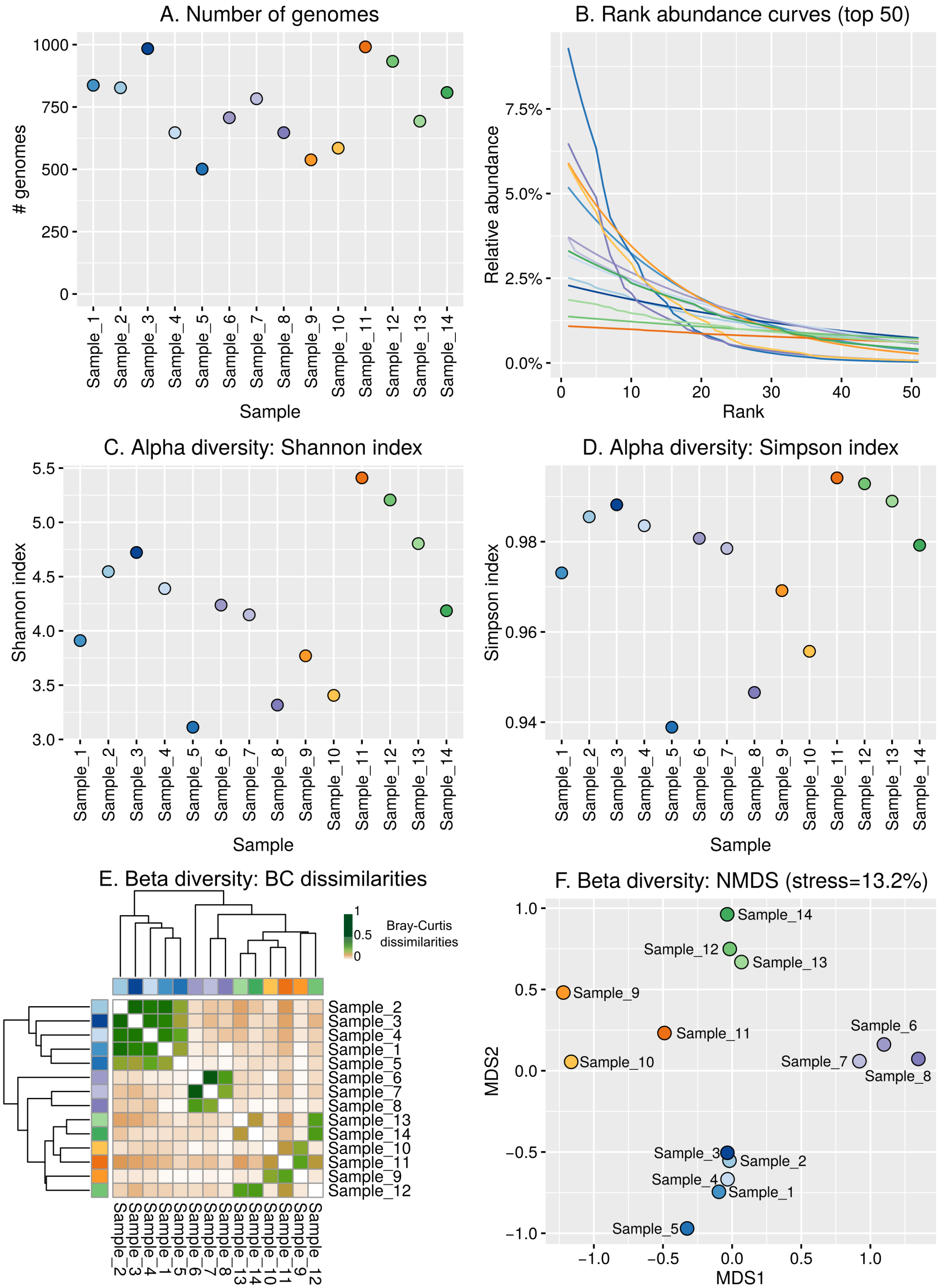
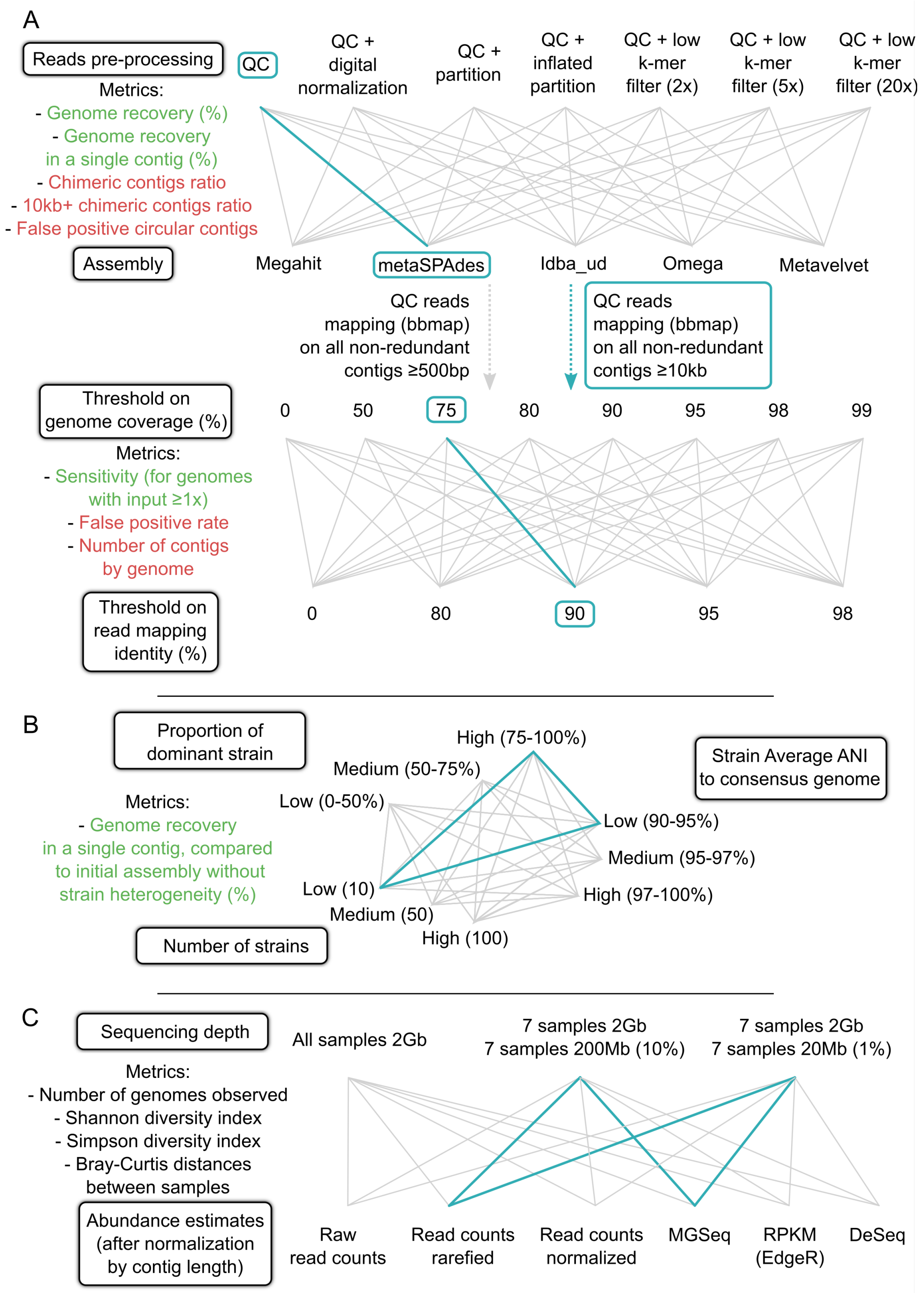
