Transmission and persistence of crAssphage, a ubiquitous human-associated bacteriophage

Fiona B. Tamburini[1], Gavin Sherlock[1], Ami S. Bhatt[1,2]*

[1]Department of Genetics, Stanford University, Stanford, CA, USA. [2]Department of Medicine, Division of Hematology, Stanford University, Stanford, CA, USA.

*To whom correspondence may be addressed.

Abstract

The recently discovered crAssphage is by far the most abundant and ubiquitous known human gut bacteriophage. It appears to be highly specific to the human gastrointestinal tract; however, the patterns of transmission and persistence of this bacteriophage are unknown. Here, we identify modes of transmission and describe long-term persistence of crAssphage in several human populations. We find that most humans harbor a single, dominant strain of crAssphage in their microbiome. This is in contrast to the bacterial microbiota, where individuals can harbor a variety of closely- or distantly-related strains of the same bacterial species. We show that crAssphage can be vertically transmitted from mother to infant, acquired through fecal microbiota transplantation, and transmitted in immunocompromised hosts in a hospital setting. We also observe that once a crAssphage strain is acquired, it persists stably within an individual over a timescale of months. These results enhance our understanding of the dynamics of crAssphage, which has emerged as one of the most successful human-associated microbes, and provide a foundation for future studies of the role of this phage in the biology of the human microbiome.

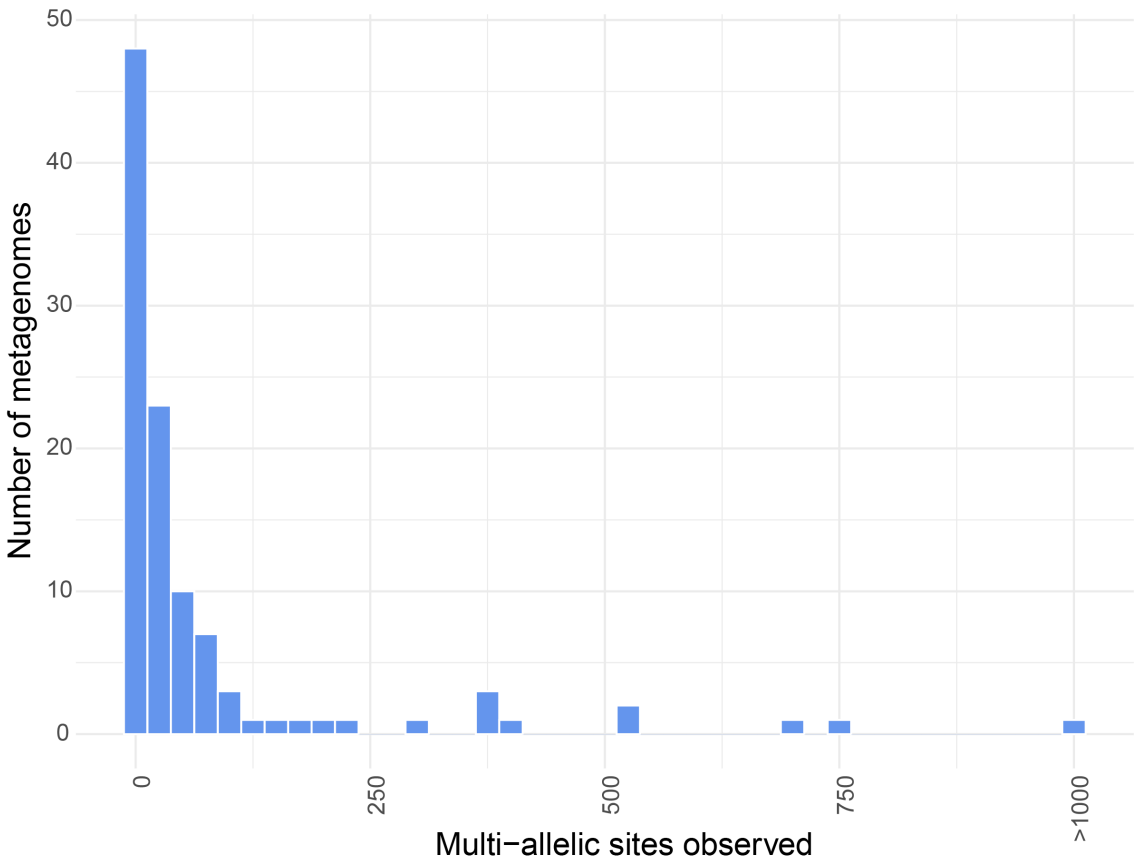
Main text

In addition to trillions of bacteria, the human gastrointestinal tract is densely populated with bacteriophages. Bacteriophages can drive bacterial community composition and mediate horizontal gene transfer[1], and alterations in the human gut virome have been associated with disease[2,3]. Yet, our knowledge of the contributions of specific bacteriophages to human biology is limited, in part due to the paucity of viral sequences represented in reference databases. High-throughput sequencing and advanced genomic tools have facilitated the *in silico* discovery and characterization of previously unknown bacteriophages. The preeminent example of such a discovery is crAssphage (cross Assembly phage), initially identified from human virome sequencing data[4]. CrAssphage is a bacteriophage with an ~97 kilobase circular, double-

