples from were downloaded from the Sequence Read Archive or European Nucleotide Archive (Table 1). Shotgun data were processed using AdapterRemoval v2 [49] 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 then screened for the presence of crAssphage.

## Reference-based mapping

The analysis-ready reads were mapped to the reference crAssphage genome (NC024711.1) using Bowtie2 [37] with the "no-unal" option to discard unmapped reads. The resulting SAM files were processed using SAMTools v1.3 [50], converted into BAM files, quality-filtered at Phred threshold 37, and duplicate reads were removed using rmdup. SAMTools mpileup and VarScan v2.4.3 [51] were used to generate a VCF file with the following parameters: minimum coverage: 10, minimum coverage of variant allele: 3; minimum average quality: 30, minimum variant allele frequency: 0.2, minimum frequency for homozygotes: 0.9, p-value: 1, and strand filter: 0. This VCF file comprised both variant and invariant sites present in the reference crAssphage genome, resulting in a total of 97,065 sites. A custom perl script was used to generate a FASTA file containing the complete genome of the crAssphage strain from the VCF file. The number heterozygous sites was used to assess presence of one or multiple crAssphage strains in the sample.

## *De novo* metagenomic assembly

Analysis-ready reads were assembled into contigs using MEGAHIT [38]. Depth of coverage was calculated by mapping analysis-ready reads to assembled contigs using Bowtie2 [37], followed by processing of resulting alignment files using SAMTools [50] and custom R scripts. Open reading frame (ORF) prediction was carried out using Prodigal [52]. A custom BLAST database was created using the collection of 90 proteins previously predicted from the crAssphage reference genome (NC024711.1) [11]. 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) ≥ 95%, 2) percent identity ≥ 95%, and 3) E value < 1e-5. A sample was considered crAss-positive if matches were recovered for at least 33 reference crAssphage proteins.

## Association of crAssphage prevalence with metadata variables

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 prevalence of crAssphage and health status, age, sex, and BMI categories were assessed using the Chi square test in R [53].

345 ## Gut bacterial taxonomic profiles

346 For the MAT dataset [31], analysis-ready reads for each individual were mapped to the
347 Greengenes database of 16S rRNA gene sequences [54] using Bowtie2 [37]. Unmapped reads
348 were removed using the --no-unal option. The resulting SAM files were converted to BAM files,
349 sorted, and duplicates were removed using SAMTools v1.3 [50]. The sequences of the reads
350 mapping to the Greengenes database were obtained from the BAM files. These sequences were
351 clustered into Operational Taxonomic Units (OTUs) using vsearch [55], employing the
352 cluster_fast algorithm and comparing to the GreenGenes database. Other parameters used
353 included a minimum sequence length of 70 bp and 97% similarity for clustering. The resulting
354 OTU table was rarefied to a depth of 10,000 sequences per sample and singleton OTUs were
355 removed using QIIME [56]. Taxonomic summaries were generated at the phylum- and genus-
356 levels. Relative abundances of genus *Bacteroides* in individuals were determined and plotted
357 using R [53].
358

359 ## Multi-gene phylogenetic analysis

360 For each assembled metagenome, the number of unique ORFs matching each crAssphage
361 reference protein was calculated from the BLASTP results (S2 Table) and used to infer the
362 number of concurrent crAssphage strains carried in the sample. Only samples identified as
363 containing one unique ORF matching each crAssphage reference protein were selected for
364 phylogenetic analysis. Depth of coverage information was used to verify that identified ORFs
365 were representative of the same viral genome.
366 A previously published study demonstrated the use of putative capsid proteins (genes 75,
367 76, 77, 78, and 79) to document diversity among crAss-like phages [11]. Multi-gene

368 phylogenetic analyses were conducted using these capsid protein-encoding genes, as well as five
369 other genes found in members of the crAssphage family. These included genes encoding
370 hypothetical proteins (gene 20 and gene 23), single-stranded DNA-binding protein (gene 21),
371 DNA-G family primase (gene 22), and PD-(D/E)XK family nuclease (gene 85). A subset of 232
372 samples and the reference crAssphage genome was included in this analysis. The gene sequences
373 were aligned separately using MUSCLE v.3.8.31 with default parameters [57] and concatenated
374 together. Sites with missing data and gaps were completely removed. A Maximum Likelihood
375 (ML) tree was built using RAXML v8.2.4 [58], using the Generalized Time-Reversible model
376 with gamma-distributed rate variation and proportion of invariable sites (GTR+G+I). Bootstrap
377 support was estimated from 100 replicates.
378

379 **Genome-based phylogenetic analysis**

380 A subset of 118 samples determined to carry only one crAssphage strain was selected. A
381 whole-genome alignment of all strains was used as input for RAXML [58], and a ML tree was
382 generated using the GTR+G+I model and 100 bootstrap replicates.
383

