Title: Global phylogeography and ancient evolution of the widespread human gut virus crAssphage

One Sentence Summary:

Bacteriophage found in human gut is part of a globally shared virome that has co-evolved with hominids for millions of years.

Authors:

Edwards, Robert.A[1,2,*], Vega, Alejandro.A[1], Norman, Holly.M[1], Ohaeri, Maria[1], Levi, Kyle[3], Dinsdale, Elizabeth.A[1], Cinek, Ondrej[4], Aziz, Ramy.K[5], McNair, Katelyn[2], Barr, Jeremy.J[6], Bibby, Kyle[7], Brouns, Stan.JJ[8], Cazares, Adrian[9], de Jonge, Patrick.A[10,8], Desnues, Christelle[11], Díaz Muñoz, Samuel.L[12,13], Fineran, Peter.C[14], Kurilshikov, Alexander[15], Lavigne, Rob[16], Mazankova, Karla[4], McCarthy, David[17], Nobrega, Franklin.L[8], Reyes Muñoz, Alejandro[18], Tapia, German[19], Trefault, Nicole[20], Tyakht, Alexander.V[21,22], Vinuesa, Pablo[23], Wagemans, Jeroen[16], Zhernakova, Alexandra[15], Aarestrup, Frank.M[24], Ahmadov, Gunduz[25], Alassaf, Abeer[26], Anton, Josefa[27], Asangba, Abigail[28], Billings, Emma[1], Cantu, Vito Adrian[2], Carlton, Jane.M[12], Cazares, Daniel[23], Cho, Gyu-Sung[29], Condeff, Tess[1], Cortés, Pilar[30], Cranfield, Mike[31], Cuevas, Daniel.A[2], De la Iglesia, Rodrigo[32], Decewicz, Przemyslaw[33], Doane, Michael.P[1], Dziewit, Lukasz[33], Elwasila, Bashir Mukhtar[34], Eren, Murat[35], Franz, Charles[29], Fu, Jingyuan[36], Garcia-Aljaro, Cristina[37], Ghedin, Elodie[12], Gulino, Kristen.M[12], Haggerty, John.M[1], Head, Steven.R[38], Hendriksen, Rene.S[24], Hill, Colin[39], Hyöty, Heikki[40], Ilina, Elena.N[41], Irwin, Mitchell.T[42], Jeffries, Thomas[43], Jofre Torroella, Juan[37], Junge, Randall.E[44], Kelley, Scott.T[1], Kowalewski, Martin[45], Kumaresan, Deepak[46], Leigh, Steven[47], Lisitsyna, Eugenia.S[48], Llagostera, Montserrat[30], Manor, Joseph[49], Maritz, Julia.M[12], Marr, Linsey.C[50], McCann, Angela[51], Mirzaei, Mohammadali Khan[52], Molshanski-Mor, Shahar[53], Monteiro, Silvia[54], Moreira-Grez, Ben[46], Morris, Megan[1], Mugisha, Lawrence[55,56], Muniesa, Maite[37], Neve, Horst[29], Nguyen, Nam-phuong[57], Nigro, Olivia.D[58], Nilsson, Anders.S[52], O'Connell, Taylor[59], Odeh, Rasha[26], Oliver, Andrew[60], Piuri, Mariana[61], Prussin II, Aaron.J[50], Qimron, Udi[62], Quan, Zhe-Xue[63], Rainetova , Petra[64], Ramírez Rojas, Adán Andrés[65], Raya, Raul[66], Rossi, Alessandro[10,67], Santos, Ricardo[54], Shimashita, John[50], Stachler, Elyse.N[68], Stene, Lars.C[19], Steward, Grieg[69], Strain, Ronan[51], Stumpf, Rebecca[28], Torres, Pedro.J[1], Twaddle, Alan[12], Ugochi Ibekwe, MaryAnn[70], Villagra, Nicolás[71], Wandro, Stephen[60], White, Bryan[28], Whitely, Andy[46], Whiteson, Katrine.L[60], Wijmenga, Cisca[15], Zambrano, Maria.M[65], Zschach, Henrike[72], and Dutilh, Bas.E[10,73,*]

* Corresponding Authors:
Dr. Robert A. Edwards

Departments of Computer Science and Biology
5500 Campanile Drive
San Diego State University, San Diego CA 92182 USA
Tel: +1 619 594 1672
Email: redwards@sdsu.edu

Dr. Bas E. Dutilh
Theoretical Biology and Bioinformatics
Utrecht University
Padualaan 8 3584 CH
Utrecht, The Netherlands
Tel: +31 30 253 4212
Email: bedutilh@gmail.com

Affiliations:

[1]Department of Biology, San Diego State University, 5500 Campanile Dr., San Diego, CA, 92182, USA
[2]Computational Sciences Research Center, San Diego State University, 5500 Campanile Dr., San Diego, CA, 92182, USA
[3]Department of Computer Science, San Diego State University, 5500 Campanile Dr., San Diego, CA, 92182, USA
[4]Department of Pediatrics, Charles University in Prague, Ovocný trh 3-5, Prague, Czech Republic
[5]Department of Microbiology and Immunology, Cairo University, Qasr El-Ainy Street, Cairo, 11562, Egypt
[6]School of Biological Sciences, Monash University, 17 Rainforest Walk, School of Biological Sciences, Clayton, VIC, 3800, Australia
[7]Civil and Environmental Engineering and Earth Sciences, University of Notre Dame, 156 Fitzpatrick Hall, Notre Dame, IN, 46556, USA
[8]Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, Van der Maasweg 9, Delft, 2629 HZ, The Netherlands
[9]Institute of Infection and Global Health, University of Liverpool, 8 W Derby St, Liverpool, L7 3EA, UK
[10]Theoretical Biology and Bioinformatics, Science4Life, Utrecht University, Padualaan 8, Utrecht, 3584 CH, The Netherlands
[11]Microbes, Evolution, Phylogeny and Infection (MEPHI), Aix-Marseille Université, IRD, AP-HM, CNRS, IHU Méditerranée Infection, 19-21 Bd Jean Moulin, Marseille, PACA, 13005, France
[12]Center for Genomics and Systems Biology & Department of Biology, New York University, 12 Waverly Place, New York, NY, 10003, USA
[13]Department of Microbiology and Molecular Genetics, University of California, Davis, One Shields Avenue, Davis, CA, 95616, USA
[14]Department of Microbiology and Immunology, University of Otago, PO Box 56, Dunedin, 9054, New Zealand
[15]Department of Genetics, University Medical Center Groningen, Hanzeplein 1, Groningen, 9713 GZ, Netherlands