stranded DNA genome. Interestingly, crAssphage sequences are found almost exclusively in human fecal metagenomes, and can be highly abundant. Initial estimates indicate that crAssphage is present in 73-77% of humans[4,5], challenging the notion that the gut virome is highly individual-specific. Subsequently, it has been shown that a wide range of crAss-like phages exists in nature[5,6]. However, whether or how crAssphage influences host biology or is involved in disease is unknown. To answer higher-order questions about the role of crAssphage in human biology, it is necessary to establish basic principles of crAssphage acquisition, persistence, and distribution. To this end, we analyzed crAssphage sequences from both published and novel datasets, finding that crAssphage typically exhibits monoclonal dominance in a given individual and can be transmitted vertically from mothers to infants as well as horizontally in adults with compromised or simplified microbiomes.

To determine whether individuals have one or many crAssphage strains in the gut metagenome, we identified sites with more than one variant present in the crAssphage genome (multi-allelic sites) from stool metagenomic sequencing data from individuals from four cohorts (Table S1)[7–9]. We aligned metagenomic sequencing reads to the crAssphage reference genome[4] and identified single-nucleotide variants (SNVs) that are present at intermediate frequency (between 10% and 90% frequency). We limited our analysis to metagenomes with 30X coverage or greater of the crAssphage genome. We find that 77 of 106 metagenomes (73%) have fewer than 50 multiallelic sites (Figure 1), suggesting that most individuals are likely colonized by a single crAssphage or near-identical crAssphages. 28 of 106 metagenomes (26%) have between 50 and 999 multiallelic sites, corresponding to polymorphism in roughly 0.05% to 0.999% of the genome. One of 106 metagenomes (1%) have 1000 or more multiallelic sites, corresponding to greater than 1% of the genome. These results suggest an exclusion principle which favors colonization of a particular strain within the gut of an individual, though notably, a minority of individuals may be simultaneously colonized by more than one crAssphage strain. The number of multi-allelic sites stays relatively stable within individuals over time (Figure S1). We observe that SNVs are relatively evenly distributed throughout the genome (Figure S2).

