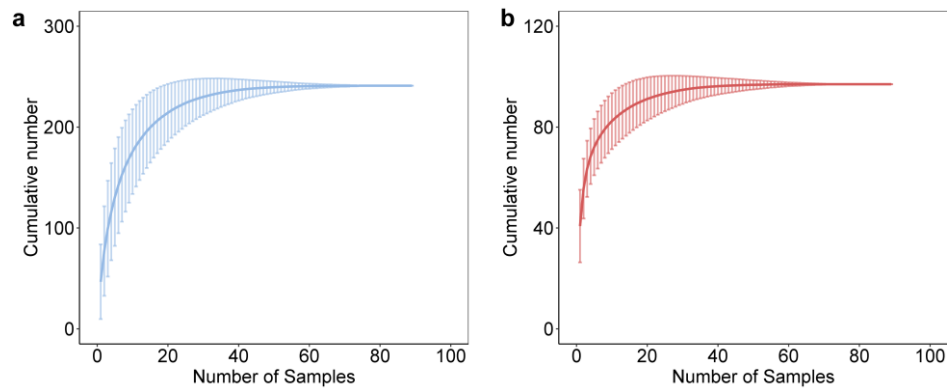
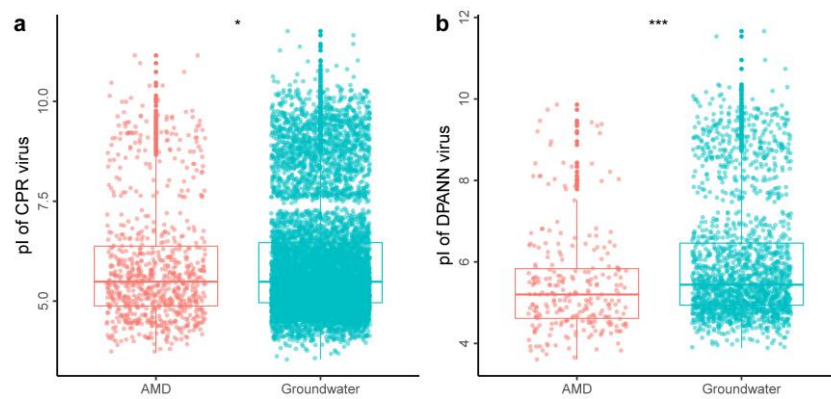


Supplementary figures



Supplementary Fig. 1 | Rarefaction analysis of CPR and DPANN vOTUs. The accumulation curves for DPANN (a) and CPR (b) vOTUs are displayed, with different colors representing distinct viral groups. Dots represent the average number of DPANN and CPR vOTUs for all combinations of a given number of samples, and error bars represent the range.



Supplementary Fig. 2 | Isoelectric points (pI) differences of DPANN and CPR viral proteins

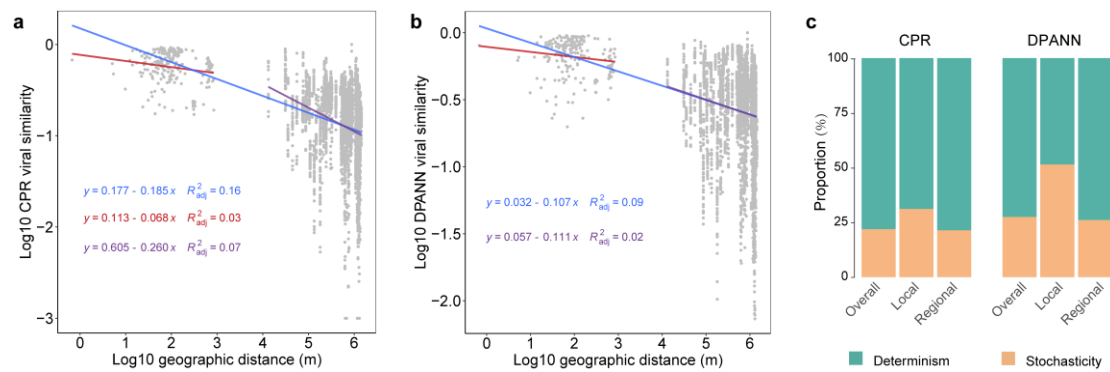
between AMD sediment and groundwater. The comparison of pI in CPR (a) and DPANN (b)

viral proteins between AMD sediment and groundwater. For each box plot, the center line

represents the median, the box limits represent the 25th and 75th percentiles (IQR), and the

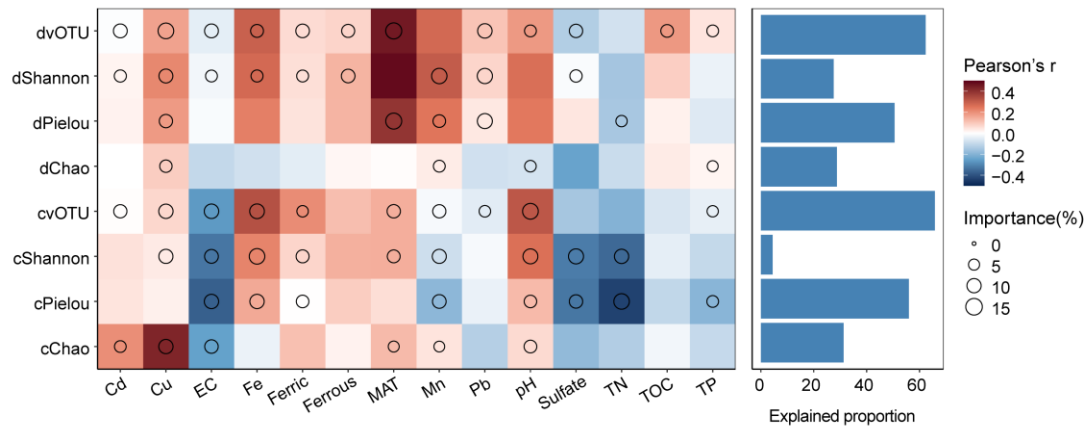
whiskers extend to 1.5×IQR. Significance levels of Wilcoxon test are indicated by asterisks: * for

$p < 0.05$ and **** for $p < 0.001$.

