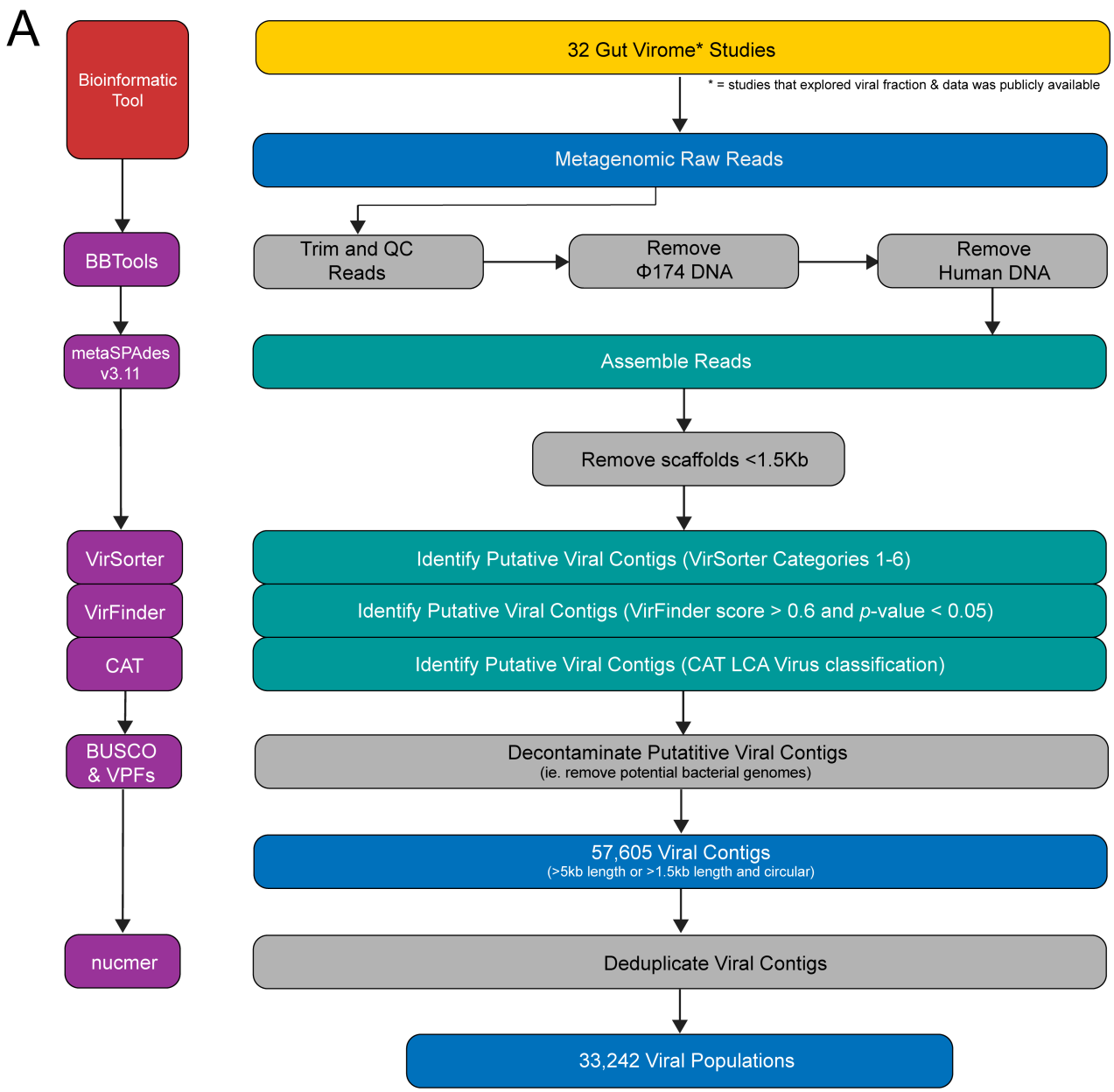
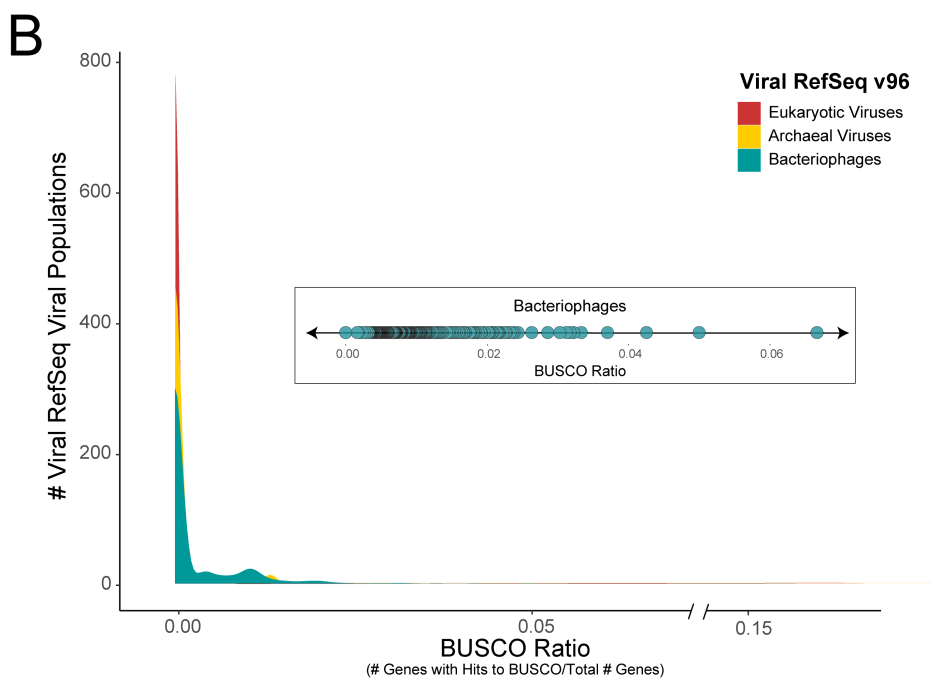
Cell Host & Microbe, Volume *28*

