Supplementary figures for

Giant viruses as reservoirs of antibiotic resistance genes

Xinzhu Yi[1,†], Jie-Liang Liang[1,†], Ping Wen[1], Pu Jia[1], Shi-wei Feng[1], Shen-yan Liu[1], Yuan-yue Zhuang[1], Yu-qian Guo[1], Jing-li Lu, Sheng-ji Zhong[1], Bin Liao[2], Zhang Wang[1], Wen-sheng Shu[1], & Jin-tian Li[1,*]

[1]Institute of Ecological Science, Guangzhou Key Laboratory of Subtropical Biodiversity and Biomonitoring, Guangdong Provincial Key Laboratory of Biotechnology for Plant Development, School of Life Sciences, South China Normal University, Guangzhou 510631, PR China

[2]School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, PR China

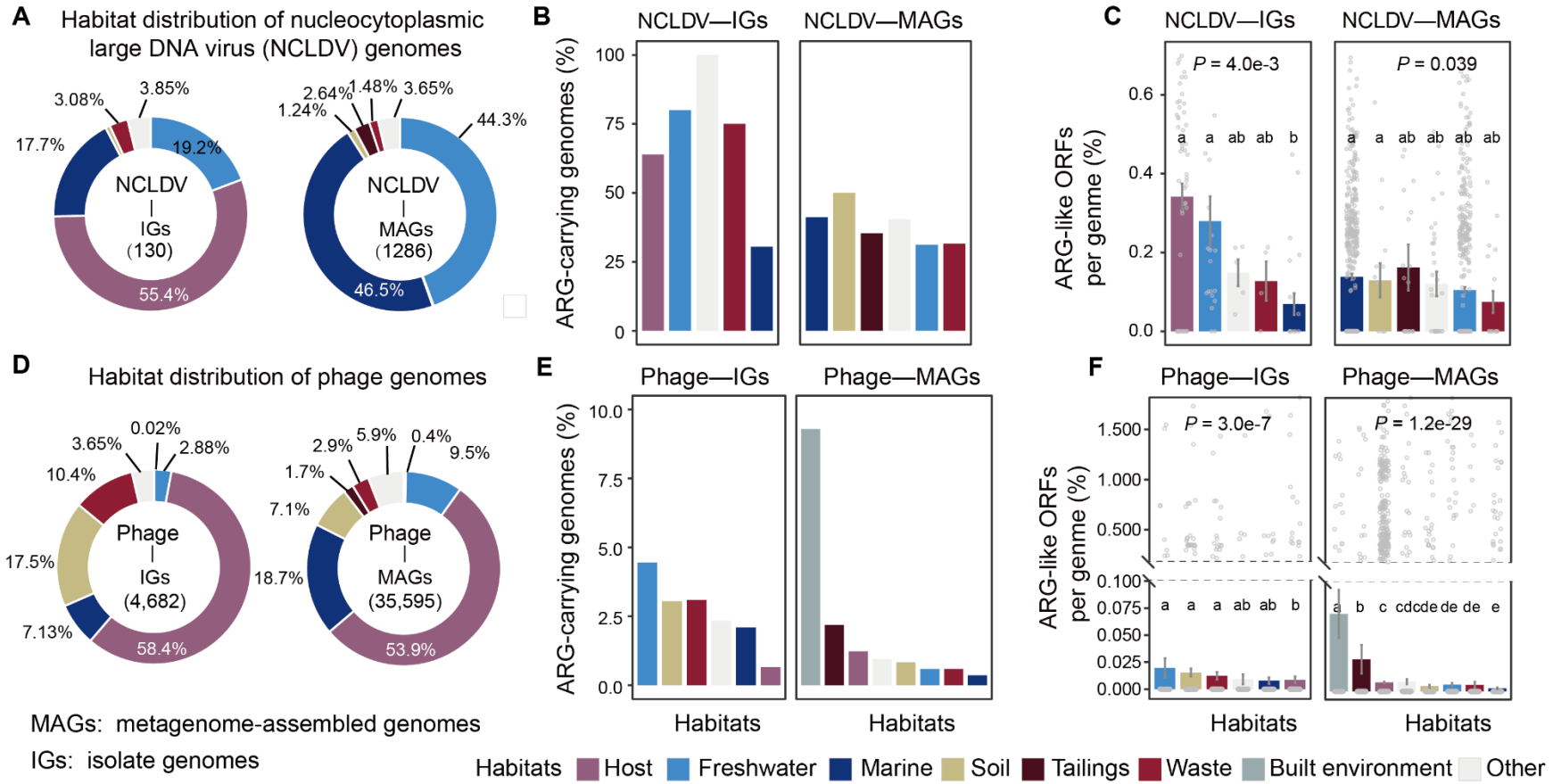
[†]These authors contributed equally to this work.