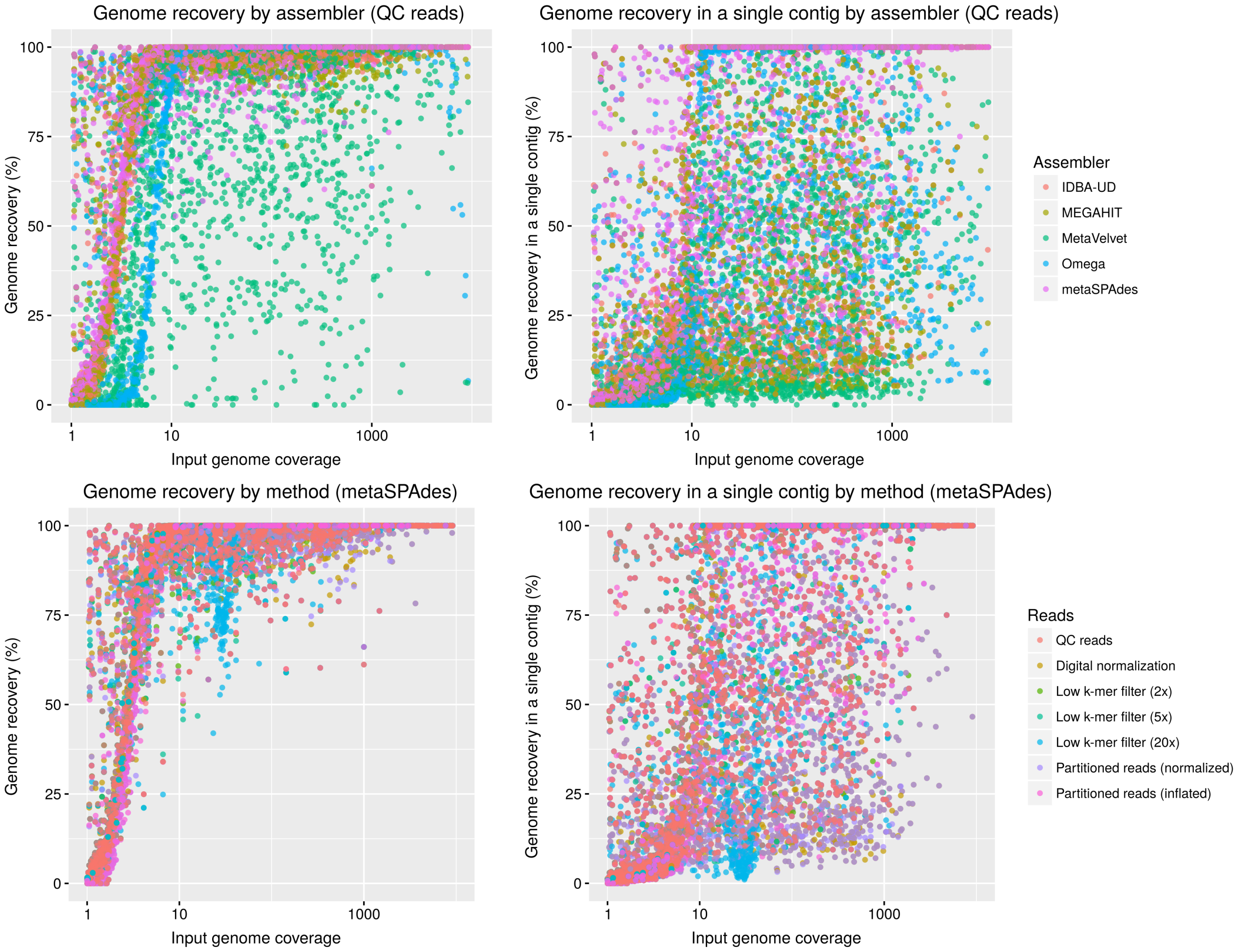