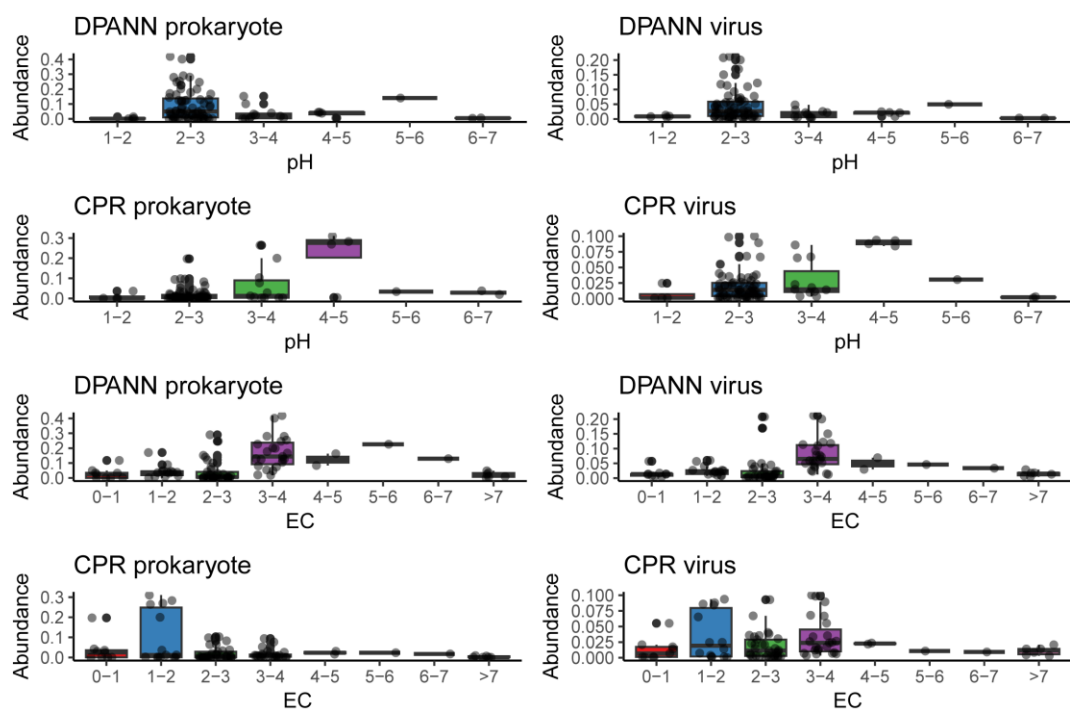


Supplementary Fig. 3 | Relationship Between Community Similarity and Distance Decay

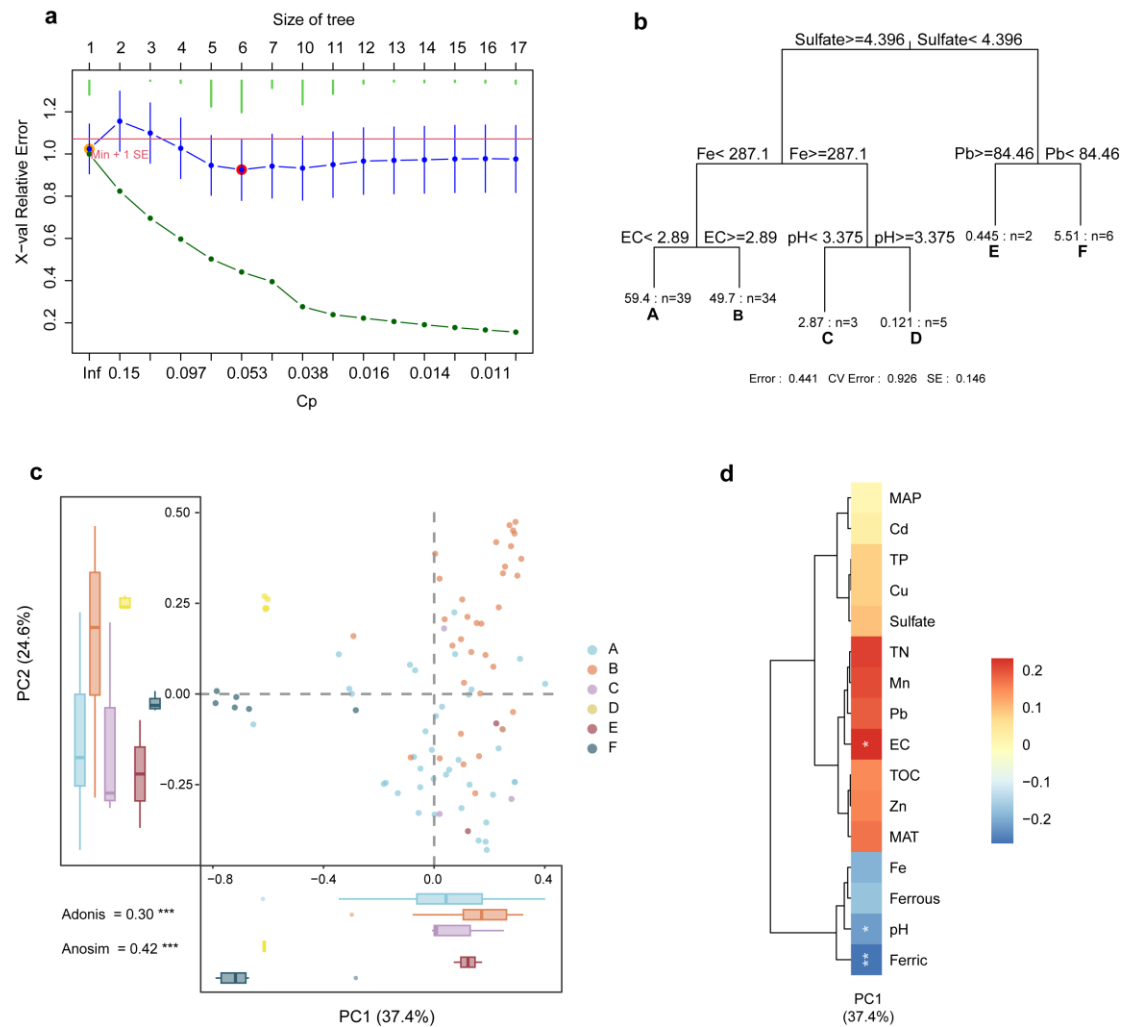
Relationship (DDR) for CPR and DPANN Viruses. a, b The relationship between community similarity and geographic distance decay relationship (DDR) for CPR (**a**) and DPANN (**b**) viruses. The blue line represents the least-squares linear regression across all spatial scales. The red line denotes the regression within samples whose distance is ≤ 1 km, while the purple line denotes the regression within samples whose distance is > 1 km. Color-coded best-fit lines and adjusted R^2 values for each DDR are presented, illustrating the spatial patterns of CPR viral communities. **c** The results of viral community assembly mechanisms.



Supplementary Fig. 4 | The important factors influencing alpha diversities of DPANN and CPR viruses. Associations of environmental factors and abundances based on correlation and random forest regression model. The circle size represents the variable importance (i.e., proportion of explained variability), and the color gradient in the heatmap denotes Pearson's correlation coefficients. EC, electronic conductivity; Ferrous, ferrous iron; Ferric, ferric iron; TOC, total organic carbon; TN, total nitrogen; TP, total phosphorus; MAT, mean annual temperature; dPielou, the Pielou index of the DPANN viruses; cPielou, the Pielou index of the CPR viruses; dShannon, the Shannon index of the DPANN viruses; cShannon, the Shannon index of the CPR viruses; dvOTU, the number of DPANN viral operational taxonomic units; cvOTU, the number of CPR viral operational taxonomic units; dChao, the Chao1 index of the DPANN viruses; cChao, the Chao1 index of the CPR viruses.



Supplementary Fig. 5 | Distribution of relative abundance of DPANN/CPR viruses and prokaryotes across varying pH and EC levels. Each box plot shows the interquartile range, with the median value marked, highlighting the variability and central tendency of microbial abundances in response to the environmental gradients of pH and EC. For each box plot, the center line represents the median, the box limits represent the 25th and 75th percentiles (IQR), and the whiskers extend to $1.5 \times \text{IQR}$.



