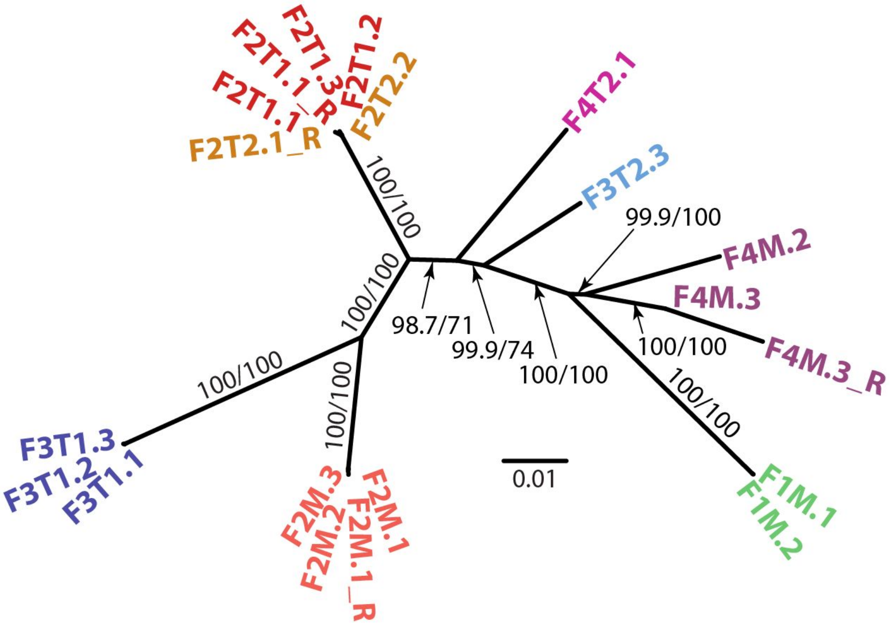
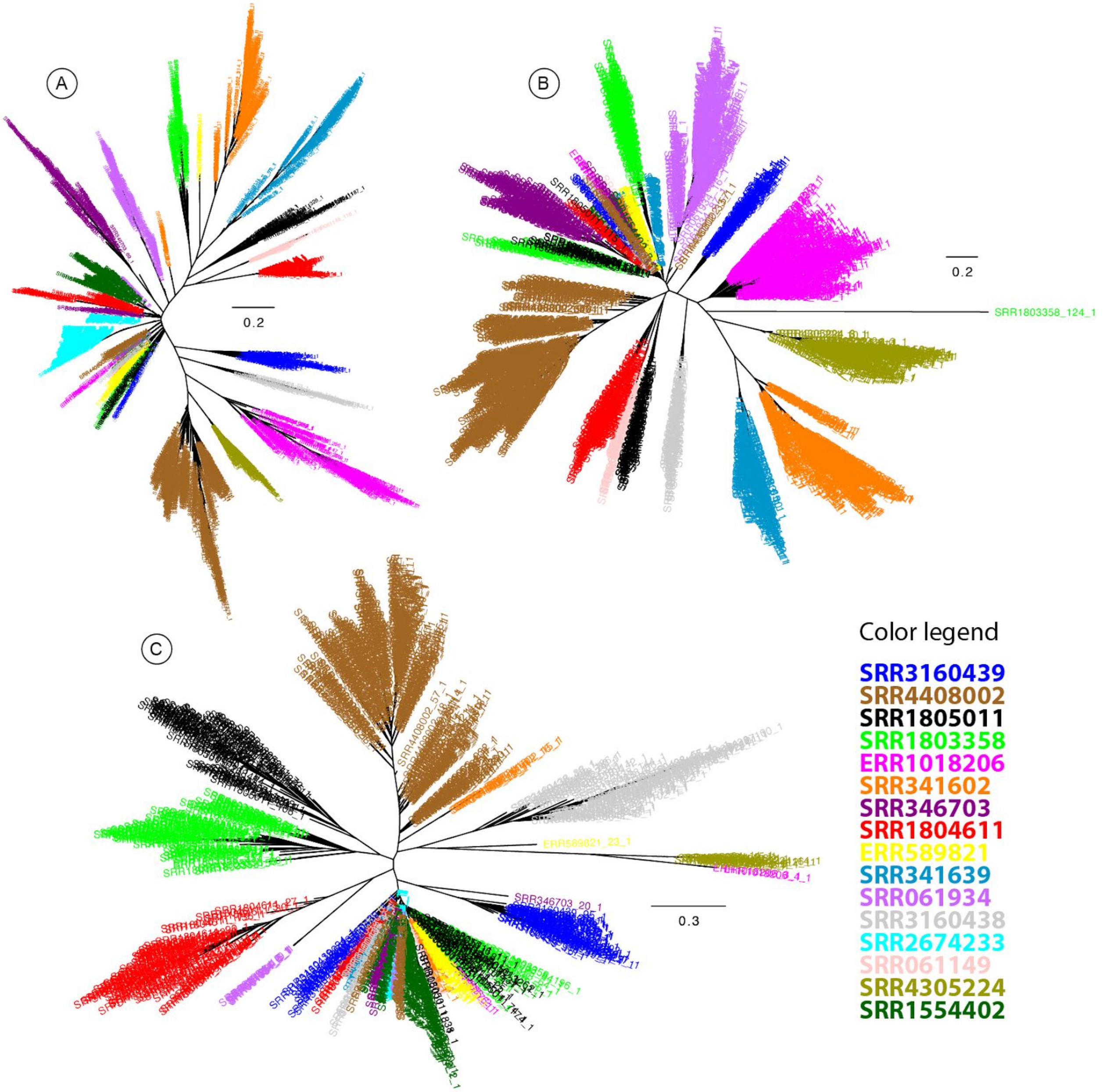
# Supplementary Figures

