

Supporting Information

Minot et al. 10.1073/pnas.1300833110

SI Methods

Sample Collection, DNA Isolation, and Sequencing. Stool samples were collected from a healthy male individual in accordance with an Internal Review Board-approved protocol. The subject was 23 y old at the start of the study and did not take antibiotics during the course of the experiment. For each virus preparation, ~1 g of stool was suspended in 40 mL of Buffer SM (1) with a Fisher Scientific PowerGen mechanical homogenizer and then was filtered at 0.22 μm . The filtrate was concentrated on a Millipore Centricon Plus-70 100K to ~0.5 mL, resuspending and re-concentrating once with 40 mL of additional Buffer SM. The concentrate was incubated with 40 μL of chloroform at room temperature for 10 min and was treated with Dnase I at 37 $^{\circ}\text{C}$ for 10 min. Then the remaining DNA was isolated using a QIAGEN DNeasy Blood and Tissue Kit (which includes a proteinase K step to degrade viral capsids). The chloroform treatment was included to disrupt cell membranes but could also have disrupted membrane-enclosed viruses, although this group appears to be rare in the gut.

Each DNA sample was amplified in triplicate with Genomiphi (GE Healthcare), and samples were pooled. Sequencing libraries were prepared with the Illumina TruSeq DNA Sample Preparation Kit v2 with one unique barcode per sample replicate. Sequencing was performed on an Illumina HiSeq2000 with 100 bp \times 2 chemistry at the Penn Genome Frontiers Institute.

Total DNA was extracted from a subset of samples using the QIAamp DNA Stool Kit. Library construction and sequencing were carried out in the same manner as the viral DNA samples, excluding the Genomiphi amplification and pooling.

Levels of contaminating human DNA were assayed using quantitative PCR (qPCR) for β -tubulin 2A. The probe/primer pairs used were from ABI (TaqMan Gene Expression Assays; Part number: 4331182, Assay ID number: Hs00742533_s1). The qPCR reaction contained 12.5 μL TaqMan Fast universal master mix, 1.25 μL DNase-free water, 1.25 μL probe/primer mix, and 10 μL genomiphi-amplified sample at 1 ng/ μL . The cycling conditions were 1 \times (20 s at 95 $^{\circ}\text{C}$) followed by 40 \times (3 s at 95 $^{\circ}\text{C}$, 30 s at 60 $^{\circ}\text{C}$). The amount of β -tubulin detected in viral DNA preparations after purification was below the limit of detection (one copy of β -tubulin per qPCR reaction).

To assay levels of contaminating bacterial DNA, qPCR was used to quantify the V1–V2 16S DNA regions using TaqMan Environmental Master Mix from ABI. Each qPCR reaction contained 1.98 μL DNase-free water, 0.62 μL probe, 0.225 μL primer F BSF8, 0.225 μL primer R BSR357, 12.5 μL 2 \times TaqMan master mix, and 10 μL of genomiphi-amplified sample at 1 ng/ μL . All oligonucleotide stocks were used at 100 μM concentration. Virus-like particle DNA preparations yielded 2.8E1–3.4E1 copies of 16S DNA per nanogram of DNA isolated.

Hybrid Sequence Assembly and Mapping. Raw reads were trimmed to Q35 with a minimum length of 80 nucleotides using FASTX. Contigs were assembled with MetaDBA (2) independently across all samples [our preferred pipeline, OPTITDBA (3), could not be used because of issues of computational feasibility with so large a data set]. Contigs were clustered across all samples using promer (4) in an iterative fashion, such that smaller contigs were removed if they aligned to a larger contig at an identity of 90% over 90% of their length. Only contigs 1 kb or longer were retained. All resulting contigs were assembled using Minimo (5) and the following flags: MIN_LEN = 1,000, ALN_WIGGLE = 15, FASTA_EXP = 1. The resulting contigs

corresponded to either a single contig from the initial round of assembly or an alignment-based consensus of multiple contigs from the first round. To estimate contig abundance and to characterize sequence diversity, reads were aligned to the resulting contigs using Bowtie2 (6). ORFs were predicted using Glimmer (7).

