

1 **Supporting Information for**

2 **Virus–pathogen interactions improve water quality along the**
3 **Middle Route of the South-to-North Water Diversion Canal**

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

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

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

9

10 **Supplemental Methods**

11 *Construction of virus–host infection network*

12 The virus–host linkages were visualized by ForceAtlas 2 algorithm in Gephi
13 v0.9.2 [1]. Modularity quantification was conducted using community detection
14 method [2]. The feature set of the virus–host infection network included average
15 degree and modularity index, which represented the number of adjacent edges
16 and the connectivity of community members [3, 4].

17

18 *Identification of auxiliary metabolic genes*

19 Removal of host contamination and identification of prophage boundaries were
20 performed via CheckV [5] in advance. The predicted protein sequences,
21 generated by Prodigal [6], were aligned to the eggNOG database [7] using
22 emapper.py v1.0.3 [8] (-m diamond; --seed_orthology_evalue 1e⁻⁵). Each
23 protein was assigned a COG annotation. AMG identification was first conducted
24 by VIBRANT [9] according to KEGG, Pfam, and VOG databases. The genes
25 with annotations “metabolic pathways” and “sulfur relay system” were regarded
26 as putative AMGs. VirSorter2 [10] provided the information of virus-associated
27 and viral hallmark genes within contigs, and generated annotation files for

28 DRAM-v [11] to perform the parallel AMG identification. The genes with M/F
29 flag assignments and auxiliary scores of ≤ 3 were regarded as putative AMGs.
30 In order to avoid false positive results, only the AMGs located between two
31 virus-associated or viral hallmark genes and those located alongside the viral-
32 associated or viral hallmark genes were selected for further analysis [12].
33 Phyre2 [13] was applied to identify tertiary protein structures with confidence >
34 90% and coverage > 70%. PROSITE [14] was used to analyze conserved
35 regions and active sites of putative AMGs based on PROSITE collection of
36 motifs. Genome maps for AMG-containing viral contigs were visualized based
37 on COG, VIBRANT, VirSorter2, and DRAM-v annotations.

38

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

41 Viral contigs with over 90% completeness were selected from the freshwater
42 sources in the IMG/VR database [15], for subsequent viral clustering analysis
43 with vOTUs in the MR-SNWDC. Each reported viral sequence was assigned to
44 a specific ecosystem subtype (lake, lentic, groundwater, sediment, wetlands,
45 river, ice, creek, lotic, pond, and drinking water). The protein sequences
46 retrieved from Prodigal v2.6.3 [6] were used for gene-sharing network analysis
47 through vConTACT2 v0.9.19 [16]. Diamond [17] was applied to estimate the
48 protein–protein similarity. Protein clusters were calculated by the Markov
49 Cluster Algorithm (MCL), with the subsequent VC generation using ClusterONE
50 [18].

51

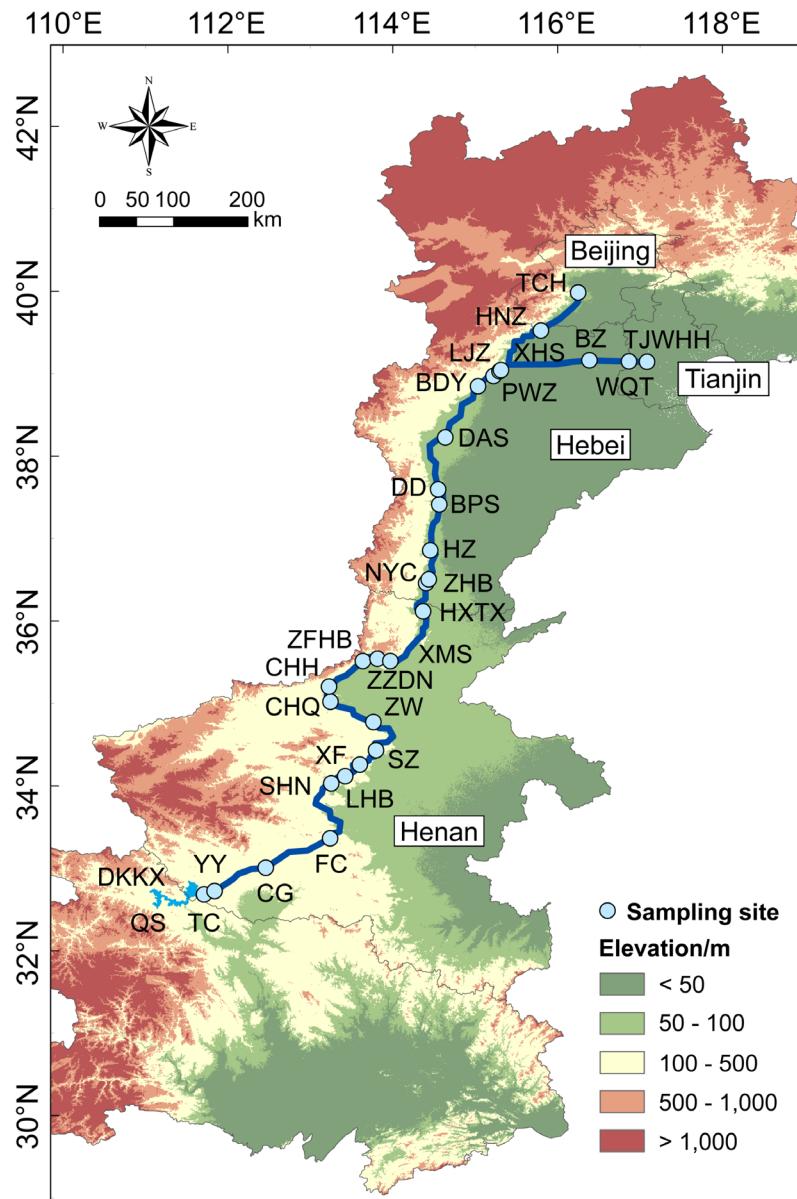
52

53 **Supplemental Results**

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

56 Gene-sharing network analysis was performed to evaluate the relationship

57 between 40,261 vOTUs in the MR-SNWDC and 37,364 viral sequences (>90%
58 completeness) from a broader diversity of freshwater ecosystems in the
59 IMG/VR database [15]. Around half of vOTUs in the MR-SNWDC were
60 assigned to 7,389 viral clusters (VCs) at the genus level, with 68.2% VCs not
61 including viruses from any other ecosystems in the IMG/VR database (Fig.
62 S10A). Only 9.1% of identified vOTUs were clustered with publicly reported
63 viruses, suggesting that the MR-SNWDC was an endemic pool of diverse and
64 novel freshwater viruses. Among 3,670 vOTUs which shared VCs with publicly
65 available viruses, over 85% were clustered with viral sequences from the lake
66 source. In addition, about one thirds of lake-derived viral genera were clustered
67 with vOTUs in the MR-SNWDC, ranking the most among all freshwater sources
68 (Fig. S10B), which highlighted the role of Danjiangkou Reservoir (lake-like) in
69 shaping the viral communities across the canal.