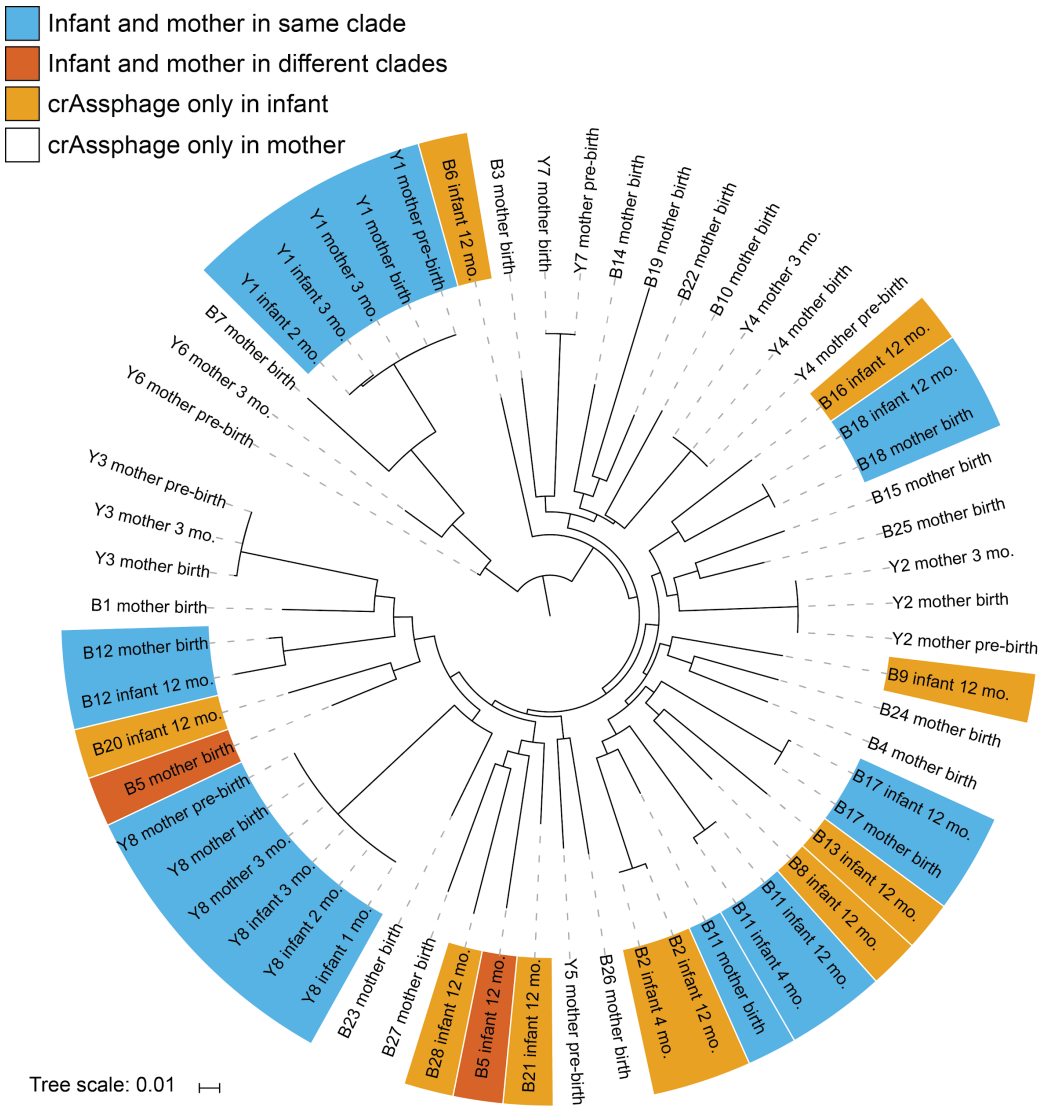
Figure 1: Multi-allelic sites in the crAssphage genome



Histogram showing distribution of the number of multi-allelic sites in the crAssphage genome in metagenomic datasets (n=106). Multi-allelic sites are defined as positions with a non-reference base between 0.1 and 0.9 frequency.

While crAssphage has been detected in infant gut metagenomes[10,11], we do not yet know from where crAssphage is acquired. Given the apparent specificity of crAssphage to the human gut as opposed to other mammals or the environment, we hypothesized that crAsspage is likely acquired through human-to-human contact. It is well-documented that infants acquire many of their first microbes, such as *Bacteroides* species, from their mother during and after delivery[7,8,12,13]. However, it has been shown that adult twins and their mothers have unique gut viromes[14]. Given that *Bacteroides* species are believed to be the bacterial host of crAssphage[4,15], we postulated that crAssphage is vertically transmitted from mother to infant, similar to what is observed for many bacterial taxa and in contrast to what is reported for other members of the human virome. To test the hypothesis that crAssphage is vertically transmitted, we examined publicly available shotgun metagenomic data from two stool microbiome datasets[7,8] from mothers and their infants (n=142 mother-infant pairs). We evaluated crAssphage presence and relatedness using StrainSifter[16], a tool that performs variant calling

86 and phylogenetic analysis of microbial genomes. We considered metagenomes to contain the
87 phage if there were reads mapping to the crAssphage reference genome with at least 5X
88 coverage. We detected crAssphage strains in 27 of the 142 mothers studied (19%) and 16 of
89 142 infants (11%) (Figure 2). Of the 27 mother-infant pairs where crAssphage strains were
90 detected in at least one maternal sample, we find that 6 pairs (22%) share an identical or highly
91 related strain of crAssphage between the mother and infant, indicating that crAssphage can be
92 vertically transmitted from mother to infant. Ten of 142 infants (7%) harbor a strain of
93 crAssphage that is not detected in the mother's stool. This could be a result of sampling during
94 a low-crAssphage state in the mother; alternatively, these infants may have acquired
95 crAssphage from another individual in their household.
96
97 It has previously been reported that birth mode does not influence crAssphage relative
98 abundance in the gut virome of Irish infants[11]. In the two mother-infant cohorts analyzed here,
99 we only detect crAssphage in the gut microbiome of vaginally-born infants. Zero of 22
100 Cesarean-born infants have crAssphage in their stool samples at 5X coverage or greater; this is
101 not statistically significantly within this sample collection (Fisher's exact test; p=0.1332).
102 However, these samples were obtained from highly heterogeneous populations in diverse global
103 regions, and studies of larger cohorts are necessary to definitively determine the relationship
104 between birth mode and crAssphage colonization. Notably, crAssphage is not sufficiently
105 abundant to be detected in meconium or shortly after birth in our analysis and is only found in
106 infant stool sampled at least one month after birth. It is important to note, however, that these
107 datasets comprise total metagenomic shotgun sequencing as opposed to enriched viral
108 particles. It is possible that crAssphage particles are transmitted to the infant during vaginal birth
109 and persist in the infant gut, but that they are not detected until host *Bacteroides* strains achieve
110 higher relative abundance later in development[7,17,18]. Alternatively, it is possible that crAssphage
111 is not sufficiently abundant to be detected in total shotgun metagenomic data in some samples.
112 Finally, we observe that crAssphage strains are maintained in samples from multiple mothers
113 over time (for up to three months of sampling), consistent with previous findings that the human
114 gut virome is stable over time[19,20].
115
116
117 **Figure 2: Vertical transmission of crAssphages**