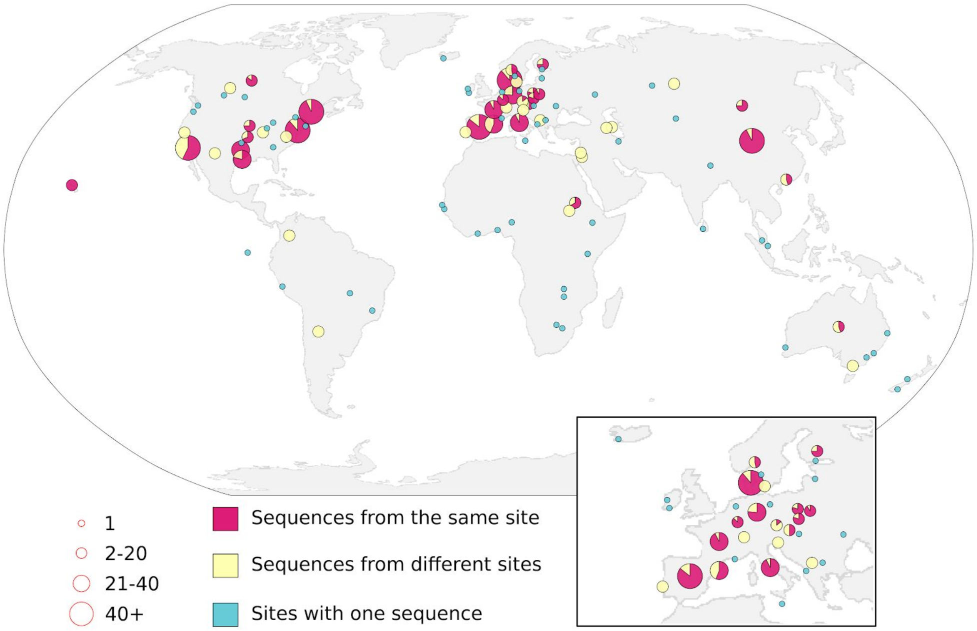


*Supplementary Figure 1. Phylogenomic tree of crAssphage genome sequences assembled from the Reyes twin study shows clustering of the strains by individual, with some samples taken up to one year apart[1] yet clustering together in the tree. Sample tags indicate the family number (F1 through F4) and mother (M) or twins (T1 and T2). We could not reconstruct complete crAssphage genomes from F1T1, F1T2, F3M, and F4T1. Branch support values are Shimodaira and Hasegawa-like approximate likelihood-ratio test (SH-aLRT) and ultrafast bootstrap (UFBoot) support, respectively (e.g. for the branch that is 98.7/71, the SH-aLRT support is 98.7 and the ultrafast bootstrap support is 71)[2]. The scale bar indicates 0.01 mutations per site of the concatenated protein alignment.*

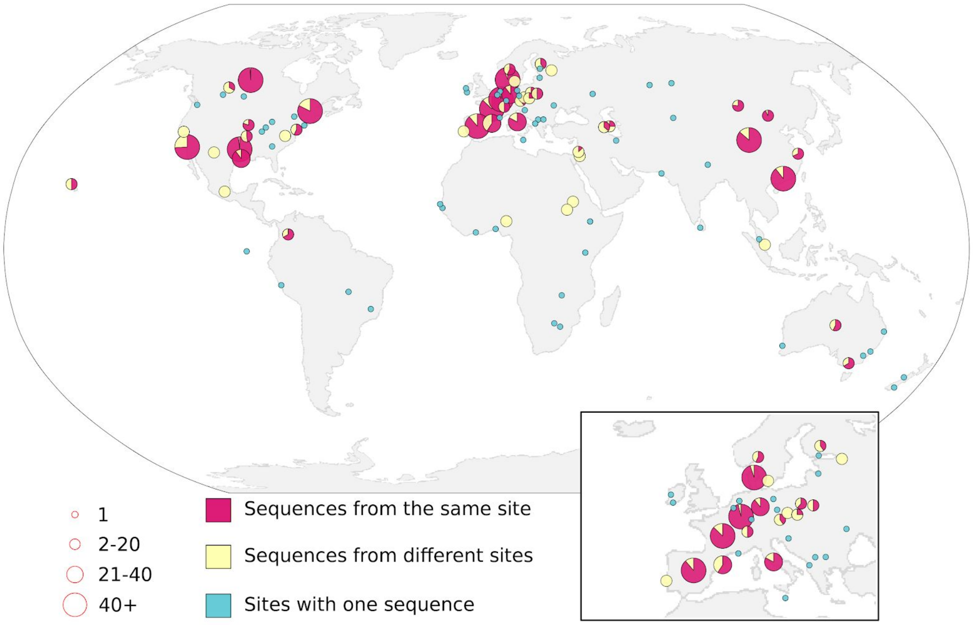


*Supplementary Figure 2.* Phylogenetic trees of the three amplicons (A, B, and C) from the sixteen samples with more than 100 strains identified for any of the amplicons (see Supplementary File 2). The leaves of the trees are colored by sample, showing the strong phylogenetic relatedness of co-occurring crAssphage strains. The color and number of strains identified for each sample are as follows (color, A, B, C): SRR3160439 (blue, 1409, 1355, 1338); SRR4408002 (brown, 702, 748, 362); SRR1805011 (black, 741, 636, 680); SRR1803358 (green, 564, 619, 680); ERR1018206 (magenta, 344, 531, 7); SRR341602 (orange, 387, 390, 16); SRR346703 (purple, 152, 318, 21); SRR1804611 (red, 154, 106, 203); ERR589821 (yellow, 196, 28, 23); SRR341639 (marine, 171, 174, 10); SRR061934 (violet, 103, 166, 10); SRR3160438 (gray, 78, 143, 164); SRR2674233 (light blue, 152, 5, 8); SRR061149 (pink, 118, 77, 10); SRR4305224 (olive, 91, 106, 16); SRR1554402 (dark green, 102, 5, 52).