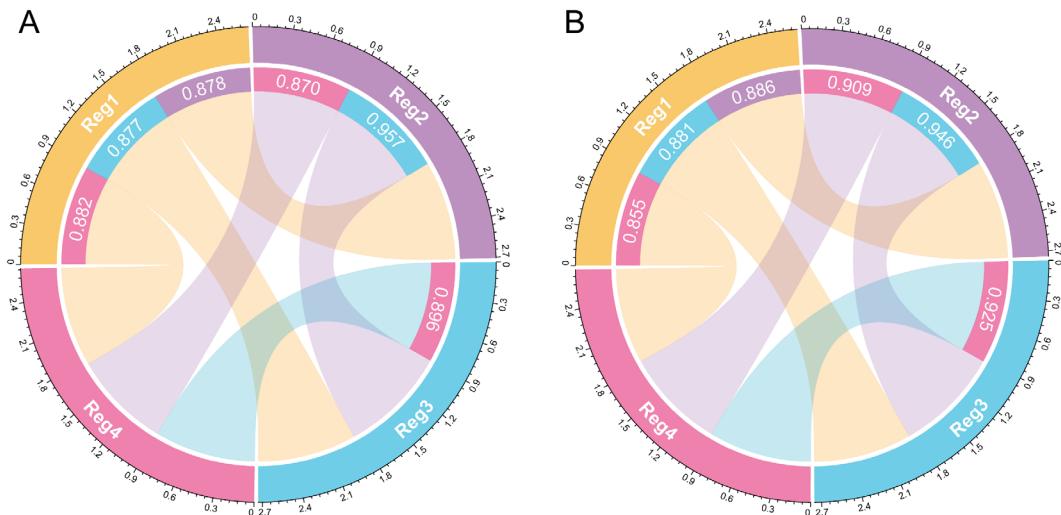
70 **Supplemental Figures**

71

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

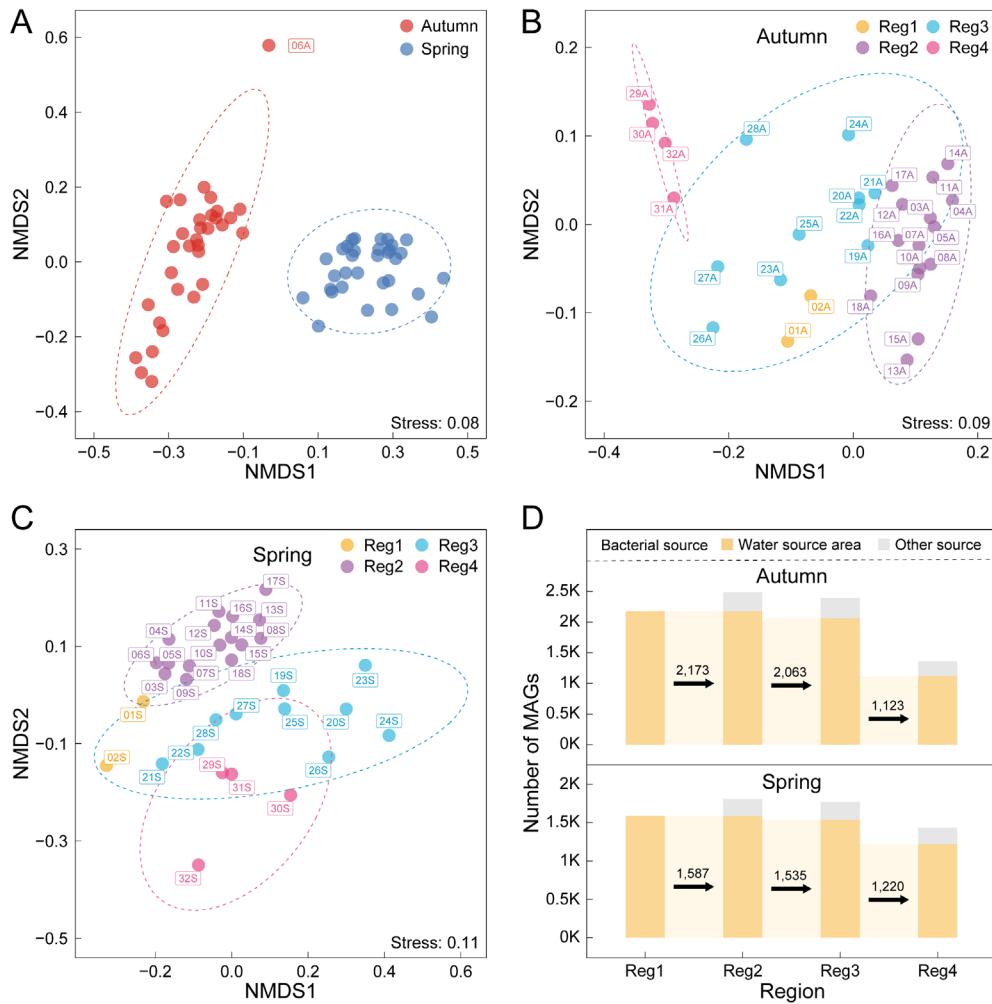
79

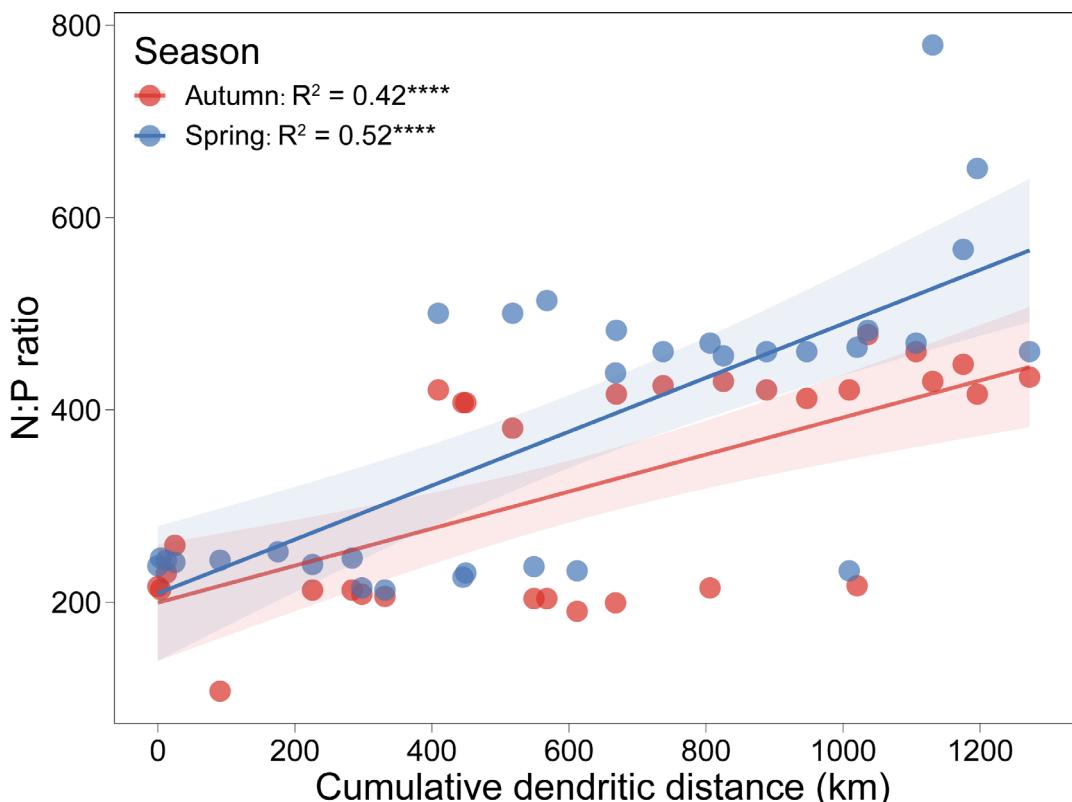
80 **Fig. S2 Regional similarity of viral communities in autumn (A) and spring**
81 **(B).** Sorenson similarity is calculated for the relative abundances of vOTUs. The
82 width of each curve represents the similarity value between the paired regions.
83 Source data are provided in the Source Data file.