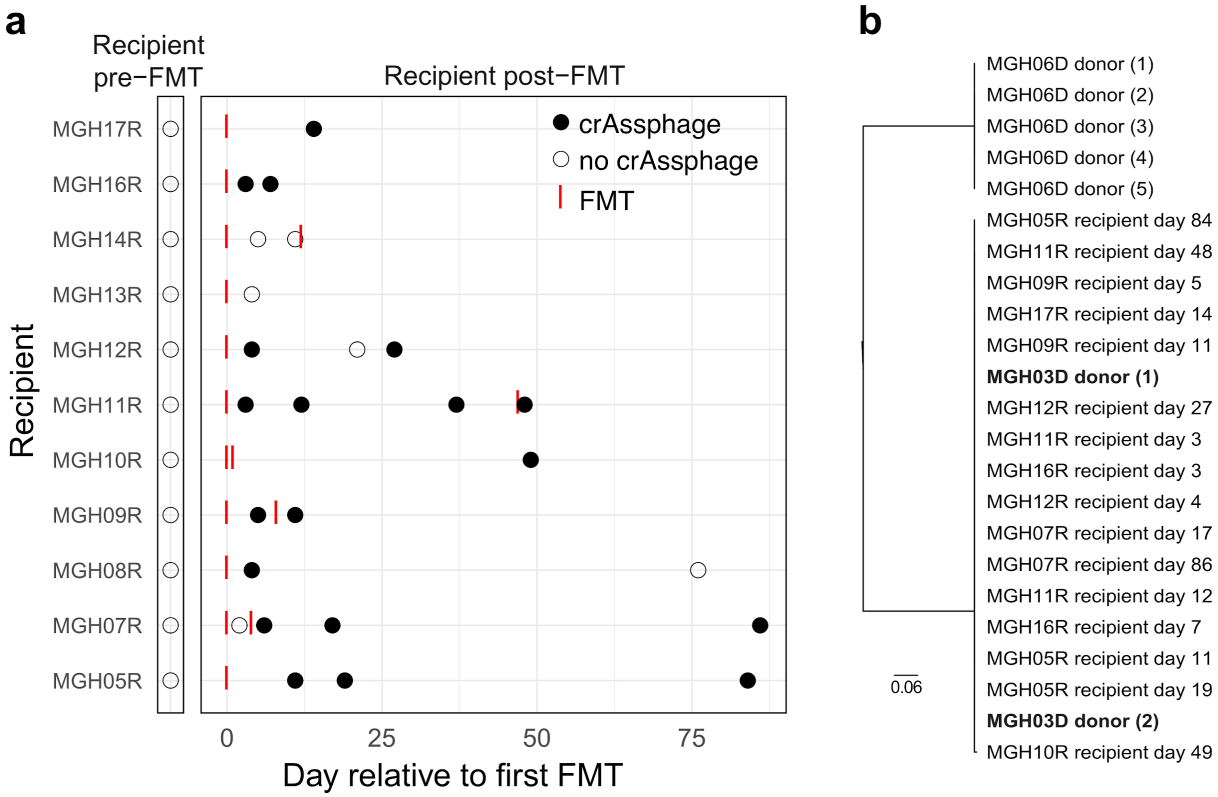


Phylogenetic tree of crAssphages in mothers and infants based on single nucleotide variants. Highlight shows mothers and infants in the same clade of the tree (blue), mothers and their infants with different strains of crAssphage (red), and infants whose corresponding maternal samples are not present on the tree due to low crAssphage abundance or crAssphage absence in the maternal sample (green). Pre-birth samples collected at 27 weeks gestation. Mo.=months. Mapping of sample labels to original sample names show in Table S2.