384 **Timing of crAssphage population expansion**

385 To estimate the timing of a potential population expansion among crAssphage sequences,
386 the multi-gene alignment was used as input for BEAST v1.8.4 [47]. Since substitution rates for
387 crAssphage or crAss-like phages remain undetermined, we used a substitution rate of $1.9 \times 10^{-4}$
388 substitutions per site per year as estimated for bacteriophages belonging to the *Siphoviridae*
389 family [48]. We used a GTR+G+I model of nucleotide substitution, an uncorrelated lognormal
390 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 100,000,000 iterations, sampling every 10,000 steps. The first 10,000,000 iterations were discarded as burn-in. Tracer [59] was used to visualize the results of the MCMC run and generate a Bayesian Skyline Plot (BSP).

# 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. Manrique P, Bolduc B, Walk ST, van der Oost J, de Vos WM, Young MJ. Healthy human gut phageome. Proceedings of the National Academy of Sciences. 2016;113(37):10400.
7. Shkoporov AN, Hill C. Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. Cell Host Microbe. 2019;25(2):195-209.
8. 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.
9. 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.
10. Group NHW, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, et al. The NIH Human Microbiome Project. Genome Res. 2009;19(12):2317-23.
11. 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.
12. Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, et al. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. Nature Communications. 2018;9(1):4781.
13. 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.
14. Holmfeldt K, Solonenko N, Shah M, Corrier K, Riemann L, VerBerkmoes NC, et al. Twelve previously unknown phage genera are ubiquitous in global oceans. Proceedings of the National Academy of Sciences. 2013;110(31):12798.

430 15. Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K, Dinsdale EA, et al. Global
431 phylogeography and ancient evolution of the widespread human gut virus crAssphage. Nature
432 Microbiology. 2019.
433 16. Ahmed W, Lobos A, Senkbeil J, Peraud J, Gallard J, Harwood VJ. Evaluation of the
434 novel crAssphage marker for sewage pollution tracking in storm drain outfalls in Tampa,
435 Florida. Water Res. 2017;131:142-50.
436 17. Garcia-Aljaro C, Balleste E, Muniesa M, Jofre J. Determination of crAssphage in water
437 samples and applicability for tracking human faecal pollution. Microb Biotechnol.
438 2017;10(6):1775-80.
439 18. Stachler E, Kelty C, Sivaganesan M, Li X, Bibby K, Shanks OC. Quantitative
440 CrAssphage PCR Assays for Human Fecal Pollution Measurement. Environ Sci Technol.
441 2017;51(16):9146-54.
442 19. Farkas K, Adriaenssens EM, Walker DI, McDonald JE, Malham SK, Jones DL. Critical
443 Evaluation of CrAssphage as a Molecular Marker for Human-Derived Wastewater
444 Contamination in the Aquatic Environment. Food and Environmental Virology. 2019.
445 20. Edwards R, Vega A, Norman H, Ohaeri MC, Levi K, Dinsdale E, et al. Global
446 phylogeography and ancient evolution of the widespread human gut virus crAssphage. bioRxiv.
447 2019:527796.
448 21. Tamburini FB, Sherlock GJ, Bhatt AS. Transmission and persistence of crAssphage, a
449 ubiquitous human-associated bacteriophage. bioRxiv. 2018:460113.
450 22. Backhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics
451 and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host
452 Microbe. 2015;17(5):690-703.
453 23. Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ, Sutton TDS, et al. Biology
454 and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. Cell
455 Host Microbe. 2018;24(5):653-64 e6.
456 24. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace
457 operational taxonomic units in marker-gene data analysis. The Isme Journal. 2017;11:2639.
458 25. Sankaranarayanan K, Ozga AT, Warinner C, Tito RY, Obregon-Tito AJ, Xu J, et al. Gut
459 Microbiome Diversity among Cheyenne and Arapaho Individuals from Western Oklahoma. Curr
460 Biol. 2015;25(24):3161-9.
461 26. Rampelli S, Schnorr Stephanie L, Consolandi C, Turroni S, Severgnini M, Peano C, et al.
462 Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. Current Biology.
463 2015;25(13):1682-93.
464 27. Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, Reid G, et al. Seasonal
465 cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science.
466 2017;357(6353):802.
467 28. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized
468 Nutrition by Prediction of Glycemic Responses. Cell. 2015;163(5):1079-94.
469 29. Karlsson FH, Tremaroli V, Nookaew I, Bergstrom G, Behre CJ, Fagerberg B, et al. Gut
470 metagenome in European women with normal, impaired and diabetic glucose control. Nature.
471 2013;498(7452):99-103.
472 30. Liu W, Zhang J, Wu C, Cai S, Huang W, Chen J, et al. Unique Features of Ethnic
473 Mongolian Gut Microbiome revealed by metagenomic analysis. Sci Rep. 2016;6:34826.

