Current Biology, Volume *32*

Supplemental Information

High viral abundance and low diversity are associated with increased CRISPR-Cas prevalence across microbial ecosystems

Sean Meaden, Ambarish Biswas, Ksenia Arkhipova, Sergio E. Morales, Bas E. Dutilh, Edze R. Westra, and Peter C. Fineran

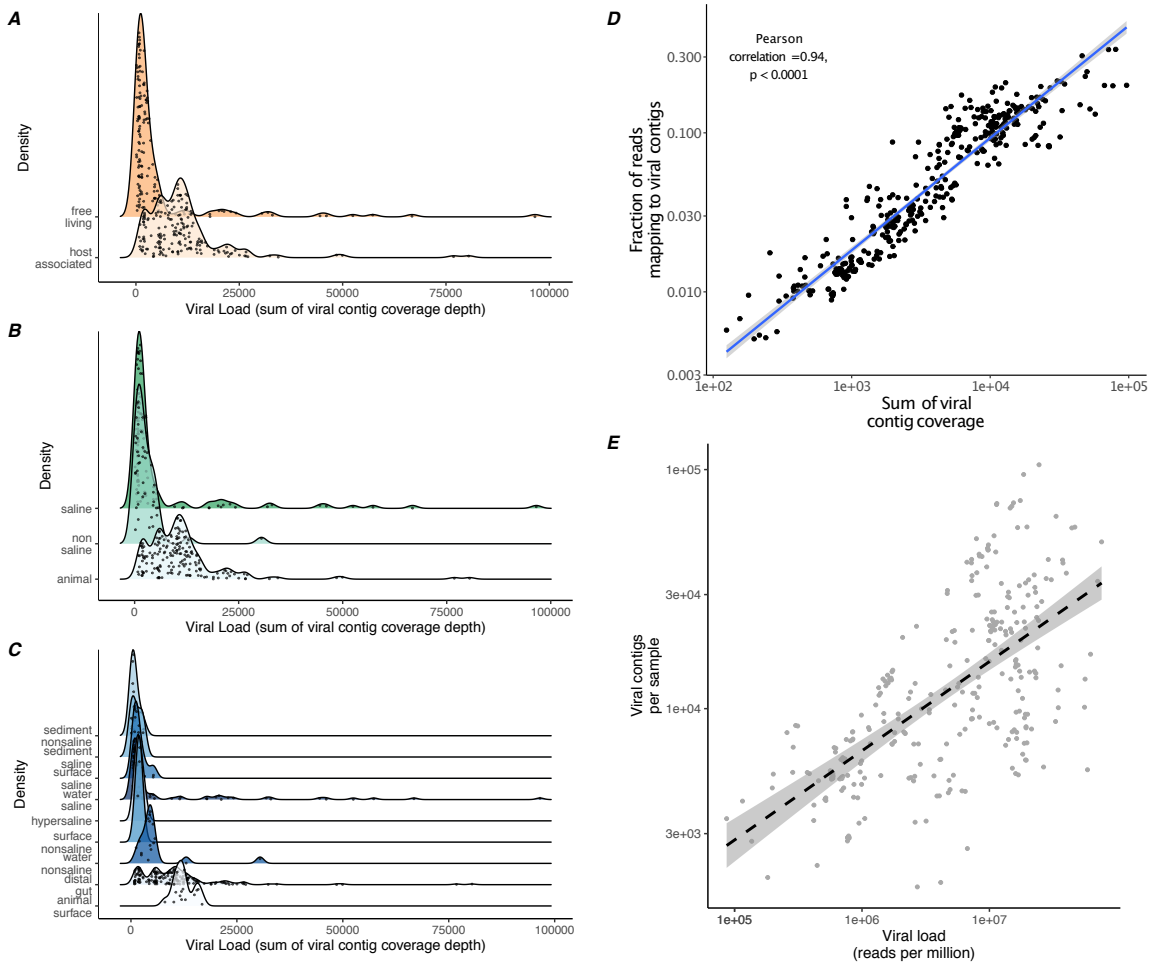

Figure S1. Viral abundance varies by environment, correlates with viral diversity and is robust to quantification method, Related to Figure 3 and STAR methods. Distributions of viral abundances (sum of viral contig coverage depth per sample) across EMPO level 1 categories (A), level 2 categories (B) and level 3 categories (C). Points represent metagenomic samples. D) Comparison of viral abundance quantification methods. Coverage-based measurements may be inflate abundance estimates if viral genomes are fragmented into many small contigs. Comparison of the sum of viral contig coverage against the fraction of reads that map to a viral contig shows a strong correlation (Pearson correlation = 0.94, $p < 0.0001$). All samples were subsampled to 1 million reads prior to mapping. E) Correlation between the number of viral contigs per sample and viral abundance (sum of viral contig coverage depth per sample). Line represents linear model fit, shaded areas denote 95% confidence intervals and $R^2 = 0.51$.