Clearly crAssphage can be acquired early in life, but we lack longitudinal datasets to determine how stable crAssphage colonization is through development into adulthood. However, the stability of crAssphage strains between serial maternal samples suggests that adults have stable crAssphage populations over at least 3 months. To evaluate whether crAssphage

130 transmission can occur beyond early life, we next asked whether new or different crAssphage
131 strains can be acquired in adulthood.
132
133 Based on the observation that adults typically have a single strain or a population of very closely
134 related strains of crAssphage, we predicted that adult individuals with the greatest likelihood of
135 acquiring a new crAssphage are those who have experienced a dramatic simplification of their
136 microbiome and subsequent exposure to crAssphage which could occupy the newly vacant
137 niche.
138
139 Two examples of such dramatic perturbations to the microbiome are (i) when individuals
140 experience infection with the gut pathogen *Clostridium difficile,* are treated with antibiotics, and
141 subsequently receive fecal microbiota transplantation (FMT)[9,21], and (ii) when individuals
142 undergo hematopoietic cell transplantation (HCT) and associated microbiome disruption from
143 drug exposures and immunosuppression[22,23].
144
145 To evaluate whether crAssphage can be transmitted to adults during fecal microbiota
146 transplantation for the treatment of *C. difficile* associated diarrhea, we examined crAssphage
147 strains in a publicly available dataset of shotgun fecal metagenomes from FMT donors and
148 recipients[9]. Recipients were treated with oral antibiotics including metronidazole, vancomycin,
149 and fidaxomicin prior to transplantation. We observe that crAssphage is present at 1X coverage
150 or greater in samples from two donors, MGH06D and MGH03D. 12 recipients received stool
151 preparations from either of those two donors, and the donor crAssphage engrafts in nine of
152 those recipients (82%) (Figure 3a). Each of those nine recipients received stool preparations
153 from the same donor (MGH03D), and the recipient strain is the same as the donor strain (Figure
154 3b). Zero recipients who received FMT from a crAssphage-negative donor acquired crAssphage
155 during the sampling period. In situations where crAssphage engrafts, it persists for days to
156 months corresponding to the duration of sampling (Figure 3a). None of the recipients in this
157 study had crAssphage in their gut prior to FMT, suggesting that crAssphage is below the
158 detection limit, not as ubiquitous as has been described, or more likely, is cleared or
159 substantially diminished in abundance when individuals are treated with drugs such as
160 metronidazole, which have high activity against the likely crAssphage hosts, *Bacteroides*
161 species. The FMT data show that crAssphage can be acquired as an adult and can stably
162 engraft on a timescale of months.
163

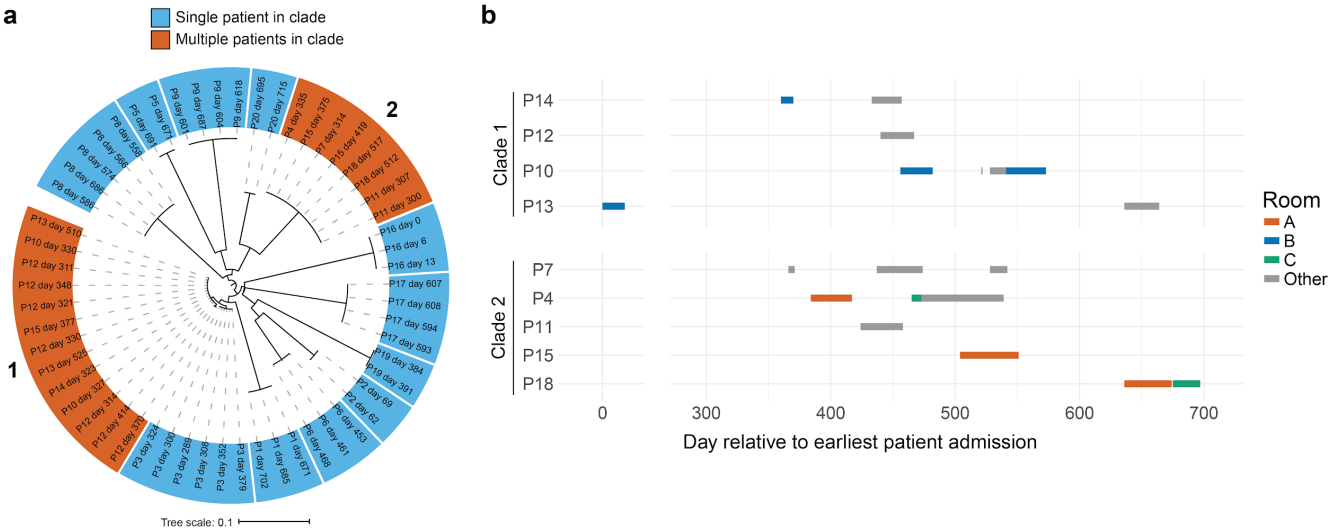
Figure 3: CrAssphage acquisition via FMT