Supplementary Fig. 6 | The relationship between DNANN and CPR viral composition and

environmental factors. a Results of the relative error (RE) and cross-validation relative error

(CVRE) for the multivariate regression tree. The red dot indicates the classification scheme with

the minimum CVRE, along with the error bars for CVRE. The green bars above the plot denote

the number of cross-validation iterations that achieved the optimal classification scheme,

highlighting the stability and reliability of the model. Cp, complexity parameter. **b** The

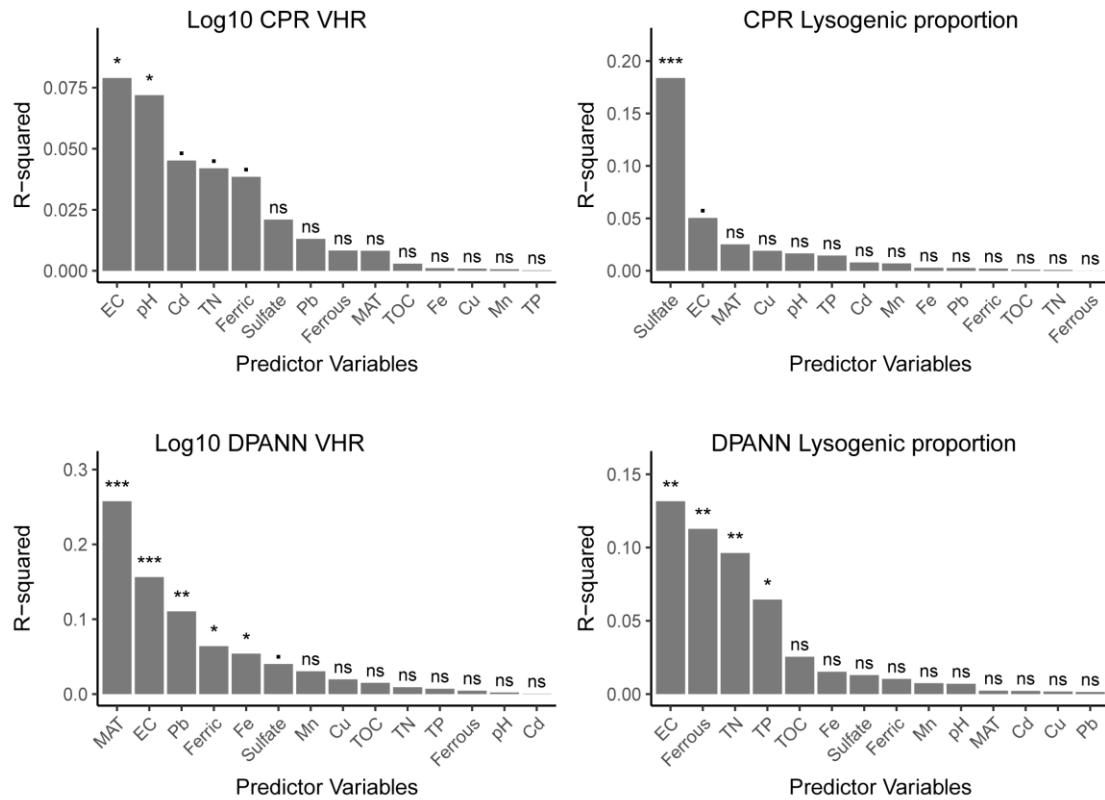
multivariate regression tree for ultra-small prokaryotic viruses explained by environmental

variables. This tree structure illustrates how different environmental factors contribute to the

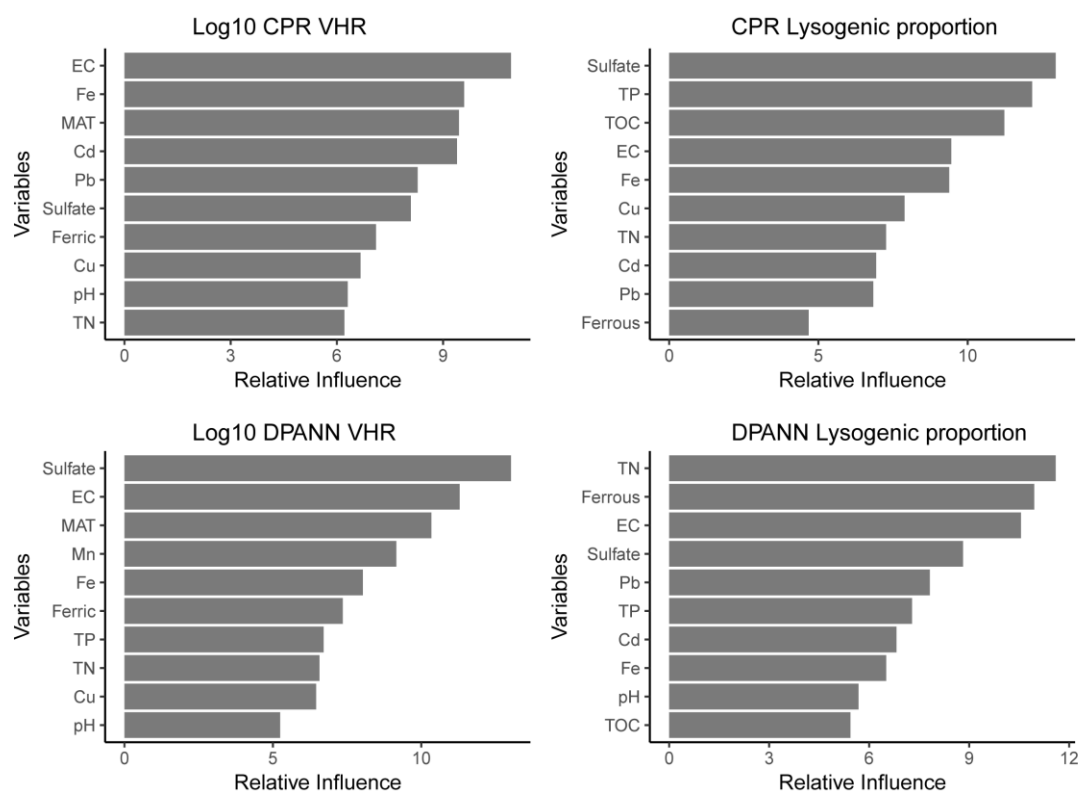
distribution and abundance of these viruses. Each node represents a split based on an

environmental variable, and the branches show the direction of the splits, leading to terminal