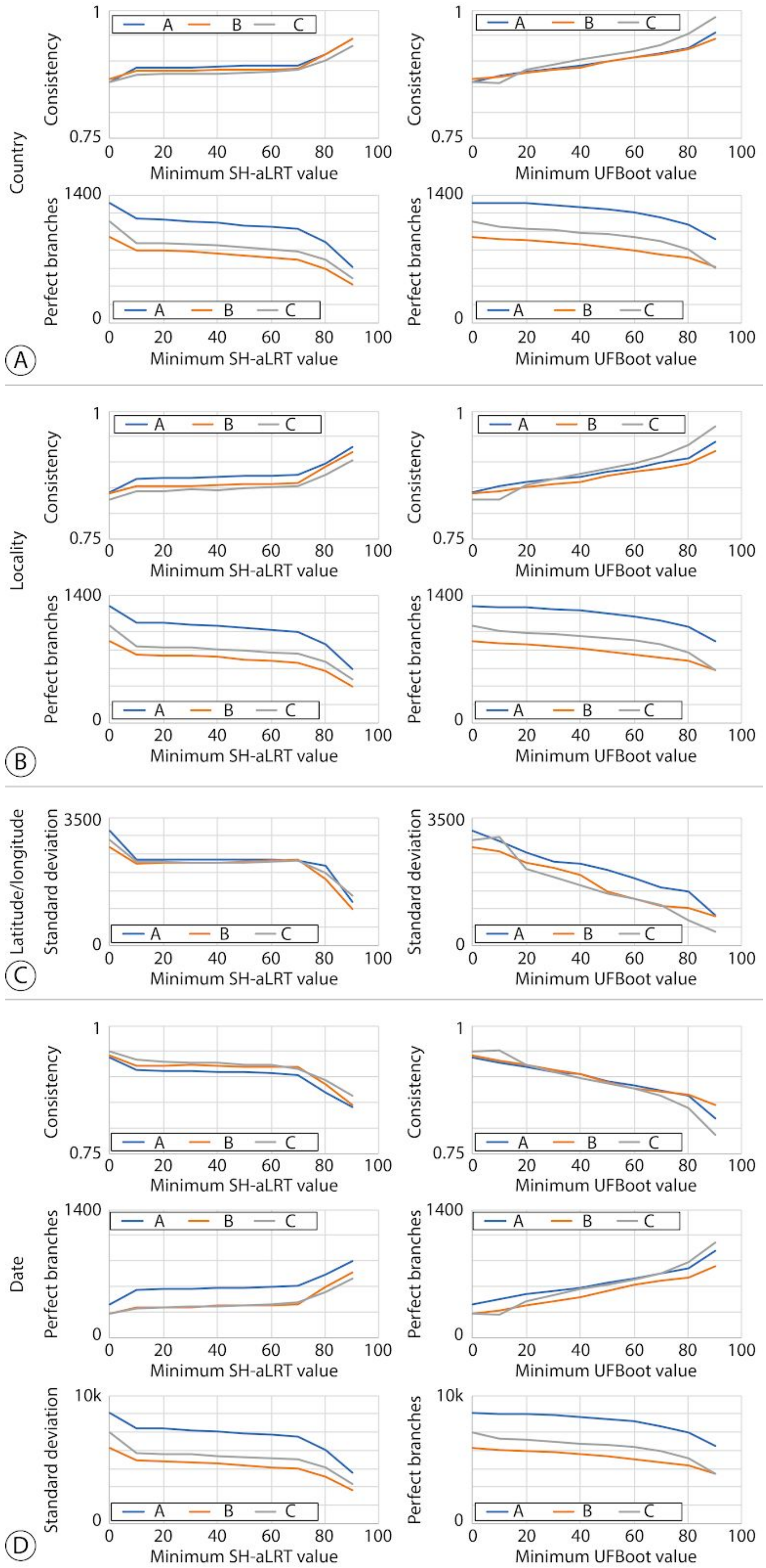
Amplicon B:



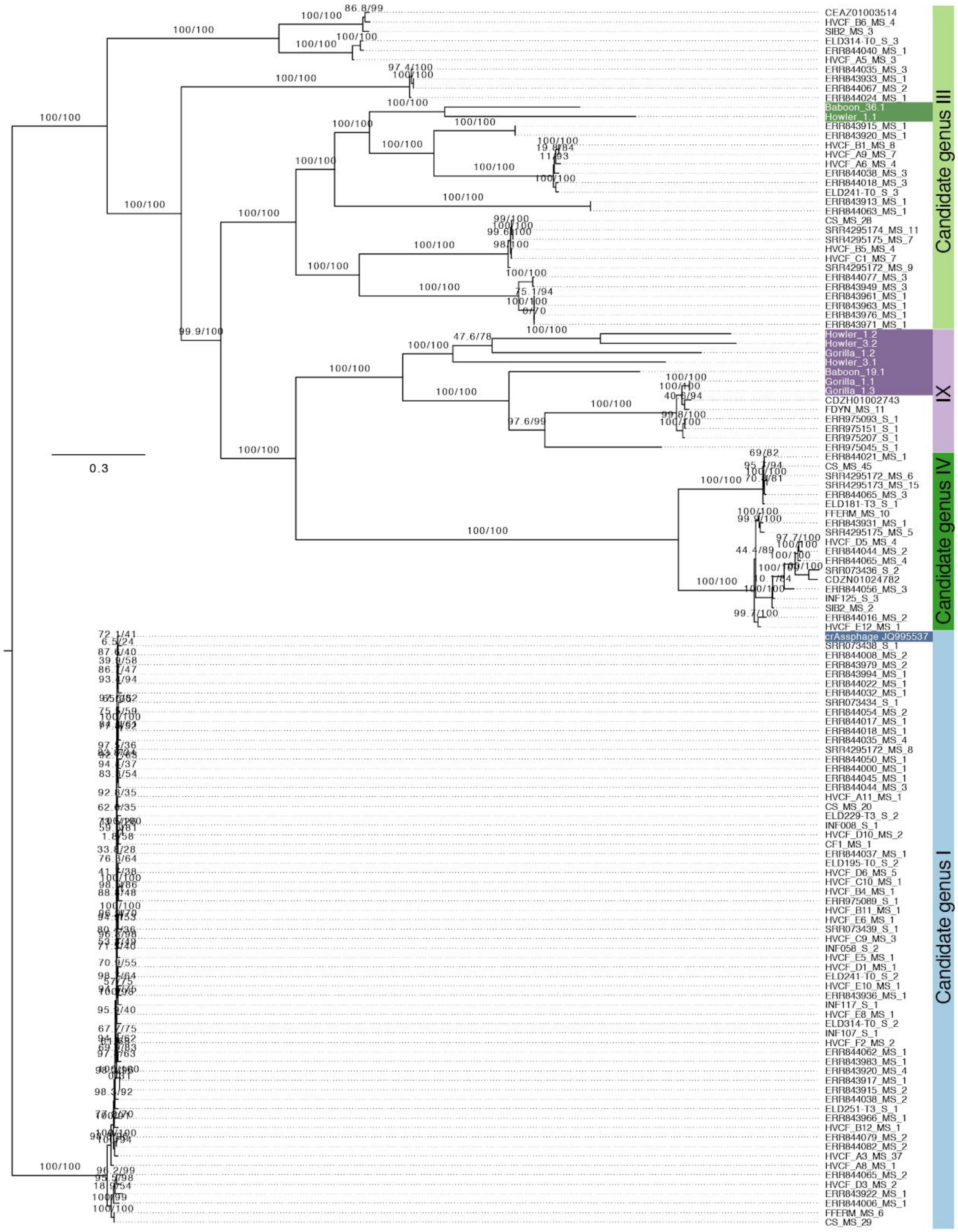
Amplicon C:



*Supplementary Figure 3. Global locations of 1,896 and 1,774 sequences from amplicons B and C, respectively. Pie diagrams indicate the fraction of most-similar strains identified at the same site (<150 km apart) and at a different site. The number of strains at each location is reflected in the size of the pie diagrams. Inset: Europe*

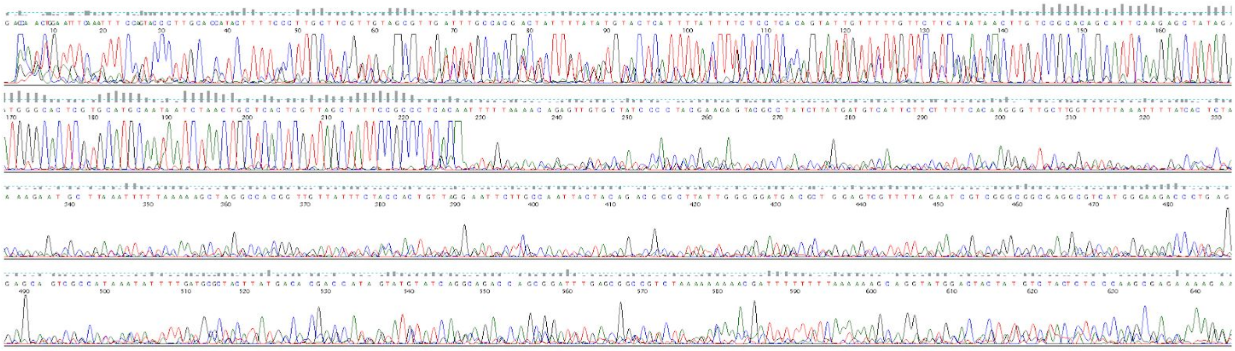

Country
Consistency
1
0.75
0
20
40
60
80
100
Minimum SH-aLRT value
Minimum UFBoot value
A
B
C
Perfect branches
1400
0
A
Locality
Consistency
Perfect branches
B
Latitude/longitude
Standard deviation
3500
0
C
Date
Consistency
Perfect branches
1400
Standard deviation
10k
Perfect branches
10k
D

*Supplementary Figure 4. Geographical (A-C) and temporal (D) 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 statistics calculated including consistency, perfect branches, and standard deviations (for details see the section "Assessment of metadata clustering" in Methods). 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 so empirically p<0.001.*