***a**, FMT recipient stool sampling relative to the date of the recipient's first FMT. Day of FMT shown as a solid red line, stool samples shown as circles (black fill indicates crAssphage detected, white fill indicates crAssphage not detected). Only recipients receiving stool preparation from donor MGH03D shown, as recipients from other donors did not acquire crAssphage. **b**, Phylogenetic tree showing strains from crAssphage-positive samples. Bold face indicates samples from the donor in **(a)**.*

The long-term stability of the dominant crAssphage strain that we observed in maternal and FMT recipient subjects suggested that this stability is a characteristic of crAssphage. To further evaluate that model, we next analyzed the persistence of crAssphage strains in a biospecimen collection of longitudinal samples from HCT recipients (Sequencing read counts in table S3). We compared phylogenetic relatedness of crAssphage strains in the gut microbiomes of these patients over the course of sampling (0 to 4 months), and found that in most cases, the dominant crAssphage strain identified in samples from a given individual is identical or near-identical over time, clustering together in a distinct clade (Figure 4a). This suggests that the majority of individuals who were studied have a dominant, stable crAssphage strain over the course of many days to months.

Figure 4: Phylogeny of crAssphages in HCT patients



***a**, Phylogenetic tree of crAssphage strains in the stool of HCT recipients. Clades containing crAssphages from a single patient shown in blue, clades containing crAssphages from unrelated patients shown in red with clades numbered 1 and 2 indicated. Days in are numbered relative to the date of collection of first stool sample in the biospecimen collection. **b**, Hospital room occupancy for patients in clades 1 and 2 from (**a**). Bars indicate the range of occupancy of a patient in a given hospital room, colors of the bars indicate which room a patient inhabited during that stay. Days are numbered relative to the earliest admission date of patients in clades 1 and 2 (P13).*

While most HCT patients with crAssphage had stable colonization over time, there were some notable exceptions. Specifically, we observed two clades which contain identical or near-identical crAssphages from multiple patients (Figure 4a). Intriguingly, these samples were collected at or around the same time, and the individuals' periods of hospitalization overlapped, suggesting potential transmission of these phages from patient to patient, or acquisition from a shared source or the built environment in the hospital. Unfortunately, sampling of the built environment was not carried out during the collection of samples from this cohort. However, we determined the specific hospital rooms that each patient was housed in during their time in the hospital. We observe that three patients from each clade occupied the same room at different times (Figure 4b), raising the possibility that individuals acquired crAssphage from a shared source within that room, or that one individual's crAssphage was left behind and persisted in the hospital room, and was acquired by subsequent occupants. These data suggest that

crAssphage may be acquired through environmental contact, and that crAssphage may be ‘viable’ outside of its likely obligately anaerobic bacterial host or the human body. Furthermore, a substantial inoculum of crAssphage in stool may not be required to transmit a new strain of crAssphage to an individual. This level of exposure is more consistent with the amount that infants experience at birth, a much less dramatic and less direct exposure than FMT.

The results herein demonstrate various modes of acquisition and transmission of crAssphage. We show that crAssphage can be vertically transmitted, and that it is usually present in the gut as early as a few months after birth. We show that crAssphage persists through sampling on the scale of months, and perhaps even longer, indicating that the host immune system tolerates this phage and does not mount an immune response. This may provide an important clue toward the evolutionary history of host and microbe, as the human immune system may have evolved tolerance to this virus. Future work remains to determine precisely whether, and how, crAssphage influences the gut ecosystem. To do so, it will be necessary to determine the host(s) of these crAssphages through isolation and culture experiments. With this knowledge, it will be possible to design and test hypotheses that will help to elucidate the role of crAssphage in the biology of the human gut.

226 **Methods**

227

228 **Hematopoietic cell transplant patient samples**

229 *Cohort selection*

230 Convenience samples were collected from autologous and allogeneic hematopoietic cell
231 transplantation (HCT) patients at Stanford University Hospital under institutional review board
232 protocol 42053 (Principal investigator: Dr. Ami Bhatt) or under protocol 37379 (Principal
233 investigator: Dr. Andrew Rezvani; co-investigator: Ami Bhatt). Informed consent was obtained
234 for all samples collected. Samples were stored at 4°C for up to 1 day after collection and were
235 subsequently aliquoted and stored at -80°C.