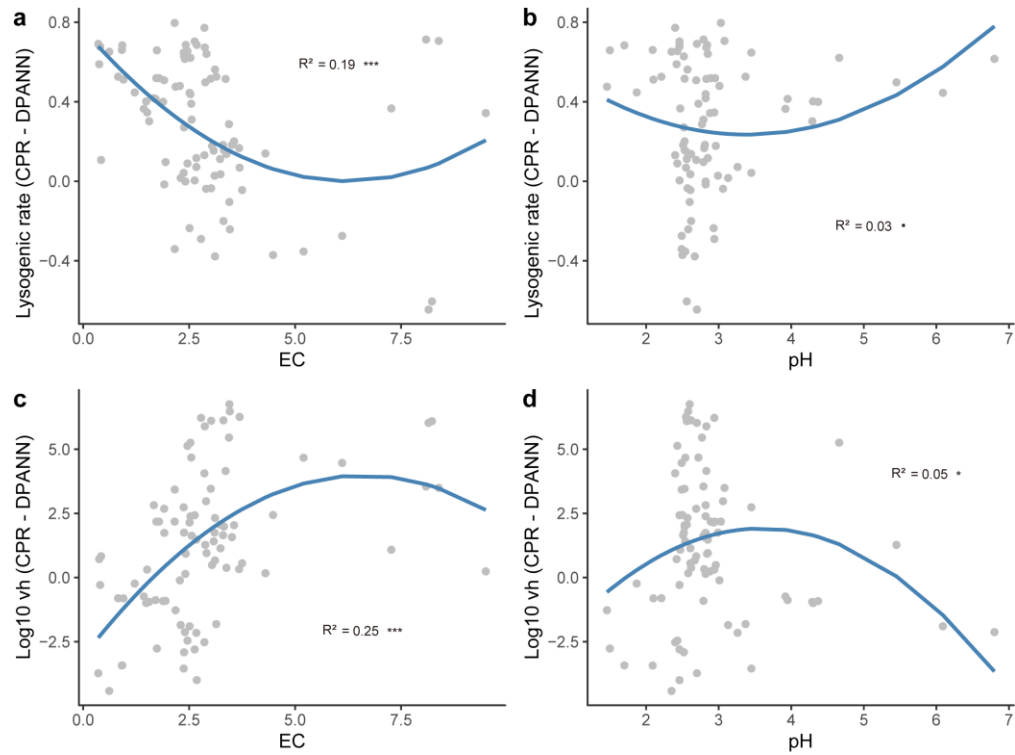
nodes that represent the final groups (labeled A through F from left to right) with their predicted values. **c** PCoA results of the total viral community structure colored by MRT-derived groups. Boxplots show distributions of different groups. **d** Correlation of viral composition (PC1) and environmental factors. The percentage in parentheses represents the proportion of total community variation explained by the PC1 from the Bray-Curtis dissimilarity-based principal coordinate analyses. The color gradient represents the strength and direction of the correlation coefficients. Significance levels are indicated by asterisks: * for $p < 0.05$ and ** for $p < 0.01$.



Supplementary Fig. 7 | Environmental drivers of DPANN and CPR virus-host dynamics. The bar plot displays the explanatory power (R-squared values) of individual environmental factors on virus-host dynamics based on Robust Linear Regression analyses. Each bar represents the proportion of variance in virus-host dynamics explained by a single environmental factor. Significance levels are indicated as follows: · $P < 0.1$; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

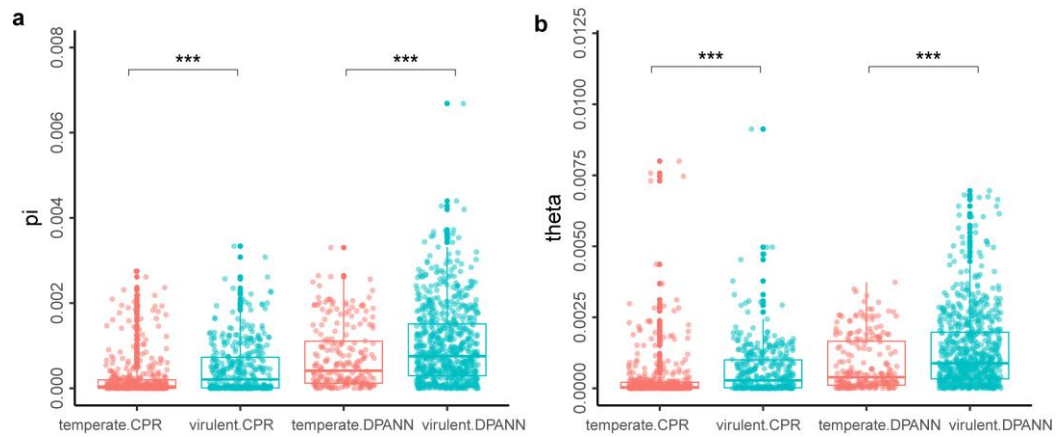


Supplementary Fig. 8 | Predictive variable importance in GBM model. A bar plot showing the relative influence of the top 10 predictive variables.



Supplementary Fig. 9 | Dynamic differences of CPR and DPANN under EC and pH

variations. a, b The differences in lysogenic rates (CPR minus DPANN) with changes in EC (**a**) and pH (**b**). The R-squared values represent the fit of the polynomial regression curves, with asterisks indicating significance. **c, d** the VHRs changes with EC (**c**) and pH (**d**) variations. • $P < 0.1$; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.



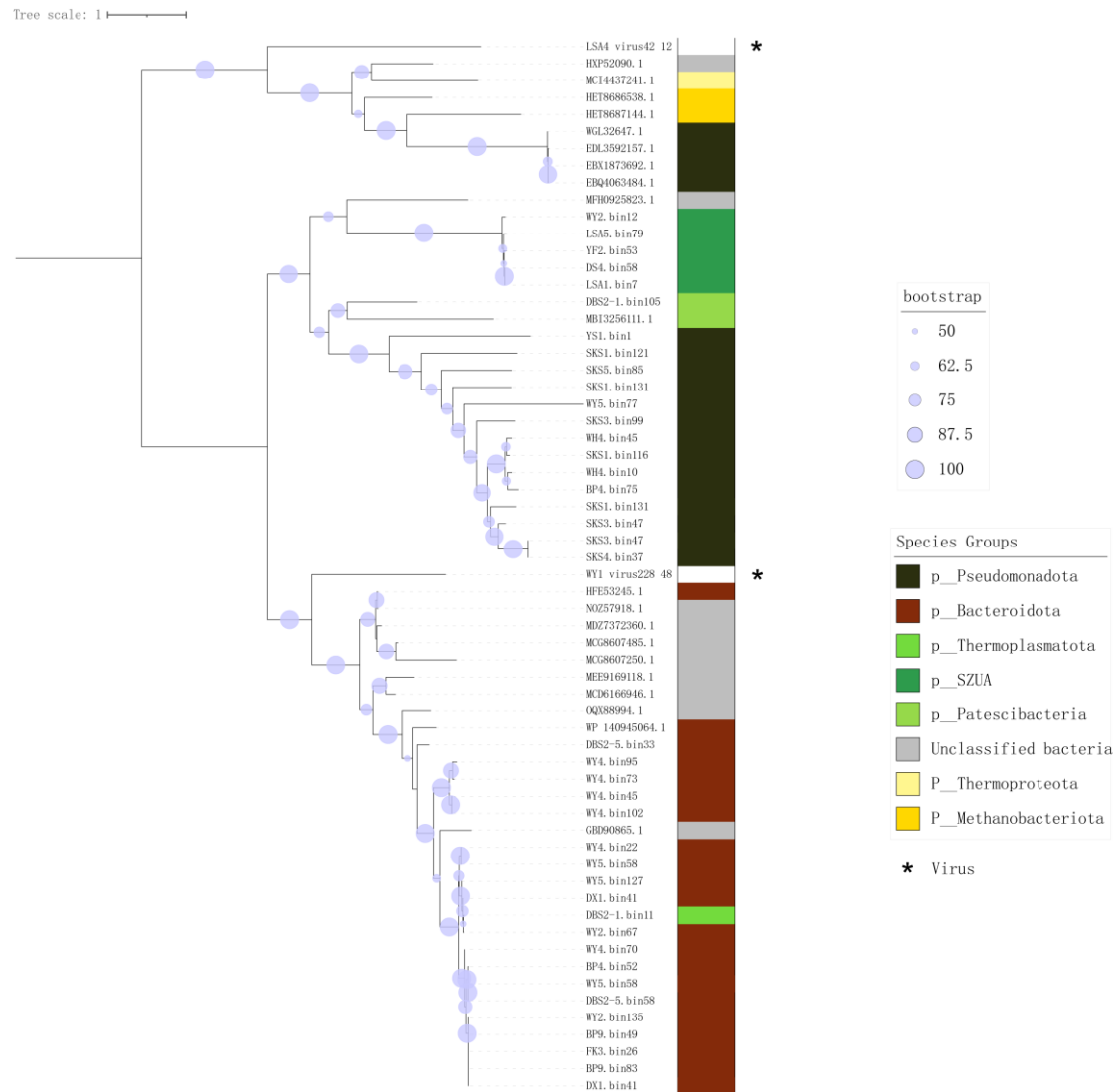
Supplementary Fig. 10 | Distribution of viral population diversity metrics across lytic, and

