Supporting Information

Minot et al. 10.1073/pnas.1119061109

SI Methods

Sample Collection. Stool samples were collected from 12 healthy adult volunteers that were enrolled in a controlled feeding study, as described in refs. 1 and 2. Subjects were at least 18 y old, had a body mass index between 18.5 and 35, were free of gastrointestinal disorders, and had not consumed antibiotics within 6 mo before sample collection. Collections proceeded according to protocols approved by the University of Pennsylvania institutional review board. Six of the 12 subjects were the same as sampled previously (2) and all 12 were studied in ref. 1. In ref. 2, multiple separate samples were sequenced for each subject, but in this study we pooled DNA from three samples per subject for sequencing.

Isolation and Sequencing of Viral DNA. Viral DNA was isolated from these stool samples as previously described (2). Approximately 0.5 g of stool was resuspended through homogenization in 40-mL SM buffer (3). Centrifugation at 4,700 × *g* was carried out for 30 min to remove large solids, and the resulting supernatant was passed through a 0.22-μm PES filter (Nalgene). The 0.22-μm filtrate was loaded onto a CsCl density step gradient (described further in refs. 3 and 4), finally extracting the middle (1.35–1.5 g/mL) fraction from the column using a sterile syringe. This study made no attempt to isolate RNA viruses or DNA viruses with densities outside this range. Chloroform was incubated with these samples for 10 min before DNase (Invitrogen) treatment for 10 min at 37 °C. Finally, viral DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen). This viral DNA was amplified with phi29 polymerase using random hexamer primers before pooling and sequencing (Genomiphi V2; GE Healthcare).

Viral DNA purity was assessed by quantifying the abundance of bacterial 16S rDNA, which is never encoded in viral genomes, via quantitative PCR. Ten separate viral extractions were quantified in triplicate, using a plasmid standard curve to determine the number of 16S copies per nanogram DNA. First, the number of 16S copies per nanogram in bacterial DNA was estimated using an average genome size of 5 Mb and an average of five 16S copies per genome. Analysis of total stool DNA shows ~10^6 16S copies per nanogram, consistent with the majority of DNA originating from bacterial genomes (1). The number of 16S copies per nanogram in the isolated viral DNA was 825 (SD = 805). The relative proportion of bacterial DNA in these viral preparations (compared with the total) was therefore estimated to be 9.0×10^{-4} (SD = 8.7×10^{-4}).

Three separate samples were pooled for each subject in this study, following isolation and amplification. Those pooled samples were randomly sheared by sonication (Covaris), and barcoded sequencing adapters were ligated using the Illumina TruSeq DNA Sample Prep Kit (Illumina). The same set of pooled samples was also prepared for Illumina sequencing using the Nextera DNA Sample Prep Kit (Epicentre). The Illumina-prepared and Nextera-prepared samples were each pooled independently and sequenced on their own single lane of a HiSEq. 2000 flow cell (Illumina). One lane of that same flow cell was set aside for Illumina control DNA isolated from the bacteriophage ΦX174.

Assembly and Mapping of Viral Sequences. Viral sequences were trimmed by quality score (cutoff at Q35 using FASTX v0.0.13), and then assembled using SOAPdenovo (v1.05) (5) (k-mer size = 63, additional flags "-p 20 –M 3 –u –G 200 –R"). We found that optimal assembly occurred when the reads from each sample were assembled separately, and when the largest possible k-mer size was used. An insert size of 300 bp was chosen based on the fragment size that was selected for sequencing. Reads were mapped back to those contigs using the Burrows-Wheeler Aligner (BWA v0.5.9-r16) (6) and the resulting alignments were visualized using the Integrative Genomics Viewer (IGV v2.0) (7). ORFs were predicted using Glimmer (v3.02) (8), and functions were predicted using RPSBLAST (9) (v2.2.20) against the Pfam and National Center for Biotechnology Information (NCBI) Conserved Domain Database (10) (accessed 3/28/2011).

The contigs generated above were compared with the RefSeq collection of viral sequences using BLASTn (9). The five genomes with the largest amount of sequence that was similar to at least one contig were selected, and raw reads were mapped to those contigs using BWA, as above. The pile-up figures were generated using IGV.

Identification of Variable Regions. Variable regions were identified using a custom R script (available upon request) that uses Rsamtools to parse the BAM alignment files output by BWA (above). We estimated the basal error rate of Illumina sequencing by mapping control reads to the ΦX174 genome (gi 9626372). After excluding positions with >0.1 polymorphism (suggesting heterogeneity in the starting population), the error rate was calculated as the proportion of bases that did not match the reference. The script scans along every contig in a 50-bp window (step size = 5 bp) and extracts the sequences that cover that region completely. For each window, we calculated the number of sequences, the proportion of those sequences that were unique (complementary to the proportion of sequences that were a duplicate of another), and the proportion of bases that did not match the consensus sequence. The criteria we chose to identify the most variable elements in this dataset were a minimum of five sequences, 0.4 unique alleles, and 0.05 polymorphic bases. Importantly, we required that nine adjacent windows (a total of 90 contiguous base pairs) fulfilled these criteria.

Manual Resequencing of Contigs. Primers were selected that flanked the target gap using Primer3 (v0.4.0) (11). The region of interest was amplified using AccuPrime Taq (Invitrogen), and the following thermocycler program: 94 °C for 15 s, 30 cycles of 94 °C for 15 s and 68 °C for 3 min, and finally 68 °C for 10 min and cool to 4 °C. The resulting PCR products were purified using a QIAquick PCR Purification Kit (Qiagen) and either sequenced directly, or cloned into a TOPO-TA vector (Invitrogen) for Sanger sequencing. The resulting Sanger reads were used in combination with shotgun reads to manually repair contigs, closing gaps with the new sequences. Those repaired contigs were then put through the analysis pipeline above, including read mapping, functional prediction, polymorphism scanning, and so forth. Contigs containing variable regions listed in Table S1 have been submitted to GenBank. Raw Solexa/Illumina data are available upon request.