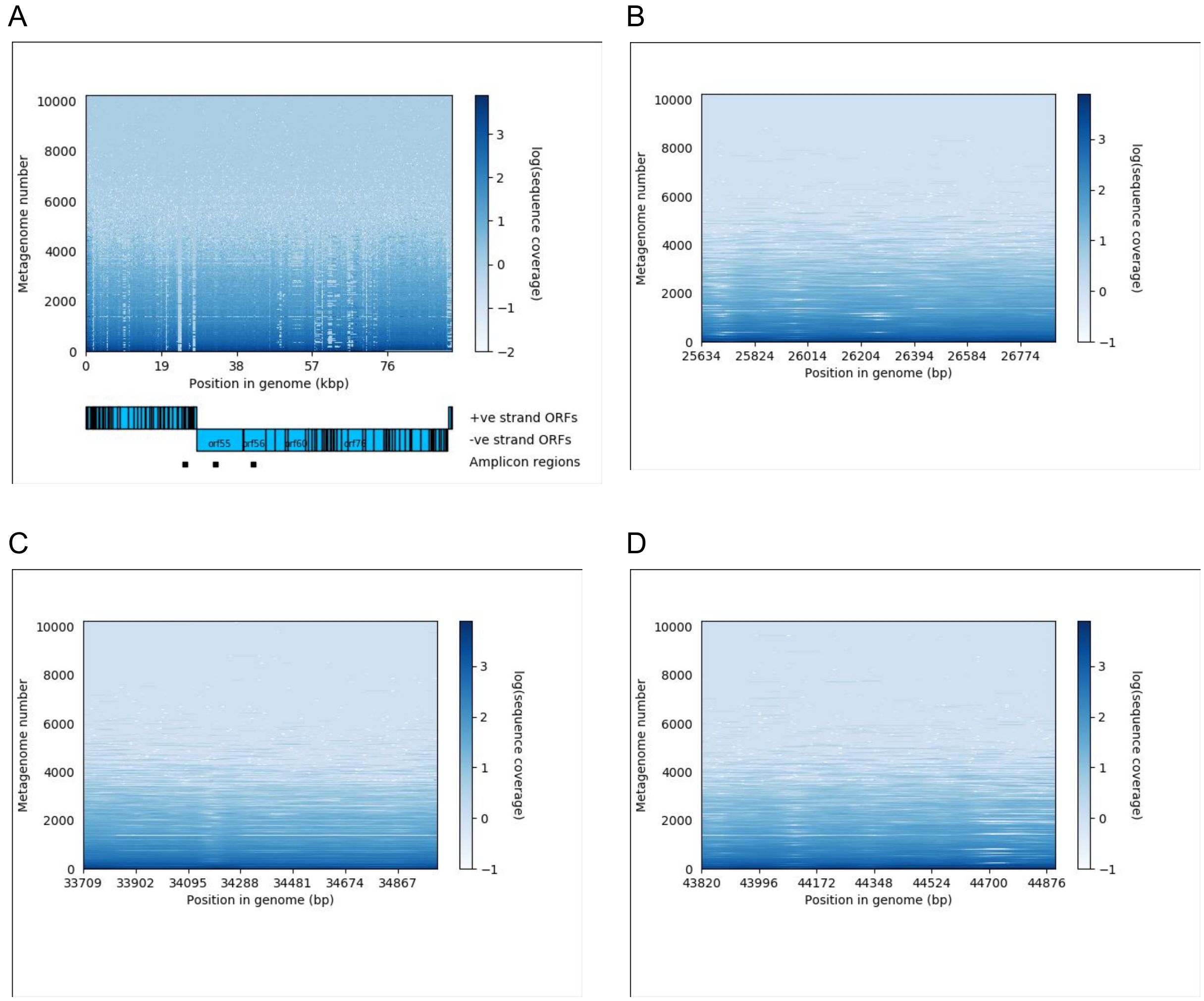


*Supplementary Figure 5. Maximum likelihood phylogeny of ten long contigs assembled from fecal metagenomes of non-human primates (highlighted) and 119 Alphacrassvirinae contigs[3].*

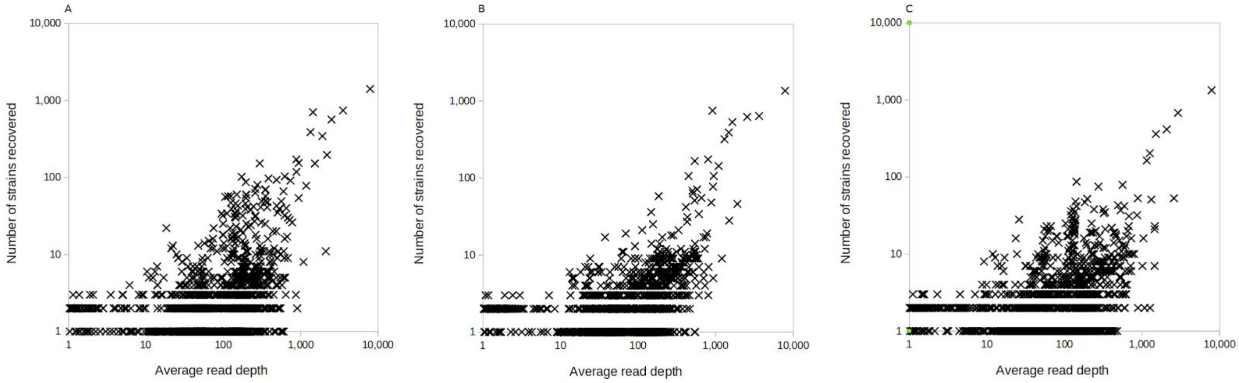
*The phylogeny is based on a 8,009 amino acid concatenated, trimmed protein alignment of 15 homologous ORFs. Branch support values are SH-aLRT and ultrafast bootstrap support[2], scale bar indicates 0.3 mutations per site. The tree is rooted as in Guerin et al.*