474 31. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, et
475 al. Subsistence strategies in traditional societies distinguish gut microbiomes. Nat Commun.
476 2015;6:6505.
477 32. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of
478 gut microbiota in type 2 diabetes. Nature. 2012;490(7418):55-60.
479 33. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of
480 human gut microbiome correlates with metabolic markers. Nature. 2013;500(7464):541-6.
481 34. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of
482 reference genes in the human gut microbiome. Nat Biotechnol. 2014;32(8):834-41.
483 35. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al.
484 Identification and assembly of genomes and genetic elements in complex metagenomic samples
485 without using reference genomes. Nat Biotechnol. 2014;32(8):822-8.
486 36. Xie H, Guo R, Zhong H, Feng Q, Lan Z, Qin B, et al. Shotgun Metagenomics of 250
487 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. Cell Syst.
488 2016;3(6):572-84 e3.
489 37. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods.
490 2012;9:357.
491 38. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node
492 solution for large and complex metagenomics assembly via succinct de Bruijn graph.
493 Bioinformatics. 2015;31(10):1674-6.
494 39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search
495 tool. Journal of Molecular Biology. 1990;215(3):403-10.
496 40. Adriaenssens E, Brister JR. How to Name and Classify Your Phage: An Informal Guide.
497 Viruses. 2017;9(4):70.
498 41. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, et al. Gut
499 microbiome of the Hadza hunter-gatherers. Nature Communications. 2014;5:3654.
500 42. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al.
501 Human gut microbiome viewed across age and geography. Nature. 2012;486:222.
502 43. Maixner F, Krause-Kyora B, Turaev D, Herbig A, Hoopmann MR, Hallows JL, et al. The
503 5300-year-old Helicobacter pylori genome of the Iceman. Science. 2016;351(6269):162.
504 44. Shkoporov A, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, et al.
505 ΦCrAss001, a member of the most abundant bacteriophage family in the human gut, infects
506 Bacteroides. bioRxiv. 2018.
507 45. Ozga AT. Viral Metagenomics and Anthropology in the Americas [Ph.D. dissertation].
508 Norman: University of Oklahoma; 2015.
509 46. Gupta VK, Paul S, Dutta C. Geography, Ethnicity or Subsistence-Specific Variations in
510 Human Microbiome Composition and Diversity. Front Microbiol. 2017;8:1162.
511 47. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees.
512 BMC Evolutionary Biology. 2007;7(1):214.
513 48. Kupczok A, Neve H, Huang KD, Hoeppner MP, Heller KJ, Franz C, et al. Rates of
514 Mutation and Recombination in Siphoviridae Phage Genome Evolution over Three Decades.
515 Mol Biol Evol. 2018;35(5):1147-59.
516 49. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming,
517 identification, and read merging. BMC Research Notes. 2016;9.
518 50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
519 Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

51. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research. 2012;22(3):568-76.
52. 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.
53. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.
54. 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.
55. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.
56. 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.
57. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32(5):1792-7.
58. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3.
59. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Systematic Biology. 2018:syy032-syy.

# Supporting information

**S1 Table.** S1 Table contains a list of all 3,341 samples included in this study, associated meta-data, and results of the crAssphage screening using reference-based mapping and *de novo* metagenomic assembly.

**S2 Table.** S2 Table shows the hits for all crAssphage reference proteins across the samples.

**S1 Fig. Relative abundance of genus *Bacteroides* among Matses individuals.** Boxplot denoting percentage relative abundance of *Bacteroides* sp. among Matses individuals. Values corresponding to crAss-positive individuals are denoted in red.

554 **S2 Fig. Multi-gene phylogenetic analysis of 233 crAssphage strains.** The Maximum
555 Likelihood tree was based on the multi-gene alignment comprising 12,642 nucleotide sites. Sites
556 with missing data were eliminated. The tree was built using the Generalized Time-Reversible
557 model with gamma-distributed rate variation and proportion of invariant sites. Bootstrap support
558 values estimated from 100 replicates are given; only values greater than 50% are shown. Strains
559 are color-coded according to the geographic location: dark green – The Americas, light green –
560 Asia, light violet – Europe, and dark purple – Middle East (Israel). The differentially-colored
561 symbols next to the taxa names are used to denote strains from the same individual.
562
563 **S3 Fig. Bayesian Skyline Plot of crAssphage strains.** The X-axis denotes time in years before
564 present (YBP) and Y-axis denotes estimated effective population size. The blue shaded region
565 denotes the 95% Highest Posterior Density interval.