[16]Department of Biosystems,, KU Leuven, Kasteelpark Arenberg 21 - box 2462, Leuven, B-3001, Belgium
[17]Environmental and Public Health Microbiology Laboratory (EPHM Lab), Civil Engineering Department, Monash University, Building 60, Wellington Road, Clayton, VIC, 3800, Australia
[18]Max Planck Tandem Group in Computational Biology, Departamento de Ciencias Biológicas, Universidad de los Andes, Carrera 1 #18A-12, Bogotá, 111711, Colombia
[19]Department of Child Health, Norwegian Institute of Public Health, Marcus Thranes gate 6, Oslo, NO-0473, Norway
[20]GEMA Center for Genomics, Ecology & Environment, Universidad Mayor, Camino La Pirámide 5750, Huechuraba, Santiago, Chile
[21]Laboratory of Bioinformatics, Federal Research and Clinical Center of Physical-Chemical Medicine, Malaya Pirogovskaya 1a, Moscow, 119435, Russia
[22]Department of Informational Technologies, ITMO University, 49 Kronverksky Pr., Saint-Petersburg, 197101, Russia
[23]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Av. Universidad s/n, Col. Chamilpa, C.P. 62210, Cuernavaca, Morelos, 62210, México
[24]National Food Institute, Research Group for Genomic Epidemiology, Technical University of Denmark, Søltofts Plads 221, Kongens Lyngby, 2800, Denmark
[25]Endocrine Centre Baku, Str. I. Hashimov 4A, Baku, AZ1114, Azerbaijan
[26]Department of Pediatrics, Jordan University Hospital, Queen Rania Street, Amman, 11942, Jordan
[27]Department of Physiology, Genetics and Microbiology, University of Alicante, Carretera San Vicente del Raspeig, 03080 Alicante, Spain
[28]Carl R. Woese Institute of Genomic Biology, University of Illinois at Urbana Champaign, 607 S Mathews Ave, Urbana, IL, 61801, USA
[29]Department of Microbiology and Biotechnology, Max Rubner-Institut, Federal Research Institute of Nutrition and Food, Hermann-Weigmann-Straße 1 , Kiel, 24103, Germany
[30]Departament de Genètica i de Microbiologia, Universitat Autònoma De Barcelona, Avinguda de l'Eix Central / Carrer dels Til·lers, Bellaterra, 8193, Spain
[31]Wildlife Health Center, University of California, Davis, One Shields Ave, Davis, CA, 95616, USA
[32]Genética Molecular y Microbiología, Pontificia Universidad Católica de Chile, Portugal 49, Santiago, Metropolitana, 8330025, Chile
[33]Department of Bacterial Genetics, Institute of Microbiology, Faculty of Biology, University of Warsaw, Miecznikowa 1, Warsaw, 02-096, Poland
[34]Department of Pediatrics and Child Health, University of Khartoum, Faculty of Medicine, El Qasr Ave, Khartoum, Sudan
[35]Knapp Center for Biomedical Discovery, University of Chicago, 900 E. 57th St., Chicago, IL, 60637, USA
[36]Department of Pediatrics, University Medical Center Groningen, Hanzeplein 1, Groningen, 9713 GZ, Netherlands
[37]Department of Genetics, Microbiology and Statistics, Universitat de Barcelona, Av. Diagonal, 643 Barcelona, 08028, Spain
[38]Next Generation Sequencing and Microarray Core Facility, The Scripps Research Institute, 10550 N Torrey Pines Rd, La Jolla, CA, 92037, USA
[39]School of Microbiology, University College Cork, BioSciences Building, Western Road, Cork, T12 YT20, Ireland
[40]Department of Virology, School of Medicine, University of Tampere, Kalevantie 4, Tampere,

FI33520, Finland
[41]Department of Molecular Biology and Genetics, Federal Research and Clinical Center of Physical-Chemical Medicine, Malaya Pirogovskaya 1a, Moscow, 119435, Russia
[42]Anthropology, Northern Illinois University, 1425 W Lincoln Hwy, DeKalb, IL, 60115, USA
[43]School of Science and Health, Western Sydney University, Locked Bag 1797, Penrith, NSW, 2751, Australia
[44]Department of Animal Health, Columbus Zoo and Aquarium, 9990 Riverside Drive, Powell, OH, 43065, USA
[45]Department Estacion Biologica Corrientes, Institution Museo Arg. Cs. Naturales-CONICET, 9 de Julio 392 8D, Corrientes, 3400, Argentina
[46]School of Earth and Environment, University of Western Australia, 35 Stirling Highway, Perth, WA, 6009, Australia
[47]Department of Anthropology, not College, University of Colorado, Boulder, 1350 Pleasant St., Boulder, CO, 80309, USA
[48]Department of Research and Development, Lytech Ltd., Malaya Semenovskaya 3a, Moscow, 107023, Russia
[49]Central Virology Laboratory, Chaim Sheba Medical Center, Tel-Hashomer Hospital, Derech Sheba 2, Tel Aviv, 52621, Israel
[50]Department of Civil and Environmental Engineering, Virginia Tech, 750 Drillfield Drive, Blacksburg, VA, 24060, USA
[51]APC Microbiome Institute, University College Cork, Room 3.39 BioSciences Building, Cork, Ireland
[52]Department of Molecular Biosciences, Stockholm University, Svante Arrhenius väg 20C, Stockholm, Sweden
[53]Clinical Microbiology & Immunology, Sackler school of Medicine, Tel-Aviv University, 55 Levanon St., Tel-Aviv, 6997878, Israel
[54]Laboratorio de Analises, Instituto Superior Tecnico, Av. Rovisco Pais, Lisboa, Lisboa, 1049-001, Portugal
[55]Conservation and Ecosystem Health Alliance (CEHA), P.O. Box, 34153, Kampala, Uganda
[56]College of Veterinary Medicine, Animal Resources & Biosecurity (COVAB), Makerere University, P.O. Box 7062, Kampala, Uganda
[57]Computer Science and Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093-0404, USA
[58]Department of Biology, Hawai'i Pacific University, 45-045 Kamehameha Highway, Ste. 206, Kaneohe, Hi, 96744-5297, USA
[59]Biological and Medical Informatics Program, San Diego State University, 5500 Campanile Dr., San Diego, CA, 92182, USA
[60]Department of Molecular Biology & Biochemistry, University of California, Irvine, 3315 McGaugh Hall, Irvine, CA, 92697, USA
[61]Departamento de Química Biológica, Ciudad Universitaria, Intendente Güiraldes 2160, Ciudad Autónoma de Buenos Aires, 1428, Argentina
[62]Department of Clinical Microbiology and Immunology, Sackler School of Medicine, Tel Aviv University, 55 Levanon St., Tel Aviv, 69978, Israel
[63]Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, Fudan University, 2005 Songhu-Road, Shanghai, 200438, P. R. China
[64]Centre of Epidemiology and Microbiology, National Institute of Public Health, Šrobárova 49/48, Prague, Czech Republic
[65]Molecular Genetics, Corporación Corpogen, Carrera 5 #66A-34, Bogotá, DC, 110231,

Colombia
[66]CERELA, Chacabuco 145 - (T4000ILC), San Miguel de Tucumán, Tucumán, Argentina
[67]Molecular medicine, University of Padova, Via Gabelli 63, Padova, Veneto, 35121, Italy
[68]Swanson School of Engineering, University of Pittsburgh, 3700 O'Hara Street, Pittsburgh, PA, 15261, USA
[69]Department of Oceanography, University of Hawai'i, Manoa, 1950 East-West Road, Honolulu, Hi, 96822, USA
[70]Department of Pediatrics, Federal Teaching Hospital Abakaliki, Ebonyi State University, Udensi Road, Abakaliki, Nigeria
[71]Lab. Patogénesis Molecular y Antimicrobianos, Universidad Andres Bello, Av. República 239, Santiago, Chile
[72]Department of Bio and Health Informatics, Technical University of Denmark, Kemitorvet 208, Kongens Lyngby, 2800, Denmark
[73]Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Centre, Geert Grooteplein 26, Nijmegen, 6525 GA, The Netherlands

