Supplemental Figure Legends

**Figure S1. Alpha diversity metrics based on 16S rRNA gene data.** Rarefied OTU statistics as represented by Chao 1 index for richness, Inverse Simpson Index for diversity and Simpson's Equitability for evenness are displayed. Samples are sorted by Chao 1 index from low to high. Error bars represent 95% high and low confidence intervals. Blue IDs are healthy traveler controls.

**Figure S2. WGS-based taxonomic profiles of prokaryotes.** Top 30 prokaryotic genera by average relative abundance are displayed. Color temperature represents the log_{10} relative abundance of reads mapped to each genus.

**Figure S3. WGS-based taxonomic profiles of non-phage viruses.** Top 50 most abundant virus species whose scientific name does not contain the substring “phage” are displayed. Color temperature represents the log_{10} relative abundance of sequences mapped to each species.

**Figure S4. Functional profiles of the metagenomes.** Identified ORFs were assigned to KEGG pathways, which fall into five major and 33 minor categories. Color temperature represents the proportion of genes (ORFs) in each category, as weighted by (gene length × host contig coverage). The assignment results to four specific features (and databases) are attached at bottom. The proportions were calculated in the same way but the color temperature is normalized within each row so as to highlight the cross-sample comparison. The dendrogram is the hierarchical clustering result based on the KEGG pathway categories only.

**Figure S5. Comparison of binning results by multiple tools.** As exemplified in sample 80129, the binning results of three automated binners: Concoct, MaxBin and MetaBat, are visualized in overlap with the k-mer signature-based scatter plot generated by VizBin. Different combinations of color and shape represent contigs assigned to different bins.

**Figure S6. Illustration of the major steps of the metagenomic assembly and binning workflow designed and used in this work.** Contigs assembled from each DNA sample are displayed in a scatter plot, with multiple properties indicated by different visual elements. Specifically, the x- and y-coordinates are the dimension reduction result computed from k-mer signatures. The area, opacity and color indicate the length, coverage and taxonomic assignment of the contig, respectively. Contigs carrying certain features (e.g., SSU gene, virulence factor)
are selectively highlighted (bordered). Clusters of contigs that may represent individual genomes ("bins") were manually selected from the scatter plot. Closely-related genomes were further separated by observing the distribution pattern of certain metrics, such as coverage (shown in inset). The quality of each candidate bin was assessed in downstream analyses (such as CheckM) and human knowledge, which in turn guided the adjustment of the bins.

**Figure S7. Cross comparison of recovered *E. coli* draft genomes.** The permutation (order) of genomic regions is visualized by colored blocks as aligned using Mauve [108]. Phylogenetically related genomes are placed in close proximity.

**Figure S8. Phylogenetic tree of 320 bins representing cellular organisms.** The positions of taxa are identical to those in Fig. 6. Taxonomic descriptions (Table S7) are marked at tips. Branch lengths are drawn to scale. Black and blue lines represent branches with Bootstrap support ≥ and <75 (out of 100), respectively.

**Figure S9. Phylogenetic tree of recovered and reference TM7 genomes.** TM7x-like strains are in green and the novel TM7z strain is in orange. Bootstrap support values (out of 100) are labeled at nodes. The top left inset shows the two TM7 bins in the scatter plot of contigs in sample 50070.

**Figure S10. Distribution of the taxonomic assignment rates of the 163 putative viral genomes.** Bin names are omitted. The 17 crAssphage genomes and the three other identified singleton viral genomes are colored. The rest of the bins cannot be assigned to any reference viral genomes. The y-axis is scaled to the square root of the actual percentage.
Top 30 prokaryotic genera

- Escherichia
- Ruminococcus
- Blautia
- Eubacterium
- Bifidobacterium
- Streptococcus
- Dorea
- Collinsella
- Clostridium
- Enterococcus
- Akkermansia
- Bacteroides
- Roseburia
- Coprococcus
- Faecalibacterium
- Veillonella
- Enterobacter
- Anaerostipes
- Shigella
- Eggerthella
- Klebsiella
- Lachnoclostridium
- Lactococcus
- Actinomyces
- Citrobacter
- Methanobrevibacter
- Lactobacillus
- Peptoclostridium
- Catenibacterium
- Tyzzerella
log_{10} rel. abund.

Sample ID (TD/HT)

Glypta fumiferanae ichnovirus
Human endogenous retrovirus K
Encephalomyocarditis virus
Pandoravirus salinus
Human herpesvirus 6A
Pandoravirus dulcis
Cotesia congregata bracovirus
Ictalurid herpesvirus 1
Cyprinid herpesvirus 3
Cercopithecine herpesvirus 2
Orgyia pseudotsugata multiple nucleopolyhedrovirus
Cyprinid herpesvirus 1
Hyposoter fugitivus ichnovirus
Human herpesvirus 6B
Pandoravirus inopinatum
Human herpesvirus 7
Cowpox virus
Megavirus chilensis
Phaeocystis globosa virus
Abalone herpesvirus Victoria/AUS/2009
Falconid herpesvirus 1
Spiroplasma kunkeli virus SKV1_CR2-3x
Ectocarpus siliculosus virus 1
Galilid herpesvirus 3
Mucacine herpesvirus 1
Suid herpesvirus 1
Chicken anemia virus
Galilid herpesvirus 2
Moumouvirus
Oryctes rhinoceros nudivirus
Acanthamoeba polyphaga mimivirus
Alcelaphine herpesvirus 2
Camelpox virus
Cercopithecine herpesvirus 5
Chimpanzee alpha-1 herpesvirus
Cyprinid herpesvirus 2
Elephantid herpesvirus 1
Gryllus bimaculatus nudivirus
Hyphantria cunea nucleopolyhedrovirus
Megavirus lba
Megavirus terra1
Meleagrid herpesvirus 1
Mimivirus terra2
Ovine herpesvirus 2
Paramecium bursaria Chlorella virus AR158
Rabbit fibroma virus
Saccharomyces cerevisiae killer virus M1
Taterapox virus
Tobacco vein clearing virus
Agritis ipsilon multiple nucleopolyhedrovirus
Taxonomic assignment: 
**Bowtie2 vs. RefSeq**

DNA sequences → de novo assembly: **IDBA-UD**

Quality assessment: **QUAST**

Dimension reduction: **VizBin**

**Functional annotation:**
- KEGG
- Resfams
- VFDB, etc.

**Cross-validation:** **BLAST**

Completeness / contamination: **CheckM**

Biological knowledge

**Related genomes**

Spatially separated, taxonomically homogeneous contig clusters (single genomes)

Standalone, circularizable contig (plasmid / phage)

Closely-related genomes

Separate by coverage

Separate, then adjust based on multiple properties