[*]Corresponding author:

School of Life Sciences, South China Normal University, Guangzhou 510631, PR China

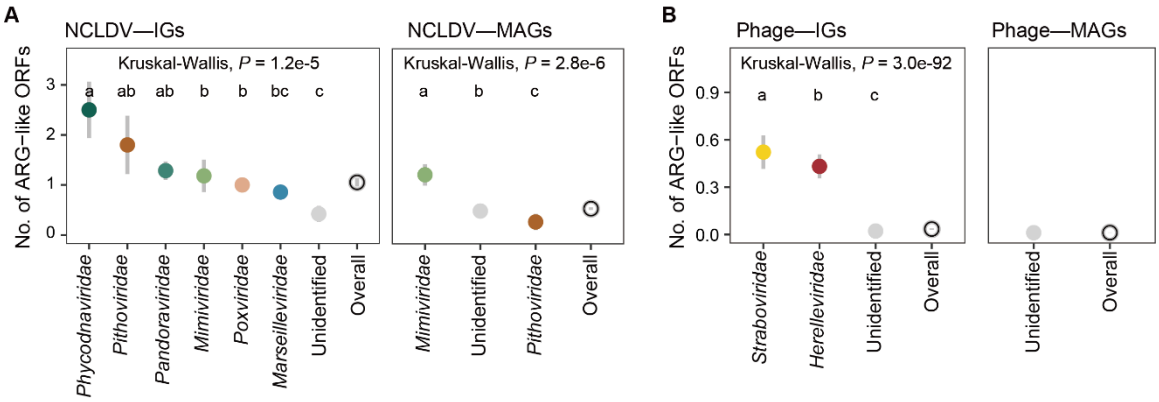
E-mail: Jin-tian Li (lijintian@m.scnu.edu.cn)

Tel.: +86 20 85211850; Fax: +86 20 85211850

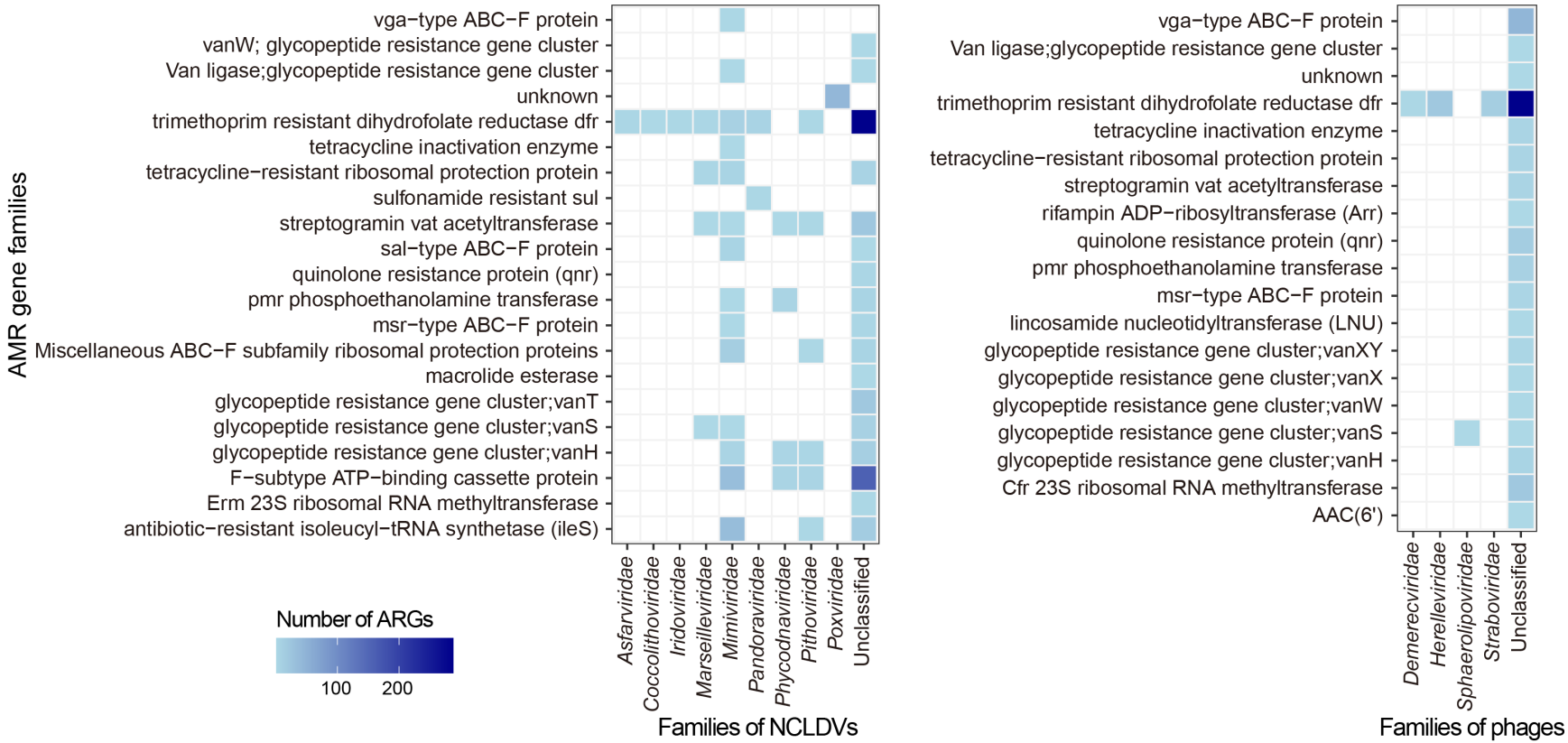


Supplementary Fig. 1 | Habitat distribution of viral genomes and their antibiotic resistance gene (ARG) contents. A, D: Habitat distribution of the genomes of A) nucleocytoplasmic large DNA viruses (NCLDVs) and D) phages analyzed in this study. **B, E:** Possibility of ARG-carriage in different habitats for B) NCLDVs and E) phages. **C, F**: Genomic potential of ARG-carriage of C) NCLDVs and F) phages in different habitats. For clarity, only seven major habitats are shown individually, with other habitats being