temperate phage groups. a Nucleotide diversity (π) quantified by average pairwise

differences. **b** Genetic diversity (θ) calculated using Watterson's estimator. Boxplots show median

(center line), interquartile range (IQR, box boundaries), and $1.5 \times \text{IQR}$ whiskers. Asterisks denote

significance levels from pairwise Wilcoxon rank-sum tests: *** $P < 0.001$.



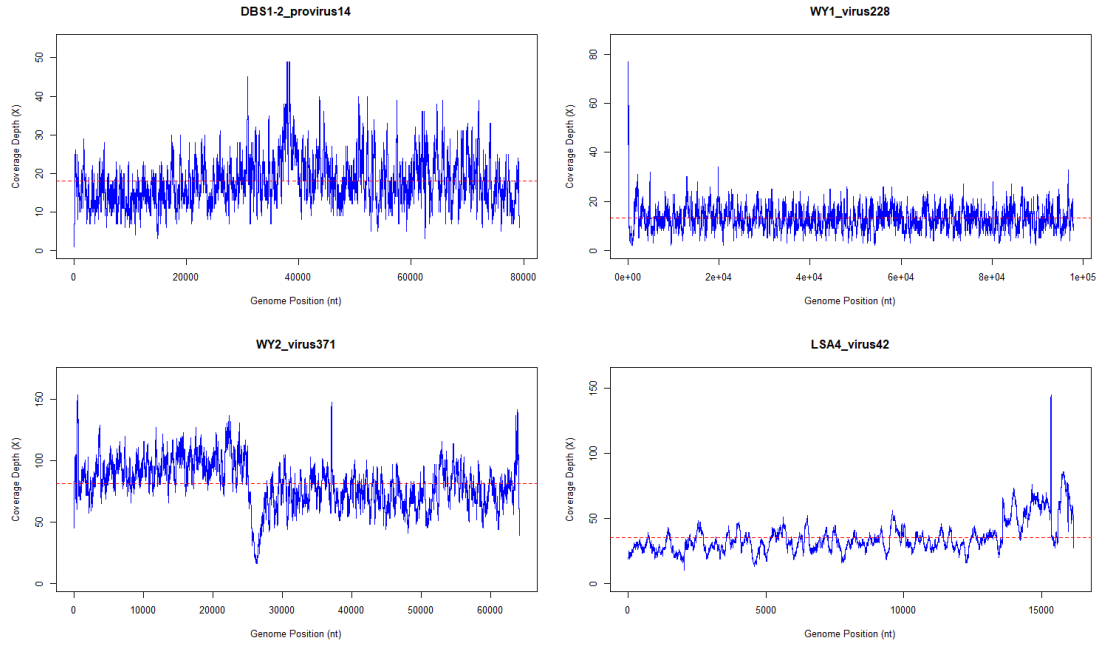
Supplementary Fig. 11 | phylogenetic analysis of viral-encoded *rpoS*. Maximum-likelihood

phylogenetic tree with *rpoS* genes from the DPANN viruses compared to homologs in

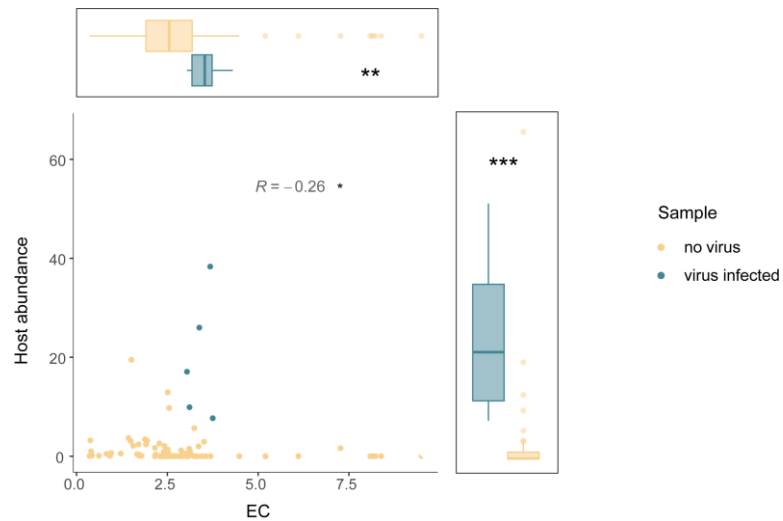
metagenome-assembled genomes (MAGs) and NCBI-nr reference sequences (colored by

bacterial/archaeal phyla). The prokaryotic proteins colored by different phyla. Viral sequences are