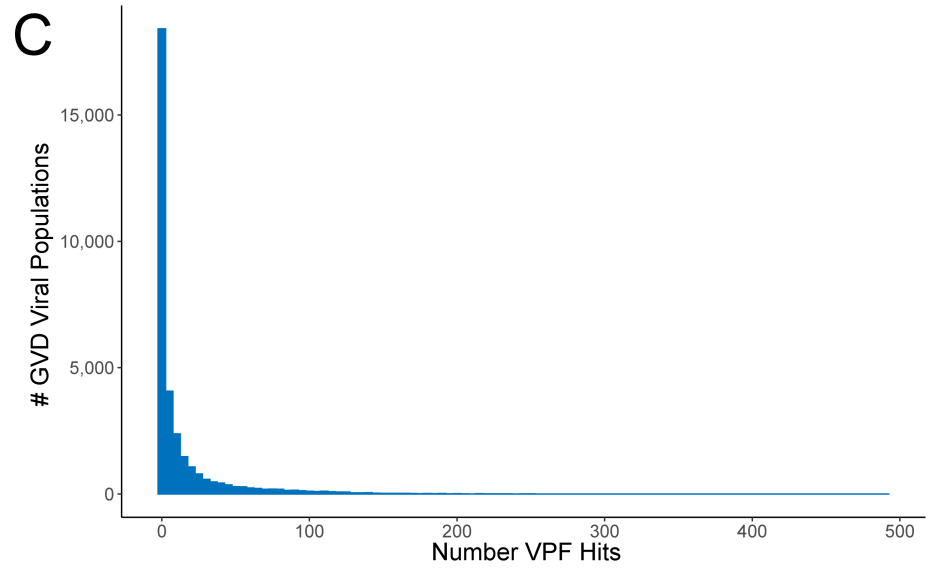
Supplemental Information

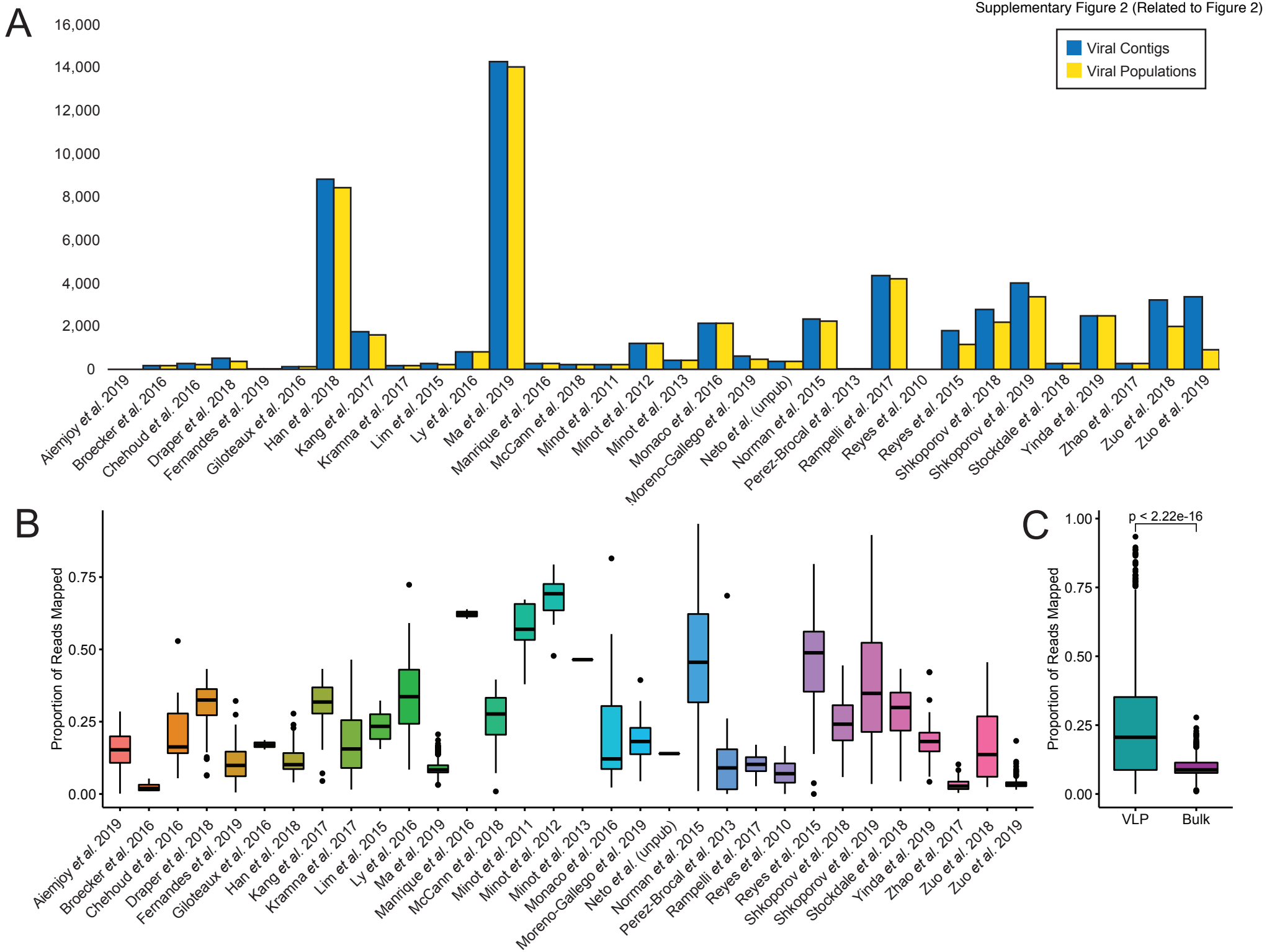
The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut

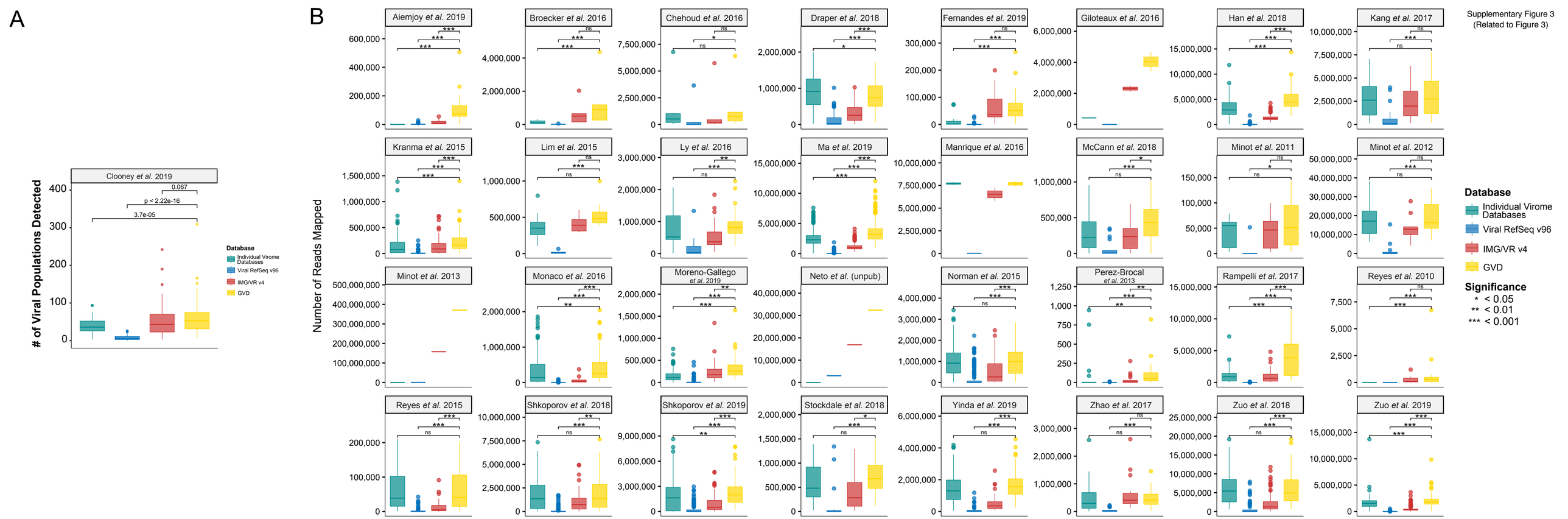
Ann C. Gregory, Olivier Zablocki, Ahmed A. Zayed, Allison Howell, Benjamin Bolduc, and Matthew B. Sullivan