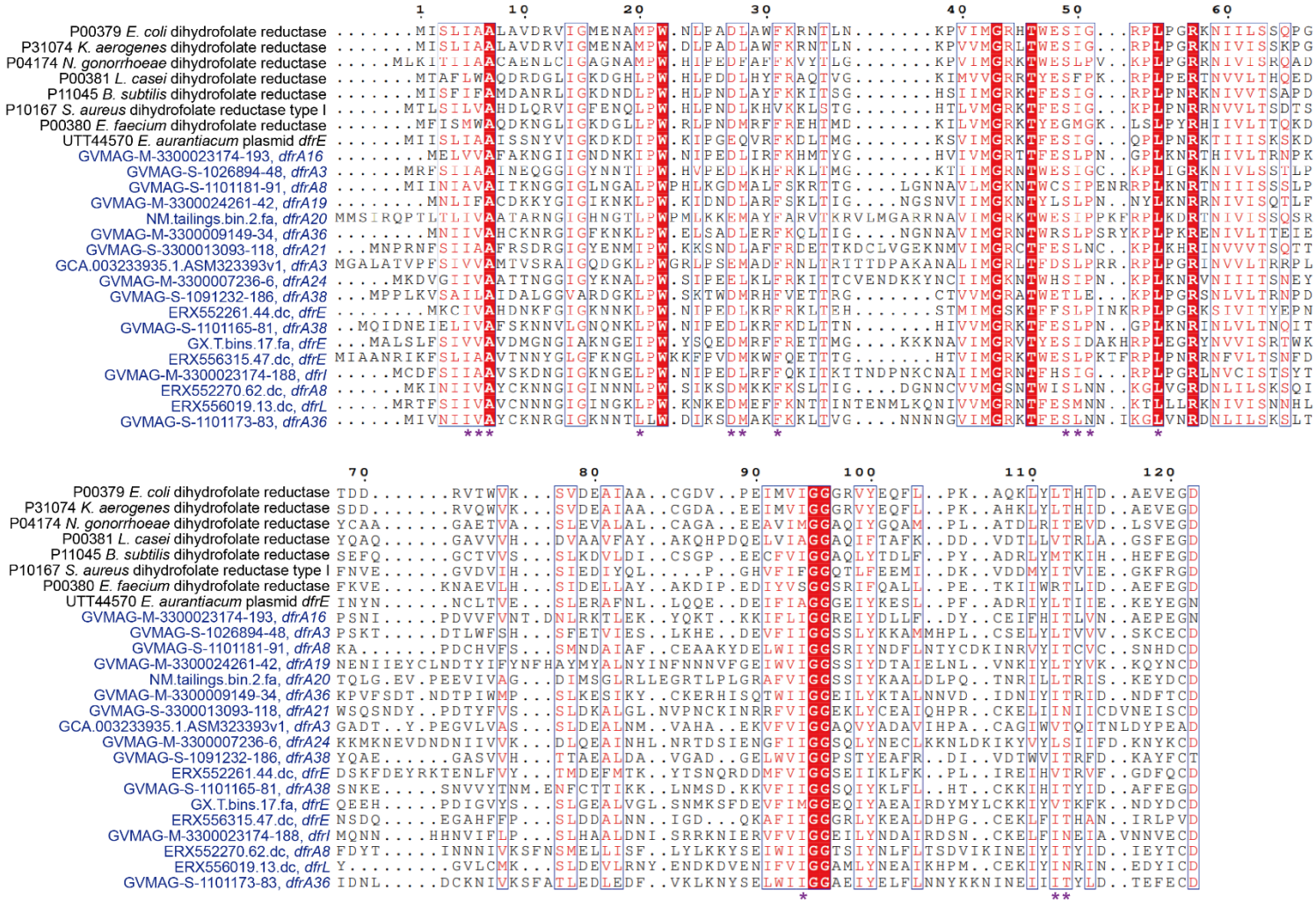
referred to as “Other” and shown as a whole. Only habitats with more than three genomes are shown in B), C), E), and F). Lower-case letters above the bars in C) and F) represent significantly different groups assessed with two-sided Wilcoxon rank-sum test, and *P*-values indicate the overall difference among all habitats assessed with Kruskal–Wallis test. Data presented in C) and F) were mean values ± standard error of the mean (SEM). Each data point represents an individual genome in the corresponding group. The number (*n*) of genomes in each group can be found in Source Data file on Github [1]. The y-axis was truncated to zoom in on values below 0.1% in panel F). ORF is the abbreviation for open reading frame, IG is for isolate genome, and MAG is for metagenome-assembled genome.