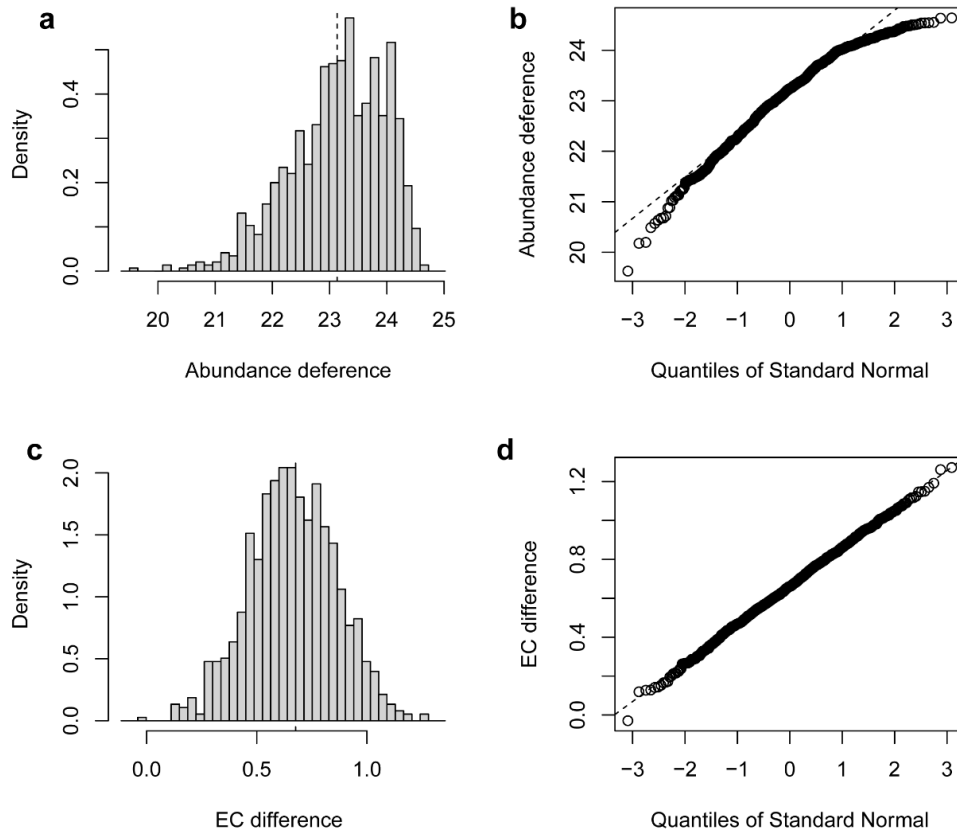
highlighted asterisks. Scale bar indicates substitutions per site.



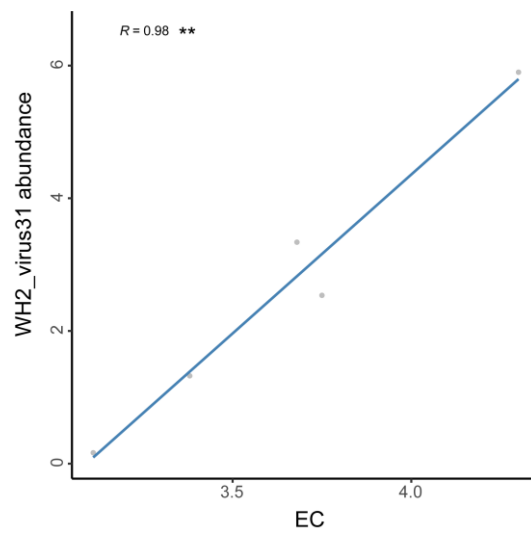
Supplementary Fig. 12 | Viral genome coverage depth. A line plot showing sequencing coverage depth (X) across the viral genome positions (nucleotide). The blue line represents raw coverage depth; the dashed red line indicates the mean coverage depth (mean x). The y-axis is scaled from zero to slightly above the maximum observed depth. Read mapping was performed using Bowtie2. To assess the uniformity of sequence coverage and to check for potential chimeric assemblies, genome-wide coverage depth was calculated using samtools depth with the -a option to include all positions, including those with zero coverage.



Supplementary Fig. 13 | Correlation of EC and the host abundance of ectoine synthase carried virus (WH2_virus31) in viral-positive (blue) vs. viral-negative (orange) samples. Scatterplot showing significant negative correlation between EC and host abundance across all samples (Pearson's $R = -0.26$, $p < 0.05$). Boxplots comparing EC distributions between groups (center line: median; box limits: 25th–75th percentiles; whiskers: $1.5 \times \text{IQR}$). and the asterisk show the wilcox test significance. Viral-positive status defined as the abundance of WH2_virus31 more than one.

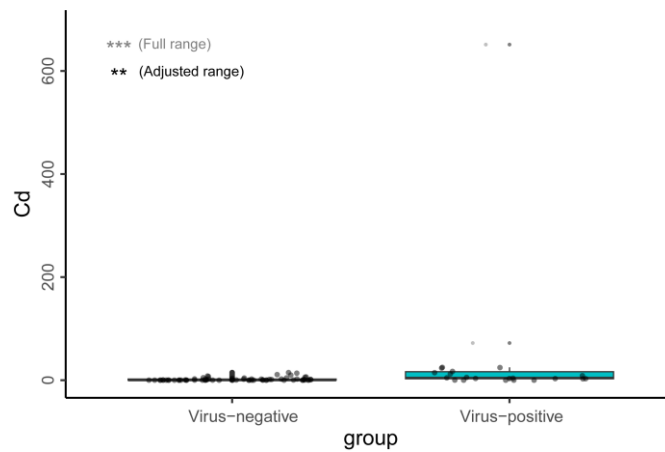


Supplementary Fig. 14 | Bootstrap analysis of Host Abundance and EC Differences Between viral-positive and viral-negative samples. **a, c** The mean differences (viral-positive samples minus viral-negative samples) in host abundance (**a**) and EC (**c**), respectively, based on 1000 bootstrap resamples. Viral-positive status defined as the abundance of WH2_virus31 more than one. **b, d** The Quantile-Quantile Plots of host abundance (**b**) and EC (**d**) for the 1000 bootstrap resamples. The dashed line represents the benchmark for comparing the quantiles of the theoretical distribution. Each point represents a comparison between a quantile of the data from one of these bootstraps resamples and the corresponding quantile in the theoretical distribution.



