PNAS
www.pnas.org

Supplementary Information for

A Catalog of Tens of Thousands of Viruses from Human Metagenomes Reveals Hidden Associations with Chronic Diseases

Authors

Michael J. Tisza and Christopher B. Buck *

Affiliation

Lab of Cellular Oncology, NCI, NIH, Bethesda, MD 20892-4263

Michael J. Tisza: 0000-0003-1168-1617

Christopher B. Buck: 0000-0003-3165-8094

*** Corresponding Author**

Name: Christopher B. Buck

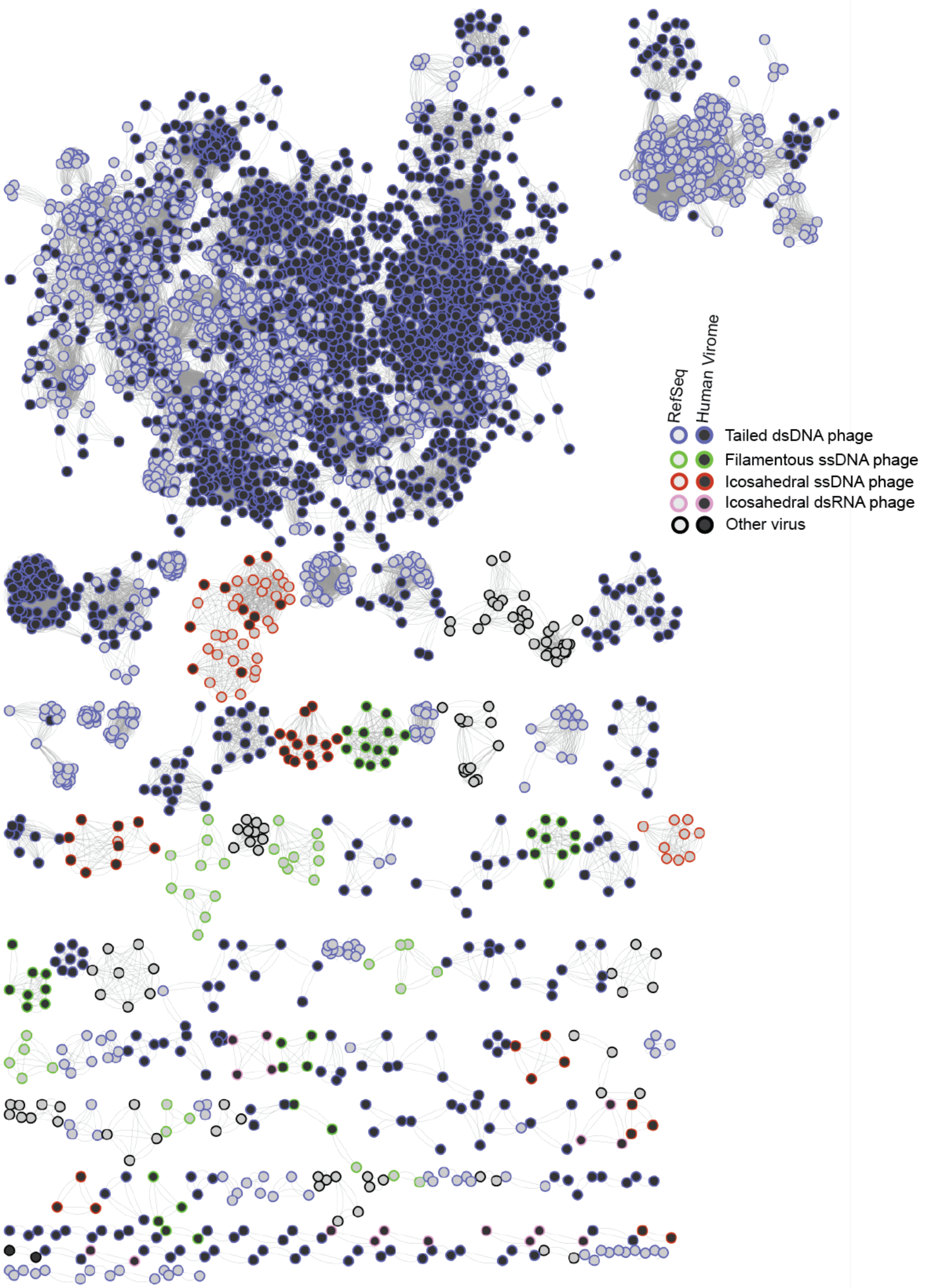
Email: buckc@mail.nih.gov

This PDF file includes:

Figures S1 to S8

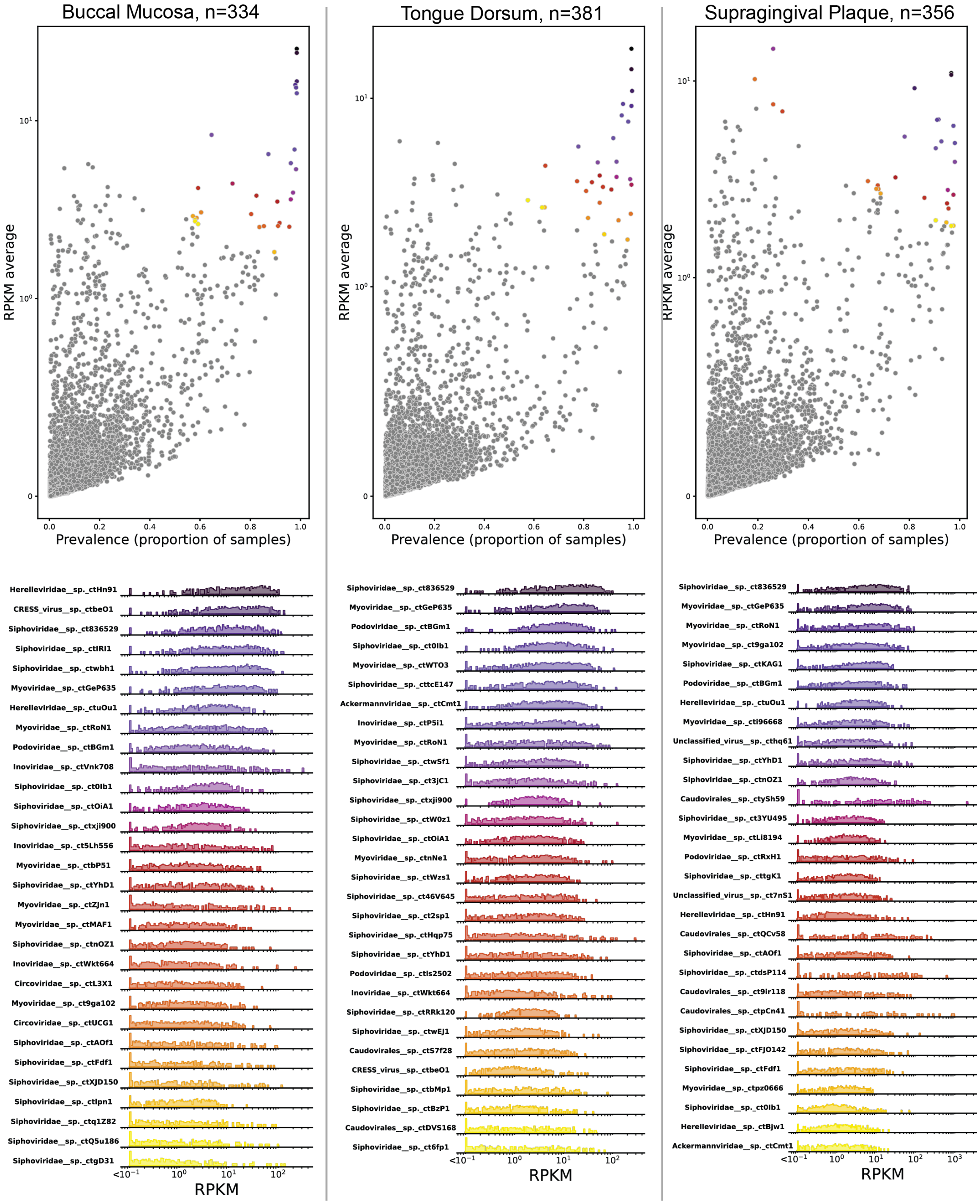
Other supplementary materials for this manuscript include the following:

Datasets S1 to S8K



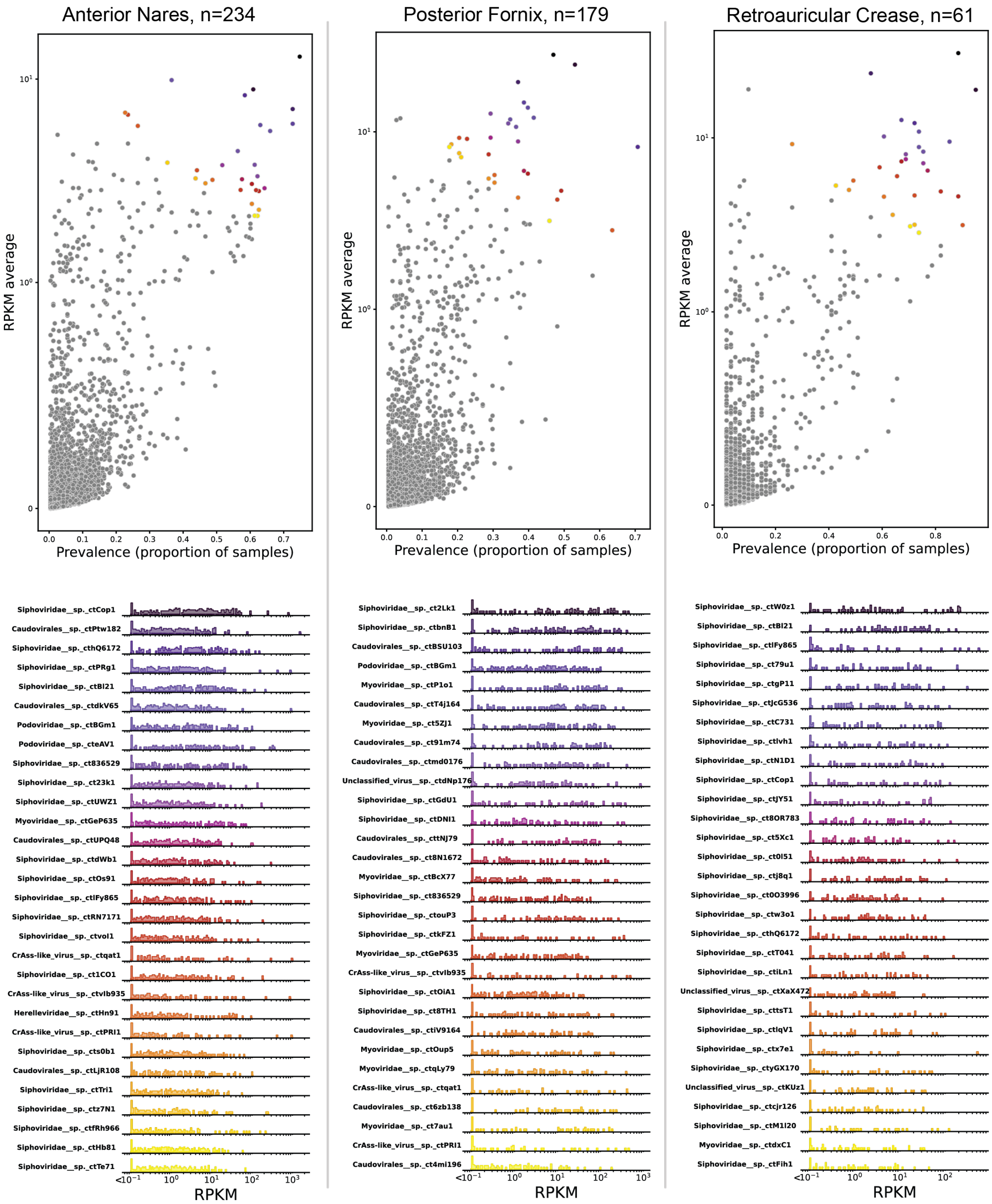
Supplemental Figure 1: Gene sharing network of RefSeq phages and unclassified OTUs
Vcontact2 was used to make a network of RefSeq bacteriophage genomes and CHVD contigs deemed "unclassified viruses." Vcontact2 is more sensitive than the taxonomy module of Cenote-

Taker 2, as Vcontact2 compares all genes encoded on each contig and Cenote-Taker 2 uses only a single hallmark gene for comparison to its taxonomy database. Edges are drawn between nodes if some proportion of their genes share protein sequence similarity. Only clusters of two or more nodes are displayed.   Each sequence cluster was given a feature label, such as "filamentous ssDNA phage," based on manual inspection of the virion hallmark gene calls made by Cenote-Taker 2 (see Supplemental Table 2) for one or more sequences in the cluster.



