**Supporting Information for**

**Virus-pathogen interactions make water-quality safer in the world’s largest water diversion canal**

Tianyi Chen, Tang Liu, Zongzhi Wu, Bingxue Wang, Qian Chen, Mi Zhang, Enhang Liang, Jinren Ni\*

\*Jinren Ni, College of Environmental Sciences and Engineering, Peking University, Beijing 100871, P. R.China

**E-mail:** jinrenni@pku.edu.cn (J.R. Ni)

Supplemental Methods

## *Comparisons of viral sequences in the MR-SNWDC and other freshwater ecosystems*

## Viral contigs with over 90% completeness were selected from the freshwater sources in the IMG/VR database [1], for subsequent viral clustering analysis with vOTUs in the MR-SNWDC. Each reported viral sequence was assigned to a specific ecosystem subtype (lake, lentic, groundwater, sediment, wetlands, river, ice, creek, lotic, pond, or drinking water). The protein sequences retrieved from Prodigal v2.6.3 [2] were used for gene-sharing network analysis through vConTACT2 v0.9.19 [3]. Diamond [4] was applied to estimate the protein-protein similarity. Protein clusters were calculated by the Markov Cluster Algorithm (MCL), with the subsequent VC generation using ClusterONE [5].

## *Construction of virus-host infection network*

## The strain-level virus-host linkages were characterized using the BiMat MatLab package [6]. Modularity quantification was conducted using leading eigenvector method, with a subsequent Kernighan-Lin tuning step applied to modularity adjustment [7]. Overlap and decreasing fill (NODF) was used for nestedness quantification. All statistical significances were performed based on 1000 random permutations using the equiprobable method. Modularity parameters included Qb 0.8491, mean 0.5511, std 0.0030, z-score 99.5177, t-score 314.7025, and percentile 100. The fraction inside module parameters corresponded to Qr 0.8254, mean 0.1834, std 0.0110, z-score 58.1052, t-score 183.7449, and percentile 100. Nestedness parameters covered nestedness value 0.0132, mean 0.0037, std 0, z-score 375.0087, t-score 3750.0874, and percentile 100.

## *Identification of auxiliary metabolic genes*

## Removal of host contamination and identification of prophage boundaries were performed via CheckV [8] in advance. The predicted protein sequences, generated by Prodigal [2], were aligned to the eggNOG database [9] using emapper.py v1.0.3 [10] (-m diamond; --seed\_orthology\_evalue 1e-5). Each protein was assigned a COG annotation. AMG identification was first conducted by VIBRANT [11] according to KEGG, Pfam and VOG databases. The genes with annotations “metabolic pathways” and “sulfur relay system” were regarded as putative AMGs. VirSorter2 [12] provided the information of virus-associated and viral hallmark genes within contigs, and generated annotation files for DRAM-v [13] to perform the parallel AMG identification. The genes with M/F flag assignments and auxiliary scores of ≤ 3 were regarded as putative AMGs. In order to avoid false positive results, only the AMGs located between two virus-associated or viral hallmark genes and those located alongside the viral-associated or viral hallmark genes were selected for further analysis [14]. Phyre2 [15] was applied to identify tertiary protein structures with confidence > 90% and coverage > 70%. PROSITE [16] was used to analyze conserved regions and active sites of putative AMGs based on PROSITE collection of motifs. Genome maps for AMG-containing viral contigs were visualized based on COG, VIBRANT, VirSorter2 and DRAM-v annotations.

Supplemental Results

## *Phosphorus dynamics in the MR-SNWDC*

## Over seven years (2015-2021) of water quality monitoring in the MR-SWNDC provided evidence of long-lasting limited total phosphorus (TP) concentration below the national standard threshold of Ⅰ class surface water (≤ 0.02 mg/L), with the annual average TP content ND~0.01 mg/L in the last three years (Fig. S1a). Importantly, a rapid decay of annual average TP content showed presence along the main canal, decreasing from 0.0112 mg/L in the upstream Henan province to 0.0065 mg/L in downstream Beijing and Tianjin municipals (Fig. S1b)