84

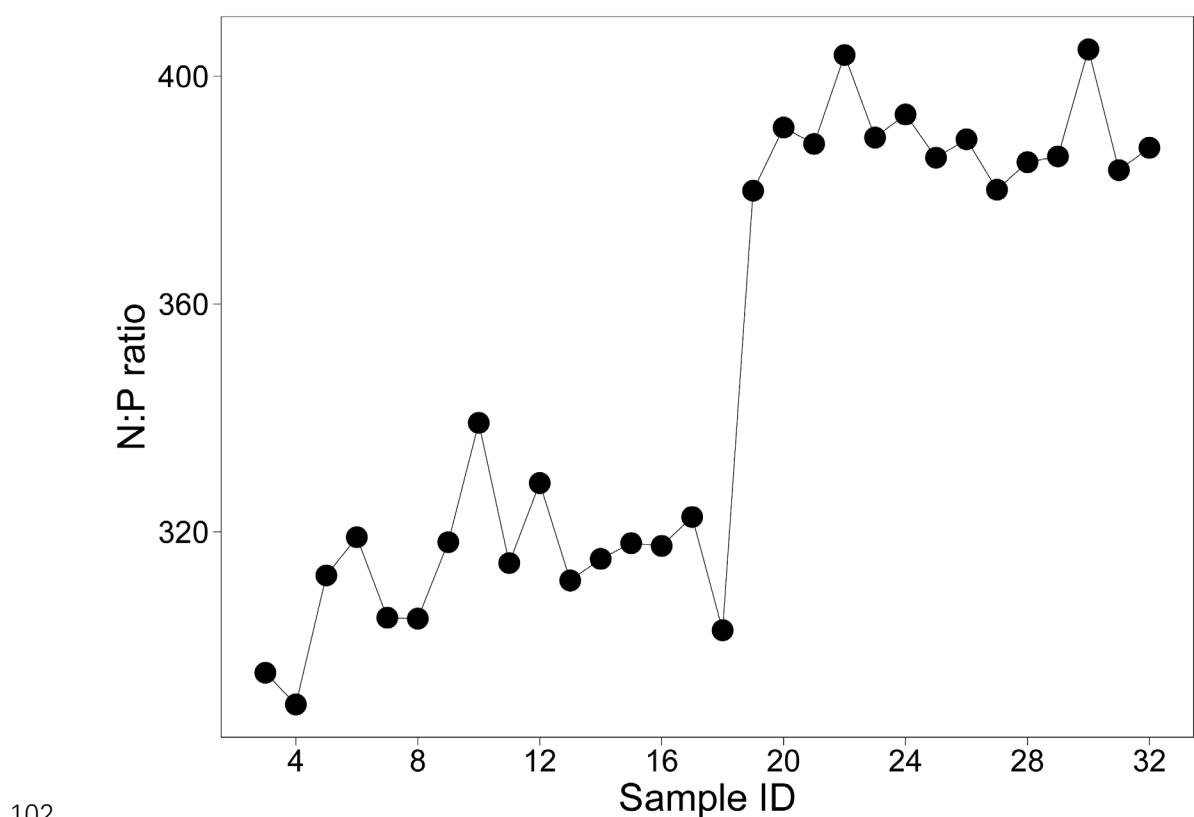
85 **Fig. S3 Spatiotemporal distribution of bacterial communities in autumn**
 86 **and spring.** Nonmetric multidimensional scaling (NMDS) analyses visualize
 87 the temporal variation of bacterial β -diversity (**A**) as well as the distinct partition
 88 of bacterial communities into four ecological regions in autumn (**B**) and spring
 89 (**C**), based on the Bray–Curtis dissimilarity matrix calculated from the relative
 90 abundances of prokaryotic MAGs. The stress value denotes the ordination
 91 fitness of each NMDS plot. Each group is encircled by an ellipse at 95%
 92 confidence interval. One outlier sample (06A) is excluded from subsequent
 93 analyses. **D** The richness of observed bacterial species transported from the
 94 water source area (Reg 1) to downstream regions (Reg 2~4) in autumn (upper
 95 panel) and spring (lower panel). Source data are provided in the Source Data
 96 file.