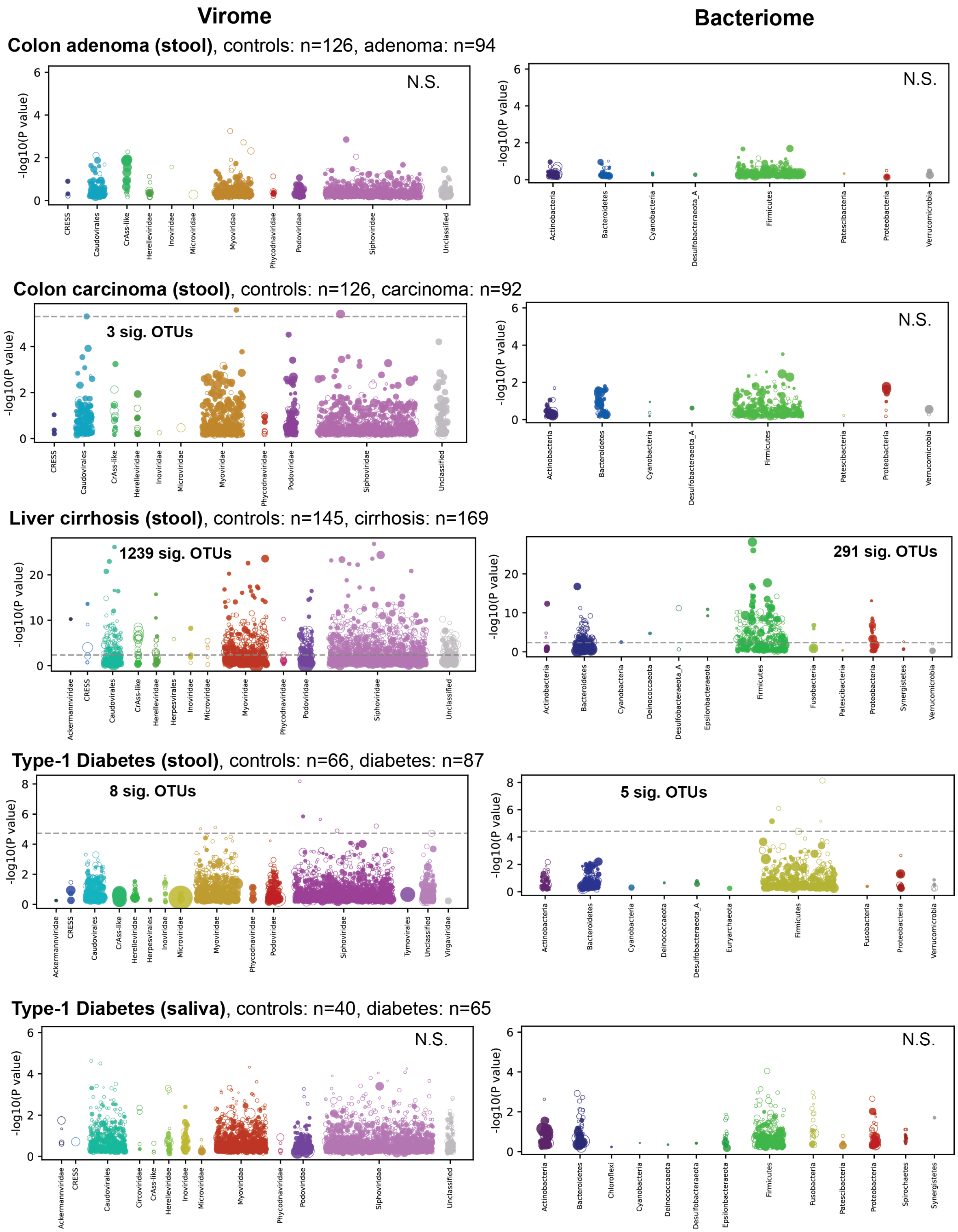
Supplemental Figure 2: Most common viruses, Buccal Mucosa, Tongue Dorsum, and Supragingival Plaque

Scatter plots of virus quantification data displayed as described in the legend of Figure 3.



Supplemental Figure 3: Most common viruses, Anterior Nares, Posterior Fornix, and Retroauricular Crease

Scatter plots of virus quantification data displayed as described in the legend of Figure 5.