Abstract:

Microbiomes are vast communities of microbes and viruses that populate all natural ecosystems. Viruses have been considered the most variable component of microbiomes, as supported by virome surveys and examples of high genomic mosaicism. However, recent evidence suggests that the human gut virome is remarkably stable compared to other environments. Here we investigate the origin, evolution, and epidemiology of crAssphage, a widespread human gut virus. Through a global collaboratory, we obtained DNA sequences of crAssphage from over one-third of the world's countries, showing that its phylogeography is locally clustered within countries, cities, and individuals. We also found colinear crAssphage-like genomes in both Old-World and New-World primates, challenging rampant viral genomic mosaicism and suggesting that the association of crAssphage with hominids may be millions of years old. We conclude that crAssphage is a benign globetrotter virus that has co-evolved with the human lineage and an integral part of the normal human gut virome.

Main Text:

Phages, viruses that infect bacteria and archaea, are considered to be the most diverse organisms in any ecosystem, including the human microbiome. They are critical for the control of bacterial populations in the human intestine, with an estimated ~5 x 10^9 phages per gram of human feces versus ~9 x 10^{10} bacteria(*1*, *2*). Phages are thought of as transient killers that decimate the most abundant bacterial hosts in a population before waning, allowing resistant strains to emerge and contributing to a dynamic and diverse microbial community(*3*–*5*).

Evolutionary and genomic studies have suggested that these dynamic phage-host interactions are reflected in phage genomes, which show high sequence diversity and mosaicism(*6*, *7*). Studies of native phages in marine aquatic ecosystems have shown that they only persist in the environment for one to two days(*8*–*10*), but those dynamics may be drastically different in the human gut microbiome, where phages can persist for over a year(*11*, *12*). Moreover, several studies have found phages that are widespread and shared among the microbiomes of different individuals(*13*, *14*), although the intestinal virome can also differ dramatically between people(*11*, *12*).

CrAssphage is stable in the human gut

We assessed the origin, evolution, and epidemiology of one of the most ubiquitous human gut viruses to understand the stability of the human gut virome. We previously recovered the crAssphage sequence from over half of 466 fecal metagenomics datasets(*13*). This data allowed us to screen the crAssphage genome for regions that were present in many different datasets, where variable segments were flanked by conserved regions suitable for targeting by PCR primers, identifying three amplicon regions of ~1.3 kilobases (see Methods). We tested fecal samples from 45 healthy individuals from four cities on two continents and found that almost half of these volunteers (21 individuals) were crAss-positive. We followed six individuals over two months, showing that crAssphage status was quite stable in time (Fig. 1). While the titers sometimes fell below the detection limit likely due to sampling bias and/or ecological dynamics, DNA sequencing revealed that crAssphage strains from one individual tend to be phylogenetically clustered. To confirm this, we recovered twenty different crAssphage genomes from the fecal viromes of three adult female twin pairs and their mothers, using the same

datasets that we originally used to discover crAssphage(*11*, *13*) and built a phylogenomic tree. Genomes sampled up to one year apart from the same individual clustered together in the tree (Fig. S1), consistent with a model of intra-individual evolution of gut virome populations that are stable in time(*11*, *12*). CrAss-positive individuals probably acquire crAssphage at a young age, and its specificity for the human gut and sewage shows that humans are the major known reservoir(*13*, *15*–*17*), although it is occasionally found in wastewater from non-human sources(*17*) and was recently identified in a termite gut metagenome from a New Orleans city park(*18*).

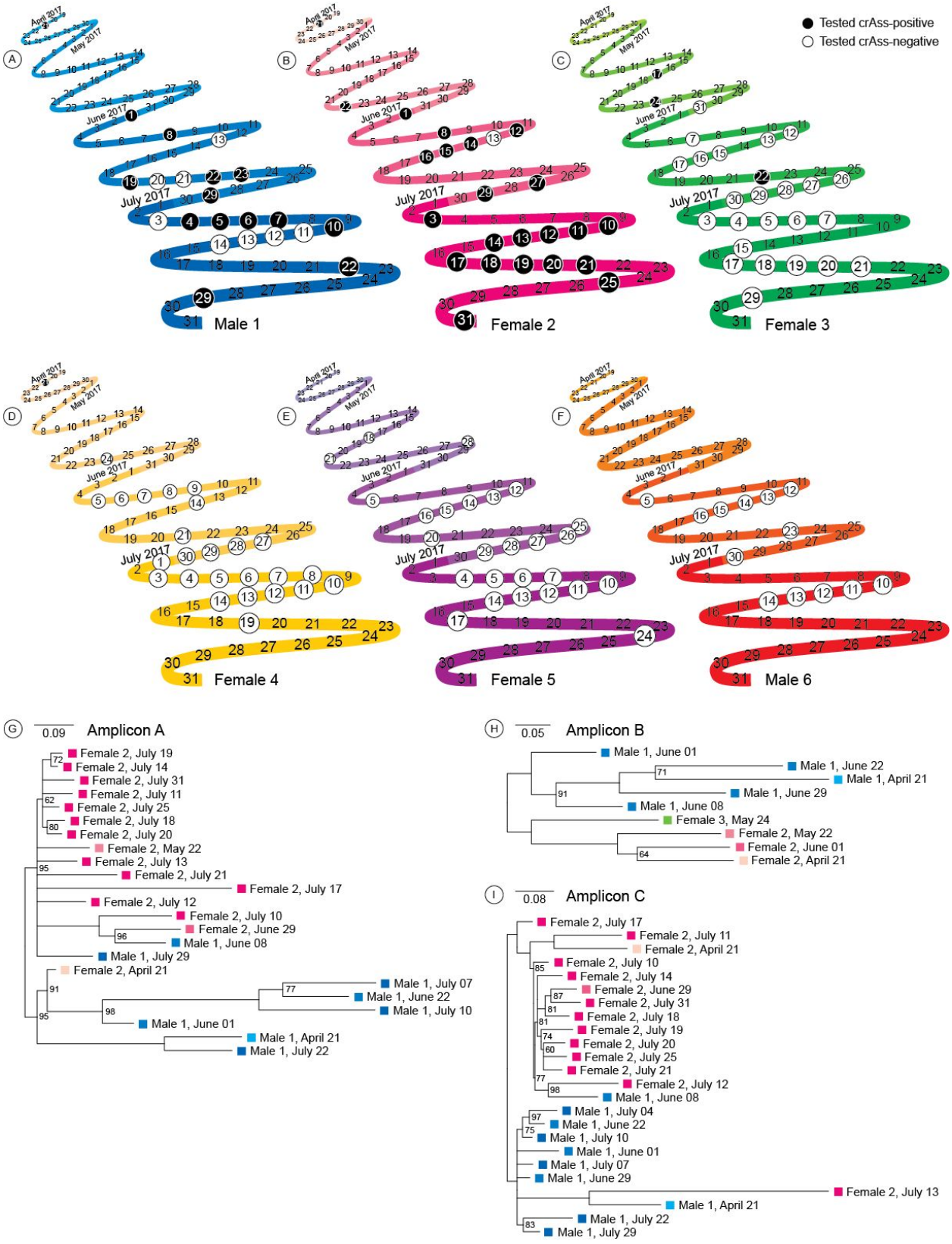
Tested crAss-positive
Tested crAss-negative
A
B
C
D
E
F
G
H
I
April 2017
May 2017
June 2017
July 2017
Male 1
Female 2
Female 3
Female 4
Female 5
Male 6
0.09
Amplicon A
Female 2, July 19
Female 2, July 14
Female 2, July 31
Female 2, July 11
Female 2, July 25
Female 2, July 18
Female 2, July 20
Female 2, May 22
Female 2, July 13
Female 2, July 21
Female 2, July 17
Female 2, July 12
Female 2, July 10
Female 2, June 29
Male 1, June 08
Male 1, July 29
Female 2, April 21
Male 1, July 07
Male 1, June 22
Male 1, July 10
Male 1, June 01
Male 1, April 21
Male 1, July 22
0.05
Amplicon B
Male 1, June 01
Male 1, June 22
Male 1, April 21
Male 1, June 29
Male 1, June 08
Female 3, May 24
Female 2, May 22
Female 2, June 01
Female 2, April 21
0.08
Amplicon C
Female 2, July 17
Female 2, July 11
Female 2, April 21
Female 2, July 10
Female 2, July 14
Female 2, June 29
Female 2, July 31
Female 2, July 18
Female 2, July 19
Female 2, July 20
Female 2, July 25
Female 2, July 21
Female 2, July 12
Male 1, June 08
Male 1, July 04
Male 1, June 22
Male 1, July 10
Male 1, June 01
Male 1, July 07
Male 1, June 29
Female 2, July 13
Male 1, April 21
Male 1, July 22
Male 1, July 29