A
Bioinformatic Tool
BBTools
metaSPAdes v3.11
VirSorter
VirFinder
CAT
BUSCO & VPFs
nucmer
32 Gut Virome* Studies
* = studies that explored viral fraction & data was publicly available
Metagenomic Raw Reads
Trim and QC Reads
Remove Φ174 DNA
Remove Human DNA
Assemble Reads
Remove scaffolds <1.5Kb
Identify Putative Viral Contigs (VirSorter Categories 1-6)
Identify Putative Viral Contigs (VirFinder score > 0.6 and p-value < 0.05)
Identify Putative Viral Contigs (CAT LCA Virus classification)
Decontaminate Putatitive Viral Contigs
(ie. remove potential bacterial genomes)
57,605 Viral Contigs
(>5kb length or >1.5kb length and circular)
Deduplicate Viral Contigs
33,242 Viral Populations

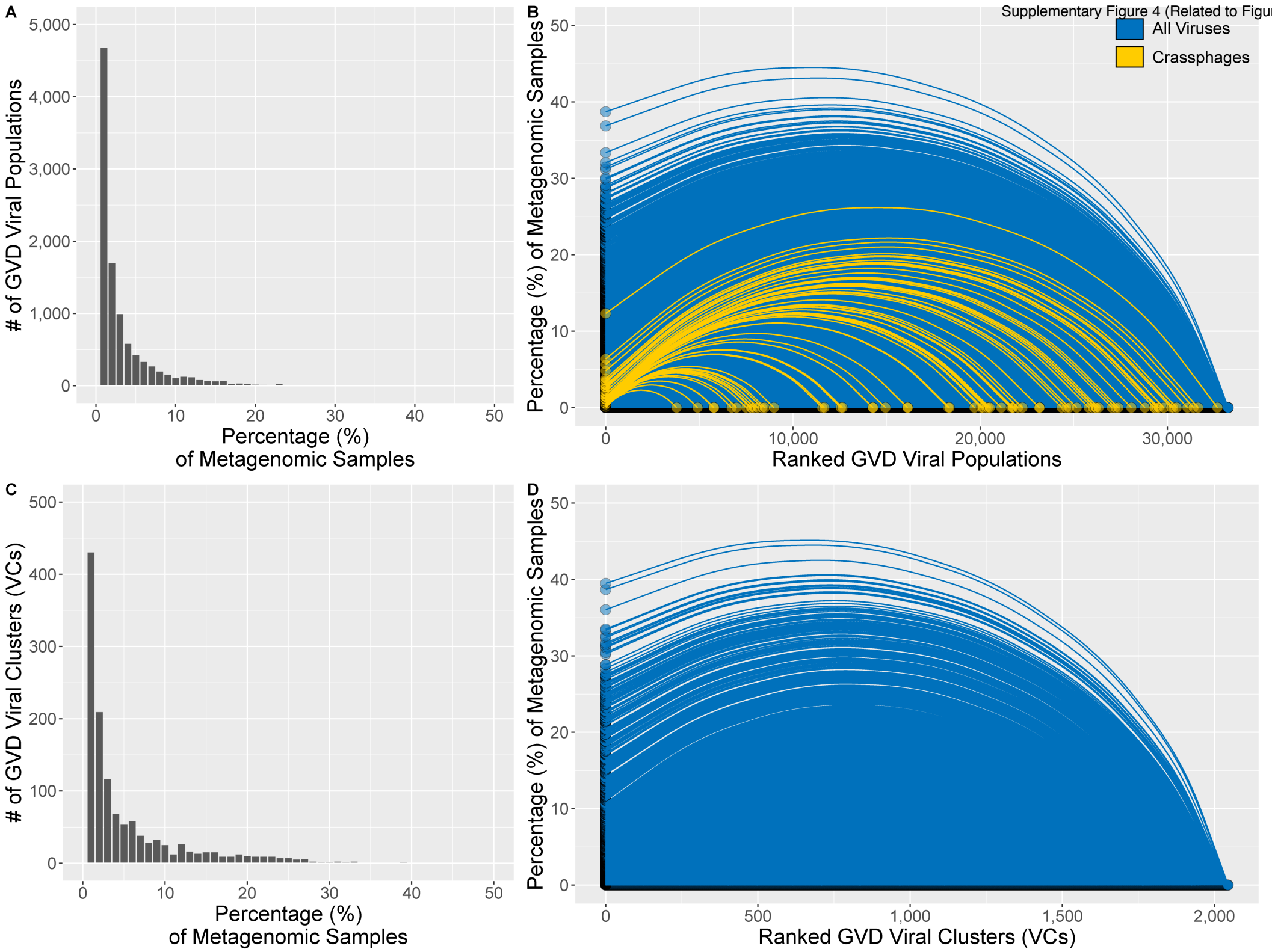

B
Viral RefSeq v96
Eukaryotic Viruses
Archaeal Viruses
Bacteriophages
Viral RefSeq Viral Populations
800
600
400
200
0
0.00
0.05
0.15
BUSCO Ratio
(# Genes with Hits to BUSCO/Total # Genes)
Bacteriophages
0.00
0.02
0.04
0.06
BUSCO Ratio