Virome

Bacteriome

Colon adenoma (stool), controls: n=126, adenoma: n=94

Colon carcinoma (stool), controls: n=126, carcinoma: n=92

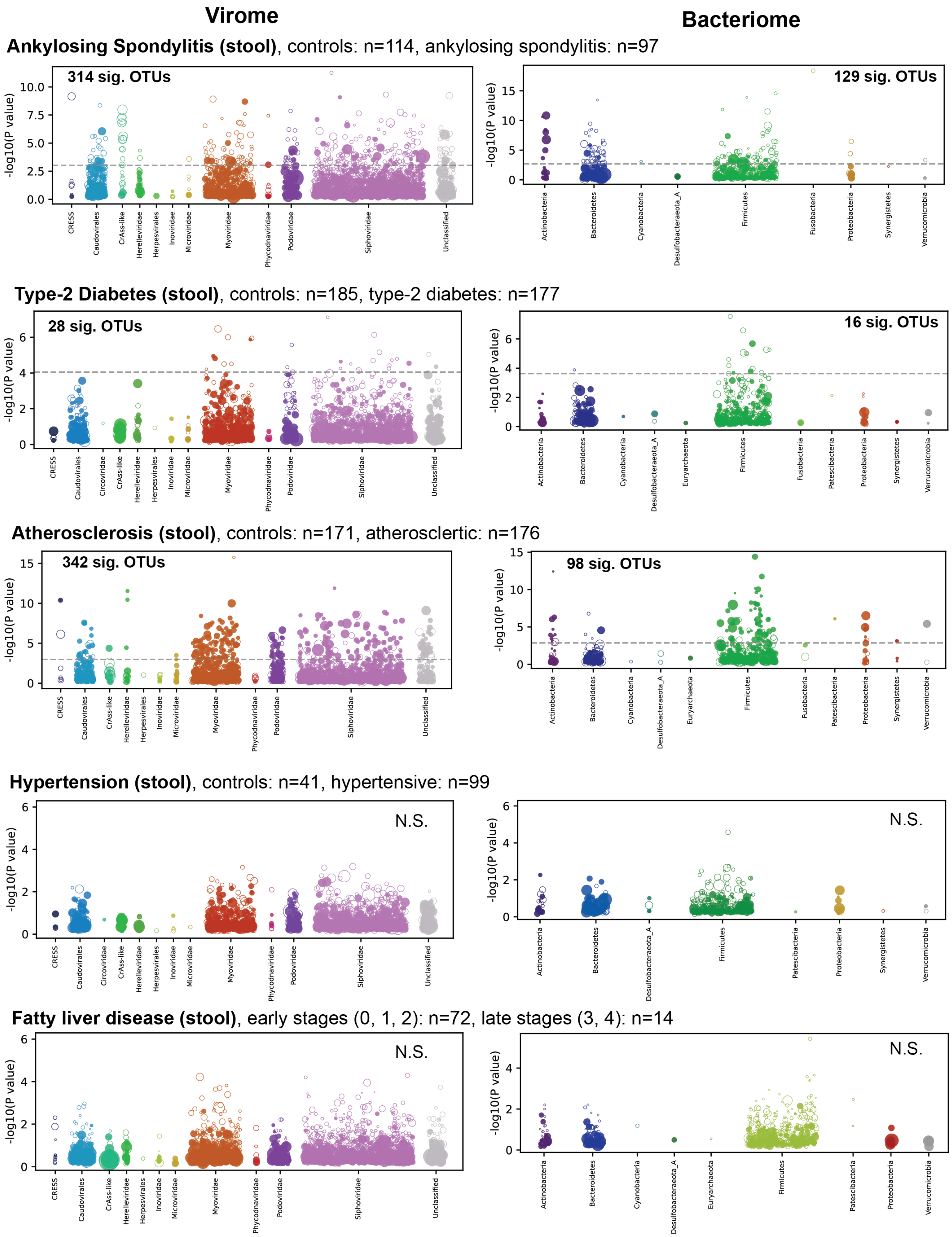
Liver cirrhosis (stool), controls: n=145, cirrhosis: n=169

Type-1 Diabetes (stool), controls: n=66, diabetes: n=87

Type-1 Diabetes (saliva), controls: n=40, diabetes: n=65

Supplemental Figure 4: Virome- and Bacteriome-wide associations with additional chronic diseases 1

Manhattan plots for viromes (left) and bacteriomes (right) are shown in the same manner as Figure 6. Accessions: Colon carcinoma and adenoma (PRJEB7774), Liver cirrhosis (PRJEB6337), Type 1 diabetes (PRJNA289586).