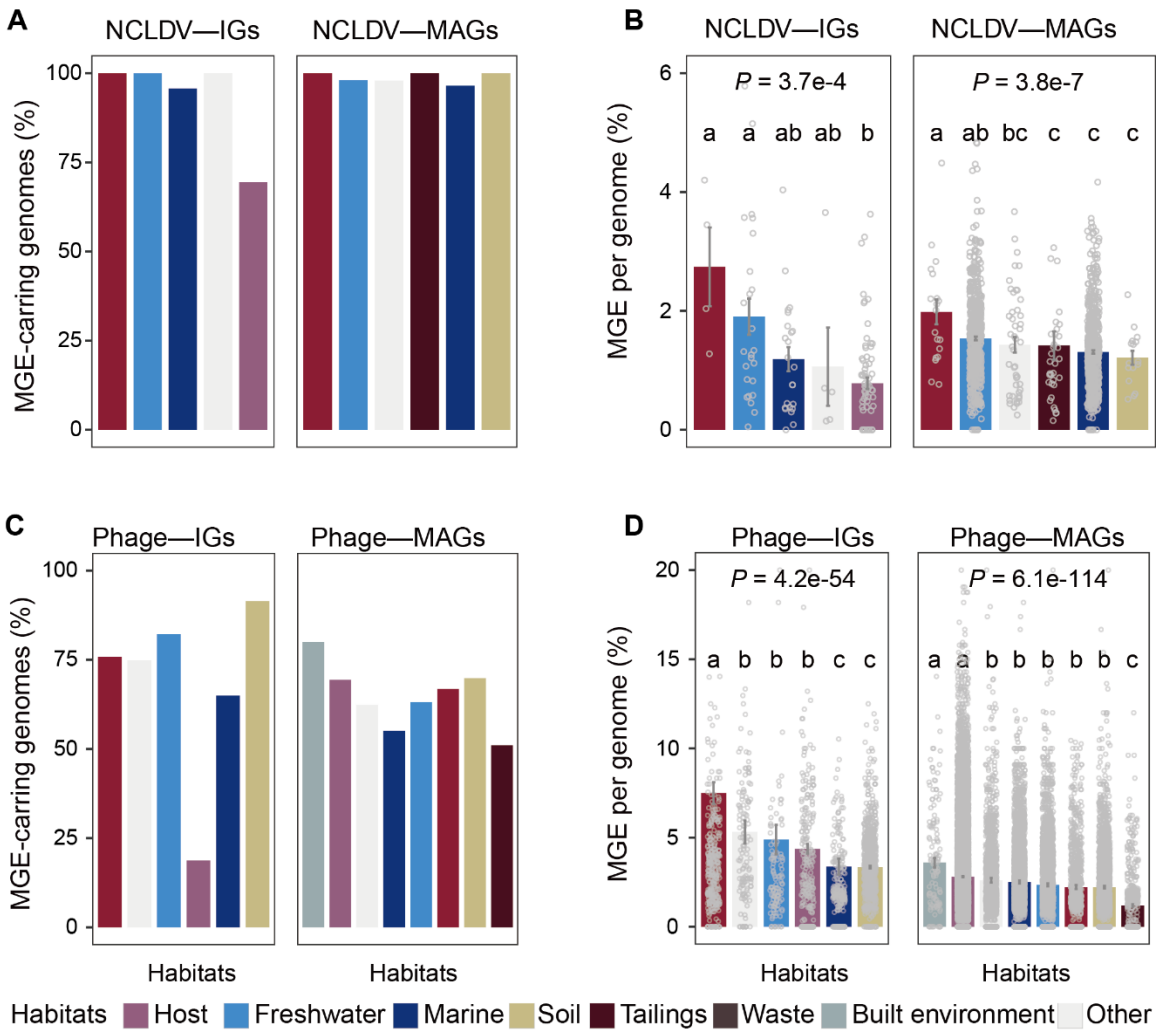
Supplementary Fig. 2 | Average numbers of ARG-like ORFs detected in different families of A) NCLDVs and B) phages. Data are presented as mean values ± SEM. The unit of study is one genome. The number (*n*) of genomes in each group can found in Source Data file on Github [1]. Lower-case letters above the bars represent significantly different groups assessed with two-sided Wilcoxon rank-sum test, and *P*-values indicate the overall difference among all families assessed with Kruskal–Wallis test.



Supplementary Fig. 3 | Distribution of antimicrobial resistance gene families in different viral families of A) NCLDVs and B) phages. All viral families with at least one ARG detected are displayed. Source data are provided as a Source Data file on Github [1].