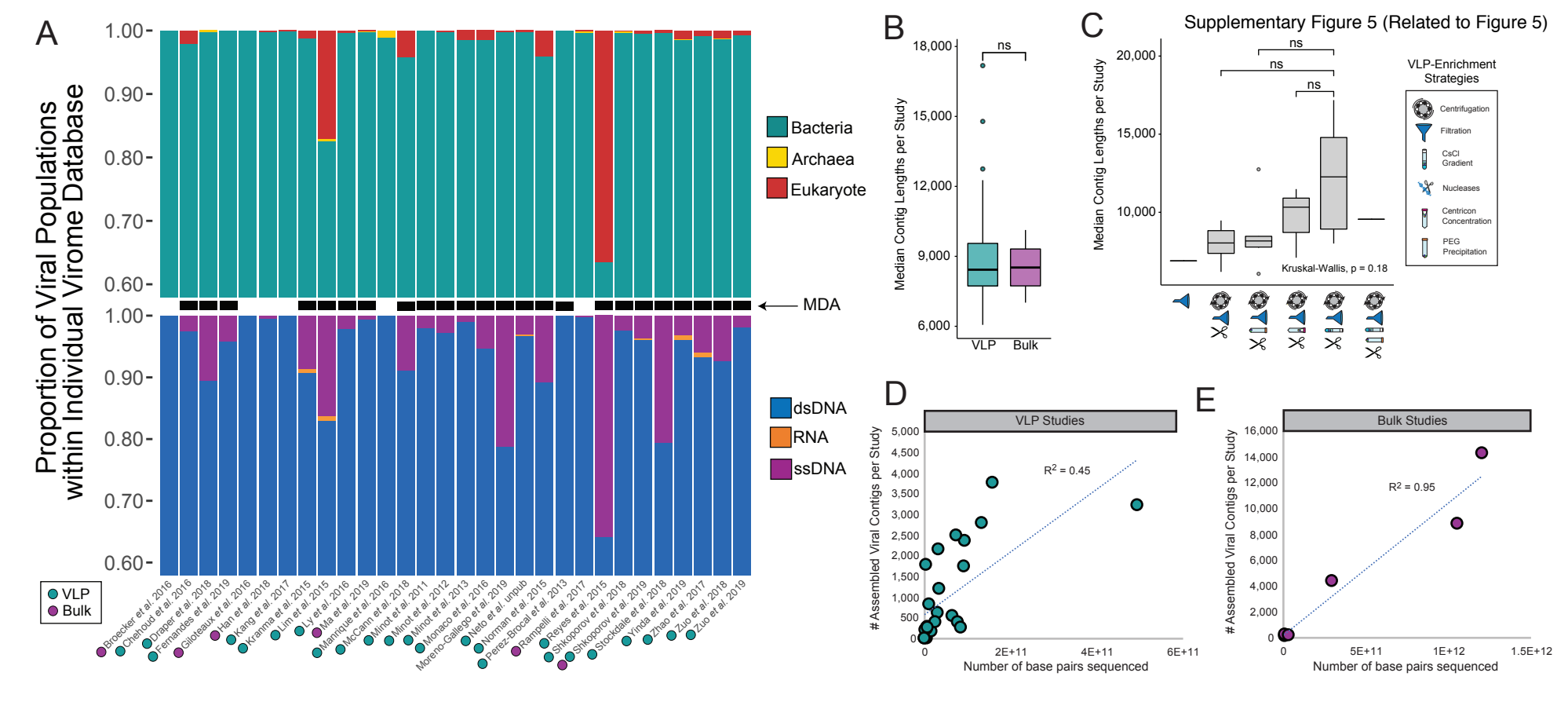

C
GVD Viral Populations
15,000
10,000
5,000
0
0
100
200
300
400
500
Number VPF Hits


A
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
Viral Contigs
Viral Populations
Aiemjoy et al. 2019
Broecker et al. 2016
Chehoud et al. 2016
Draper et al. 2018
Fernandes et al. 2019
Giloteaux et al. 2016
Han et al. 2018
Kang et al. 2017
Kramna et al. 2017
Lim et al. 2015
Ly et al. 2016
Ma et al. 2019
Manrique et al. 2016
McCann et al. 2018
Minot et al. 2011
Minot et al. 2012
Minot et al. 2013
Monaco et al. 2016
Moreno-Gallego et al. 2019
Neto et al. (unpub)
Norman et al. 2015
Perez-Brocal et al. 2013
Rampelli et al. 2017
Reyes et al. 2010
Reyes et al. 2015
Reyes et al. 2018
Shkoporov et al. 2019
Shkoporov et al. 2018
Stockdale et al. 2018
Yinda et al. 2019
Zhao et al. 2017
Zuo et al. 2018
Zuo et al. 2019
B
Proportion of Reads Mapped
0.75
0.50
0.25
0.00
C
p < 2.22e-16
Proportion of Reads Mapped
1.00
0.75
0.50
0.25
0.00
VLP
Bulk