Fig. 1. CrAssphage presence/absence status is stable over time in the human gut. A-F: Timelines showing the crAssphage status of six volunteers between April and July 2017. Circled dates were tested, black and white circles indicating crAss-positive and crAss-negative samples, respectively. CrAssphage status was always consistent between amplicons A, B, and C. G-I: Unrooted maximum likelihood phylogenies of amplicons A-C show clustering of the sequences by volunteer (note: not all crAss-positives could be sequenced). Branches with <60% bootstrap support were collapsed, values <100% are displayed.

CrAssphage is globally distributed and locally clustered

The phylogenies in Fig. 1G-I and Fig. S1 suggested that individuals have a dominant and stable crAssphage population in their gut microbiome, but these results might be skewed by PCR amplification or metagenome assembly. While higher order groups including species and genera remain controversial in viral taxonomy(*19*), strains can readily be defined as unique sequences(*20*). To analyze how many strains could co-occur within one sample, we downloaded 95,552 metagenomics datasets from all environments from the Sequence Read Archive(*21*). Using a strain-resolved bioinformatics pipeline developed for this analysis(*22*) (see Methods), we extracted the three amplicon regions from 2,216 datasets, most of which contained only a single crAssphage strain (Fig. 2). One strain of amplicon C was independently identified up to 104 times in different datasets (listed in Supplementary File 1), showing the ubiquity of some strains around the world. It has been suggested that crAssphage is not acquired early in life(*23*), but our global analysis identified crAssphage in at least 134 infant samples (26 with locality information, see Supplementary File 2), confirming recent incidental findings(*23*, *24*). Interestingly, the two samples with the most diverse crAssphage populations

are from young individuals, including a healthy USA child(*25*) containing up to 1,409 strains and a one-year old Finnish infant(*26*) containing up to 748 strains (Fig. 2; Supplementary File 3). Still, our phylogenomic tree based on the twin study(*11*) suggests that crAssphage is not always vertically transmitted since none of the daughter strains cluster with their mothers (Fig. S1).

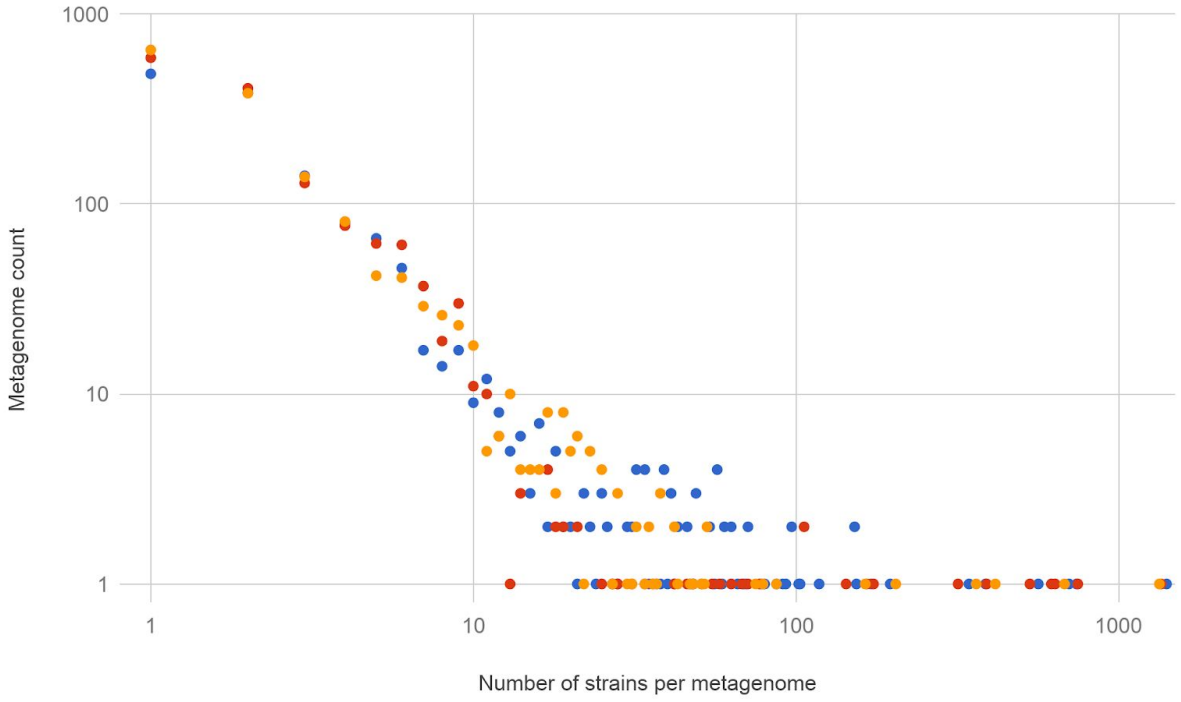

Fig. 2. Diversity of crAssphage strains in metagenomic samples. Strains for three amplicon regions A, B, and C were detected with Gretel(22) in 2,216 metagenomes (see Supplementary File 3).