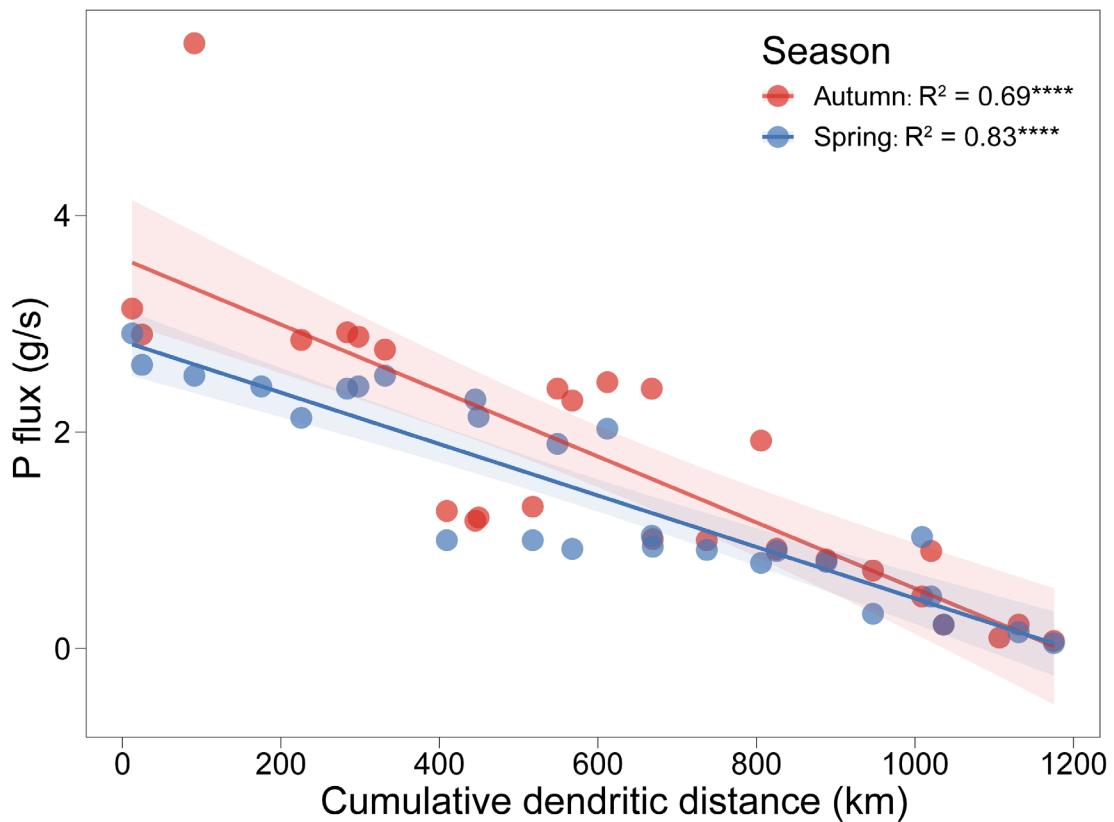


97

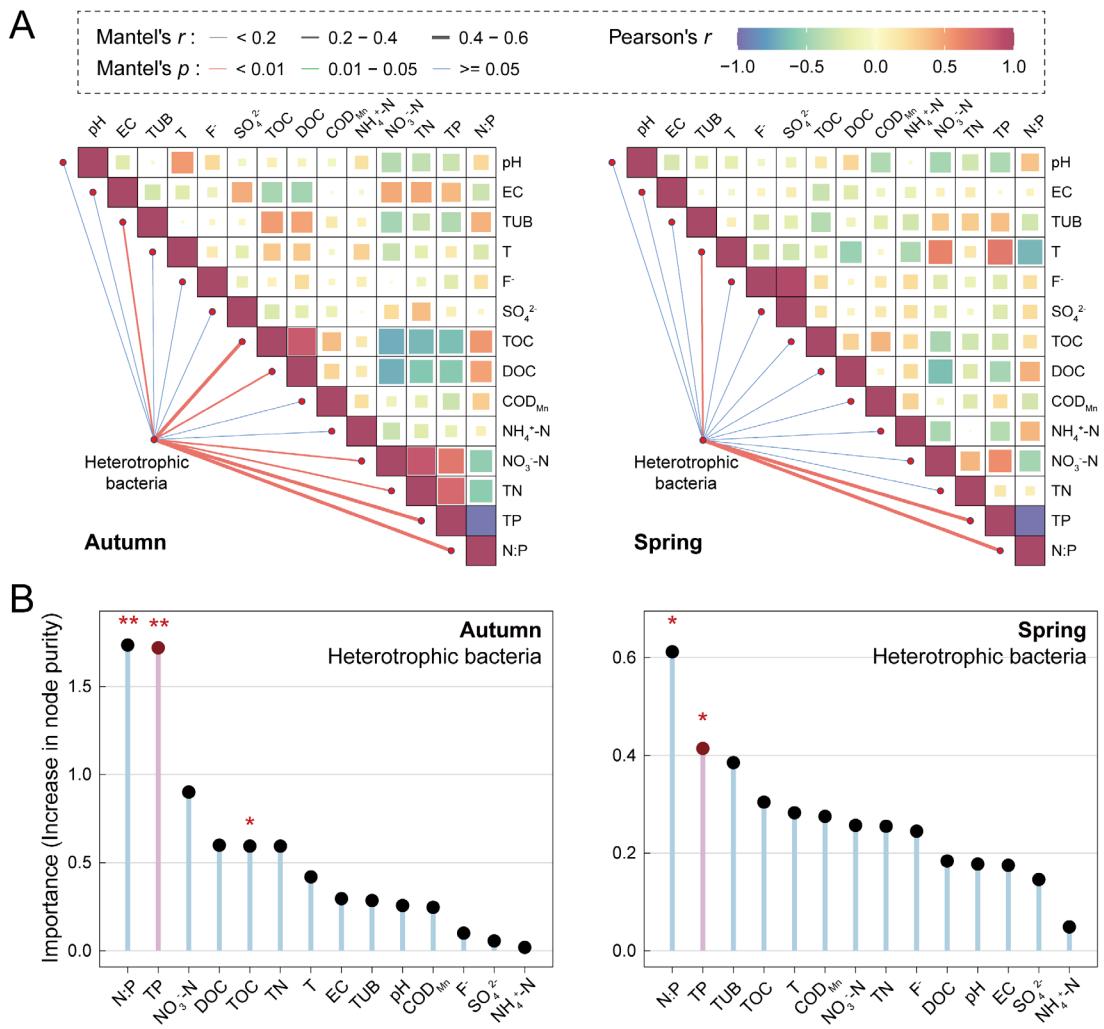
98 **Fig. S4 Changes in the N:P ratio (molar) along the canal in autumn and**
 99 **spring.** The goodness-of-fit R^2 value and the statistical significance are
 100 presented for each linear regression (***: < 0.0001). Source data are provided
 101 in the Source Data file.



102
103 **Fig. S5 Changes in the N:P ratio (molar) along the main canal during
104 2015~2021.** Source data are provided in the Source Data file.

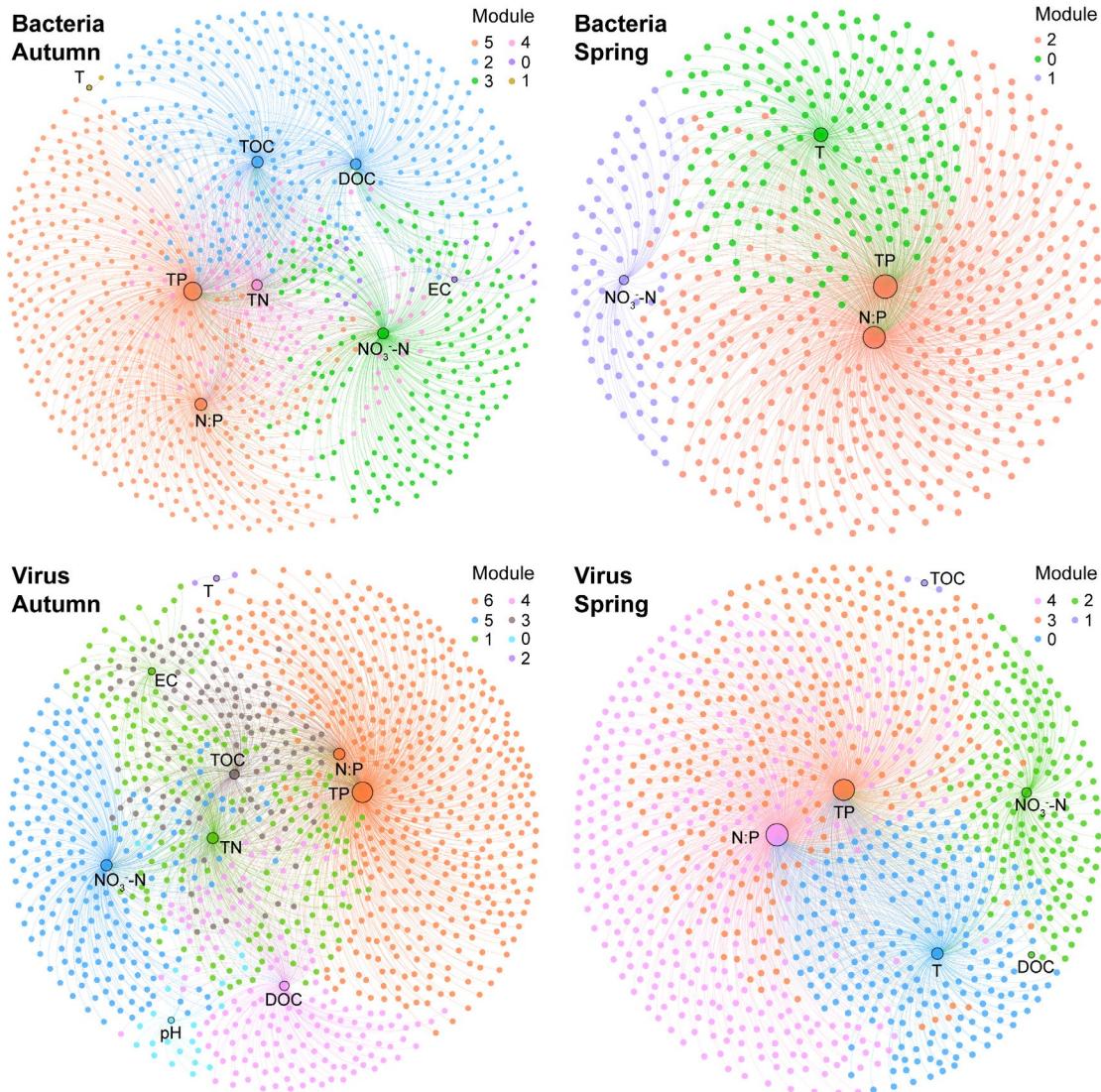


105
106 **Fig. S6 Changes in P flux along the canal in autumn and spring**
107 **(concentration × flow rate, g/s).** Each linear regression is denoted by the
108 goodness-of-fit R^2 value and the significance level of p value ($^{****}: < 0.0001$).
109 Source data are provided in the Source Data file.