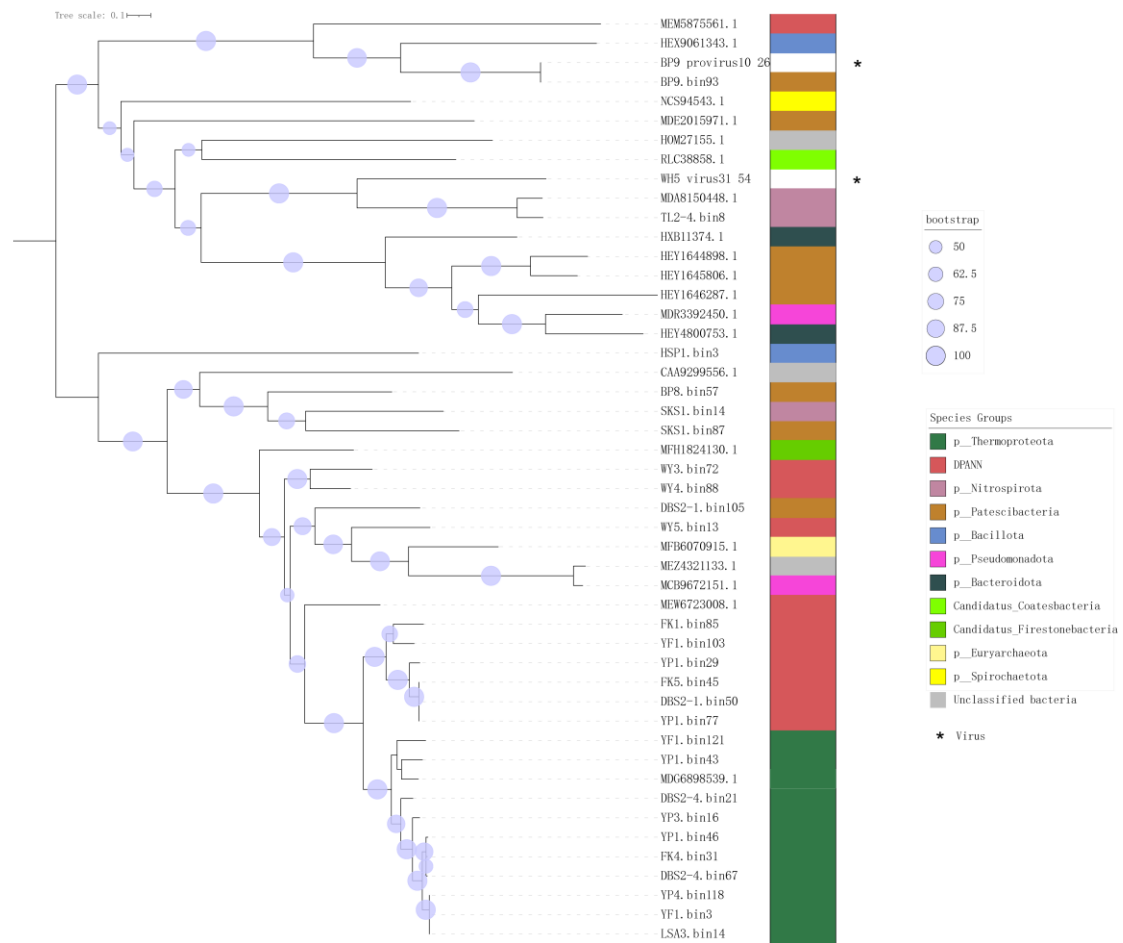
Supplementary Fig. 15 | Correlation Between WH2_virus31 abundance and EC in MAS

Samples. The relationship between the abundance of WH2_virus31 and electrical conductivity (EC) in MAS samples. The blue line represents the linear regression. The R value denotes the Pearson correlation coefficient, and asterisks indicate the significance level (** $p < 0.01$).



Supplementary Fig. 16 | Association between SmtB-carrying viral occurrence and Cd

concentration. Box plots comparing cadmium concentrations (mg/kg) in virus-negative (absence of the *SmtB*-carrying virus) versus virus-positive samples. Asterisks above boxes indicate significance levels from Wilcoxon rank-sum tests: gray asterisks represent analysis including all data points (full range, outliers retained; *** $P < 0.001$); black asterisks represent analysis after outlier removal (adjusted range; ** $P < 0.01$).

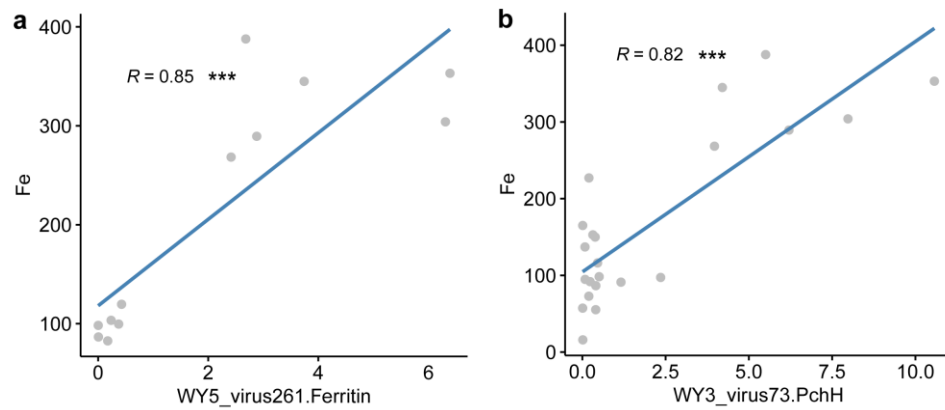


Supplementary Fig. 17 | phylogenetic analysis of viral-encoded *purD*. Maximum-likelihood

phylogenetic tree with *purD* genes from the CPR viruses compared to homologs found in our

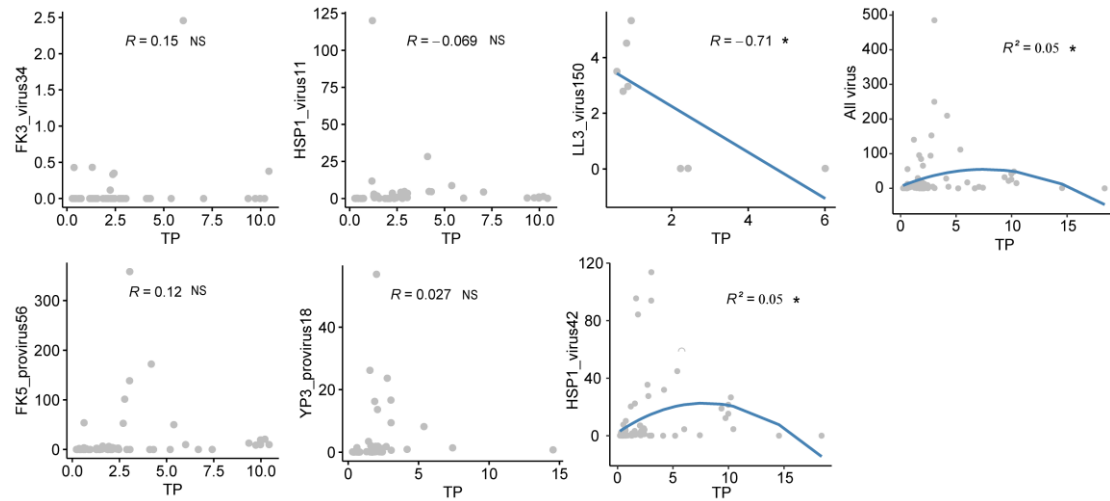
MAG sets and nr database. The prokaryotic proteins colored by different phyla. Viral sequences

are highlighted asterisks. Scale bar indicates substitutions per site.



Supplementary Fig. 18 | Correlation Between Fe (g/kg) and the abundance of each viruses

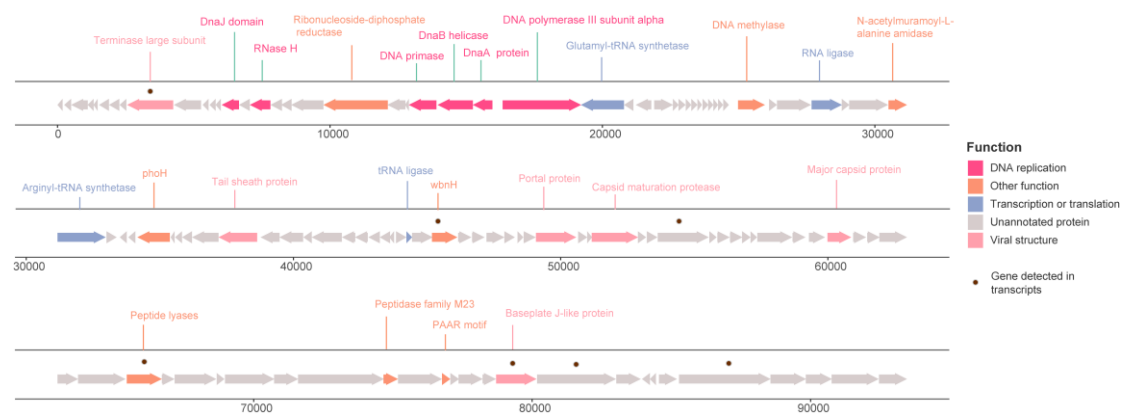
carrying iron-related AMGs. a The correlation between Fe concentration and the abundance of WY5_virus261 with ferritin related AMG (PF00210). **b** The correlation between Fe concentration and the abundance of WY3_virus73 with *PchH*. The blue line represents the linear regression. The R value denotes the Pearson correlation coefficient, and asterisks indicate the significance level.