To investigate the global phylogeography of crAssphage, we collected data about the three amplicon regions from various sources and combined them in a large-scale phylogenetic analysis, providing the first worldwide overview of the evolution of an epitome of the human gut virome (Table S1). First, we launched a global collaboratory to amplify and sequence the three

regions of the crAssphage genome from local sites. To obtain the highest expected rate of detection, collaborators sampled wastewater treatment plants. We combined these sequences with data from the COMPARE sewage sampling project (http://www.compare-europe.eu/), and the sequences from our metagenomics searches and individual volunteers found above. Together, we analyzed 32,273 different crAssphage sequences from at least 67 countries on six continents (34% of the countries in the world, see Fig. 3, Fig. S2, and Supplementary File 2). We reconstructed phylogenetic trees for the subset of strains with locality information and assessed the distribution of associated sampling metadata by using permutation statistics(*27*). Sequences from the same country, location, and sampling date are significantly clustered in the phylogeny (p<0.001, see Fig. S3-S4), and the genetically most similar other strain tends to be geographically close (Fig. 3). Thus, crAssphage is a cosmopolitan inhabitant of the human gut the world-over, with a geographically and temporally local sequence signature that may prove useful in future forensic applications of fecal contamination identification and detection(*15*–*17*, *28*).

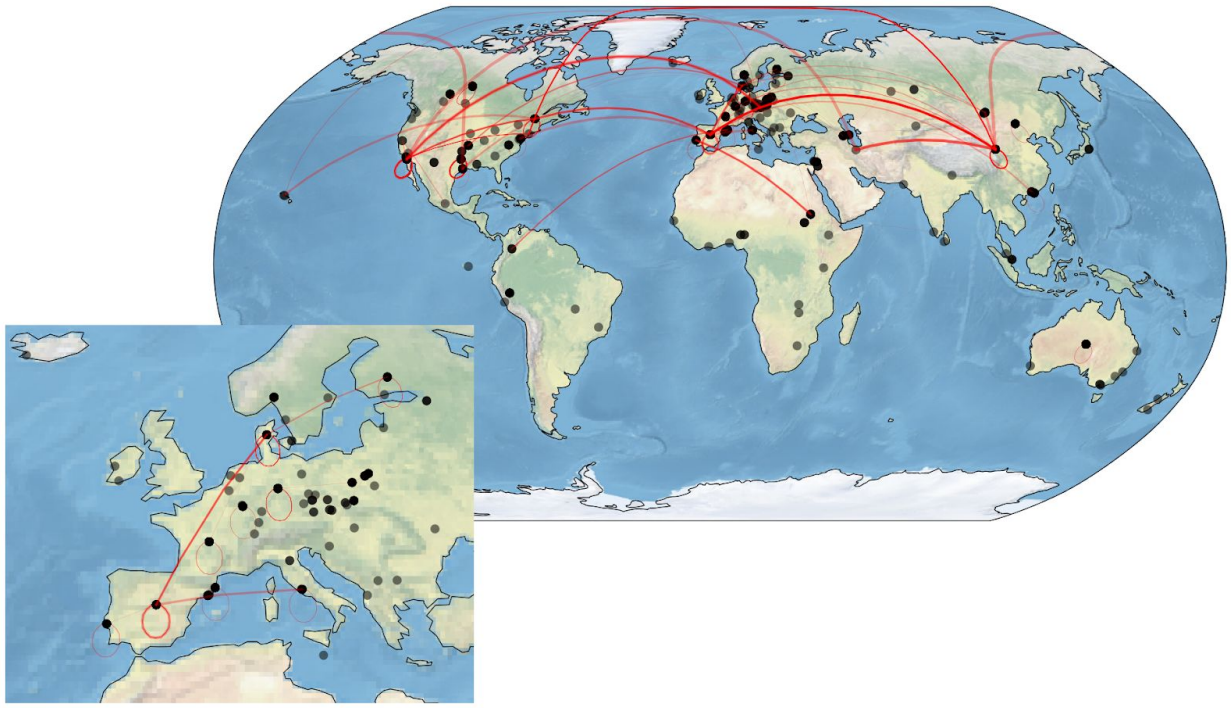

Fig. 3. Global locations of 2,424 crAssphage strains (amplicon A, see Fig. S2 for amplicons B and C). The number of samples at each location is reflected in the intensity of the black markers. For each strain, a link to the genetically most similar other strain is indicated with a red line, the intensity of which indicates similarity: circles indicate the most similar other strain at the same location, lines indicate links to different locations. Inset: samples from Europe.

CrAssphage has evolved with humans

The global distribution of crAssphage led us to ask the question whether this virus was present in early humans and has evolved with us as we spread out and colonized the planet. Alternatively, and consistent with the view of viruses as rapidly evolving entities, it is possible that crAssphage emerged recently, perhaps through recombination of other viruses, and spread

around the world either via factors related to the human host, e.g. the global food supply chain or international travel, or via the epidemiology of our intestinal bacteria.

To assess the possible ancient association of crAssphage-like phages with the human lineage, we screened the datasets from our global survey for remote human populations. We found a few crAssphage-like sequences in fecal samples from rural Malawi and from the Amazonas of Venezuela(*29*) (see Table S2). In contrast, mummified gut samples from three pre-Columbian Andean mummies(*30*) and the European iceman(*31*) were all crAss-negative. While this could suggest that these individuals were crAss-negative, it is also likely that any crAssphage DNA has degraded over thousands of years, in the absence of viable gut bacteria to maintain their titers.

Next, we sequenced and assembled fifteen fecal metagenomes from five species of non-human primates to search for crAssphage in our most distant relatives. None of the assembled nucleotide sequences matched the amplicon regions used above, as only short stretches of nucleotide homology were identified to the crAssphage genome(*32*). Surprisingly, many short homologous regions were found in several long sequences of ~90,000 nucleotides, and when displayed in a dot-plot, revealed a range of near-complete, distant crAssphage relatives in apes, Old-World monkeys, and New-World monkeys (Fig. 4). While those genomes were distantly related to crAssphage they were clearly colinear, showing the long-term genomic stability of this widespread gut virus. Clustering and alignment of translated protein sequences allowed us to reconstruct a phylogenomic tree of these genomes. This tree does not reflect the phylogeny of the hominids, instead reflecting the presence of multiple crAssphage-like species in the gut

virome of non-human primates, consistent with their higher gut microbiome diversity(*33*) and the coevolution of multiple populations of the likely crAssphage hosts(*13*, *34*).