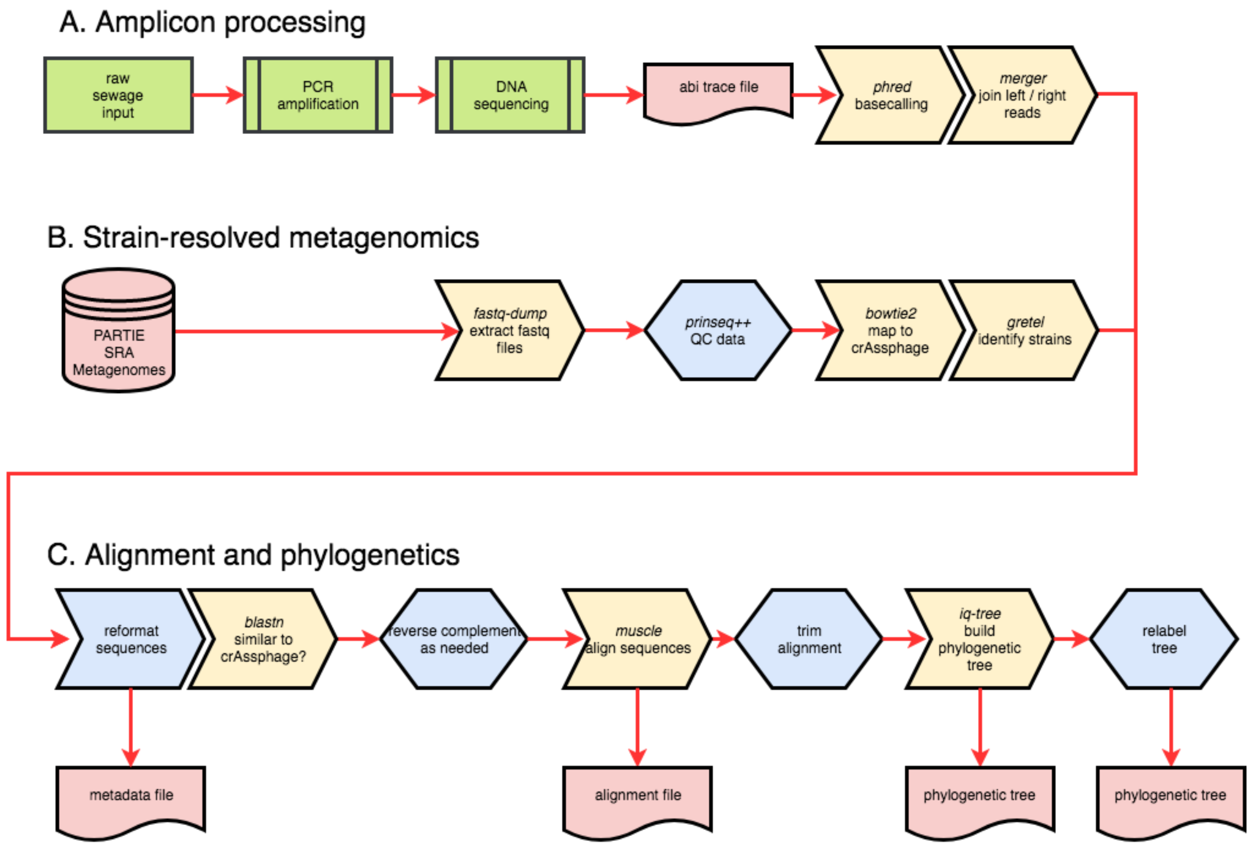
*Supplementary Figure 6. Sequencing trace of amplicon B from the wastewater treatment plant in Leuven, Belgium (sample 52GJ06_G04_B_F, see [https://github.com/linsalrob/crAssphage/blob/master/Global_Survey/Sequences/raw_data/Lavigne/52GJ06_G04_B_F.ab1](https://github.com/linsalrob/crAssphage/blob/master/Global_Survey/Sequences/raw_data/Lavigne/52GJ06_G04_B_F.ab1)). 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.*



*Supplementary Figure 7. Coverage of the crAssphage genome in 10,260 metagenomes. A. Coverage across the entire genome. The predicted ORFs are shown below the genome position (x-axis) and the metagenomes are on the y-axis. Positions of the three amplicon regions are also shown. Each position represents the log of the average sequence coverage over a 1kb window as shown in the scale bar. B-D, Coverage across the three amplicon regions (regions A-C, respectively). Each position represents the log of the sequence coverage as shown in the scale bar.*



*Supplementary Figure 8. The relationship between the average per-base read depth as reported by samtools depth (including zero-coverage bases) and the number of strains recovered for each amplicon using ordinary least squares. Amplicon A (n=11,054; Pearson's $r^2$=0.658; p=0.00), Amplicon B (n=11,054; Pearson's $r^2$=0.629; p=0.00), and C (n=11,054; Pearson's $r^2$=0.612; p=0.00).*



*Supplementary Figure 9. Flow chart of the sequencing analysis. Biological sample processing is 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