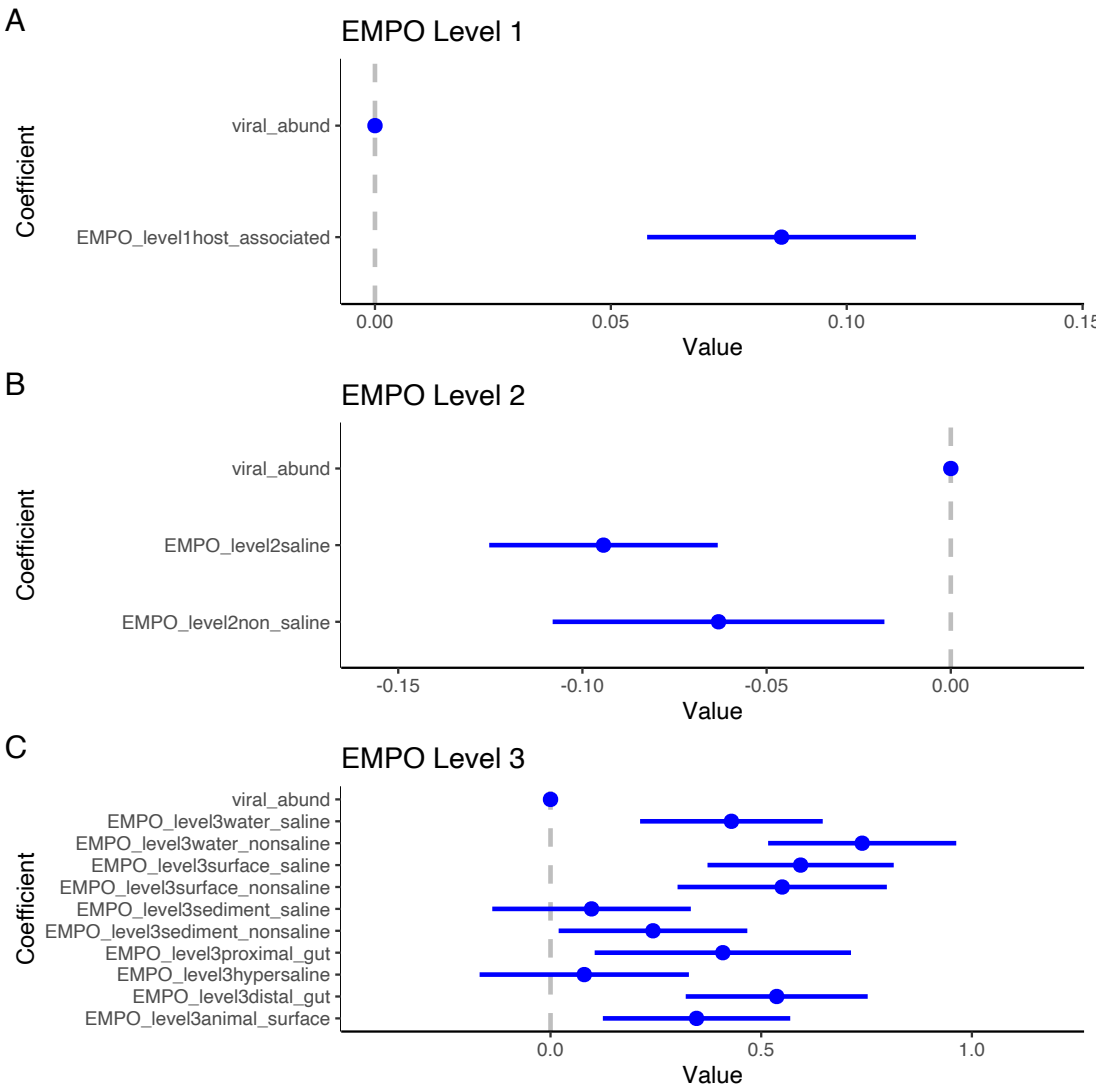

Figure S2. Environmental effects when controlling for viral abundance, Related to Figure 2. Model coefficients from general linear models that model how environmental classification affects CRISPR abundance. Panels represent A) EMPO level 1, B) EMPO level 2 and C) EMPO level 3. Points and 95% confidence intervals represent how much each level contributes to the increase (or decrease) in CRISPR abundance, relative to the first level of the factor (e.g. 'free-living' for A, 'host associated' for B and 'animal secretion' for C. Note that in almost all cases the environmental classifications do not overlap with the intercept, despite inclusion of viral abundance in the model.