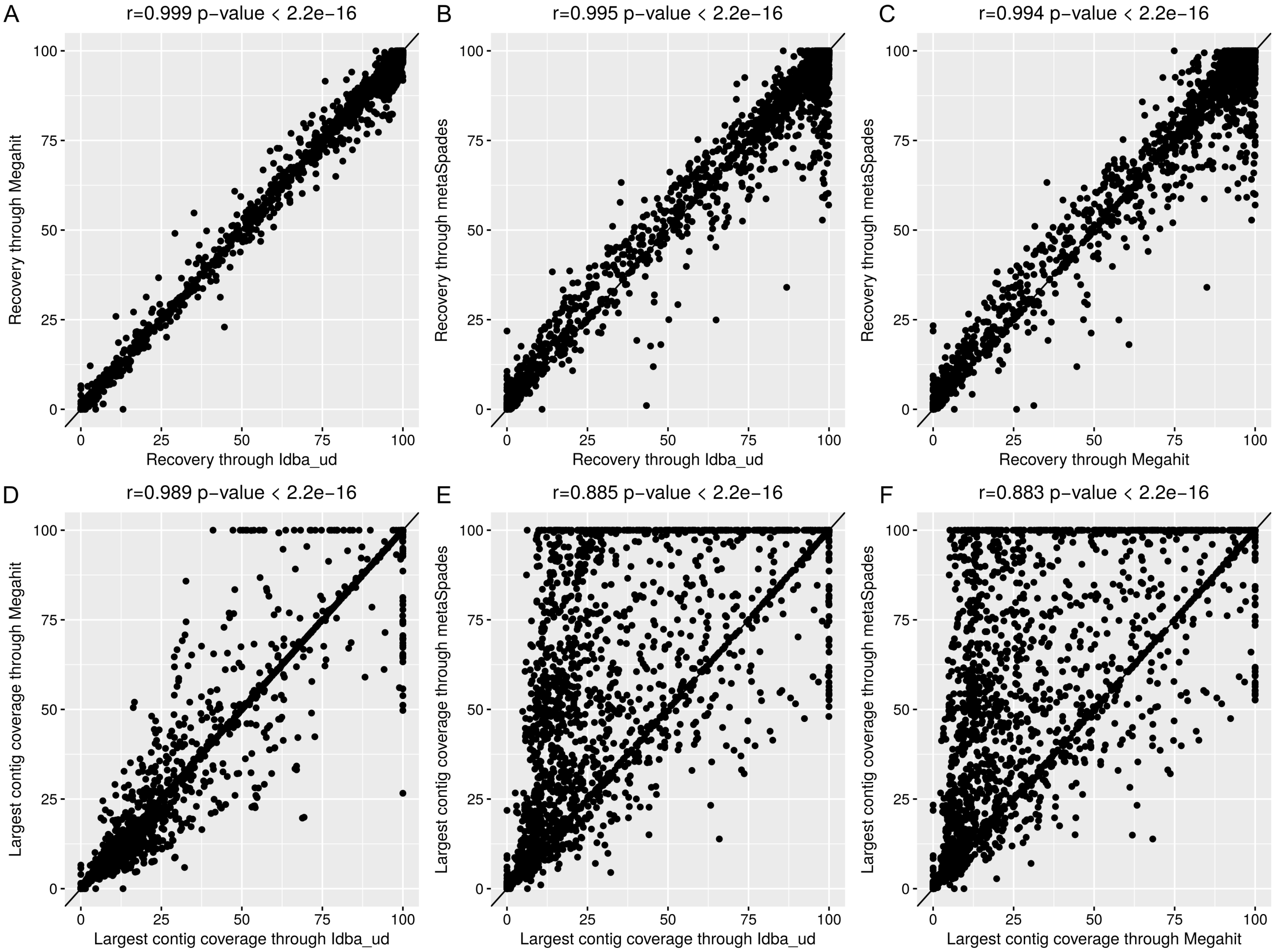
Supplementary Figure 1. Characteristics of mock viral communities. Mock communities generated had different number of genomes (panel A) and population distribution (panel B). This led to a range of alpha diversity, as illustrated with Shannon diversity index (C) and Simpson index (D). Mock communities were also designed to display a beta diversity pattern with 4 groups of samples (E and F, BC: Bray-Curtis). The 14 mock communities were designed to cluster into 4 distinct groups, and are colored across all panels according to these 4 groups of significantly similar communities (panel F, PerMANOVA p-value < 0.001).