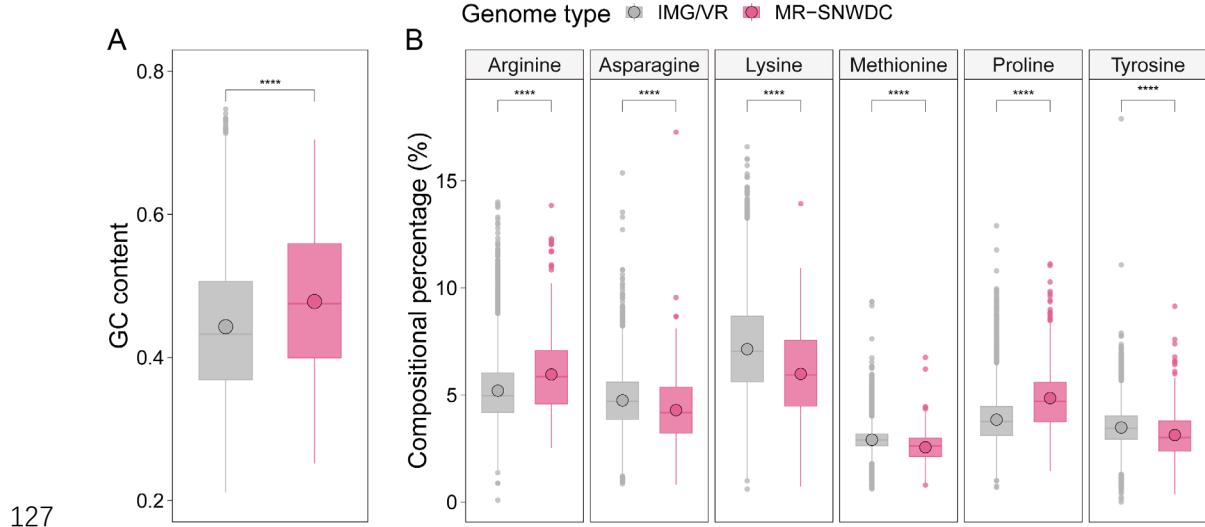
110

111 **Fig. S7 Relevance of environmental factors and heterotrophic bacterial
112 communities. A** Correlation between environmental factors and heterotrophic
113 bacterial communities in autumn and spring. Pairwise Pearson's coefficients
114 are denoted by color gradients. Edge width demonstrates the Mantel's r
115 correlation coefficients. Edge color represents the significance level of p value
116 based on 999 permutations. **B** Random forest importance of each
117 environmental factor for heterotrophic bacterial communities in two seasons.
118 All environmental factors are brought into a ranking by their importance index
119 represented by the increase in node purity. The significance of each
120 environmental factor is shown in asterisks (**: < 0.01; *: < 0.05). Source data
121 are provided in the Source Data file.



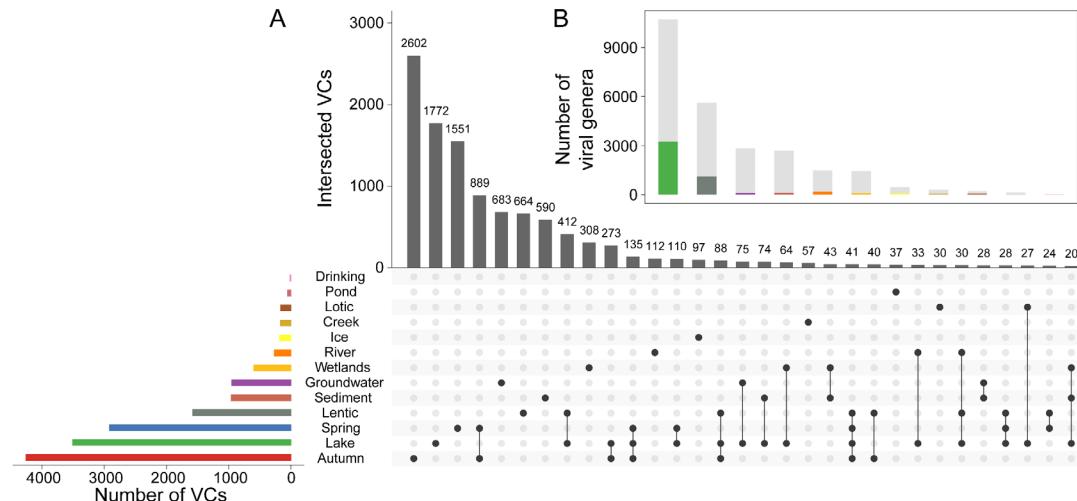
122

123 **Fig. S8 Co-occurrence network of environmental factors and**
 124 **MAGs/vOTUs.** The size of each dot marking environmental factors is
 125 proportional to the number of connections. Dots with different colors denote
 126 different modules in networks. Source data are provided in the Source Data file.