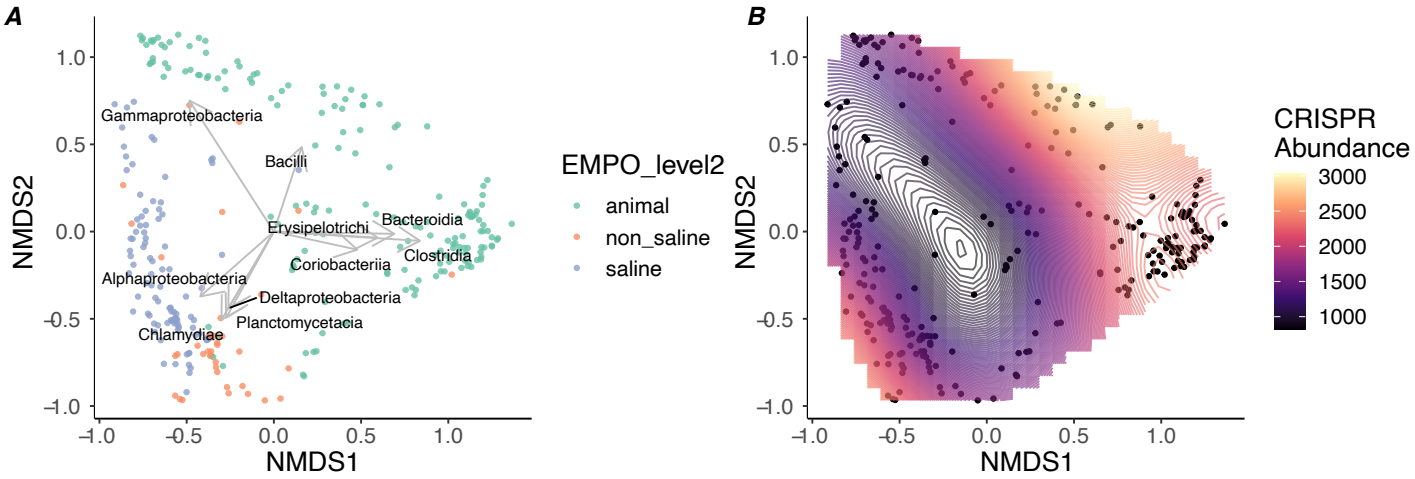

Figure S3. Phylogenetic ordinations of sample community composition, Related to STAR methods. Ordinations of sample community compositions at the Class level. Points represent individual metagenomic samples and are clustered using non-metric multidimensional scaling (NMDS) on Bray-Curtis dissimilarity scores. (*A*) Colors represent the sample classification at EMPO level 2 and arrows represent the top 10 species loadings. (*B*) The same ordination of samples with CRISPR abundance used in a generalized additive model (GAM) to fit a surface predicting CRISPR abundance.