236

237 *DNA sequencing*

238 DNA was extracted from HCT patient stool samples using the QIAamp DNA Stool Mini Kit
239 (QIAGEN) according to the manufacturer's instructions, plus a bead-beating step prior to
240 extraction consisting of 7 rounds of 30 seconds of bead-beating followed by 30 seconds of
241 cooling on ice, using the Mini-Beadbeater-16 (BioSpec Products) and 1 mm diameter
242 Zirconia/Silica beads (BioSpec Products). DNA concentration was measured using Qubit
243 Fluorometric Quantitation DS-DNA High Sensitivity Assay (Life Technologies). DNA sequencing
244 libraries were prepared using the Nextera XT DNA Library Prep Kit (Illumina). Final library
245 concentration was measured using Qubit Fluorometric Quantitation and library size distributions
246 was analyzed with the Bioanalyzer 2100 system (Agilent). Libraries were multiplexed and 100-
247 base pair paired-end reads were generated on the HiSeq 4000 platform (Illumina).

248

249 **Computational methods**

250 *Sequence read preprocessing*

251 All sequence data were preprocessed as follows: PCR and optical duplicates were removed
252 with Super Deduper v1.4[24] with the start location at 5 base pairs (-s 5) and minimum read length
253 of 50 base pairs (-l 50). Deduplicated reads were trimmed using Trim Galore v0.4.4[25] with a
254 minimum quality score of 30 for trimming (-q 30), minimum read length of 50 (--length 50) and
255 the "--nextera" flag to remove Illumina Nextera adapter sequences.

256

257 *CrAssphage SNV analysis*

258 Metagenomic reads were aligned to the crAssphage reference genome (NCBI RefSeq
259 NC_024711.1) and variants were identified using snippy with default parameters[26]. For multi-

allelic site analysis, the raw vcf output from snippy was filtered with bcftools[27] to include only SNVs with frequency between 0.1 and 0.9.

Mother-baby and HCT phylogeny

Phylogenetic trees were constructed using the StrainSifter pipeline as previously described[16]. Briefly, reads are aligned to the crAssphage reference genome using the Burrows-Wheeler Aligner v0.7.10[28] and to include only high-confidence alignments with mapping quality of 60 using the SAMtools[29] view and filtered using BamTools[29,30] filter (v2.4.0) to include only reads with 5 or fewer mismatches. Samples in which reads cover at least 50% of the genome at a depth of 5X were included. Pileup files are created from BAM files using SAMtools mpileup, and SNVs with at least 0.8 frequency are identified and concatenated into a fasta file, from which a multiple sequence alignment is created using MUSCLE[29–31] and a phylogenetic tree is computed using FastTree v2.1.7[32]. Phylogenetic trees were midpoint rooted and visualized using the iTOL web tool[33].

FMT analysis

Phylogenetic trees of crAssphage-containing samples were generated using StrainSifter with default parameters. Coverage of reads mapping to the crAssphage genome was determined from StrainSifter output. Plots were generated using the R programming language (v3.4.0) using the ggplot2 v2.2.1[34], reshape2 v1.4.3[35], and dplyr v0.7.4[36]. Sample MGH06R was excluded from the FMT cohort analysis as it could not be definitively determined whether sample designated as pre-FMT was actually collected prior to transplantation (personal communication with authors).

Data availability

Publicly available datasets analyzed herein are show in Table S1. Sequence data unique to this manuscript will be deposited in the NCBI SRA at the time of publication.

Acknowledgements

We thank Tessa Andermann, Joyce Kang, Ekaterina Tkachenko, and Ryan Brewster for collecting and sequencing HCT patient stool samples. We also thank the patients and nurses on the Blood and Marrow Transplantation service for their enthusiastic participation in collecting HCT samples for this project. We thank Dylan Maghini for helpful comments on the manuscript. We thank Ryan deGive for his assistance in accessing patient location information. This work

was supported in part by the National Science Foundation Graduate Research Fellowship and the Stanford CEHG Pre-Doctoral Fellowship (F. B. T.) as well as the National Cancer Institute NIH K08 CA184420, the Damon Runyon Clinical Investigator Award, and the Amy Strelzer Manasevit Award (A. S. B.).

Author contributions

F. B. T. performed computational analysis. G. S. designed the multi-allelic site analysis and edited the manuscript. F. B. T. and A. S. B designed the study and wrote and edited the manuscript.