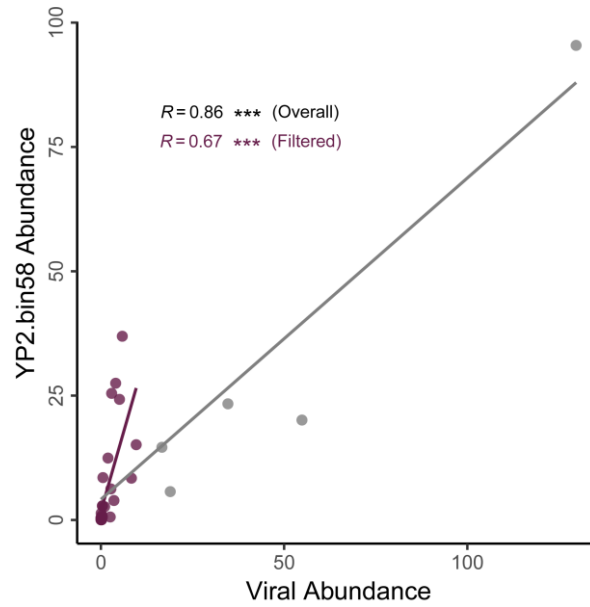
Supplementary Fig. 19 | Correlation Between TP (g/kg) and the abundance of each viruses

carrying *phoH*. The blue line indicate best-fitting polynomial functions. The R value denotes the Pearson correlation coefficient, and asterisks indicate the significance level. NS, no significance.

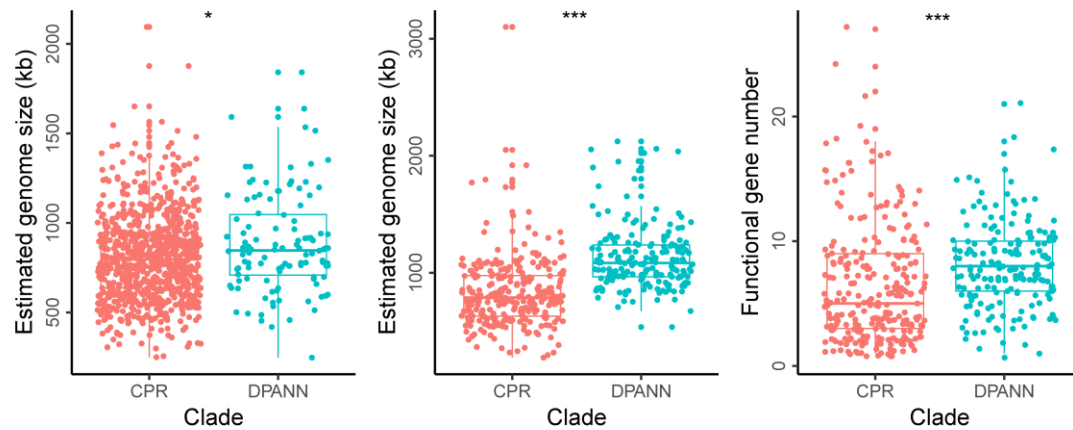
The R^2 value (coefficient of determination) reflects the goodness-of-fit for the best-fitting polynomial model, with asterisks indicating the statistical significance of the polynomial regression analysis.



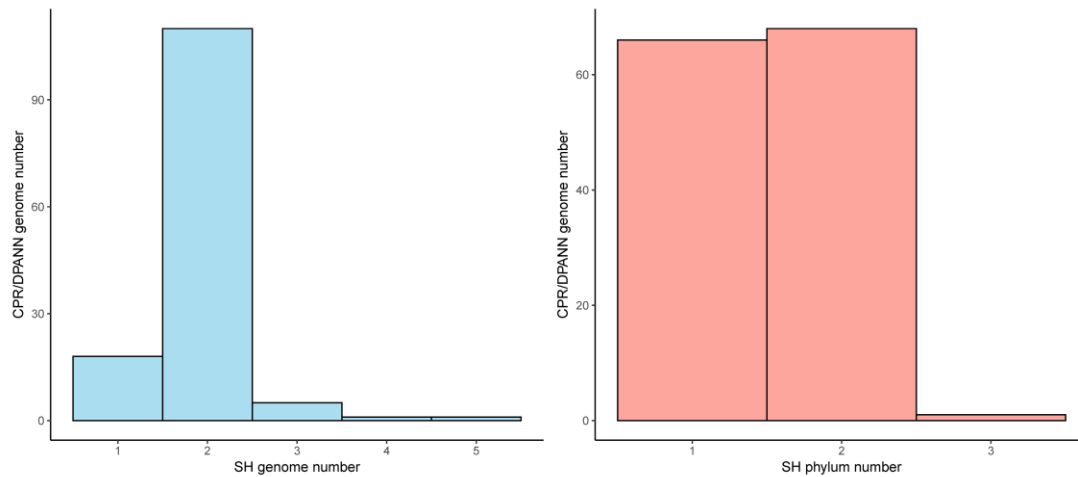
Supplementary Fig. 20 | Genomic map of the archaea virus (YP3_provirus18). Different functional categories are distinguished using various colors. Genes detected in transcription profiles are represented by red circles.



Supplementary Fig. 21 | Correlation between the abundance of YP2.bin58 and the abundance of YP4_provirus51. Gray points: viral abundance outliers ($1.5 \times \text{IQR}$, $n=5$). Purple line: significant Pearson correlation after outlier exclusion; gray line: overall trend. Asterisks indicate the statistical significance (** $P < 0.001$).



Supplementary Fig. 22 | Comparative genomics of CPR and DPANN lineages. Box plots with jittered data points comparing functional gene numbers and estimated genome sizes between CPR and DPANN clades. (Left) Estimated genome sizes (kb) of metagenome-assembled genomes from Castelle et al. (2018). (Center) Estimated genome sizes (kb) of metagenome-assembled genomes from our datasets. (Right) Functional gene numbers of genomes from our datasets. Asterisks above boxes indicate significance levels from Wilcoxon rank-sum tests: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.



Supplementary Fig. 23 | Host range analysis of parasitic genomes. (Left) Histogram showing the distribution of SH genome counts per CPR/DPANN genome. The x-axis indicates the number of distinct SH genomes, and the y-axis shows the count of CPR/DPANN genomes associated with each host count. (Right) Histogram displaying the distribution of SH phylum diversity per CPR/DPANN genome. The x-axis represents the number of distinct host phyla.