A
Clooney et al. 2019
0.067
p < 2.22e-16
3.7e-05
# of Viral Populations Detected
Database
Individual Virome Databases
Viral RefSeq v96
IMG/VR v4
GVD
B
Number of Reads Mapped
Aiemjoy et al. 2019
Broecker et al. 2016
Chehoud et al. 2016
Draper et al. 2018
Fernandes et al. 2019
Giloteaux et al. 2016
Han et al. 2018
Kang et al. 2017
Kranma et al. 2015
Lim et al. 2015
Ly et al. 2016
Ma et al. 2019
Manrique et al. 2016
McCann et al. 2018
Minot et al. 2011
Minot et al. 2012
Minot et al. 2013
Monaco et al. 2016
Moreno-Gallego et al. 2019
Neto et al. (unpub)
Norman et al. 2015
Perez-Brocal et al. 2013
Rampelli et al. 2017
Reyes et al. 2010
Reyes et al. 2015
Shkoporov et al. 2018
Shkoporov et al. 2019
Stockdale et al. 2018
Yinda et al. 2019
Zhao et al. 2017
Zuo et al. 2018
Zuo et al. 2019
Supplementary Figure 3
(Related to Figure 3)
Database
Individual Virome Databases
Viral RefSeq v96
IMG/VR v4
GVD
Significance
* < 0.05
** < 0.01
*** < 0.001

Supplementary Figure 4 (Related to Figure 4)




Supplementary Figure 5 (Related to Figure 5)
A
Proportion of Viral Populations within Individual Virome Database
1.00
0.90
0.80
0.70
0.60
Bacteria
Archaea
Eukaryote
MDA
dsDNA
RNA
ssDNA
VLP
Bulk
Broecker et al. 2016
Chehoud et al. 2016
Draper et al. 2018
Fernandes et al. 2019
Gilioteaux et al. 2016
Han et al. 2018
Kang et al. 2017
Kramna et al. 2015
Lim et al. 2015
Ly et al. 2016
Ma et al. 2019
Manrique et al. 2016
McCann et al. 2018
Minot et al. 2011
Minot et al. 2012
Minot et al. 2013
Monaco et al. 2016
Moreno-Gallego et al. 2019
Nelo et al. unpub
Norman et al. 2015
Perez-Brocal et al. 2013
Rampelli et al. 2017
Reyes et al. 2015
Shkoporov et al. 2018
Shkoporov et al. 2019
Stockdale et al. 2018
Yinda et al. 2019
Zhao et al. 2017
Zuo et al. 2018
Zuo et al. 2019
B
Median Contig Lengths per Study
18,000
15,000
12,000
9,000
6,000
ns
VLP
Bulk
C
Median Contig Lengths per Study
20,000
15,000
10,000
ns
ns
ns
Kruskal-Wallis, p = 0.18
VLP-Enrichment Strategies
Centrifugation
Filtration
CsCl Gradient
Nucleases
Centricon Concentration
PEG Precipitation
D
VLP Studies
Assembled Viral Contigs per Study
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
R2 = 0.45
0
2E+11
4E+11
6E+11
Number of base pairs sequenced
E
Bulk Studies
Assembled Viral Contigs per Study
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
R2 = 0.95
0
5E+11
1E+12
1.5E+12
Number of base pairs sequenced

A

All Individuals across Studies

Viral Populations per bp Sequenced

1.0e-05
7.5e-06
5.0e-06
2.5e-06
0.0e+00

75% Quantile
25% Quantile

Aiemjoy *et al.* 2019
Broecker *et al.* 2016
Chehoud *et al.* 2016
Draper *et al.* 2018
Fernandes *et al.* 2019
Giloteaux *et al.* 2016
Han *et al.* 2018
Kang *et al.* 2017
Kranma *et al.* 2015
Lim *et al.* 2015
Ly *et al.* 2016
Ma *et al.* 2019
Manrique *et al.* 2016
McCann *et al.* 2018
Minot *et al.* 2011
Minot *et al.* 2012
Minot *et al.* 2013
Monaco *et al.* 2016
Moreno-Gallego *et al.* 2019
Neto *et al.* (unpub)
Norman *et al.* 2015
Perez-Brocal *et al.* 2013
Rampelli *et al.* 2017
Reyes *et al.* 2010
Reyes *et al.* 2015
Shkoporov *et al.* 2018
Shkoporov *et al.* 2019
Stockdale *et al.* 2018
Yinda *et al.* 2019
Zhao *et al.* 2017
Zuo *et al.* 2018
Zuo *et al.* 2019