303 **References**

304 1. Touchon, M., Moura de Sousa, J. A. & Rocha, E. P. Embracing the enemy: the
305 diversification of microbial gene repertoires by phage-mediated horizontal gene transfer.
306 *Curr. Opin. Microbiol.* **38,** 66–73 (2017).
307 2. Cadwell, K. *et al.* Virus-plus-susceptibility gene interaction determines Crohn's disease
308 gene Atg16L1 phenotypes in intestine. *Cell* **141,** 1135–1145 (2010).
309 3. Zhao, G. *et al.* Intestinal virome changes precede autoimmunity in type I diabetes-
310 susceptible children. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E6166–E6175 (2017).
311 4. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences
312 of human faecal metagenomes. *Nat. Commun.* **5,** 4498 (2014).
313 5. Guerin, E. *et al.* Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant
314 Virus in the Human Gut. *Cell Host Microbe* (2018). doi:10.1016/j.chom.2018.10.002
315 6. Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most
316 abundant viruses from the human gut. *Nat Microbiol* **3,** 38–46 (2018).
317 7. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the
318 First Year of Life. *Cell Host Microbe* **17,** 852 (2015).
319 8. Yassour, M. *et al.* Strain-Level Analysis of Mother-to-Child Bacterial Transmission during
320 the First Few Months of Life. *Cell Host Microbe* **24,** 146–154.e4 (2018).
321 9. Smillie, C. S. *et al.* Strain Tracking Reveals the Determinants of Bacterial Engraftment in
322 the Human Gut Following Fecal Microbiota Transplantation. *Cell Host Microbe* **23,** 229–
323 240.e5 (2018).
324 10. Liang, Y. Y., Zhang, W., Tong, Y. G. & Chen, S. P. crAssphage is not associated with
325 diarrhoea and has high genetic diversity. *Epidemiol. Infect.* **144,** 3549–3553 (2016).
326 11. McCann, A. *et al.* Viromes of one year old infants reveal the impact of birth mode on
327 microbiome diversity. *PeerJ* **6,** e4694 (2018).
328 12. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes

329 the Developing Infant Gut Microbiome. *Cell Host Microbe* **24,** 133–145.e5 (2018).
330 13. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the
331 initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U. S. A.*
332 **107,** 11971–11975 (2010).
333 14. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers.
334 *Nature* **466,** 334–338 (2010).
335 15. Shkoporov, A. *et al.* ΦCrAss001, a member of the most abundant bacteriophage family in
336 the human gut, infects Bacteroides. *bioRxiv* 354837 (2018). doi:10.1101/354837
337 16. Tamburini, F. B. *et al.* Precision identification of diverse bloodstream pathogens in the gut
338 microbiome. *Nat. Med.* (2018). doi:10.1038/s41591-018-0202-8
339 17. Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic
340 treatment on bacterial strain diversity and stability. *Sci. Transl. Med.* **8,** 343ra81 (2016).
341 18. Chu, D. M. *et al.* Maturation of the infant microbiome community structure and function
342 across multiple body sites and in relation to mode of delivery. *Nat. Med.* **23,** 314–326
343 (2017).
344 19. Minot, S. *et al.* Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U. S. A.*
345 **110,** 12450–12455 (2013).
346 20. Minot, S. *et al.* The human gut virome: Inter-individual variation and dynamic response to
347 diet. *Genome Res.* **21,** 1616–1625 (2011).
348 21. Moss, E. L. *et al.* Long-term taxonomic and functional divergence from donor bacterial
349 strains following fecal microbiota transplantation in immunocompromised patients. *PLoS*
350 *One* **12,** e0182585 (2017).
351 22. Taur, Y. *et al.* Intestinal domination and the risk of bacteremia in patients undergoing
352 allogeneic hematopoietic stem cell transplantation. *Clin. Infect. Dis.* **55,** 905–914 (2012).
353 23. Shono, Y., Docampo, M. D., Peled, J. U., Perobelli, S. M. & Jenq, R. R. Intestinal
354 microbiota-related effects on graft-versus-host disease. *Int. J. Hematol.* **101,** 428–437

(2015).

24. Petersen, K. R., Streett, D. A. & Gerritsen, A. T. Super deduper, fast PCR duplicate detection in fastq files. *Proceedings of the 6th* (2015).

25. Krueger, F. Trim Galore! Available at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore.

26. Seeman, T. *snippy: fast bacterial variant calling from NGS reads*. (2015).

27. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

28. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* (2009).

29. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

30. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27,** 1691–1692 (2011).

31. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

32. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5,** e9490 (2010).

33. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44,** W242–5 (2016).

34. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2009).

35. Wickham, H. reshape2: Flexibly reshape data: a reboot of the reshape package, version 1.4. 2. *See https://cran. r-project. org/web/packages/reshape2/index. html* (2016).

36. Wickham, H., Francois, R., Henry, L. & Müller, K. *dplyr: A Grammar of Data Manipulation*. (2017).