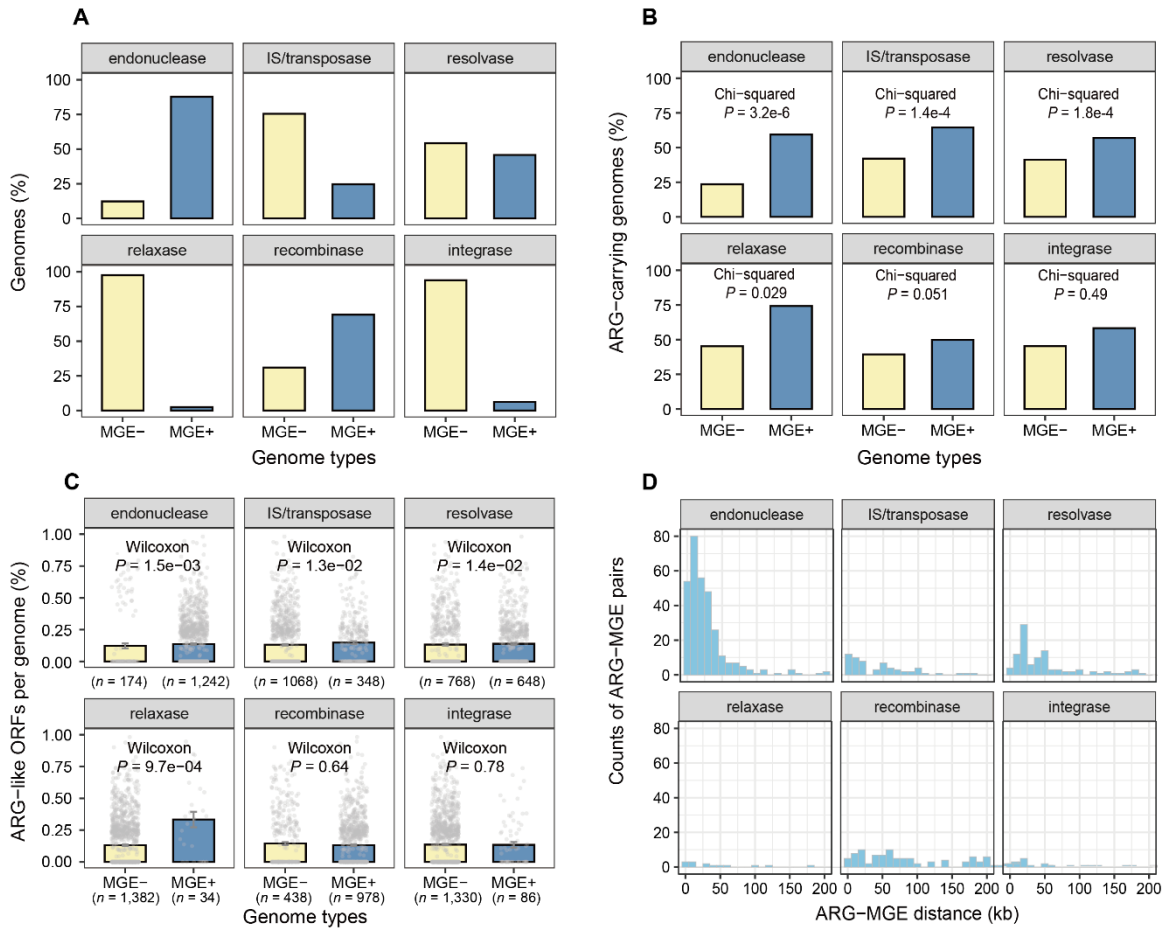

P00379 E. coli dihydrofolate reductase
P31074 K. aerogenes dihydrofolate reductase
P04174 N. gonorrhoeae dihydrofolate reductase
P00381 L. casei dihydrofolate reductase
P11045 B. subtilis dihydrofolate reductase
P10167 S. aureus dihydrofolate reductase type I
P00380 E. faecium dihydrofolate reductase
UTT44570 E. aurantiacum plasmid dfrE
GVMAG-M-3300023174-193, dfrA16
GVMAG-S-1026894-48, dfrA3
GVMAG-S-1101181-91, dfrA8
GVMAG-M-3300024261-42, dfrA19
NM.tailings.bin.2.fa, dfrA20
GVMAG-M-3300009149-34, dfrA36
GVMAG-S-3300013093-118, dfrA21
GCA.003233935.1.ASM323393v1, dfrA3
GVMAG-M-3300007236-6, dfrA24
GVMAG-S-1091232-186, dfrA38
ERX552261.44.dc, dfrE
GVMAG-S-1101165-81, dfrA38
GX.T.bins.17.fa, dfrE
ERX556315.47.dc, dfrE
GVMAG-M-3300023174-188, dfrI
ERX552270.62.dc, dfrA8
ERX556019.13.dc, dfrL
GVMAG-S-1101173-83, dfrA36
1
10
20
30
40
50
60
70
80
90
100
110
120
* Trimethoprim binding sites

Supplementary Fig. 4 | Sequence alignment of the putative dihydrofolate reductase proteins in NCLDVs and reference sequences in bacteria. The binding residues of trimethoprim (indicated by purple stars below the alignments) and the accessions of DFR sequences from bacteria were obtained from Coque *et al*. [2]. NCLDV sequence labels are marked blue.