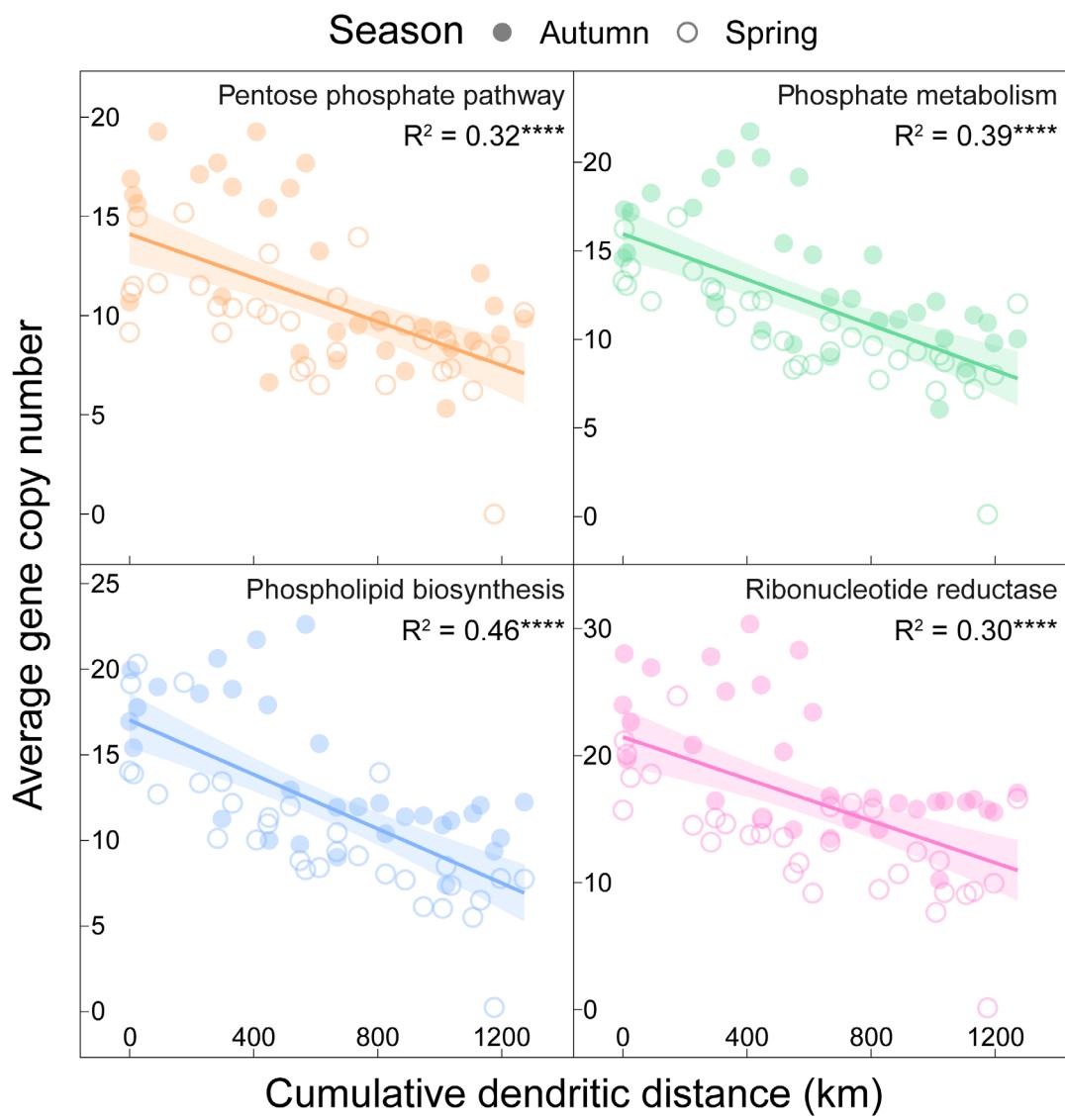


128 **Fig. S9 Molecular properties of viral genomes in the MR-SNWDC and the**
 129 **IMG/VR database.** Differences in GC content (**A**) and specific amino acid
 130 frequencies (**B**) are estimated by Bonferroni-adjusted Wilcoxon test. The
 131 statistical significance is marked by asterisks (****: ≤ 0.0001). Source data are
 132 provided in the Source Data file.

133

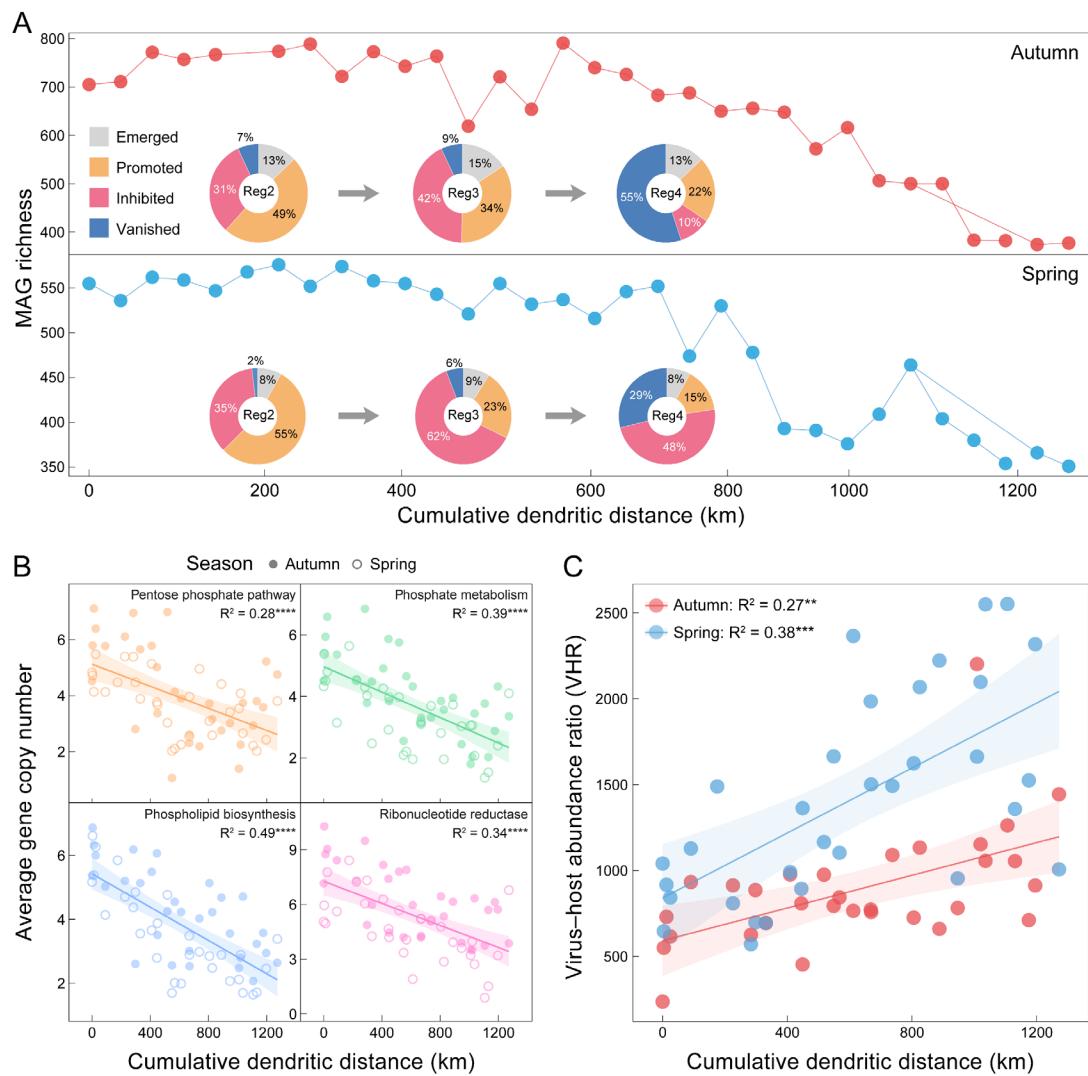


134 **Fig. S10 Comparison of viral species in the MR-SNWDC and the IMG/VR**
135 **database. A** Shared viral clusters (VCs) among different datasets. Viral
136 sequences from 11 freshwater ecosystems are selected from the IMG/VR
137 database. Each source of VCs is defined as a set. The bars on the left represent
138 the total number of VCs in each set. Dots with interconnecting vertical black
139 lines represent the intersections, where black dots represent sets that were
140 within the intersection and unfilled light gray dots represent sets that were not
141 part of the intersection. The bars on the top right represent the number of VCs
142 within the intersection. **B** Proportional number of viral genera from diverse
143 freshwater sources in the IMG/VR database which are clustered with vOTUs in
144 the MR-SNWDC. Source data are provided in the Source Data file.