A
Reads pre-processing
Metrics:
- Genome recovery (%)
- Genome recovery in a single contig (%)
- Chimeric contigs ratio
- 10kb+ chimeric contigs ratio
- False positive circular contigs
QC
QC + digital normalization
QC + partition
QC + inflated partition
QC + low k-mer filter (2x)
QC + low k-mer filter (5x)
QC + low k-mer filter (20x)
Assembly
Megahit
metaSPAdes
Idba_ud
Omega
Metavelvet
QC reads mapping (bbmap) on all non-redundant contigs ≥500bp
QC reads mapping (bbmap) on all non-redundant contigs ≥10kb
Threshold on genome coverage (%)
Metrics:
- Sensitivity (for genomes with input ≥1x)
- False positive rate
- Number of contigs by genome
0
50
75
80
90
95
98
99
Threshold on read mapping identity (%)
0
80
90
95
98
B
Proportion of dominant strain
High (75-100%)
Medium (50-75%)
Low (0-50%)
Strain Average ANI to consensus genome
Low (90-95%)
Medium (95-97%)
High (97-100%)
Metrics:
- Genome recovery in a single contig, compared to initial assembly without strain heterogeneity (%)
Low (10)
Medium (50)
High (100)
Number of strains
C
Sequencing depth
All samples 2Gb
7 samples 2Gb
7 samples 200Mb (10%)
7 samples 2Gb
7 samples 20Mb (1%)
Metrics:
- Number of genomes observed
- Shannon diversity index
- Simpson diversity index
- Bray-Curtis distances between samples
Abundance estimates (after normalization by contig length)
Raw read counts
Read counts rarefied
Read counts normalized
MGSeq
RPKM (EdgeR)
DeSeq