B

All Individuals by Seq Type

0.76
9.5e-06
p < 2.22e-16

1e-05
5e-06
0e+00

454
Illumina
Ion Torrent

C

Non-454 Individuals by Type

p < 2.22e-16

6e-06
4e-06
2e-06
0e+00

Bulk
VLP

D

Non-454, VLP Individuals by MDA status

0.12

6e-06
4e-06
2e-06
0e+00

No MDA
MDA

E

Non-454, VLP, MDA Individuals by Western status

8.6e-07

6e-06
4e-06
2e-06
0e+00

Non-Western
Western

F

Non-454, VLP, MDA, Western Individuals by Health status

8.5e-13

3e-06
2e-06
1e-06
0e+00

Indeterminant
Diseased
Healthy

G

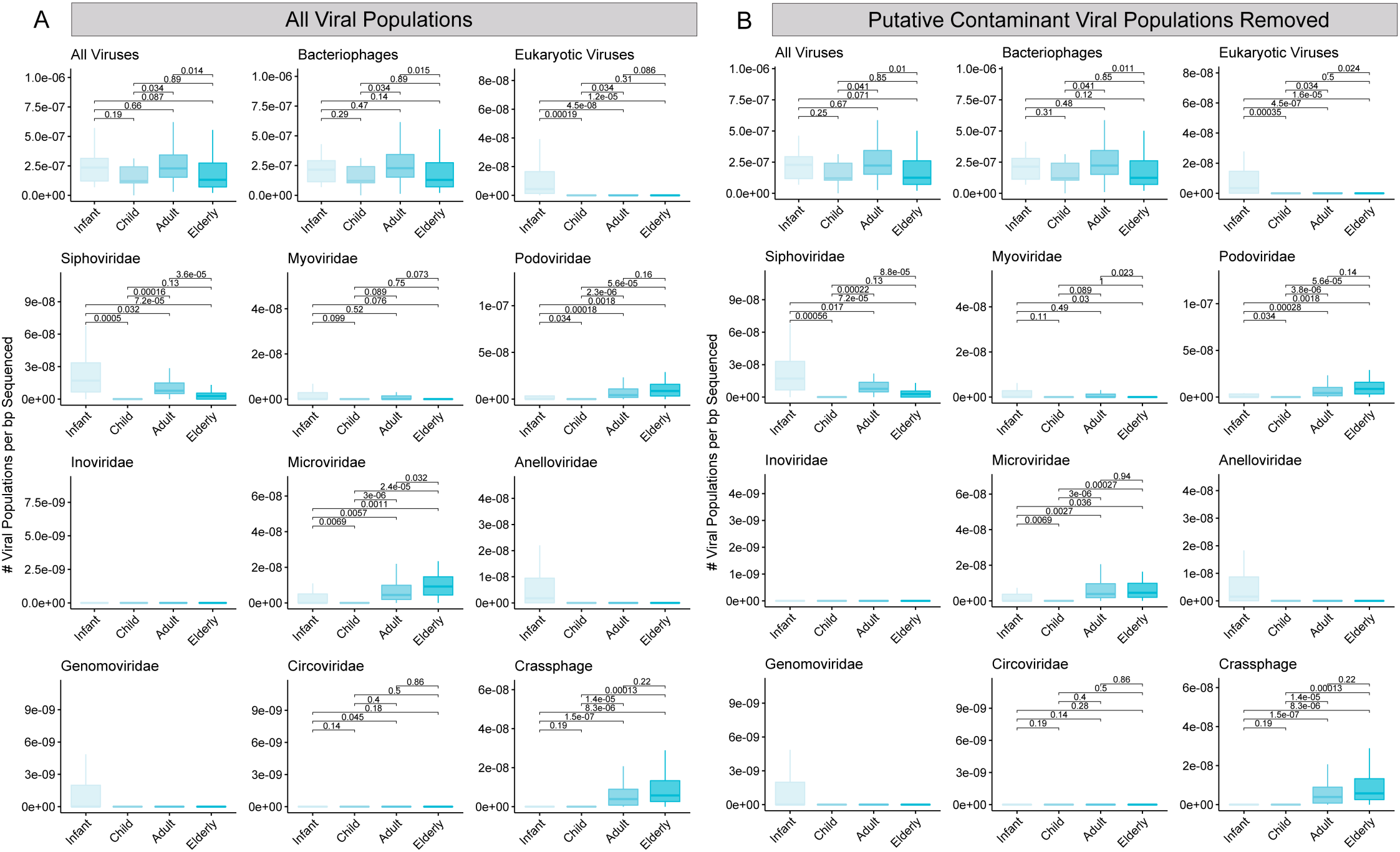
Non-454, VLP-enriched, MDA, Healthy Western Individuals by Study

Viral Populations per bp Sequenced

7.5e-07
5.0e-07
2.5e-07
0.0e+00

75% Quantile
25% Quantile

Chehoud *et al.* 2016
Draper *et al.* 2018
Fernandes *et al.* 2019
Lim *et al.* 2015
Ly *et al.* 2016
McCann *et al.* 2018
Minot *et al.* 2012
Minot *et al.* 2013
Moreno-Gallego *et al.* 2019
Shkoporov *et al.* 2019
Stockdale *et al.* 2018


A
All Viral Populations
B
Putative Contaminant Viral Populations Removed
All Viruses
Bacteriophages
Eukaryotic Viruses
Siphoviridae
Myoviridae
Podoviridae
Inoviridae
Microviridae
Anelloviridae
Genomoviridae
Circoviridae
Crassphage
Viral Populations per bp Sequenced
Infant
Child
Adult
Elderly