Taxonomic Classification of Variable Contigs. Each variable contig was compared with the viral proteins in RefSeq using BLASTx with a cutoff of $e \leq 10^{-40}$. The taxonomic classification of each RefSeq genome was retrieved from the NCBI Web site. When hits overlapped, the hit with the lowest e-value was retained. The taxonomic classifications of those nonoverlapping hits were recorded in Table S1 using the following abbreviations: P represents Podoviridae, S represents Siphoviridae, and M represents Myoviridae. When one contig resembled reference genomes from multiple viral families, all of the matching families were recorded (e.g., S/M).

Sequence Structure Adjacent to Variable Repeat. Recent work has identified short hairpins (8-bp stem, 4-bp loop, 20-bp total) located in the initiation of mutagenic homing region at the 3′ end of the variable repeat (VR) as essential for diversity-generating retroelement function (12). We identified short hairpins in this dataset using a custom R script that scanned in 26-bp windows looking for hairpins with either even or odd numbers of bases in the loop, and at least 7 bp in the stem. An additional characteristic of the initiation of mutagenic homing region is a 14-bp GC-only sequence. We identified GC-only sequences using a custom R script that scanned each contig in 12-bp windows, identifying each GC-only window, and then merging overlapping windows. The R script correctly called the experimentally verified signals in *Bordetella* bacteriophage BPP-1 (12).

Structural Prediction of Variable ORFs. The amino acid sequence of ORFs covering hypervariable regions was generated using custom scripts. The structure of those amino acid sequences was predicted using Phyre2 (13), which uses homology of the input to sequences with known structures to generate the output. A threshold of 95% confidence was used to evaluate the output models.

Phylogenetic Analysis of Reverse-Transcriptase Sequences. Reverse-transcriptase (RT) sequences were identified using homology to the Pfam PF00078 (RPSBLAST; e-value < 0.00001), and the amino acid sequences were found using a custom script. Reference RT sequences were selected from two previous analyses of RTs associated with diversity-generating systems (14, 15), as well as representatives from each family of retroviruses (LTR-group RTs). The RT sequences from this dataset were aligned along with the reference sequences by individual alignment against the HMM contained in PF00078. A master alignment that preserved the position of each sequence relative to that HMM was generated by hmmalign (HMMER v3.0) (16). This method does not involve any comparisons between the selected sequences, but rather relies on their similarity to conserved elements within the curated position-specific scoring matrix that constitutes the Pfam PF00078. An approximate maximum-likelihood tree was generated by FastTree (17). The figure was generated using FigTree (http://tree.bio.ed.ac.uk/software/figtree/), coloring internal branches according to the confidence estimates generated by FastTree. Branch tips were adjusted and circles were added to indicate tips corresponding to novel RT sequences from this dataset. Overlapping circles were merged to avoid overplotting. A distance matrix was also calculated by MEGA using the Poisson-corrected distance, using only the subset of sequences described in (14).

To compare the relative abundance of different clades of RT sequences, we used the sequences (above) that fell into the hypervariation, group II intron, retron, and putatively novel clades to compare against a variety of nucleotide databases using BLASTx (9). The three genome databases that we used were (*i*) the collection of complete and partial viral genomes generated in this study, (*ii*) all of the phage RefSeq genomes, and (*iii*) all of the bacterial RefSeq genomes. For each of the genomes in those databases we recorded the clade that most closely resembled their RT sequences.

1. Wu GD, et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334:105–108.
2. Minot S, et al. (2011) The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Res* 21:1616–1625.
3. Sambrook J, Russell DW (2001) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab Press, Cold Spring Harbor, NY).
4. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc* 4:470–483.
5. Li R, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20:265–272.
6. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
7. Robinson JT, et al. (2011) Integrative genomics viewer. *Nat Biotechnol* 29:24–26.
8. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23:673–679.
9. Camacho C, et al. (2009) BLAST+: Architecture and applications. *BMC Bioinformatics* 10:421.
10. Marchler-Bauer A, et al. (2011) CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 39(Database issue):D225–D229.
11. Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, eds Krawetz S, Misener S (Humana Press, Totowa, NJ), pp 365–386.
12. Guo H, et al. (2011) Target site recognition by a diversity-generating retroelement. *PLoS Genet* 7:e1002414.
13. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the Web: A case study using the Phyre server. *Nat Protoc* 4:363–371.
14. Doulatov S, et al. (2004) Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* 431:476–481.
15. Simon DM, Zimmerly S (2008) A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res* 36:7219–7229.
16. Finn RD, Clements J, Eddy SR (2011) HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res* 39(Web Server issue):W29–W37.
17. Price MN, Dehal PS, Arkin AP (2010) FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490.

Fig. S1. DNA sequences in hypervariable regions. See Table S1 for the location of hypervariable regions on contigs. Each panel depicts a separate contig, with the contig name indicated at the top of each panel. The black and red boxes underneath indicate the position of the VR on the contig. The size of the region shown is indicated between the arrows beneath those boxes. Immediately beneath the scale bar for the region is a bargraph of sequencing depth at each position, with the maximum value indicated in the top left corner of that bargraph. The solid gray arrows above the reference sequence (at bottom) show individual Illumina reads that align to this region of the contig. The sequences are the same as the reference contig unless otherwise indicated. Insertions are shown with purple vertical brackets, and deletions are shown with black dashes.

Fig. S1

Table S1. Summary of contigs containing template repeat/variable repeat pairs and RTs

Table S1