*Supplementary Table 1. The 104 samples containing the most ubiquitous crAssphage strain.*

| Project | Title | Samples | Location | Lat/Lon |
|---|---|---|---|---|
| | Seasonal Dynamics of DNA and RNA Viral Bioaerosol Communities in a Daycare Setting | Daycare study sample AP-DNA-10 | Virginia, USA | |
| ERP008729 | Gut microbiome development along the colorectal adenoma-carcinoma sequence | ERR688473 | | |
| SRP059928 | Non-human sequence from stool, colon biopsy, ileum resection, kefir, and artificial bacterial mixtures | SRR2082443 | Canada | 53.520710 N 113.524239 W |
| SRP065270 | Functional dynamics of the elderly gut microbiome during probiotic consumption | SRR2857970 | USA | 42.3601 N 71.0589 W |
| ERP002061 | A method for identifying metagenomic species and variable genetic elements by exhaustive co-abundance binning | ERR210123;<br>ERR210122;<br>ERR210052;<br>ERR209575;<br>ERR209644 | | |
| ERP005860 | Liver cirrhosis occurs as a consequence of many chronic liver diseases that are prevalent worldwide. Previous studies have shown an association between the gut microbiota and liver complications such as cirrhosis and other liver injuries. We therefore undertook a whole gut microbiome wide association study of stool samples from 98 liver cirrhosis patients and 83 healthy controls to characterise the faecal microbial communities and their functional composition. In total, we generated 860 Gb of high-quality sequence data and built a reference gene set for the liver cirrhosis cohort containing 2.69 million genes, 36.1% of which was not covered by previously published gene catalogues. | ERR527052 | | |
| SRP083099 | Gut microbiota and metagenomic diversity of omnivore, vegetarian and | SRR4074354 | Italy | |

| | | | | |
|---|---|---|---|---|
| | vegan healthy subjects | | | |
| ERP000108 | A human gut microbial gene catalog established by deep metagenomic sequencing | ERR011190 | | |
| DRP000700 | Metgenomic analysis of human gut microbiome in patients with multiple sclerosis (MS). | DRR002666 | | |
| SRP056641 | Human Microbiome Environment Metagenome | SRR2175726 | | |
| SRP066479 | Antibiotic resistance exchange between microbiota in resource-poor settings in Latin America | SRR2938428 | El Salvador | |
| ERP005534 | Potential of fecal microbiota for early stage detection of colorectal cancer | ERR480821; ERR479008; ERR480516; ERR479525; ERR480673; ERR479524; ERR480711 | | |
| SRP029441 | Fiji COMP | SRR2195841; SRR2249814; SRR2222814; SRR2250644; SRR2189708 | | |
| SRP049045 | Abundance of antibiotic resistance genes and structure of the microbial community in wastewater of medical facilities besides hospitals | SRR1616987 | Germany | 48 N 8 E |
| SRP072561 | Human Gut Microbiome in a Multiplex Family Study of Type 1 Diabetes Mellitus | SRR3313047 | Luxembourg | 49.5 N 6.2 E |
| ERP013562 | Gut microbial dysbiosis in young adults with obesity | ERR1190645; ERR1190689; ERR1190633 | | |
| SRP060568 | Hospital Air Samples Metagenome | SRR2183670 | | |
| SRP066514 | Human gut Metagenome | SRR2940957 | USA | |
| ERP013933 | Reproducibility of associations between the human gut microbiome and colorectal cancer assessed in a patient population from Washington, DC, USA | ERR1293299; ERR1293522 | | |
| ERP012929 | Towards personalized nutrition by prediction of glycemic responses | ERR1137395; ERR1137041; ERR1136988 | | |
| DRP000446 | Comprehensive Detection of Possible Pathogens Associated with Kawasaki Disease | DRR014146 | | |

| | | | | |
|---|---|---|---|---|
| SRP002163 | Human Microbiome Project (HMP) Metagenomic WGS Projects, deeper sequencing of the human microbiome samples: Production Phase | SRR528262;<br>SRR1804206;<br>SRR532466;<br>SRR1804707;<br>SRR532351;<br>SRR514308;<br>SRR063906;<br>SRR539598;<br>SRR1804115;<br>SRR549428;<br>SRR532027 | | |
| SRP051174 | NIBSC_BSRI Metagenome | SRR1714192 | USA | 38.98 N<br>77.11 W |
| SRP064913 | Library preparation methodology can influence genomic and functional predictions in human microbiome research | SRR2726666 | | |
| ERP009422 | Temporal and technical variability of human gut metagenomes | ERR748434;<br>ERR748433;<br>ERR748174;<br>ERR748184;<br>ERR748319;<br>ERR748477 | | |
| SRP000319 | A core gut microbiome in obese and lean twins | SRR029696 | | |
| SRP058816 | Methanogenic digester communities Raw sequence reads | SRR2043640 | | |
| SRP040146 | C.diff FMT | SRR1492958;<br>SRR1437940;<br>SRR1491454;<br>SRR1437798;<br>SRR1462693;<br>SRR1437716;<br>SRR1437790;<br>SRR1461800;<br>SRR1491724;<br>SRR1490908;<br>SRR1461818;<br>SRR1490972;<br>SRR1490923 | | |
| ERP013563 | Gut microbiome-dependent stratification of patients for anti-diabetic treatment | ERR1190879;<br>ERR1190804 | | |
| SRP056054 | A prospective, longitudinal analysis of the developing gut microbiome in infants en route to type 1 diabetes | SRR1918833;<br>SRR1910622 | | |
| SRP031463 | Microbiome analysis of stool samples | SRR1012404 | | |