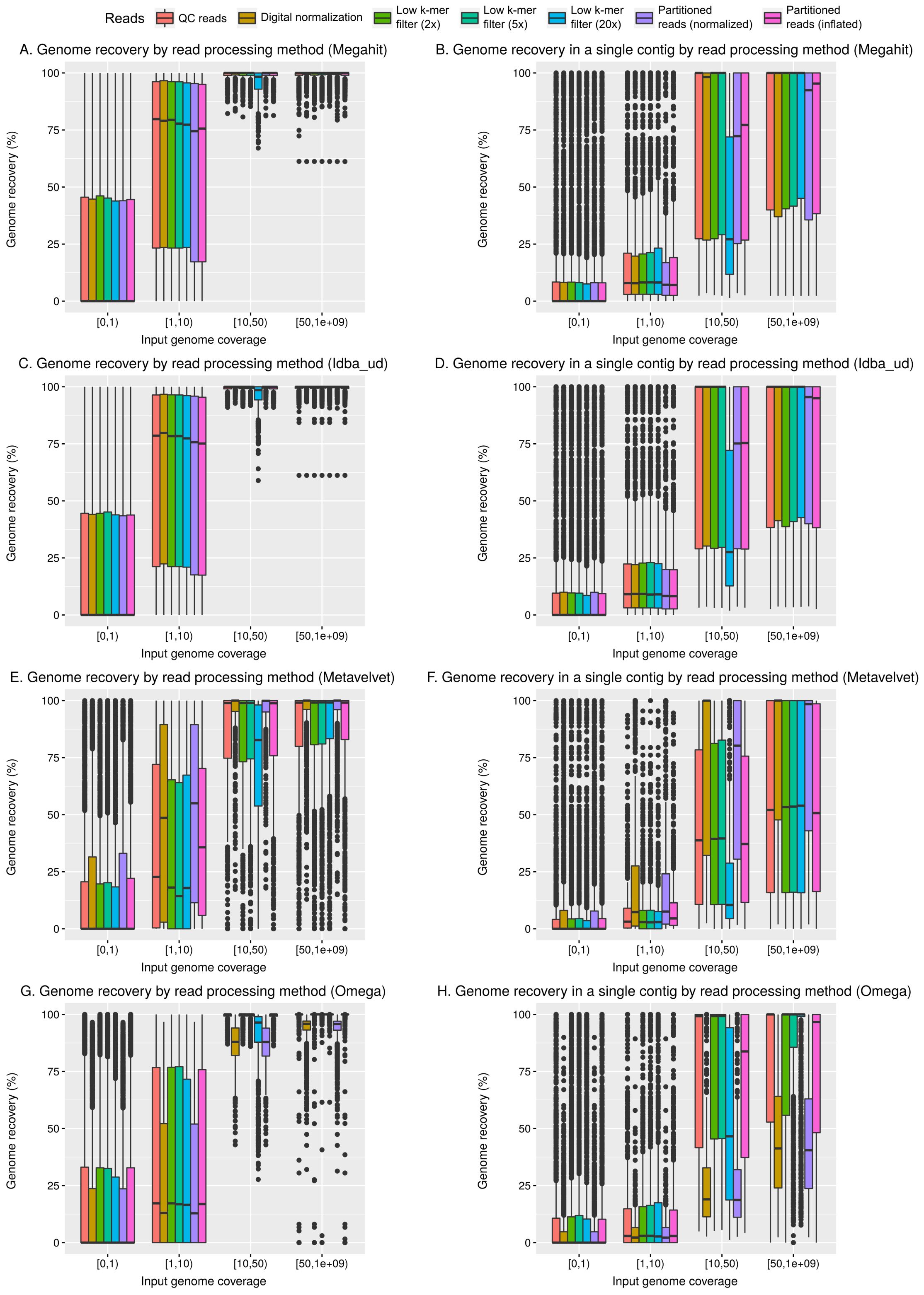
Supplementary Figure 2. Schematic of the methods evaluated in this study. A. Benchmarking of the assemblers, read pre-processing methods, and thresholds on genome coverage and read mapping identity used to calculate abundance matrices. B. Estimation of the impact of strain heterogeneity on the assembly efficiency. Reference genomes were replaced by populations composed of a set of related strains controlled by 3 parameters. C. Evaluation of the different normalization methods across the three types of datasets, with varying differences in sequencing depth across samples. For all panels, the different methods tested are indicated for each step, and the method and/or threshold chosen or optimal are highlighted in blue (other tests are colored in gray). The metrics used to identify the optimal methods/thresholds are indicated on the left, in green for metrics to maximize, red for metrics to minimize, and black for metrics to compare to "true" data based on whole communities. QC: quality-controlled.