## *Relationship between MR-SNWDC vOTUs and publicly reported viral sequences in the IMG/VR database*

## Gene-sharing network analysis was performed to evaluate the relationship between 40,261 vOTUs in the MR-SNWDC and 37,364 viral sequences (>90% completeness) from a broader diversity of freshwater ecosystems in the IMG/VR database [1]. Around half of vOTUs in the canal were assigned to 7,389 viral clusters (VCs) at the genus level, with 68.2% VCs not including viruses from any other ecosystems in the IMG/VR database (Fig. S6a). Only 9.1% of identified vOTUs were clustered with publicly reported viruses, which contributed to the expansion of existing virosphere and proved the MR-SNWDC as an endemic pool of diverse and novel freshwater viruses. Among 3,670 vOTUs which shared VCs with publicly available viruses, over 85% were clustered with viral sequences from the lake source. In addition, about one thirds of lake-derived viral genera were clustered with MR-SNWDC vOTUs, ranking the most among all freshwater sources (Fig. S6b), which highlighted the role of Danjiangkou Reservoir (lake-like) in shaping the viral communities across the canal.

Supplemental Figures



Fig. S1 Long-term dynamics of total phosphorus concentration in the MR-SNWDC. a Monthly and annual average TP concentrations across 7 years from 2015 to 2021. The gray area displays standard deviation of monthly average TP content. b Changes in annually TP concentration along the main-canal. The error bar corresponds to the 95% confidence interval.



**Fig. S2 Sketch map of the MR-SNWDC.** Sampling sites are distributed at 32 monitoring stations along the water canal (see Table S1). The length of the canal (1,432 km) is measured by the sum of the dendritic distances of each two sampling sites from upstream to downstream, as an indication of canal network density, rather than the straight-line distance between the water source area and the canal end. Sampling campaigns are carried out at the same sites in August 2020 and March 2021, respectively.



**Fig. S3 Regional similarity of viral communities in autumn (a) and spring (b).** Sorenson similarity is calculated for relative abundances of vOTUs. The width of each curve represents the similarity value between paired regions. Source data are provided in the Source Data file.



**Fig. S4 Spatiotemporal distribution of bacterial communities in autumn and spring.** Non-metric multidimensional scaling (NMDS) analyses visualize the seasonal shift of bacterial β-diversity (**a**) as well as the distinct partition of bacterial communities into four ecological regions in autumn (**b**) and spring (**c**), based on the Bray-Curtis dissimilarity matrix calculated from relative abundances of prokaryotic MAGs. The stress value denotes the ordination fitness of each NMDS plot. Each group is encircled by an ellipse at 95% confidence interval. One outlier sample (06A) is excluded from subsequent analyses. **d** The richness of observed bacterial species transported from the water source area (Reg1) to downstream regions (Reg2-4) in autumn (Upper panel) and spring (Lower panel). Source data are provided in the Source Data file.



**Fig. S5 Molecular properties of viral genomes in the MR-SNWDC and the IMG/VR database.** Differences in GC content (**a**) and specific amino acid frequencies (**b**) are estimated by Bonferroni-adjusted Wilcoxon test. The statistical significance is marked by asterisks (\*\*\*\*: ≤ 0.0001). Source data are provided in the Source Data file.



**Fig. S6 Comparison of viral species in the MR-SNWDC and the IMG/VR database.** **a** Shared viral clusters (VCs) among different datasets. Viral sequences from 11 freshwater ecosystems are selected from the IMG/VR database. Each source of VCs is defined as a set. The bars on the left represent the total number of VCs in each set. Dots with interconnecting vertical black lines represent the intersections, where black dots represent sets that were within the intersection and unﬁlled light gray dots represent sets that were not part of the intersection. The bars on the top right represent the number of VCs within the intersection. **b** Proportional number of viral genera from diverse freshwater sources in the IMG/VR database which are clustered with vOTUs in the MR-SNWDC. Source data are provided in the Source Data file.



**Fig. S7 Changes in averaged copies of bacteria-encoded genes involved in key P-associated metabolism processes along the canal.** The Pearson correlation coefficient and significant level of *p*-value are presented for each linear regression (\*\*\*\*: ≤ 0.0001). Source data are provided in the Source Data file.