Reproducibility of Contig Detection. The correlation coefficients for detection between replicate samples shown in Fig. S1. Day – R² are

0–0.9855
3–0.9914
11–0.9743
12–0.989
13–0.9912
21–1
22–0.999
23–0.9982

Taxonomic Assignment of Contigs. The complete collection of viral contigs with assigned taxonomy was downloaded from the National Center for Biotechnology Information and annotated by Family. Each contig from this study was compared with this taxonomically defined group using Blastp. Taxonomy was assigned using a voting system. For each ORF, the best-hit taxonomy was used, and the taxonomy of the entire contig was taken as the majority assignment for all of the constituent ORFs. A minimum threshold of one ORF per 10 kb was taken to exclude contigs with only limited regions of similarity.

MetaPhlan (8) was used to assign bacterial taxonomy to samples using unassembled reads. Reads were compared with the MetaPhlan database using Bowtie2, and MetaPhlan was run on the output using standard settings.

Quantification of Base Substitutions. Substitutions were quantified by parsing the pileup files generated by SAMTOOLS from the Bowtie2 mapping BAM output. For each contig, the proportion of base substitutions between any two time points was calculated as the mean proportion of bases that are different across every position, normalized for sequencing depth, with a minimum sequencing depth of 10-fold. Note that diversity was assessed by analyzing the raw reads, not consensus sequences. The rate of substitution was taken as the proportion of nucleotide changes per time unit (day) separating each pair of samples. For each contig, the substitution rate was normalized to the basal rate of variation accountable to sampling and sequencing error, which was estimated as the substitutions between technical replicates. A variety of models were used to fit the observed distribution of substitutions per time unit of separation, and a linear fit was found to have the best support.

Consensus genomes were found by taking the majority aligned nucleotide at each position, for each sample, with a minimum required depth of 10-fold.

Phylogenetic Analysis of Microviridae Microviridae contigs from this study and ref. 9 and finished genomes were compared phylogenetically by aligning the major capsid protein (F) and constructing

a neighbor joining tree in MEGA (10). Bootstrap replicates were used to quantify support for nodes.

Clustered Regularly Interspaced Short Palindromic Repeats Prediction and Analysis. Contigs were generated from total shotgun DNA contig sequences using MetaIDBA (maxk = 80). Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) were predicted in contigs using PILER-CR (11) and were validated manually. Closely related repeats potentially differentiated because of sequencing error were combined manually. Because of the difficulty of correctly assembling CRISPR arrays from metagenomic samples, spacer sequences were extracted from unaligned reads in the following manner. Spacers were extracted from bacterial reads if they were flanked by copies of a single repeat using custom scripts. The potential targets of each spacer were found by comparing those spacers with the collection of viral contigs using Blastn (12). For each of the targeted regions, the reads aligning at each time point were extracted from the BAM alignment files generated by Bowtie2 using the Rsamtools package. The transcriptional direction of the spacer that targets each viral contig is not known.

Diversity-Generating Retroelements. Hypervariable regions were found as described previously (9), using an R script that uses a sliding window to find regions of viral contigs with high proportions of unique alleles (Settings: minimum depth, 10; minimum contig length, 2,000 bp; unique allele frequency, 0.25; window size, 50 bp; minimum region size, 110 bp; minimum SNP frequency, 0.01). Those contigs also were compared with curated collections of reverse-transcriptase (RT) protein sequences using Hidden Markov Model matrix PF00078. Diversity-generating repeats (DGRs) were found by manual inspection of the contigs with both an annotated RT ORF and a hypervariable region. The location of template repeat/variable repeat pairs was determined by finding all significant local alignments of each contig to itself using BLASTN. For each ORF that contains a hypervariable region, the predicted protein fold was found using the Phyre2 server (13). Those ORFs also were compared with the Conserved Domain Database of protein motifs using RPSBLAST. Contigs inferred to contain a DGR by the above method were 231_106, 38, 42, 032_43621, 166, and 90.