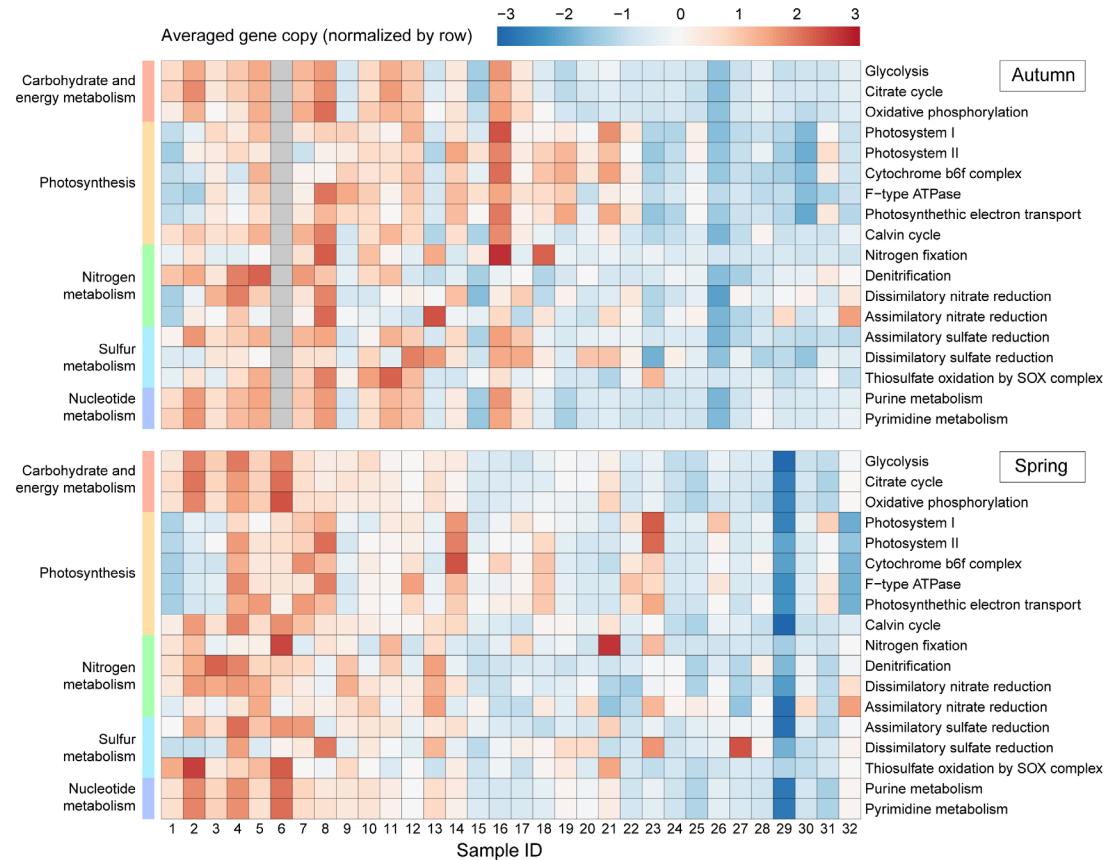
145

146 **Fig. S11 Changes in average copy number of bacteria-encoded genes**
 147 **involved in key P-associated metabolic processes along the canal.** The
 148 goodness-of-fit R^2 value and the significance level of p value are presented for
 149 each linear regression (****: ≤ 0.0001). Source data are provided in the Source
 150 Data file.



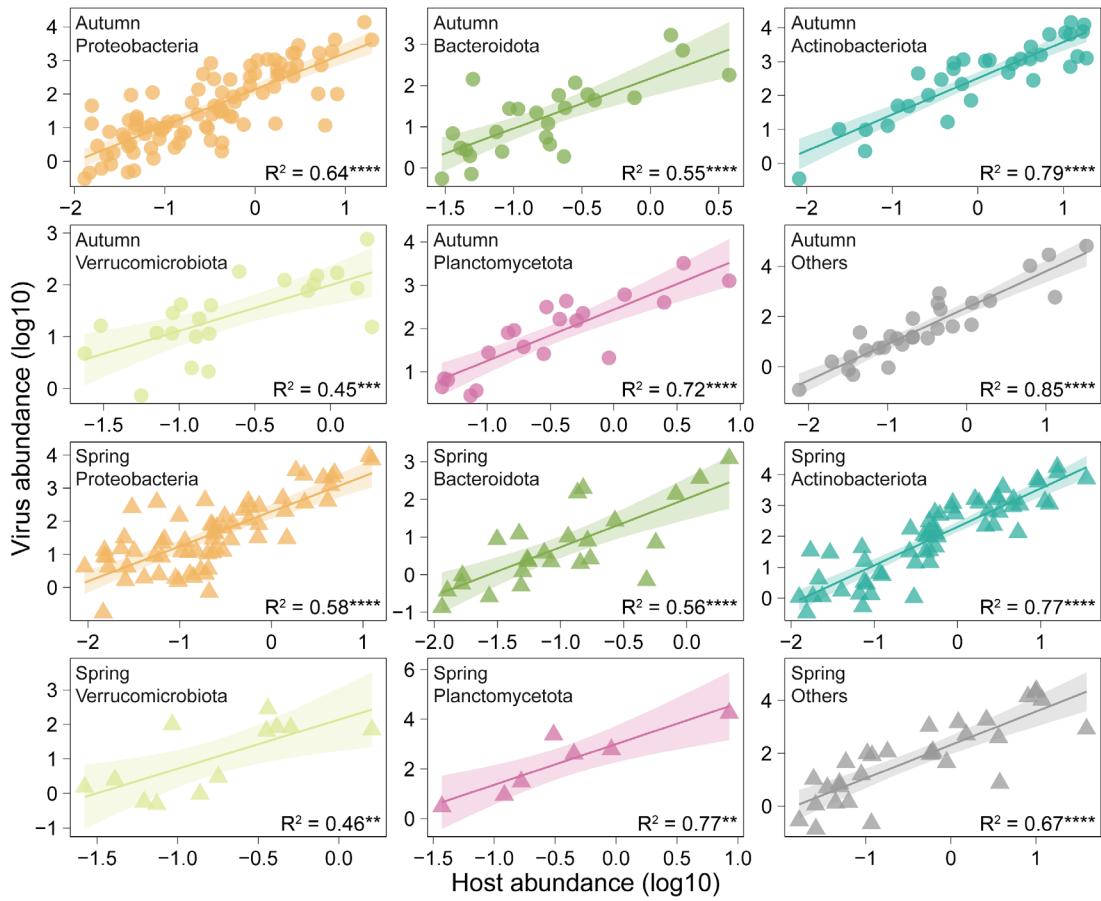
151

152 **Fig. S12. Dynamics and P-associated functions of bacteria with high-**
 153 **quality genomes (completeness > 90%, contamination < 5%), as well as**
 154 **their relationships with viruses. A** Changes in the richness (line charts) and
 155 growth potential (pie charts, see Materials and Methods) of bacteria along the
 156 canal in autumn and spring. **B** Changes in average copy number of key
 157 bacteria-encoded genes of four metabolic processes associated with P
 158 acquisition and utilization. **C** Virus–host abundance ratios display notable
 159 increase with water flow in both seasons. Each linear regression is denoted by
 160 the goodness-of-fit R^2 value and the significance level of p value. Source data
 161 are provided in the Source Data file.