| | | | | |
|---|---|---|---|---|
| | from African Americans with colon polyps | | | |
| SRP040765 | Microbiome study of the RISK cohort | SRR1765589 | | |
| SRP064400 | Intestinal microbiota dynamics in hospitalized patients | SRR2565987;<br>SRR2565536;<br>SRR2566055 | Canada | 45.50 N<br>73.63 W |
| SRP011011 | A Metagenome-Wide Association Study of gut microbiota identifies markers associated with Type 2 Diabetes | SRR413683 | | |
| ERP016813 | Integrated metabolomics and metagenomics analysis of plasma and urine identified microbial metabolites associated with coronary heart disease | ERR1578695 | | |
| ERP009131 | The initial state of the human gut microbiome determines its reshaping by antibiotics | ERR719489;<br>ERR719406;<br>ERR719401;<br>ERR719424;<br>ERR719642 | | |
| ERP003612 | Richness of human gut microbiome correlates with metabolic markers | ERR321165 | | |
| ERP005989 | Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life | ERR525816 | | |
| SRP080787 | Mongolian Metagenome | SRR3992959 | China | 43.95 N<br>116.16 E |
| SRP059392 | Ecological reactor Metagenome | SRR2062623 | USA | 43.727094 N<br>72.425964 W |
| SRP008047 | A Metagenome-Wide Association Study of gut microbiota identifies markers associated with Type 2 Diabetes | SRR341594 | | |
| DRP003048 | Metagenomics of Japanese gut microbiomes | DRR042632;<br>DRR042593;<br>DRR042410 | | |
| SRP002523 | Metagenomic analysis of viruses in the fecal micorobiota of monozygotic twins and their mothers | SRR073436;<br>SRR073432 | | |
| ERP003671 | Deep Illumina-based shotgun sequencing reveals dietary effects on the structure and function of the faecal microbiome of growing kittens | ERR318688 | USA | |
| ERP014654 | microbial diversity and function | ERR1333182 | | |
| ERP004605 | An integrated catalog of reference genes in the human gut microbiome | ERR414735;<br>ERR414539 | Spain:Madrid | 40.463667,-3.74922 |

Notes:

1. The sample from the daycare study is not yet available from the SRA.
2. Location, latitude, and longitude are provided when they are known.

*Supplementary Table 2. 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 1.*

| Source | Amplicon A | Amplicon B | Amplicon C |
|---|---|---|---|
| Global collaboratory | 192 / 180 / 192 | 172 / 159 / 172 | 208 / 198 / 208 |
| Volunteer sequences | 27 / 27 / 27 | 11 / 11 / 11 | 25 / 25 / 25 |
| COMPARE project | 56 / 56 / 56 | 64 / 64 / 64 | 62 / 62 / 62 |
| Daycare Study | 4 / 4 / 4 | 2 / 2 / 2 | 6 / 6  / 6 |
| Strain resolved metagenomics (SRA) | 12,392 / 11,985 / 2,145 | 10,038 / 9,851 / 1,647 | 9,013 / 8,794 / 1,473 |
| Total | 12,671 / 12,252 / 2,424 | 10,287 / 10,087 / 1,896 | 9,314 / 9,083 / 1,774 |

*Supplementary Table 3. Number of crAssphage reads in fecal metagenomes from rural Malawi and the Amazonas of Venezuela[4].*

| Sample ID | MG-RAST ID | Country | CrAssphage reads | Total reads |
|---|---|---|---|---|
| h47b.1 | mgm4461164.3 | Malawi | 118 | 184,366 |
| h47a.1 | mgm4461163.3 | Malawi | 115 | 254,363 |
| amzc5chldm | mgm4461140.3 | Venezuela | 4 | 213,186 |

*Supplementary Table 4. Primer sequences. Primer A, expected product size: 1,331 bp. Primer B: 1,354 bp. Primer C: 1,238 bp.*

| Primer | Expected product size | Sequence | Position on RefSeq NC_024711 |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Primer A Fwd | 1,331 bp | CTGATAGTATGATTGGTAAT | 25,634 .. 25,653 |
| Primer A Rev | | ATAAGTTCTCCAACTATCTT | Complement (26,945 .. 26,964) |
| Primer B Fwd | 1,354 bp | CCAGTATCTCCATAAGCATC | 33,709 .. 33,728 |
| Primer B Rev | | GTGAGGGCGGAATAGCTA | Complement (35,045 .. 35,062) |
| Primer C Fwd | 1,238 bp | GCAACAGGAGTAGTAAAATCTC | 43,820 .. 43,841 |
| Primer C Rev | | GCTCCTGTTAATCCTGATGTTA | Complement (45,036 .. 45,057) |

*Supplementary Table 5. PCR reaction mixture.*

| Reagent | Volume |
|---|---:|
| DNA template | 7.0 µl |
| 2x Master Mix | 25.0 µl |
| Forward primer (10 µM) | 2.0 µl |
| Reverse primer (10 µM) | 2.0 µl |
| DNAse free water | 14.0 µl |
| Total volume | 50.0 µl |

*Supplementary Table 6. PCR amplification protocols.*

| Primer A | Primers B & C |
|---|---|
| Denature 95°C for 3 minutes | Denature 95°C for 3 minutes |
| Then 30 cycles of:<br>Denature 95°C for 45 seconds<br>Annealing 42.6°C for 30 seconds<br>Extension 68°C for 90 seconds | Then 30 cycles of:<br>Denature 95°C for 45 seconds<br>Annealing 50°C for 30 seconds<br>Extension 68°C for 90 seconds |
| Final extension 68°C for 5 minutes | Final extension 68°C for 5 minutes |

# Supplementary References

1. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).
2. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
3. Guerin, E. *et al.* Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* **24**, 653–664.e6 (2018).
4. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).