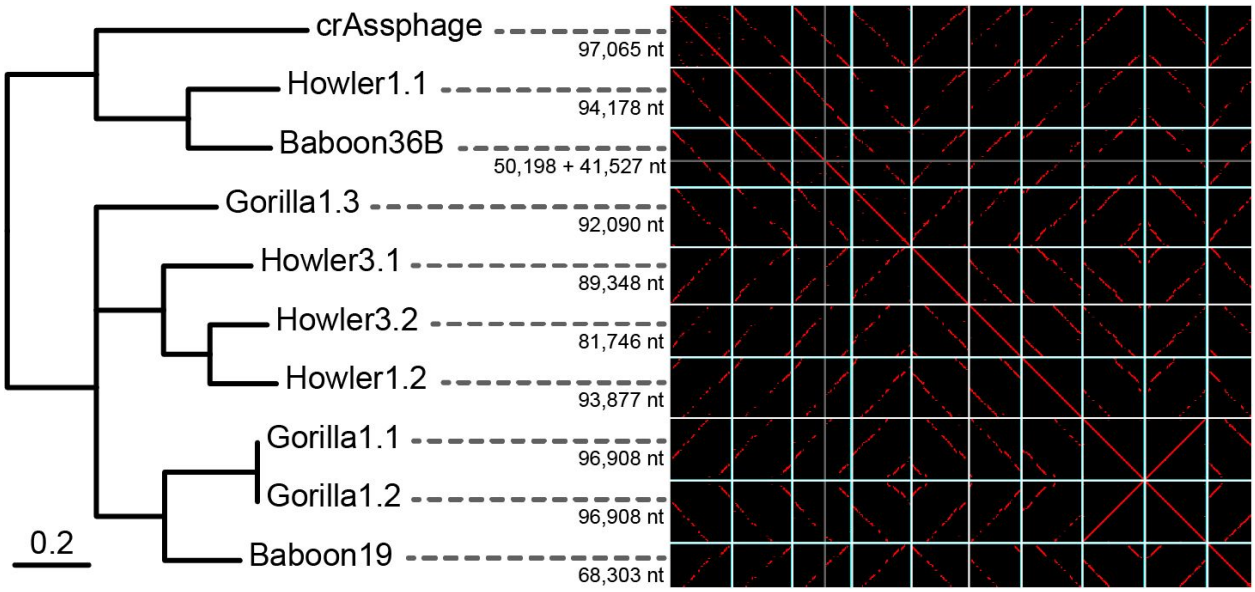

Fig. 4. Unrooted maximum likelihood phylogeny and dotplots showing full genomic colinearity between crAssphage and ten long contigs assembled from fecal metagenomes of different hominids. The phylogeny is based on a concatenated protein alignment of homologous ORFs. All branches had 100% bootstrap support, one exception <50% was collapsed. Dotplots are based on high-scoring segment pairs (blastn E-value <0.001) between all contigs. The figure is to scale, numbers to the left of the dotplot indicating genome or contig lengths. Note that circular permutation of some genomes leads to apparently broken diagonals in some dotplots.

CrAssphage belongs to the normal human virome

To investigate the association of crAssphage with characteristics of the human host and the human microbiome, we investigated the correlation between fecal crAssphage abundance and

a range of host factors and microbial taxa. By exploiting shotgun metagenomes and host metadata from the LifeLines-DEEP cohort(*35*, *36*), we correlated the abundance of crAssphage across 1,135 individuals with 207 exogenous and intrinsic human variables, including 78 dietary factors, 41 intrinsic factors, 39 diseases, 44 drug groups, 5 smoking categories (Supplementary File 4), and 490 microbial taxa (Supplementary File 5). We found significant but weak correlations with several diet categories (Benjamini-Hochberg <5% false discovery rate), including protein, carbohydrates, and caloric intake, basic food groups that are probably related to the dietary preferences of the crAssphage host bacteria(*13*, *35*, *37*–*39*). The most significant correlations of crAssphage with microbial taxa in the LifeLines-DEEP cohort included the family *Prevotellaceae*, consistent with our previous prediction that crAssphage infects bacteria of the *Bacteroidetes* phylum(*13*). Diverse dietary associations have been observed for different *Bacteroidetes* members, including the genus *Bacteroides* that was linked to a long-term Western diet rich in animal protein and sugars(*40*), while *Prevotella* and *Paraprevotella* were linked to low protein and high fiber(*41*). The most reliable computational phage-host signal to date(*42*) is a 100% matching CRISPR spacer in *Porphyromonas* sp. 31_2 isolated from human feces (Eugene Koonin, pers. comm.), another species within the *Bacteroidetes*. Given the potentially family-scale taxonomic diversity of crAssphages(*18*), it is likely that they infect a range of hosts throughout the *Bacteroidetes* phylum, leading to poor abundance correlations between crAssphage and different host groups, given (i) the taxonomic resolution that can be measured by metagenomic analysis and (ii) the rate of phage host-range evolution. The LifeLines-DEEP cohort did not reveal a significant relationship between crAssphage and any human health or disease parameters, consistent with a previous study showing absence of an association with diarrhea(*24*). As crAssphage abundance is not related to any health-related variables, we conclude that it is a part of the normal human virome(*43*).

Conclusions

The human gut virome mainly consists of phages that infect the abundant and diverse bacteria living in our gut. Phages are generally thought of as transient entities in the environment, whose fast infection cycle and relatively error-prone replication machinery enable rapid co-evolution with their hosts, which in turn would be reflected in highly diverse viral (meta-)genome sequences(*6*, *7*). Indeed, we found thousands of crAssphage strains throughout human feces-associated environments around the world. These strains are geographically and temporally clustered, consistent with rapid evolution and local dispersion. However, we also identified identical strains in up to 104 different samples from e.g. Denmark, France, Germany, Israel, Italy, Japan, and USA (Supplementary File 1). We suggest that this conservation primarily reflects recent spread by human global migration, although a crAssphage strain with potentially high fitness or environmental stability cannot be ruled out. Moreover, we identified highly divergent but fully colinear genome sequences in all major groups of hominids, suggesting that crAssphage has had a stable genome structure for millions of years, and a stable association with the hominid lineage and its microbiome(*34*) since our early ancestors began their great migration out of Africa.

Recently, the extent of gene flux and genomic mosaicism has been proposed to differ between temperate and virulent phages(*44*). Virulent phages tend to be genomically stable, while temperate phages fall into either high or low gene flux modes. Thus, crAssphage may be virulent or temperate based on its genomic stability. Our results challenge the notion of rampant

genomic mosaicism in viruses, showing that phage genome structure can be remarkably conserved in the stable environment provided by the human gut. Based on our observations, components of the human gut virome may be remarkably stable over millions of years, reflecting the environmental stability of its niche. This high stability of the hominid gut also limits the ability of its specialized microbes and viruses to escape to other environments. Indeed, this specificity makes crAssphage one of the strongest human fecal contamination markers to date(*15*–*17*). Taken together, our results provide the first global overview of the phylogeography of one of the most abundant and widespread viruses in the human gut, with evidence of both an ancient evolution and ongoing local dispersion.