Supplementary Figure 3. Influence of assembly software and read curation on genome recovery – dotplots (underlying data for boxplots presented in Figure 1). In these plots, each dot represents the assembly of a single genome in a single sample. A. Genome recovery (i.e. genome coverage by all contigs, y-axis) by genome coverage (x-axis) for different assemblers (colors). B. Genome recovery in a single contig (i.e. genome coverage by the largest assembled contig, y-axis) by genome coverage (x-axis) for different assemblers (colors). C. Genome recovery (i.e. genome coverage by all contigs, y-axis) by genome coverage (x-axis) for different read curation methods (colors). B. Genome recovery in a single contig (i.e. genome coverage by the largest assembled contig, y-axis) by genome coverage (x-axis) for different read curation methods (colors).



Supplementary Figure 4. Correlation between assembly results of different assemblers. Top panels display the correlations of genome recovery (i.e. genome coverage by all contigs) for each genome between Megahit and Idba_ud (A), metaSPAdes and Idba_ud (B) and metaSPAdes and Megahit (C). Bottom panels display the correlations of genome recovery in a single contig (i.e. genome coverage by the largest assembled contig) for each genome between Megahit and Idba_ud (D), metaSPAdes and Idba_ud (E) and metaSPAdes and Megahit (F).



A. Genome recovery by read processing method (Megahit)

B. Genome recovery in a single contig by read processing method (Megahit)

C. Genome recovery by read processing method (Idba_ud)

D. Genome recovery in a single contig by read processing method (Idba_ud)

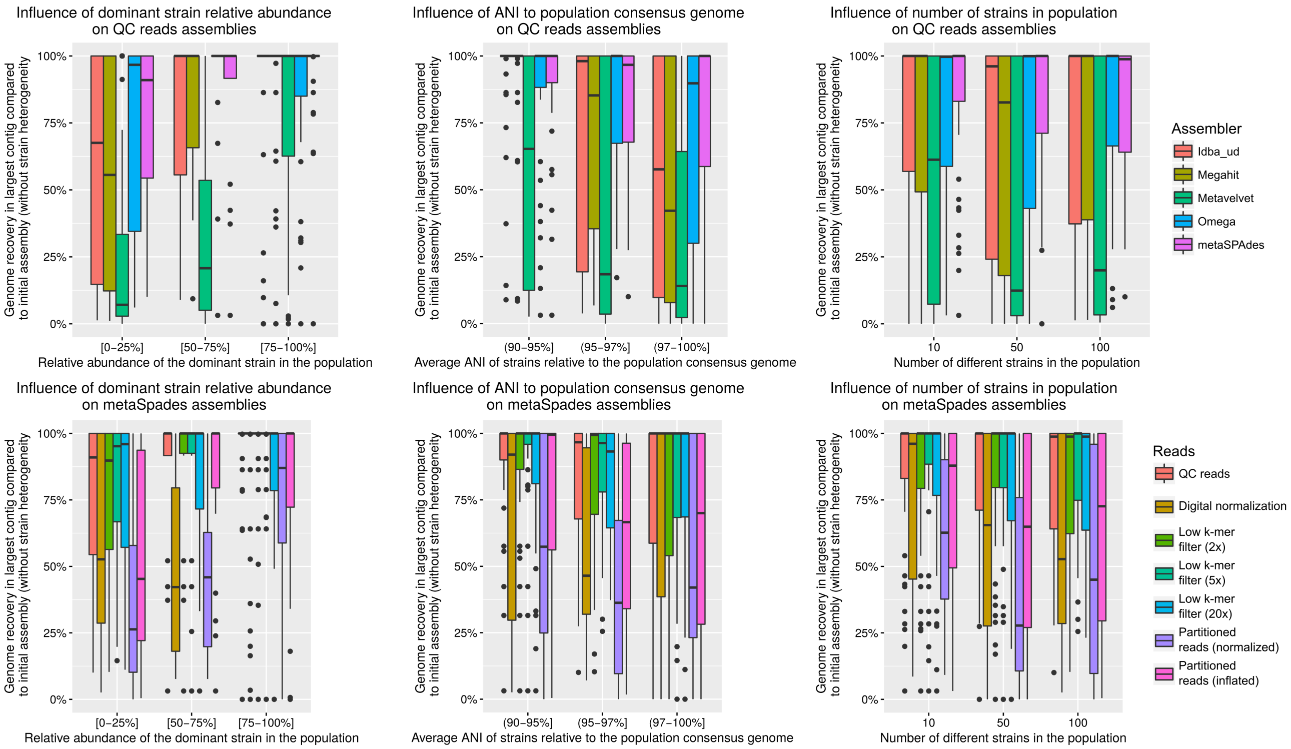
E. Genome recovery by read processing method (Metavelvet)