Ankylosing Spondylitis (stool), controls: n=114, ankylosing spondylitis: n=97

Type-2 Diabetes (stool), controls: n=185, type-2 diabetes: n=177

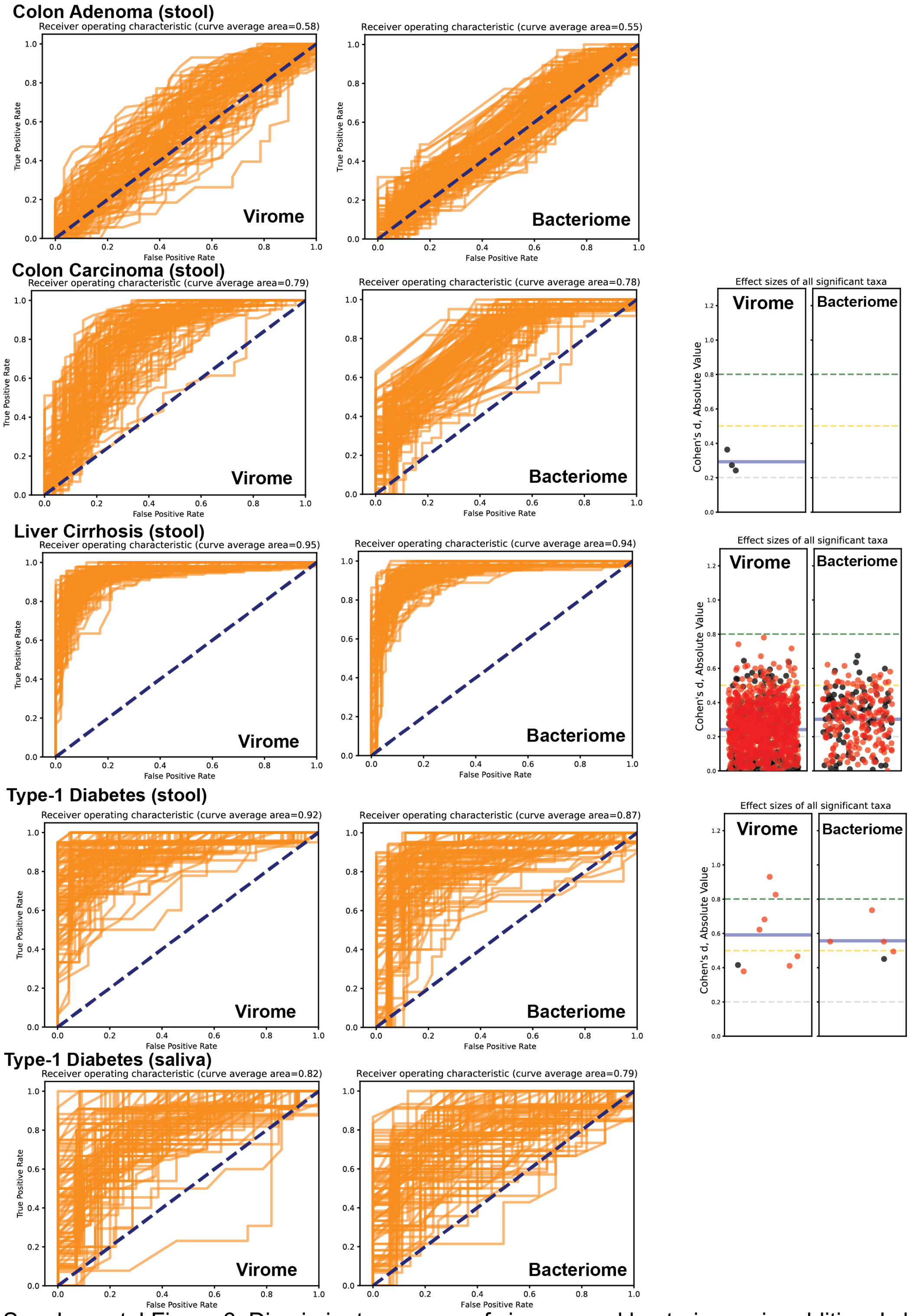
Atherosclerosis (stool), controls: n=171, atheroscleritic: n=176

Hypertension (stool), controls: n=41, hypertensive: n=99

Fatty liver disease (stool), early stages (0, 1, 2): n=72, late stages (3, 4): n=14

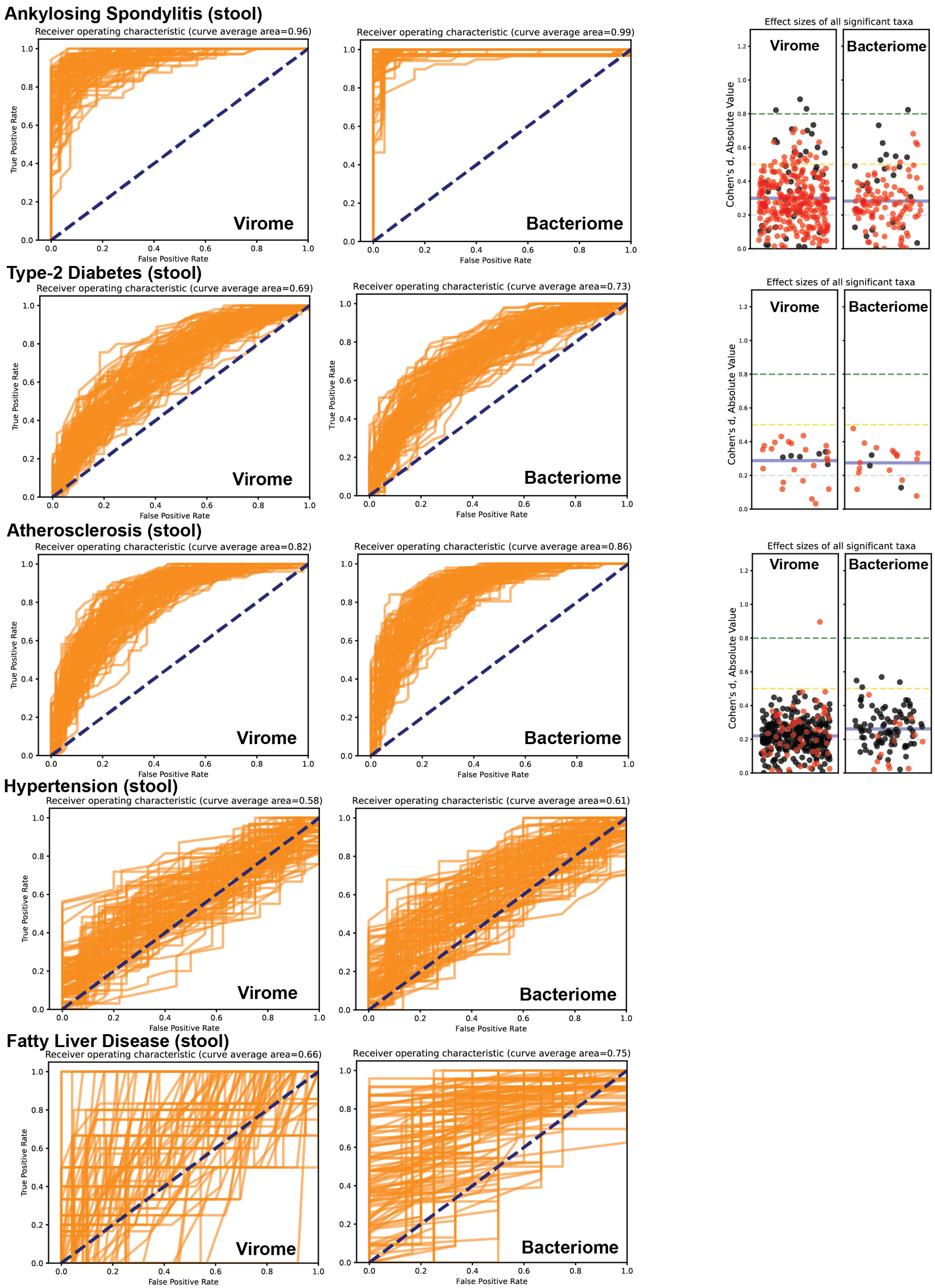
Supplemental Figure 5: Virome- and Bacteriome-wide associations with additional chronic diseases 2