References and Notes

1. R. Sender, S. Fuchs, R. Milo, Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell*. **164**, 337–340 (2016).
2. S. Nguyen *et al.*, Bacteriophage Transcytosis Provides a Mechanism To Cross Epithelial Cell Layers. *MBio*. **8** (2017), doi:10.1128/mBio.01874-17.
3. B. Rodriguez-Brito *et al.*, Viral and microbial community dynamics in four aquatic environments. *ISME J.* **4**, 739–751 (2010).
4. F. Rodriguez-Valera *et al.*, Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* **7**, 828–836 (2009).
5. T. Frede Thingstad, in *Eutrophication in Planktonic Ecosystems: Food Web Dynamics and Elemental Cycling* (Springer, Dordrecht, 1998), *Developments in Hydrobiology*, pp. 59–72.
6. S. Paterson *et al.*, Antagonistic coevolution accelerates molecular evolution. *Nature*. **464**, 275–278 (2010).
7. M. L. Pedulla *et al.*, Origins of highly mosaic mycobacteriophage genomes. *Cell*. **113**, 171–182 (2003).
8. M. Heldal, G. Bratbak, Production and decay of viruses in aquatic environments. *Mar. Ecol. Prog. Ser.* **72**, 205–212 (1991).
9. M. Breitbart, L. Wegley, S. Leeds, T. Schoenfeld, F. Rohwer, Phage community dynamics

in hot springs. *Appl. Environ. Microbiol.* **70**, 1633–1640 (2004).

10. G. F. Steward, D. C. Smith, F. Azam, Abundance and production of bacteria and viruses in the Bering and Chukchi Seas. *Mar. Ecol. Prog. Ser.* **131**, 287–300 (1996).

11. A. Reyes *et al.*, Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*. **466**, 334–338 (2010).

12. S. Minot *et al.*, Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12450–12455 (2013).

13. B. E. Dutilh *et al.*, A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).

14. A. Stern, E. Mick, I. Tirosh, O. Sagy, R. Sorek, CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* **22**, 1985–1994 (2012).

15. E. Stachler, K. Bibby, Metagenomic Evaluation of the Highly Abundant Human Gut Bacteriophage CrAssphage for Source Tracking of Human Fecal Pollution. *Environ. Sci. Technol. Lett.* **1**, 405–409 (2014).

16. E. Stachler *et al.*, Quantitative CrAssphage PCR Assays for Human Fecal Pollution Measurement. *Environ. Sci. Technol.* **51**, 9146–9154 (2017).

17. C. García-Aljaro, E. Ballesté, M. Muniesa, J. Jofre, Determination of crAssphage in water samples and applicability for tracking human faecal pollution. *Microb. Biotechnol.* (2017), doi:10.1111/1751-7915.12841.

18. N. Yutin *et al.*, Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol* (2017), doi:10.1038/s41564-017-0053-y.

19. J. Barylski *et al.*, Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Bacteriophages. *bioRxiv* (2018), p. 220434.

20. B. J. Callahan, P. J. McMurdie, S. P. Holmes, Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).

21. NCBI Resource Coordinators, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **44**, D7–19 (2016).

22. S. M. Nicholls *et al.*, Probabilistic Recovery Of Cryptic Haplotypes From Metagenomic Data. *bioRxiv* (2017), p. 117838.

23. E. S. Lim *et al.*, Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* **21**, 1228–1234 (2015).

24. Y. Y. Liang, W. Zhang, Y. G. Tong, S. P. Chen, crAssphage is not associated with diarrhoea and has high genetic diversity. *Epidemiology & Infection*. **144**, 3549–3553 (2016).

25. H. G. Piper *et al.*, Severe Gut Microbiota Dysbiosis Is Associated With Poor Growth in Patients With Short Bowel Syndrome. *JPEN J. Parenter. Enteral Nutr.* **41**, 1202–1212 (2017).

26. T. Vatanen *et al.*, Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell*. **165**, 842–853 (2016).

27. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

28. W. Ahmed *et al.*, Evaluation of the novel crAssphage marker for sewage pollution tracking in storm drain outfalls in Tampa, Florida. *Water Res.* **131**, 142–150 (2017).

29. T. Yatsunenko *et al.*, Human gut microbiome viewed across age and geography. *Nature*. **486**, 222–227 (2012).

30. T. M. Santiago-Rodriguez *et al.*, Natural mummification of the human gut preserves bacteriophage DNA. *FEMS Microbiol. Lett.* **363**, fnv219 (2016).

31. F. Maixner *et al.*, The 5300-year-old Helicobacter pylori genome of the Iceman. *Science*. **351**, 162–165 (2016).

32. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

33. A. H. Moeller *et al.*, Rapid changes in the gut microbiome during human evolution. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 16431–16435 (2014).

34. A. H. Moeller *et al.*, Cospeciation of gut microbiota with hominids. *Science*. **353**, 380–382 (2016).

35. A. Zhernakova *et al.*, Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*. **352**, 565–569 (2016).

36. E. F. Tigchelaar *et al.*, Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open*. **5**, e006772 (2015).

37. L. A. David *et al.*, Diet rapidly and reproducibly alters the human gut microbiome. *Nature*. **505**, 559–563 (2014).

38. P. J. Turnbaugh *et al.*, The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1**, 6ra14 (2009).

39. R. K. Singh *et al.*, Influence of diet on the gut microbiome and implications for human health. *J. Transl. Med.* **15**, 73 (2017).

40. C. De Filippo *et al.*, Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14691–14696 (2010).

41. P. Kovatcheva-Datchary *et al.*, Dietary Fiber-Induced Improvement in Glucose Metabolism Is Associated with Increased Abundance of Prevotella. *Cell Metab.* **22**, 971–982 (2015).

42. R. A. Edwards, K. McNair, K. Faust, J. Raes, B. E. Dutilh, Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.* **40**, 258–272 (2016).

43. P. Manrique *et al.*, Healthy human gut phageome. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 10400–10405 (2016).

44. T. N. Mavrich, G. F. Hatfull, Bacteriophage evolution differs by host, lifestyle and genome. *Nat Microbiol*. **2**, 17112 (2017).

Acknowledgements:

- Special thanks to the following individuals who provided access to wastewater treatment samples: Robert Matthews, Mitchell Wright, John Alexander, Susie Arredondo, Nicki Branch, Doug Campbell, Ravy Chea, Dawn McDougle, Jeff Parks, and Vasana Vipatapat.
- We thank the Mountain Gorilla Veterinary Project, and the Maryland Zoo for collecting the gorilla fecal samples in Rwanda. We thank Gillian Britton for collecting the baboon fecal samples in Ethiopia. We thank the Chimpanzee Sanctuary and Wildlife Conservation Trust CSWCT), the Uganda Wildlife Authority (UWA), and the Uganda National Council for Science and Technology (UNCST) for collecting the chimpanzee fecal samples in Uganda.
- We thank the COMPARE and LifeLines-DEEP projects for sharing data.

Funding

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

- This work used the Extreme Science and Engineering Discovery Environment (XSEDE) Jetstream resources at Indiana University and Texas Advanced Computing Center through allocation MCB170036 to RAE, which is supported by National Science Foundation grant number ACI-1548562.
- Some of this work was supported by San Diego State University Grants Programs to RAE including the Summer Undergraduate Research Program.
- This work was supported by National Science Foundation grant numbers MCB-1441985 to RAE and DUE-1323809 to EAD.
- This work was supported by the Department of Energy Lawrence Livermore National Laboratory grant B618146 to RAE.
- PAJ and BED were supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004.
- FLN was supported by the Netherlands Organization for Scientific Research (NWO) Veni grant 016.Veni.181.092.
- SJJB was supported by European Research Council Stg grant [638707] and the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.11.005.
- OC and KM were supported by Ministry of Health of the Czech Republic grants nr. 15-31426A and 15-29078A.
- PCF was supported by a Rutherford Discovery Fellowship from the Royal Society of New Zealand.