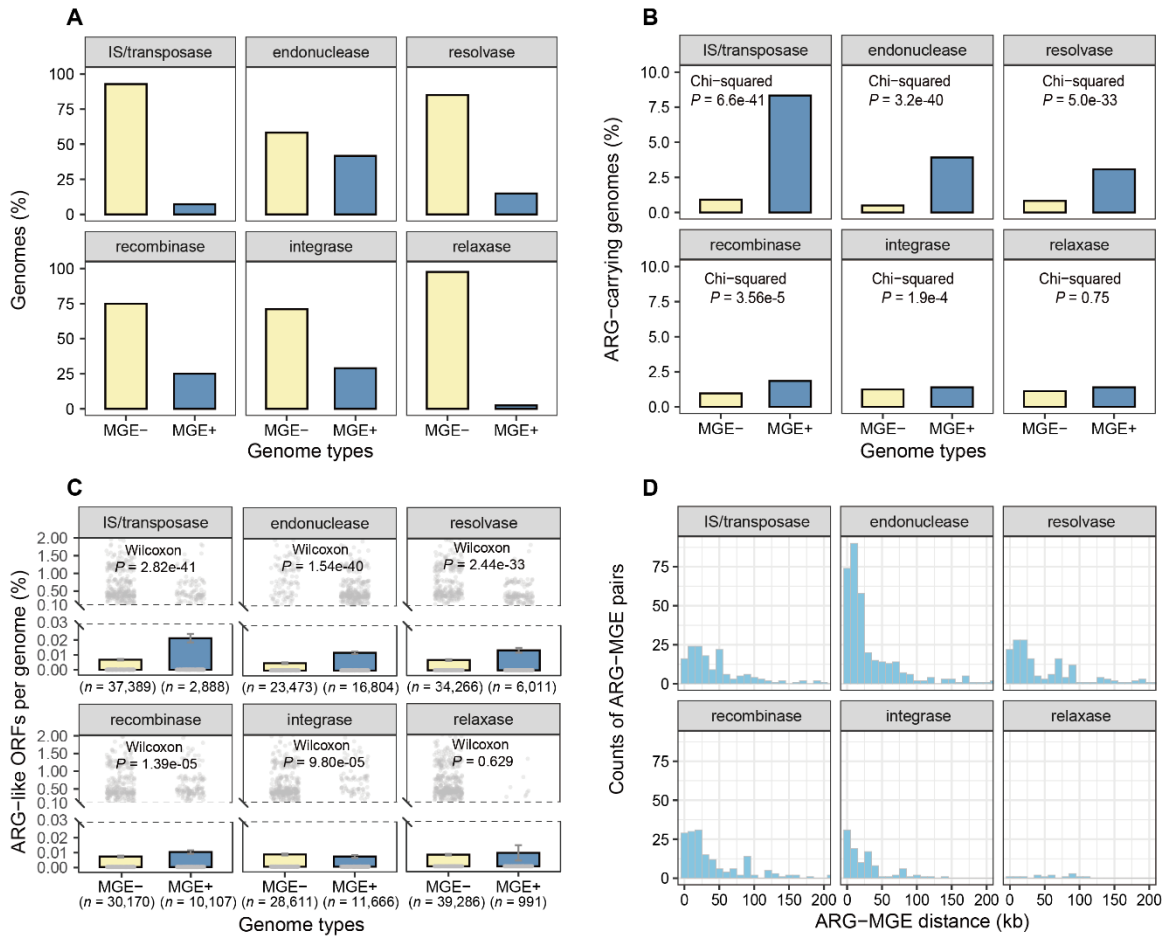
Supplementary Fig. 5 | Mobile genetic elements (MGEs) carried by NCLDVs and phages from different habitats. A, C: Possibility of MGE-carriage of A) NCLDVs and C) phages in different habitats. **B, D**: Genomic potential of ARG-carriage of B) NCLDVs and D) phages in different habitats. Lower-case letters above the bars represent significantly different groups assessed with two-sided Wilcoxon rank-sum test, and *P*-values indicate the overall difference among all families assessed with Kruskal–Wallis test. Data presented in C) and F) were mean values ± SEM. Each data point represents an individual genome in the corresponding group. The number (*n*) of genomes in each group can be found in the Source Data file on Github [1]. For clarity, only seven major habitats are

shown individually, with other habitats being referred to as “Other”. Only habitats with at least three genomes are shown.



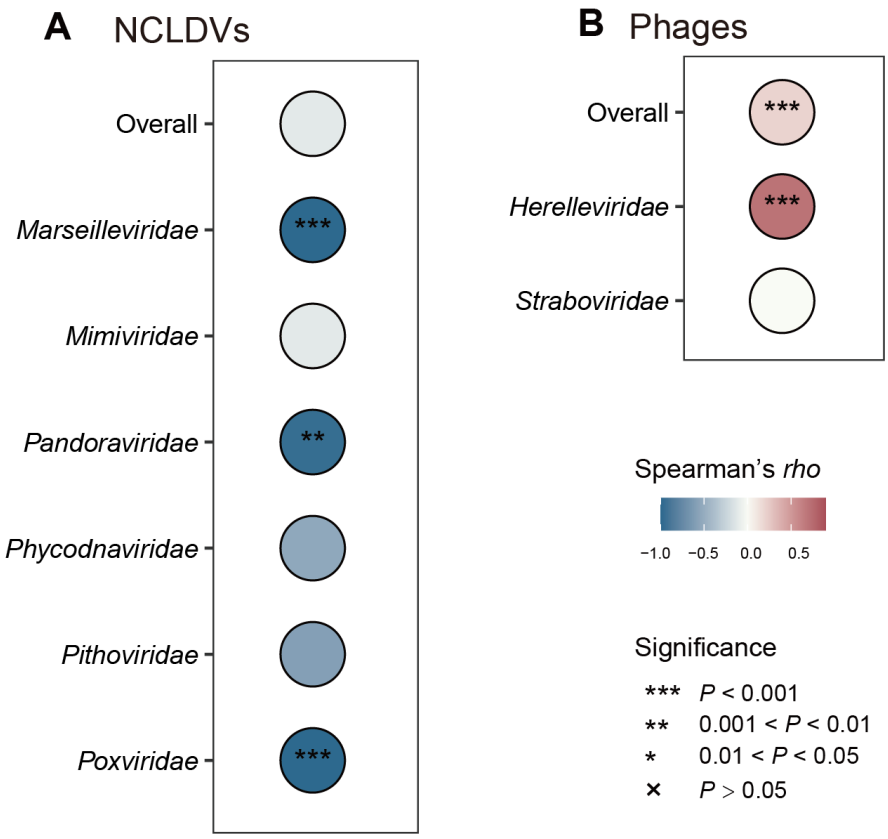
Supplementary Fig. 6 | Dependency between ARG-carriage and the carriage of different MGE types in the NCLDV genomes. A: Overall incidence of individual MGE types. **B:** Possibility of ARG-carriage in genomes with or without a given MGE type. Two-sided Chi-squared test was performed to test dependency between carriage of a given MGE type and ARG-carriage. **C:** Genomic potential of ARG-carriage in genomes with or without a given MGE type. Two-sided Wilcoxon rank-sum test was performed to evaluate the statistical significance of the differences in ARG content between genomes with or without a given MGE type. MGE+: genomes with at least one MGE belonging to a specific MGE type, MGE-: genomes without any MGE belonging to a specific MGE type. Data are presented as mean values ± SEM. Each data point is one genome. **D:** Histogram of distance