F. Genome recovery in a single contig by read processing method (Metavelvet)

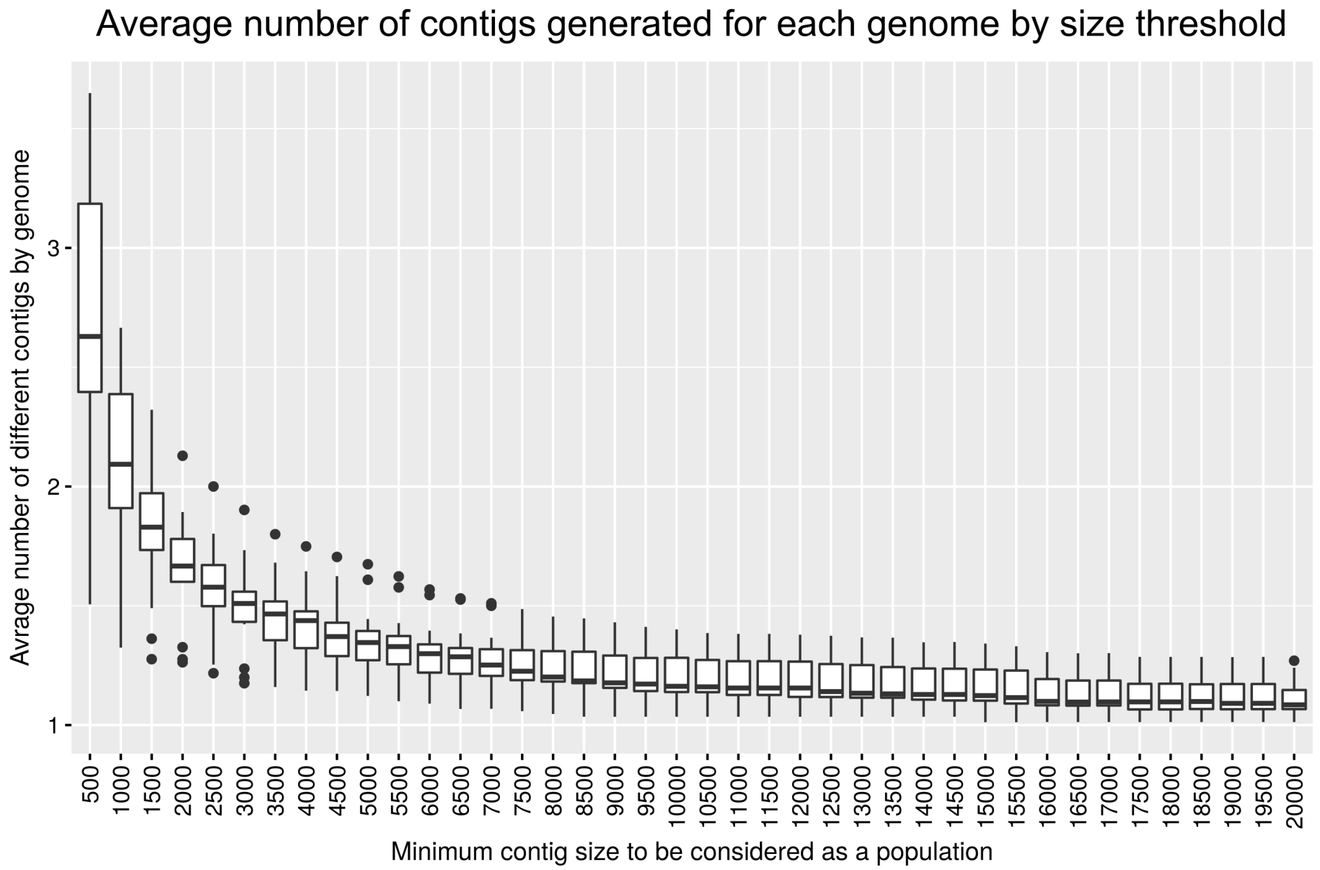
G. Genome recovery by read processing method (Omega)