Two different methods were used to detect longitudinal DGR activity. In the first, we analyzed only those contigs with adequately deep sampling on either side of the 22-mo gap between sampling points, which is the longest gap in the study. The two contigs that met this criterion (42 and 032_43621) were compared by calculating distances for DGR sequence collections between all pairs of time points using a custom script, including the duplicate within-time point samples. The collections of distances then were compared. The distances between pairs spanning the 22-mo gap were compared with distances between pairs separated by shorter times. Distances also were compared between within-time point replicates. For contig 42, the distances over the 22-mo gap were significantly larger than over shorter periods ($P < 0.0001$) or for the within-time point replicates ($P < 0.0001$), but results for contig 032_43621 did not achieve significance. Each element also was tested for a significant difference in the set of alleles found at the beginning and end of the sampling period.

The difference in distribution of alleles was used to calculate the Chi-statistic, and significance was assessed by comparison with 10,000 random permutations of the data, using a threshold of 0.05. By this measure, DGRs on contigs 42 and contig 38 were active. However, inspection of data for contig 38 did not show a clear longitudinal pattern, and deep sequence data were available only over a 5-d time window, so detection of activity for contig 38 must be taken as tentative.

Confirmation of the Sequence of Contig 122_321 by the Sanger Method. Contig 122_321 was chosen for sequence confirmation using the Sanger method. Primer pairs were designed and used to amplify a nearly complete genome 6.5 kb in size. Sequence was acquired from the day 0 time point. The size of the 6.5-kb amplification product was as predicted from the assembled Illumina data. Further confirmation was provided by the sizes of amplicons used to validate sequence variation described in the next section. Sanger sequence was acquired from the 6.5-kb amplicon using primers described in Table S5. The consensus from the Sanger sequence analysis closely matched the consensus from the day 0 time point with more than 96% identity. Thus, we conclude that the sequence determined from Illumina short reads followed by deBruijn graph assembly yielded an accurate picture of the contig 122_321 genome.

Confirmation of Longitudinal Sequence Diversification in Microviridae Contigs 122_321 and 001_39 Using the Sanger Method. For contig 122_321, a 463-bp region on the contig that was observed to have high base substitution (4.5%) over time in the Illumina metagenomic data was analyzed using Sanger sequencing to confirm the longitudinal changes. The day 0 and day 883 time points were sequenced in triplicate and quadruplicate, respectively. The predominant peak on each sequencing chromatogram was used to determine the bases present.

Levenshtein distances between the time points were calculated to evaluate base substitution. Base substitution of 6.7% (31 bases) occurred between day 0 and day 883. The time 0 Sanger reads diverged from the Illumina consensus for time 0 by up to 2%. All four of the day 883 Sanger sequences were identical to the day 883 consensus in the Illumina data set. Thus, the variation in contig 122_321 inferred from the Illumina data paralleled the Sanger data.

For contig 001_39, a 598-bp region was studied. At day 0, the consensus sequences from the Sanger and Illumina sequencing methods were identical. There were not enough reads available to generate a consensus sequence for the day 883 sample. Sanger data for the samples from day 0 and day 883 were compared and found to differ by 51 bases (8.5%).

Thus, Sanger sequencing showed extensive variation in Microviridae over the time course studied, paralleling the Illumina data.

Comparison with T7. The bacteriophages annotated as podophage were compared with the well-known phage T7 to assess similarity within this group. None of the phages from this study were close in sequence; for alignments over the capsid region, the best match showed 35% identity over a region of 40 aa, which had an eval of 0.02.

1. Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning, a Laboratory Manual* (Cold Spring Harbor Lab Press, Cold Spring Harbor, NY).
2. Peng Y, Leung HC, Yiu SM, Chin FY (2012) IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28(11):1420–1428.
3. Minot S, Wu GD, Lewis JD, Bushman FD (2012) Conservation of gene cassettes among diverse viruses of the human gut. *PLoS ONE* 7(8):e42342.
4. Delcher AL, Salzberg SL, Phillippy AM (2003) Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics* Chapter 10:Unit 10.13.
5. Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M (2011) Next generation sequence assembly with AMOS. *Curr Protoc Bioinformatics* Chapter 11:Unit 11.18.
6. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.
7. Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res* 40(1):e9.
8. Segata N, et al. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 9(8):811–814.
9. Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD (2012) Hypervariable loci in the human gut virome. *Proc Natl Acad Sci USA* 109(10):3962–3966.
10. Tamura K, et al. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10):2731–2739.

11. Edgar RC (2007) PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* 8:18.
 12. Camacho C, et al. (2009) BLAST+: Architecture and applications. *BMC Bioinformatics* 10:421.

13. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the Web: A case study using the Phyre server. *Nat Protoc* 4(3):363–371.

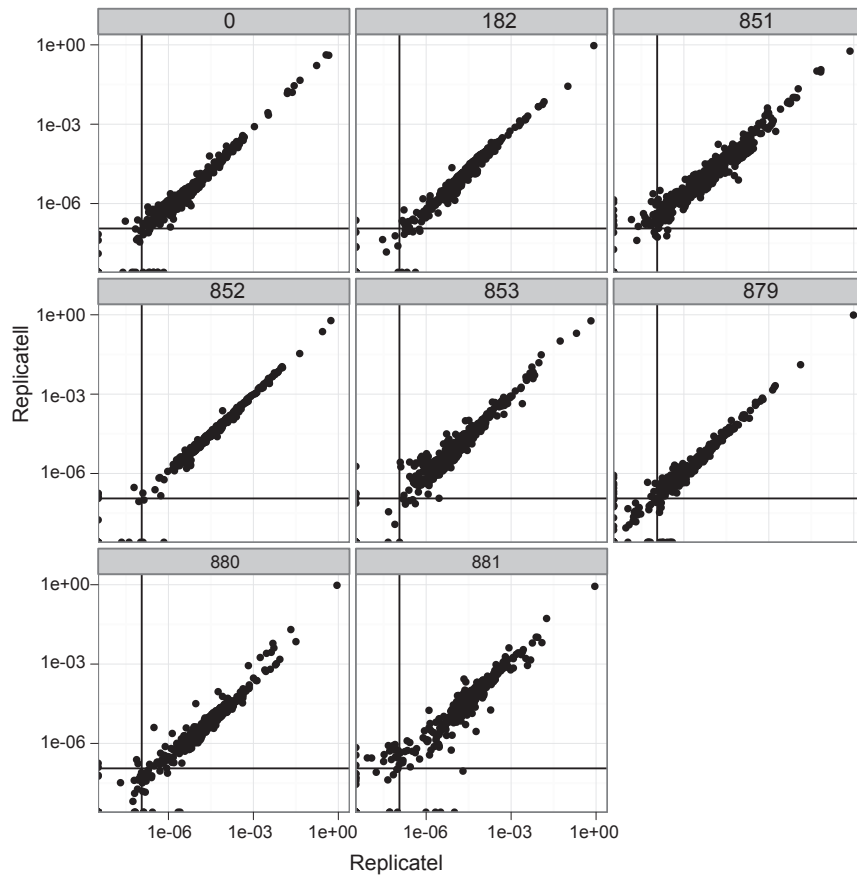


Fig. S1. Reproducibility between replicates. Each point represents the normalized abundance of a contig in a pair of replicate virome samples from the same time point. All contigs and pairs of technical replicates are represented in the figure.