Manhattan plots for stool viromes (left) and bacteriomes (right) are shown in the same manner as Figure 6. Accessions: Ankylosing spondylitis (PRJNA375935), Type 2 diabetes (PRJNA422434), Atherosclerosis (PRJEB21528), Hypertension (PRJEB13870), Fatty liver disease (PRJNA373901).



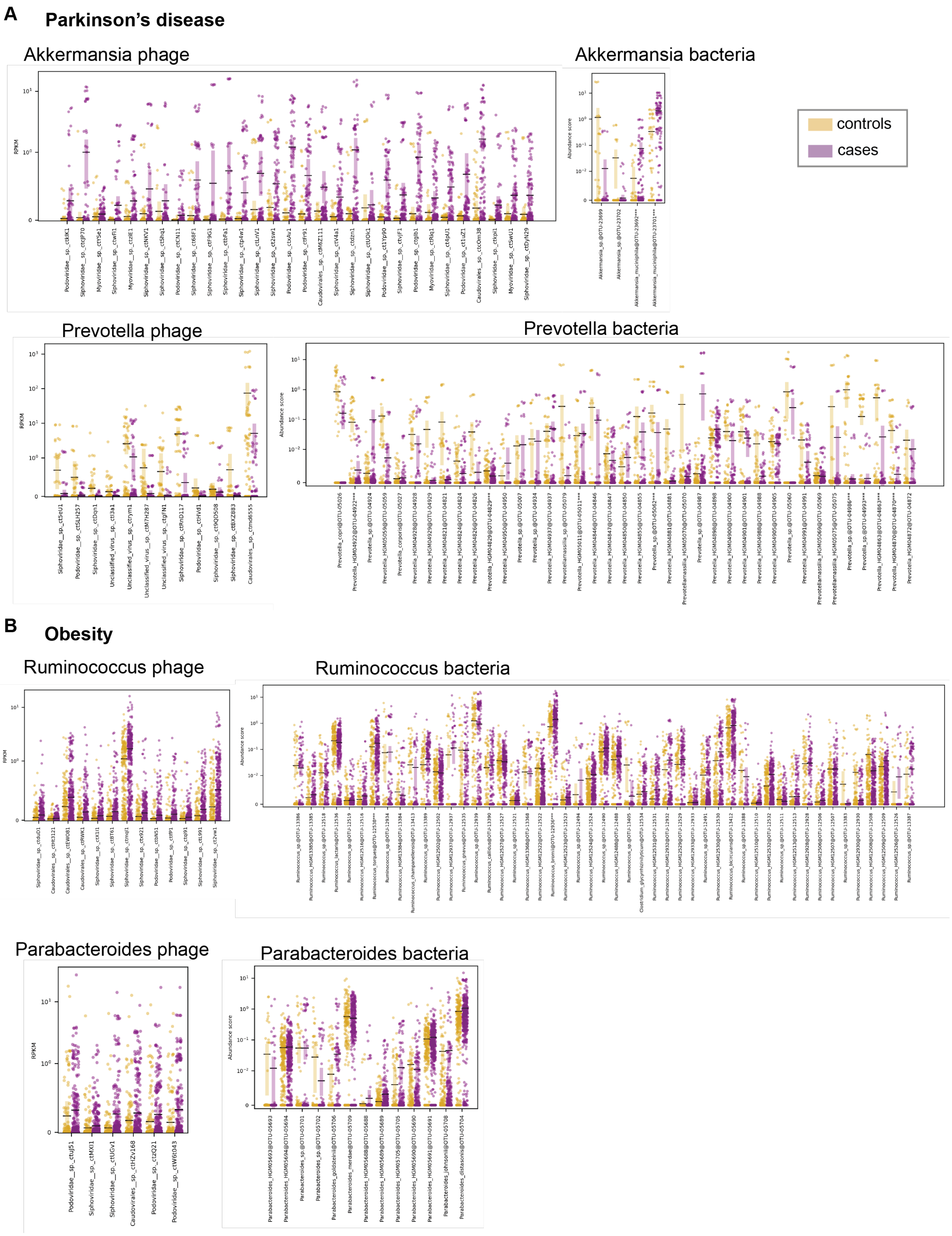
Supplemental Figure 6: Discriminatory power of viromes and bacteriomes in additional chronic diseases 1