(kb) between ARGs and different types of MGEs. Source data are provided as a Source Data file on Github [1].



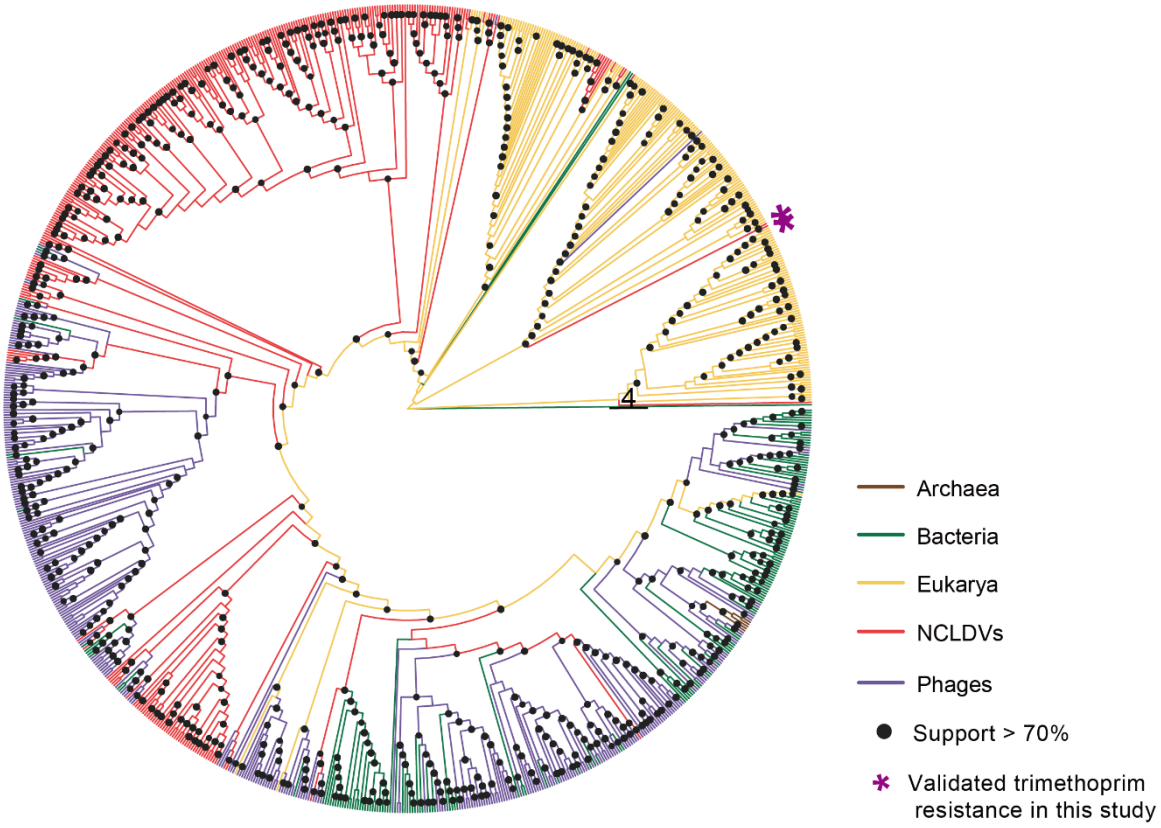
Supplementary Fig. 7 | Dependency between ARG-carriage and the carriage of different MGE types in the phage genomes. A: Overall incidence of individual MGE types. **B:** Possibility of ARG-carriage in genomes with or without a MGE type. Two-sided Chi-squared test was performed to test dependency between carriage of a MGE type and ARG-carriage. **C:** Genomic potential of ARG-carriage in genomes with or without a MGE type. Two-sided Wilcoxon rank-sum test was performed to evaluate the statistical significance of the differences in ARG content between genomes with or without a MGE type. MGE+: genomes with at least one MGE belonging to a specific MGE type, MGE-: genomes without any MGE belonging to a specific MGE type. Data are presented as mean values ± SEM. Each data point is one genome. **D:** Histogram of distance (kb) between

ARGs and different types of MGEs. Source data are provided as a Source Data file on Github [1].