Table S1. Virus like particle sequence sample characteristics

Sampling day	Replicate	Read	Aligned reads	Percent aligned
0	1	20,312,322	18,331,267	90.25
0	2	24,435,114	22,289,142	91.22
180	1	24,875,744	24,691,054	99.26
181	1	23,959,436	23,841,743	99.51
182	1	15,505,820	15,208,373	98.08
182	2	16,812,218	16,706,701	99.37
183	1	17,774,454	17,264,914	97.13
184	1	19,893,608	19,702,285	99.04
851	1	25,101,704	25,084,614	99.93
852	2	27,714,314	27,697,853	99.94
852	1	28,434,716	28,387,692	99.83
852	2	29,751,810	29,692,193	99.80
853	1	15,903,684	15,896,899	99.96
853	2	25,144,920	25,123,536	99.92
854	1	29,634,128	29,623,353	99.96
855	1	22,257,952	22,255,753	99.99
879	1	16,071,666	16,070,799	99.99
879	2	16,405,346	16,402,036	99.98
880	1	21,052,128	21,046,503	99.97
880	2	19,138,984	19,136,500	99.99
881	1	29,389,988	29,372,515	99.94
881	2	35,751,748	35,724,342	99.92
882	1	29,211,340	29,205,227	99.98
883	1	39,238,778	39,234,718	99.99

Table S2. Assignment of phage contigs to bacterial hosts

Contig	Length (bp)	Bacterial (GI)	Bacterial species	Connection
111_52	36,084	60491031	<i>Bacteroides fragilis</i> NCTC 9343	CRISPR
132_57	7,156	291556121	<i>Eubacterium siraeum</i> V105c8a	CRISPR
232_308	5,336	291541372	<i>Ruminococcus bromii</i> L2-63	CRISPR
111_107	5,222	224485451	<i>Bacteroides</i> sp. 2_1_7	Prophage
221_131	4,177	224485451	<i>Bacteroides</i> sp. 2_1_7	Prophage
231_217	5,118	224485451	<i>Bacteroides</i> sp. 2_1_7	Prophage
021_4	37,938	319644666	<i>Bacteroides</i> sp. 3_1_40A	Prophage
031_147	4,924	319644666	<i>Bacteroides</i> sp. 3_1_40A	Prophage
231_103	11,455	319644666	<i>Bacteroides</i> sp. 3_1_40A	Prophage
231_91	13,236	319644666	<i>Bacteroides</i> sp. 3_1_40A	Prophage
38	20,452	256402715	<i>Blautia hansenii</i> DSM 20583	Prophage
44	5,669	256402715	<i>Blautia hansenii</i> DSM 20583	Prophage
231_106	26,844	256402715	<i>Blautia hansenii</i> DSM 20583	Prophage
74	10,031	331640228	<i>Lachnospiraceae bacterium</i> 3_1_46FAA	Prophage
232_270	6,065	331640228	<i>Lachnospiraceae bacterium</i> 3_1_46FAA	Prophage
232_349	4,323	331640228	<i>Lachnospiraceae bacterium</i> 3_1_46FAA	Prophage
011_27	27,157	291541372	<i>Ruminococcus bromii</i> L2-63	Prophage
117	15,472	291541372	<i>Ruminococcus bromii</i> L2-63	Prophage
107	36,432	291541372	<i>Ruminococcus bromii</i> L2-63	Prophage

Shown are viral contigs assigned to bacterial hosts by both CRISPR spacer matches and annotation as prophage in sequenced genomes.

Table S3. Variation in DGR contigs over time

Contig	Significant change over time	ORF length	CDD (hit - bit score - evalue)	Phyre2
231_106	No	381	164824 - MTD - 104-5e-25	Clec (MTD)
38	No	381	164824 - MTD - 108-2e-26	Clec (MTD)
42	Yes	351	32846 - FxsA - 28.2-3.7	Clec
032_43621	No	603	48198 - GlucD - 34.5-0.15	Clec (MTD)
166	No	592	145488 - Big_2 - 37.3-0.001	Ig superfamily (α -amylase)
90	No	365	164824 - MTD - 80.2-1e-16	Clec (MTD)

Contigs queried for significant variation and gene types affected. CDD, conserved domains database; MTD, major tropism determinant.