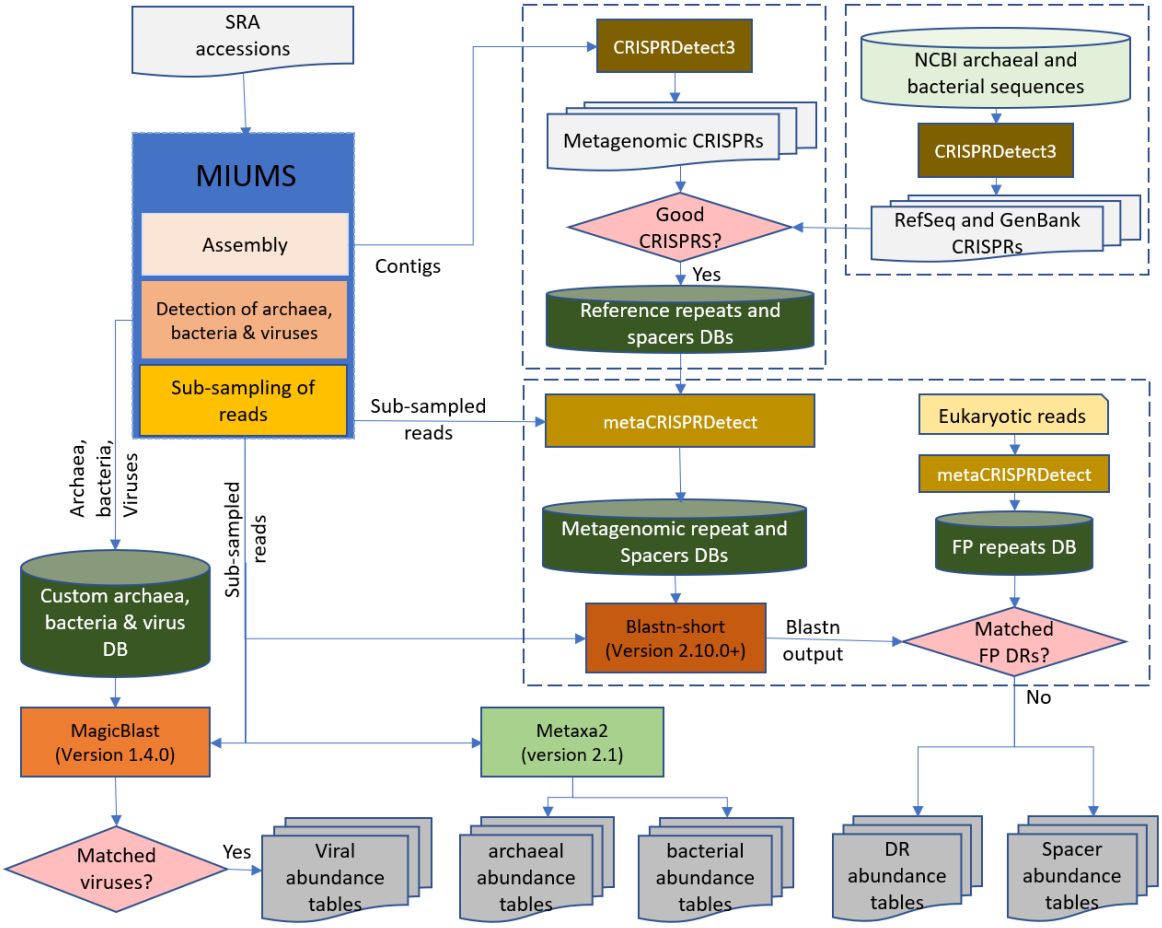

Figure S4. Overview of computational pipeline, Related to STAR methods. Overview of the computational pipeline used to generate CRISPR abundance tables, microbial and viral community abundance tables. In each stage, only the case (yes/no) where data is retained is presented.