**Fig. S8 Relative abundance patterns of viruses and their predicted hosts in the MR-SNWDC.** **a** Chord Diagram showing the taxonomically virus-host linkages. The thickness of each ribbon represents the relative abundance of viruses. **b** Spearman correlation between abundances of viruses and their hosts (calculated by logarithmic transformation of normalized mean coverage depth, reads per kilobase mapped reads: RPKM). **c** Virus/host abundance ratios (VHR) for all bacterial phyla. All mentioned viral families and host phyla are denoted in the right legend. Source data are provided in the Source Data file.



**Fig. S9 The nested-modular structure of virus-host interaction networks.** Strain-level prokaryotic hosts and viruses are represented as rows and columns, respectively. Allvirus-host linkages are reorganized and colored according to BiMat leading eigenvector modules. Four largest modules are highlighted as examples. Source data are provided in the Source Data file.



**Fig. S10 Genomic context and protein structure of selected virus-encoded AMGs.** **a** Genome map of representative AMG-encoding viral contig. Each contig is marked by its genome length. **b** Tertiary structures of selected AMGs based on structural modelling using Phyre2.rpiB: ribose 5-phosphate isomerase B; pstS: phosphate transport system substrate-binding protein; purA: adenylosuccinate synthase; dcd: dCTP deaminase; DNMT: DNA (cytosine-5)-methyltransferase 1; galE: UDP-glucose 4-epimerase; DUT: dUTP pyrophosphatase; thyA: thymidylate synthase. Source data are provided in the Source Data file.



**Fig. S11 Schematic description of the natural phage therapy in the MR-SNWDC analogous to the target phage therapy used in human body.** Highly specific infections of viruses are expected to facilitate precise treatments of pathogen-induced human diseases in clinical practice. The viral predation effects in the MR-SNWDC serve as a natural phage therapy to eliminate the waterborne pathogens.

Supplemental Tables

**Table S1.** Sequencing depth for each of the 64 samples in autumn and spring.

**Table S2.** The number and averaged length of vOTUs within different quality levels in autumn and spring.

**Table S3.** Permutational multivariate analysis of variance (PERMANOVA) for statistical significances of viral and bacterial communities spatiotemporally.

**Table S4.** Summary of virus-encoded auxiliary metabolism genes (AMGs) identified in the MR-SNWDC.

Supplemental References

1. Roux S, Páez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses*.* Nucleic Acids Res. 2020;49:D764-D775.

2. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification*.* BMC Bioinformatics. 2010;11:119.

3. Jang HB, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks*.* Nat Biotechnol. 2019;37:632-639.

4. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND*.* Nat Methods. 2015;12:59-60.

5. Nepusz T, Yu HY, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks*.* Nat Methods. 2012;9:471-472.

6. Flores CO, Poisot T, Valverde S, Weitz JS. BiMat: a MATLAB package to facilitate the analysis of bipartite networks*.* Methods Ecol Evol. 2016;7:127-132.

7. Kauffman KM, Chang WK, Brown JM, Hussain FA, Yang J, Polz MF, et al. Resolving the structure of phage–bacteria interactions in the context of natural diversity*.* Nat Commun. 2022;13:372.

8. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes*.* Nat Biotechnol. 2021;39:578-585.

9. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses*.* Nucleic Acids Res. 2018;47:D309-D314.

10. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper*.* Mol Biol Evol. 2017;34:2115-2122.

11. Kieft K, Zhou ZC, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences*.* Microbiome. 2020;8:90.

12. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses*.* Microbiome. 2021;9:37.

13. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa Sabina L, Solden LM, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function*.* Nucleic Acids Res. 2020;48:8883-8900.

14. Pratama AA, Bolduc B, Zayed AA, Zhong ZP, Guo JR, Vik DR, et al. Expanding standards in viromics: in silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation*.* PeerJ. 2021;9:e11447.

15. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis*.* Nat Protoc. 2015;10:845-858.

16. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE*.* Nucleic Acids Res. 2012;41:D344-D347.