H. Genome recovery in a single contig by read processing method (Omega)

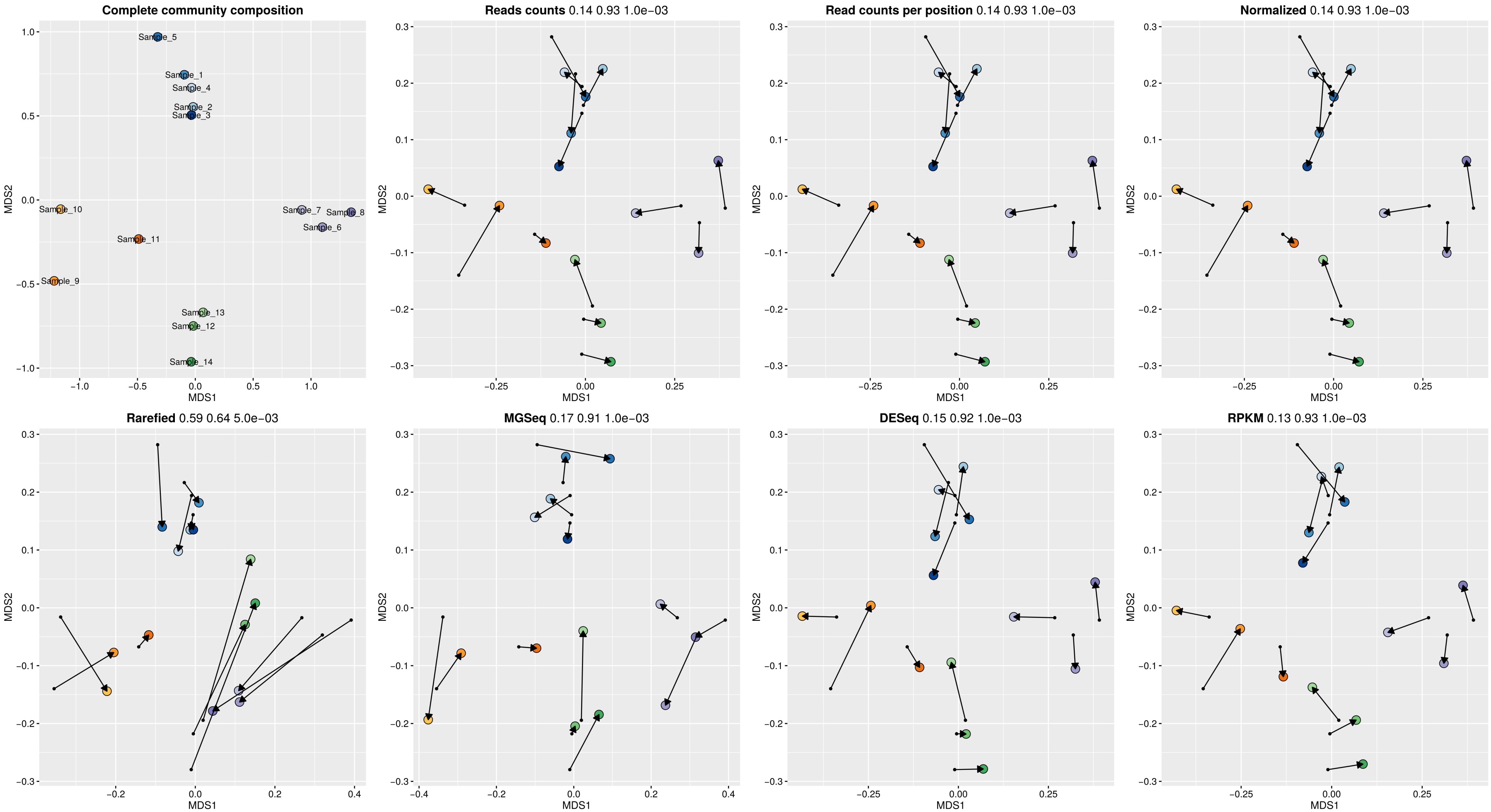
Supplementary Figure 5 Influence of read curation on genome recovery for different assemblers. The assemblers used here are Megahit (A, B), Idba_ud (C, D), Metavelvet (E, F), and Omega (G, H). For each assembler, the genome recovery (i.e. genome coverage by all contigs, y-axis) by genome coverage (x-axis) for different read curation methods (panels A, C, E, and G) as well as genome recovery in a single contig (i.e. genome coverage by the largest assembled contig, y-axis) by genome coverage (x-axis) for different read curation methods (panels B, D, F, and H) are displayed. Similar data are displayed for metaSPAdes in Fig. 1 panels C and D.



Supplementary Figure 6. Influence of strain-level diversity on assembly efficiency. These tests were computed on one mock community (Sample_1), where each reference genome was replaced with a set of related strains with varying divergence and relative abundances. In each plot, the y-axis represents the ratio between the largest contig assembled for a genome when strain heterogeneity is introduced and the same parameter without strain heterogeneity (i.e. previous assemblies of the same Sample_1). Plots on the top row display the differences in QC reads assemblies between assemblers, while plots on the bottom row show differences between different reads processing for metaSPAdes assemblies. Populations are grouped based on the different parameters controlling strain heterogeneity, i.e. relative abundance of the dominant strain (left), divergence of the strains (middle), and number of strains in the population (right).



Supplementary Figure 7. Number of population contigs detected for each input genome depending on the minimum contig size threshold for inclusion in population pool. The threshold on contig size used for inclusion in population contigs pools (in bp) is displayed on the x-axis. The distribution of average number of contigs for a given genome (across the samples where this genome was covered ≥ 1x) is displayed on the y-axis.



Supplementary Figure 8. Comparison of NMDS based on viral population counts and the reference mock community composition (reference in top left panel). The different NMDS were computed from the viral population count matrices normalized with the different methods tested in the manuscript (in bold), from the dataset including 6 samples strongly under-sequenced (subset 1%). For each NMDS, the sum of square difference, scaling factor, and significance of the correlation to the reference NMDS is indicated in the plot title (calculated with the function protest from the R package vegan). Samples are colored as in Fig. S1, and an arrow is used to illustrate the difference between the original sample placement in the reference NMDS and the new placement in the NMDS derived from population contigs.