Table S4. Nucleotide divergence among Microviridae from the International Committee on Taxonomy of Viruses

	Chp2	Alpha3	St-1	ID18	WA13	phiX174	G4	ID2 Moscow/ID/2001	Chp4	PhiCPG1	Chp3
<i>Chlamydia</i> phage Chp2		0.0	0.0	0.0	0.0	0.0	0.0	0.0	90.8	91.5	96.9
Enterobacteria phage alpha3	0.0		90.1	0.0	63.6	5.0	0.0	0.0	0.0	0.0	0.0
Enterobacteria phage St-1	0.0	90.1		0.0	62.2	5.0	0.0	0.0	0.0	0.0	0.0
Enterobacteria phage ID18	0.0	0.0	0.0		4.1	5.6	44.3	70.7	0.0	0.0	0.0
Enterobacteria phage WA13	0.0	63.6	62.2	4.1		5.3	0.0	3.9	0.0	0.0	0.0
Enterobacteria phage phiX174	0.0	5.0	5.0	5.6	5.3		0.0	6.4	0.0	0.0	0.0
Enterobacteria phage G4	0.0	0.0	0.0	44.3	0.0	0.0		47.8	0.0	0.0	0.0
Enterobacteria phage ID2 Moscow/ID/2001	0.0	0.0	0.0	70.7	3.9	6.4	47.8		0.0	0.0	0.0
<i>Chlamydia</i> phage 4	90.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0		94.8	90.6
<i>Chlamydia</i> phage PhiCPG1	91.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	94.8		91.0
<i>Chlamydia</i> phage 3	96.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	90.6	91.0	

Entries in the matrix show the identity between the isolates compared.

Table S5. Oligonucleotides used in this study

Position relative to 122_321	Orientation	Name	Sequence	Comments
408	F	122_321_408_F	TTCGCTAGCCAACAGTCCTT	Sequencing
427	R	122_321_427_R	AAGGACTGTTGGCTAGCGAA	Sequencing/ amplify 122_321
496	F	122_321_496_F	TGTACTTCGGCAGCATTGAG	Sequencing/ amplify 122_321
918	F	122_321_918_F	CGCCGTTTGTCCGTAAGTAT	Sequencing
937	R	122_321_937_R	ATACTTACGGACAAACGGCG	Sequencing
987	F	122_321_987_F	AGGAGCAGTTGCGTTTCCTA	Sequencing
1,299	F	122_321_1299_F	AGAAGCAGCACCTTTTCCAA	Sequencing
1,318	R	122_321_1318_R	TTGGAAAAGGTGCTGCTTCT	Sequencing
2,154	F	122_321_2154_F	AGACCGGAGAATGTTTCGATG	Sequencing
2,173	R	122_321_2173_R	CATCGAACATTCTCCGGTCT	Sequencing
3,238	F	122_321_3238_F	ATTTGGGGCGTGTATTACCA	Sequencing
3,257	R	122_321_3257_R	TGGTAATACACGCCCAAAT	Sequencing
4,072	F	122_321_4072_F	CGGGGTTAATGCGTAAAGAA	Sequencing
4,091	R	122_321_4091_R	TTCTTTACGCATTAACCCCG	Sequencing
4,653	F	122_321_4653_F	GACGAGCATAAACACGAGCA	Sequencing
4,672	R	122_321_4672_R	TGCTCGTGTATGCTCGTC	Sequencing
6,121	F	122_321_6121_F	GGCACGAAAAGACCATTGTT	Sequencing
6,140	R	122_321_6140_R	AACAATGGTCTTTTCGTGCC	Sequencing

F, forward; R, reverse.