- JJB was supported by the Australian Research Council (ARC) Discovery Early Career Researcher Award (DE170100525).
- SLDM was supported by an NIH Pathway to Independence Fellowship (1K99AI119401-01A1).
- DTM thanks the Australian Research Council's Linkage Project LP160100408, Melbourne Water and EPA Victoria for funding the collection of samples in Melbourne.
- KB was supported by award number 1510925 from the United States National Science Foundation.
- MTI was supported by National Geographic Society (CRE) and NSERC.
- CD was supported by Agence Nationale de la Recherche JCJC grant #ANR-13-JSV6-0004 and Investissements d'Avenir Méditerranée Infection #10-IAHU-03.
- The LifeLines-DEEP sample collection and analysis was funded by the Netherlands Heart Foundation (IN-CONTROL CVON grant 2012-03 to AZ and JF), by the Top Institute Food and Nutrition, Wageningen, the Netherlands (TiFN GH001 to CW), by the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.13.013 to JF, NWO Vidi grant 016.178.056 to AZ, NWO Spinoza Prize SPI 92-266 to CW), and by the European Research Council (ERC) FP7/2007-2013/ERC Advanced Grant agreement 2012-322698 to CW, ERC Starting Grant 715772 to AZ. AZ also holds a Rosalind Franklin Fellowship from the University of Groningen.
- The COMPARE data collection was funded by The Novo Nordisk Foundation (NNF16OC0021856) to HZ.

Author Contributions

BED, RAE conceived of the study, performed the experiments and bioinformatics, and wrote the paper with input from all authors. AAV performed the volunteer experiments and sampled San Diego wastewater treatment plants. FLN, HMN, MO, PAJ performed human volunteer experiments. AR, AVT, DAC, JMH, KL, KMcN, ME, TOC, VAC performed bioinformatics analysis. AARR, AAl, ACz, AMcC, AO, ARM, ASN, AW, BM, BME, CD, CF, CH, DC, DK, DMcC, EAD, EB, ENI, ENS, ESL, GA, GCA, GS, GSC, GT, HH, HN, JAB, JJB, JJT, JM, JMC, JMM, JW, KB, KLW, KM, LCS, LD, MAUI, MKM, ML, MMZ, MMo, MMu, MP, MPD, NT, NV, OC, ODN, PC, PCF, PD, PR, PV, RI, RKA, RL, RO, RR, RSa, RSr, SJJB, SLDM, SM, SMM, SW, TC, TJ, UQ, ZXQ performed sampling, PCR, and sequencing. AK, AZ, CW, JF performed the Lifelines analysis. FMA, HZ, RSH provided and analysed COMPARE project data. AAs, BW, NPN, RSu, SL analyzed and provided the non human primate sequences. MC collected Gorilla samples. AT, EG, KMG performed the NYC sewage sampling and data analysis. AJP, JS, LCM, PJT, SRH, STK examined crAssphage transfer among infants. MTI, REJ collected lemur sample. MK collected Howler monkey samples. LM collected chimpanzee samples.

Competing interests

None declared.

Data and materials availability:

All sequence data and scripts have been released on GitHub under the MIT license. The data and code may be found at https://github.com/linsalrob/crAssphage. The current release is version 2.0 and has DOI: 10.5281/zenodo.1230436.

List of Supplementary materials:

- **Materials and Methods**
- **Supplementary Figures**
    - Fig. S1. Phylogenomic tree of crAssphage genome sequences assembled from the Reyes twin study shows clustering of the strains by individual, with some samples taken up to one year apart1 yet clustering together in the tree. Sample tags indicate family number (F1 through F4) and mother (M) or twins (T1 and T2). All branches separating different individuals have bootstrap support >90%, except F2T1 and F2T2 that are not monophyletic. The scale bar indicates 0.01 mutations per site of the concatenated protein alignment.
    - Fig. S2. Global locations of 1,896 and 1,774 sequences from amplicons B and C, respectively. The number of samples at each location is reflected in the intensity of the black markers. For each strain, a link to the genetically most similar other strain is indicated in red: circles indicate the most similar other strain at the same location, lines indicate links to different locations. Inset: samples from Europe.
    - Fig. S3. Unrooted maximum likelihood phylogeny of crAssphage sequences collected from different sources. Branches are colored by country (A-C) and by date (D-F) for amplicons A (A/D, 1,900 sequences after alignment trimming), B (B/E, 1,368 sequences), and C (C/F, 1,621 sequences). Note that some sequences were deleted after trimming the MUSCLE alignment. Trees were visualized with iTOL.
    - Fig. S4. Geographical and temporal clustering statistics in the global phylogenetic trees of amplicon regions A (1,900 leaves), B (1,368 leaves), and C (1,621 leaves). Branches with increasing bootstrap values were collapsed (IQ-tree provides SH-aLRT and UFBoot bootstrap values, see left and right panels, respectively) and the statistics calculated. Next, statistics were also calculated based on 1,000 permutations of the leaf labels in the phylogenetic tree, but these statistics were never higher than with the original leaf labels.
    - Fig. S5. Sequencing trace of amplicon B from the wastewater treatment plant in Leuven, Belgium (sample 52GJ06_G04_B_F, see https://github.com/linsalrob/crAssphage/blob/master/Global_Survey/Sequences/raw_data/Lavigne/52GJ06_G04_B_F.ab1). The trace contains a single sequence for the first 227 nucleotides and then more than one sequence (presumably through an indel), rendering the trace unreadable.
    - Fig. S6. Coverage of the crAssphage genome in 10,260 metagenomes. The predicted ORFs are shown below the genome position (x-axis) and the metagenomes are on the y-axis. Each position represents the log of the average sequence coverage over a 1kb window as shown in the scale bar.

- Fig. S7. Relationship between the average per-base read depth as reported by samtools depth (including zero-coverage bases) and the number of strains recovered for amplicon A (Pearson's r2=0.683; p<.001), B (Pearson's r2=0.655; p<0.001), and C (Pearson's r2=0.640; p<0.001).
- Fig. S8. Flow chart of the sequencing analysis. Biological sample processing are shown in green, files and databases in red, external software in yellow, and software developed for this project in blue. Hexagons indicate decision steps. Amplicon sequencing starts with generating the sequences, while the metagenomics pipeline starts with publicly available sequence data. Both pipelines use the same downstream processing steps to generate the trees.

- **Supplementary Tables**
  - Table S1. All crAssphage sequences collected from different sources. The numbers indicate: (i) total sequences identified, (ii) unique sequences, and (iii) sequences with locality information. The information per strain is provided in Supplementary File 2.
  - Table S2. Number of crAssphage reads in fecal metagenomes from rural Malawi and the Amazonas of Venezuela.
  - Table S3. Primer sequences. Primer A, expected product size: 1,331 bp. Primer B: 1,354 bp. Primer C: 1,238 bp.
  - Table S4. PCR reaction mixture.
  - Table S5. PCR amplification protocols.