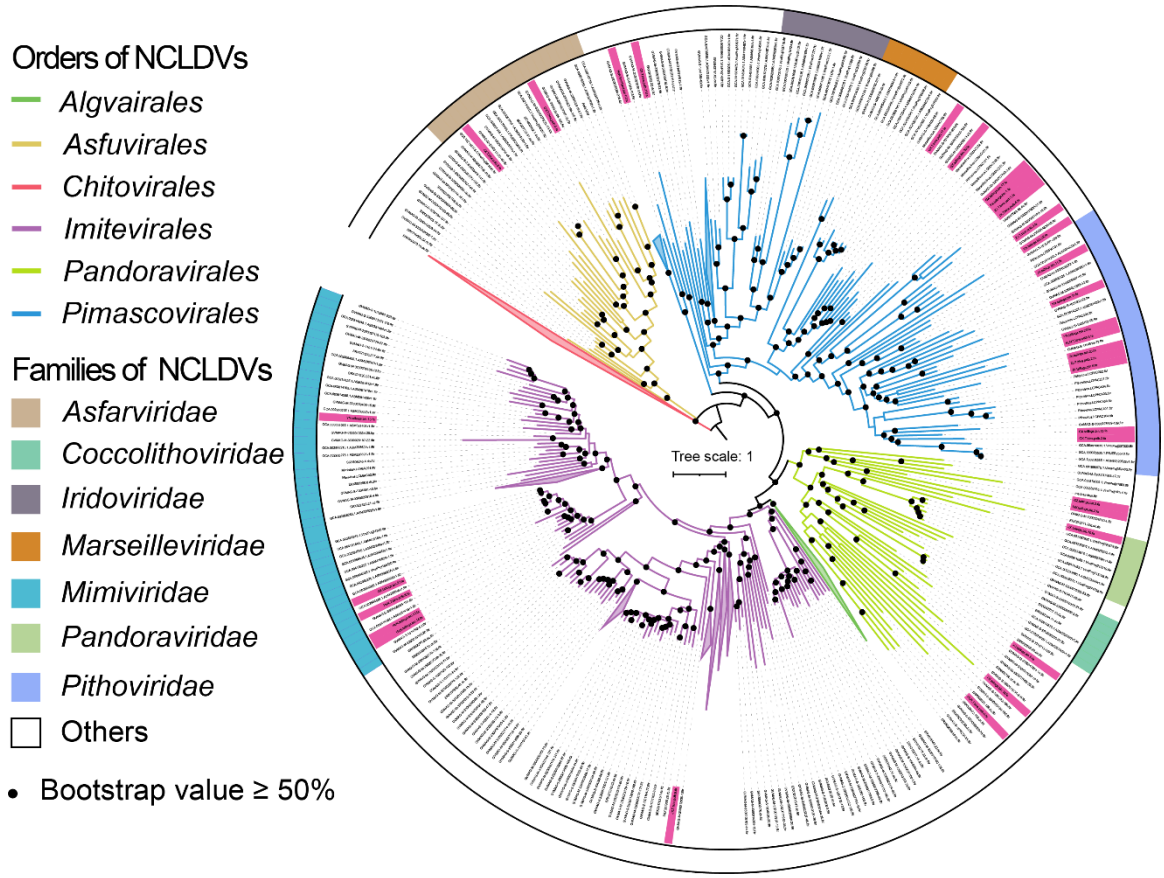


Supplementary Fig. 8 | Correlations between genomic potential of ARG carriage of isolated viruses and their genome size. Spearman coefficients are shown as heatmap in **A)** for isolated NCLDVs and **B)** for isolated phages. Colors represent Spearman's *rho*. Stars and cross symbols indicate significance of the Spearman correlation, which are FDR-adjusted for multiple comparisons. Only viral families with more than three ARG-carrying genomes and the overall patterns of NCLDVs and phages are displayed. Exact *n* and *P*-values for individual groups can be found in the Source Data file on Github [1].

Figure S9



Supplementary Fig. 9 | Phylogenetic tree of dihydrofolate reductases from NCLDVs, eukaryotes, prokaryotes, and phages. This tree is the same as that in Figure 3A in the main text, with trimethoprim sequences from NCLDVs functionally validated in this study marked with stars.



Supplementary Fig. 10 | Maximum-likelihood phylogenetic tree of NCLDVs. Reference genome sequences were adopted from a previous publication [3] (https://faylward.github.io/GVDB). The tree was constructed using concatenated alignments of seven marker genes SFII, RNAPL, PolB, TFIIB, TopoII, A32 and VLTF3, employing the LG+I+F+G4 model and rooted with *Poxviridae*. Branches in different colors represent taxonomic information of the giant viruses at the order level. The giant virus genomes assembled in this study are highlighted in pink font. The outermost circle represents the taxonomic information of the giant viruses at the family level. The marker genes comprise: SFII (DEAD/SNF2-like helicase), RNAPL (DNA-directed RNA polymerase alpha subunit), PolB (DNA polymerase family B), TFIIB (transcription

initiation factor IIB), TopoII (DNA topoisomerase II), A32 (Packaging ATPase), and VLTF3 (Poxvirus late transcription factor VLTF3). For better presentation, branches not involving any genomes in this study are folded.

Supplementary references

1 Yi, X., *et al*. Giant viruses as reservoirs of antibiotic resistance genes. *Github.* https://doi.org/10.5281/zenodo.13234118 (2024).

2 Coque, T. M., Singh, K. V., Weinstock, G. M. & Murray, B. E. Characterization of dihydrofolate reductase genes from trimethoprim-susceptible and trimethoprim-resistant strains of *Enterococcus faecalis*. *Antimicrob. Agents Chemother.* **43**, 141-147 (1999).

3 Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. V. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol.* **19**, e3001430 (2021).