Receiver operating characteristic plots for virome (left panels) and bacteriome (middle panels) data, as described in Figure 6. Summary of effect size data for significant OTUs (right panel), as described in Figure 6.



Supplemental Figure 7: Discriminatory power of viromes and bacteriomes in additional chronic diseases 2

Receiver operating characteristic plots for virome (left panels) and bacteriome (middle panels) data, as described in Figure 6. Summary of effect size data for significant OTUs (right panel), as described in Figure 6.



Supplemental Figure 8: Qualitative comparison of differentially abundant phage and their prospective hosts

Plots include all differentially abundant virus OTUs that were statistically significant after multiple testing correction, and all bacterial OTUs within each given genus that were abundant enough to evaluate (see methods). Bacterial OTUs with differential abundance that remained statistically significant after multiple testing correction are marked with "***" in the x-axis labels. Black lines are median values of the population with transparent purple or gold bands showing 90%

confidence intervals. Bacterial host of virus OTUs was determined at the genus level per Figure 2. (A) Representation of Akkermansia phage and bacteria (top) and Prevotella phage and bacteria (bottom) in Parkinson's Disease cohort.(B) Representation of Ruminococcus phage and bacteria (top) and Parabacteroides phage and bacteria (bottom).

Dataset S1 (separate file). Information on Bioprojects used for production of the Cenote Human Virome Database.

Dataset S2 (separate file). Cenote Human Virus Database virus OTU master table.

Dataset S3 (separate file). Cenote Human Virus Database virus OTU vs. GenBank complete genome dataset mash results.

Dataset S4 (separate file). Cenote Human Virus Database virus OTU vs. Gut Virome Database virus OTU mash results.

Dataset S5 (separate file). Read alignment/recruitment percent and ViromeQC scores of virus-enriched/virus-like particle (VLP) against CHVD99.

Dataset S6 (separate file). Phage-encoded CRISPR spacer matches to other phages within the CHVD and putative bacterial host information of source and target of CRISPR spacer.

Dataset S7 (separate file). Information on "Cosmopolitan" virus OTUs which were detected in multiple body sites in the Human Microbiome Project dataset.

Dataset S8A (separate file). Case-control accession numbers for reads associated with Parkinson's disease study.

Dataset S8B (separate file). Case-control accession numbers for reads associated with fatty liver disease study.

Dataset S8C (separate file). Case-control accession numbers for reads associated with hypertension study.

Dataset S8D (separate file). Case-control accession numbers for reads associated with atherosclerosis study.

Dataset S8E (separate file). Case-control accession numbers for reads associated with type-II diabetes study.

Dataset S8F (separate file). Case-control accession numbers for reads associated with ankylosing spondylitis study.

Dataset S8G (separate file). Case-control accession numbers for reads associated with type-I diabetes (stool) study.

Dataset S8H (separate file). Case-control accession numbers for reads associated with type-I diabetes (saliva) study.

Dataset S8I (separate file). Case-control accession numbers for reads associated with liver cirrhosis study.

Dataset S8J (separate file). Case-control accession numbers for reads associated with colon carcinoma study.

Dataset S8K (separate file). Case-control accession numbers for reads associated with colon adenoma study.

Dataset S8L (separate file). Case-control accession numbers for reads associated with obesity study.