Supplementary Figure 1 (*Related to Figure 2).* Pipeline for the selection, processing and virus validation of human gut datasets. **(a)** Pipeline workflow showing that datasets were processed individually. Reads were filtered for quality, trimmed, and reads that mapped to Φx174 and the human genome were removed. The remaining reads were assembled into scaffolds, filtered for lengths ≥1.5kb, and run through tools that collectively utilize homology to viral reference databases, probabilistic models on viral genomic features, and viral *k*-mer signatures to identify putative viral genomes. The putative set of viral genomes were decontaminated using the BUSCO and VPF hmm models. Viral genomes were then deduplicated to get a total of 33,242 viral populations. **(b)** Density plot showing the number of BUSCO hits per total number of genes (BUSCO ratio) for all viruses in Viral Refseq v96. The **inset** dot plot showing the distribution of the BUSCO ratio for bacteriophages, with the highest value being 0.067. This value was the max value allowed for all GVD bacteriophages. **(c)** Histogram showing the number of GVD viral populations with different numbers of viral protein family (VFP) hits.

Supplementary Figure 2 (*Related to Figure 2).* (a) Barplot showing the number of assembled viral genomes versus the number of deduplicated viral populations per study. **(b&c)** Box plots showing median and quartiles of proportion of the total reads sequenced that mapped to the GVD viral contigs **(b)** per study and **(c)** and between VLP and bulk metagenomes.

Supplementary Figure 3 (*Related to Figure 3).* GVD improves read recruitment. Boxplots showing median and quartiles of **(a)** the number of viral populations detected in a study not included in GVD, and **(b)** the number of reads mapped per study to the individual virome, Viral Refseq v96, JGI IMG/VR, or GVdb databases. All pairwise comparisons were performed using Mann-Whitney U-tests.

Supplementary Figure 4 (*Related to Figure 4).* There are no core viral populations or viral clusters across GVD samples. (a & c) Histogram showing the number of (**a**) viral populations and (**c**) viral clusters (VCs) present in different percentages of GVD samples. The vast majority of viral populations and VCs are found in <10% of the individuals. **(b & d)** Hive plot showing the percentage of GVD samples each (**b**) viral population and (**d**) VC is detected within. The dots on the x-axis represent each GVD viral population or VC in ascending order of the percentage of GVD samples that they are found within. The y-axis is the percentage of GVD samples that each viral population or VC is detected within. CrAssphage viral populations are highlighted in yellow in plot **(b)**.

Supplementary Figure 5 (*Related to Figure 5).* (a) Barplots showing the proportion of those viruses that are bacteriophages, archaeal viruses, or eukaryotic viruses (top) and the proportion of those viruses that are dsDNA, ssDNA, or RNA viruses (bottom). The total number of assembled viral contigs and viral populations per study are available in **Supplementary Fig. 2a** and further details in **Supplementary Table 6**. Studies that used VLP-enriched or bulk metagenomes or used multiple displacement amplification (MDA) were indicated by marking in between the barplots. No viral contigs ≥1.5kb were assembled from the Aiemjoy *et al.* 2019 and Reyes *et al.* 2010 studies. Boxplots showing median and quartiles of the median contig length per study **(b)** of VLP and bulk metagenomes and **(c)** of the different VLP-enrichment methodologies across the studies. Scatter plots with linear regressions lines showing the number of assembled viral contigs per base paired sequenced per study in **(d)** only VLP and **(e)** only bulk metagenome studies.

Supplementary Figure 6 (*Related to Figure 7)*. Removing confounding variables for cross-study viral diversity analyses. (a) Box plots showing median and quartiles of the number of viral populations per base pair sequenced across the different studies. The dashed black and red lines represent the 75% and 25% quantiles, respectively, of the number of viral populations per base pair sequenced across all individuals in each study. **(b-f)** Boxplots showing median and quartiles of the number of viral populations per base pair sequenced comparing sequencing platform, enrichment type (VLP vs. bulk), MDA status, geographic origin (Western vs. non-Western), and health status all sequentially and additively tested. All pairwise comparisons were performed using Mann-Whitney U-tests. **(g)** Boxplots showing median and quartiles of the number of viral populations per base pair sequenced per study for the remaining non-454, VLP-enriched, MDA, healthy Western individuals. The dashed black and red lines represent the 75% and 25% quantiles, respectively, of the number of viral populations per base pair sequenced across all remaining individuals in each study.

Supplementary Figure 7 (*Related to Figure 7)*. Statistical differences across viral groups assessed across the life stages. (a & b) Box plots showing median and quartiles of the number of viral populations per base pair sequenced for all GVD viruses, bacteriophages, eukaryotic viruses, and different viral families including crAssphage across the life stages across healthy Western individuals for **(a)** all viral populations and **(b)** with putative contaminant viral populations removed. All pairwise comparisons were performed using Mann-Whitney U-tests.