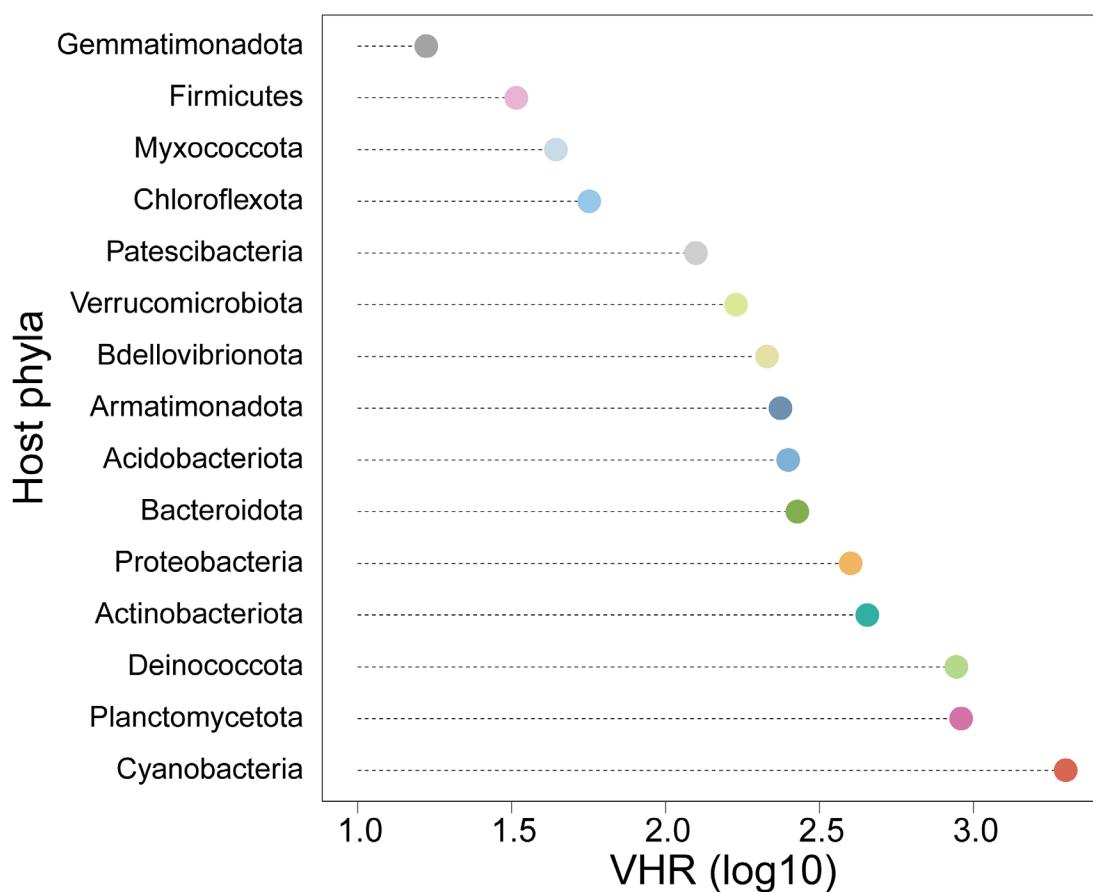
162

163 **Fig. S13 Changes in average copy number of functional genes involved
164 in carbohydrate, energy, nitrogen, sulfur, and nucleotide metabolism
165 along the canal in autumn and spring.** The average gene copy is normalized
166 by each KEGG pathway/module. Source data are provided in the Source Data
167 file.



168

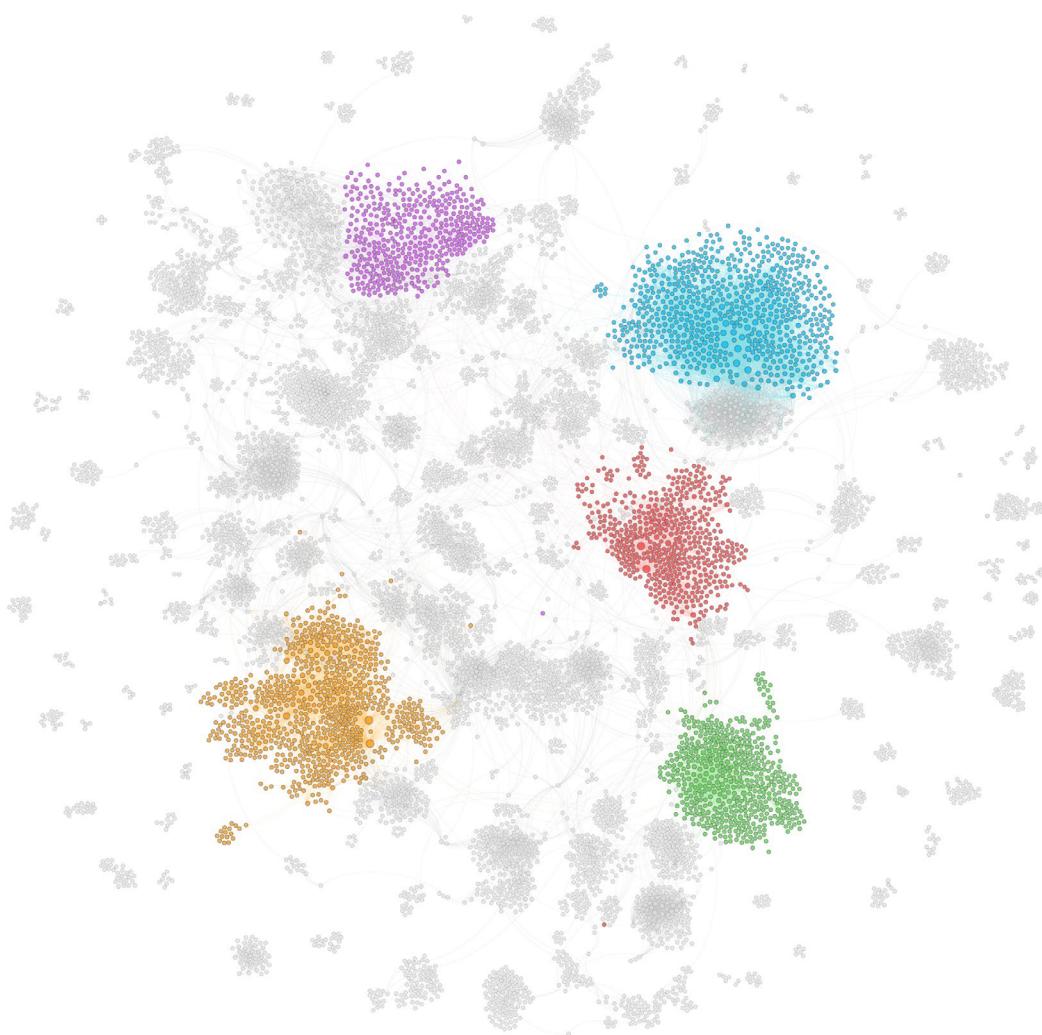
169 **Fig. S14 Correlation between abundances of viruses and their hosts for**
170 **each phylum in autumn and spring.** Host phyla linked to relatively fewer
171 viruses are categorized into “Others”. Each linear regression is denoted by the
172 goodness-of-fit R^2 value and the significance level of p value. Source data are
173 provided in the Source Data file.



174

175 **Fig. S15 Virus–host abundance ratios (VHR) for each bacterial phylum.**

176 Source data are provided in the Source Data file.



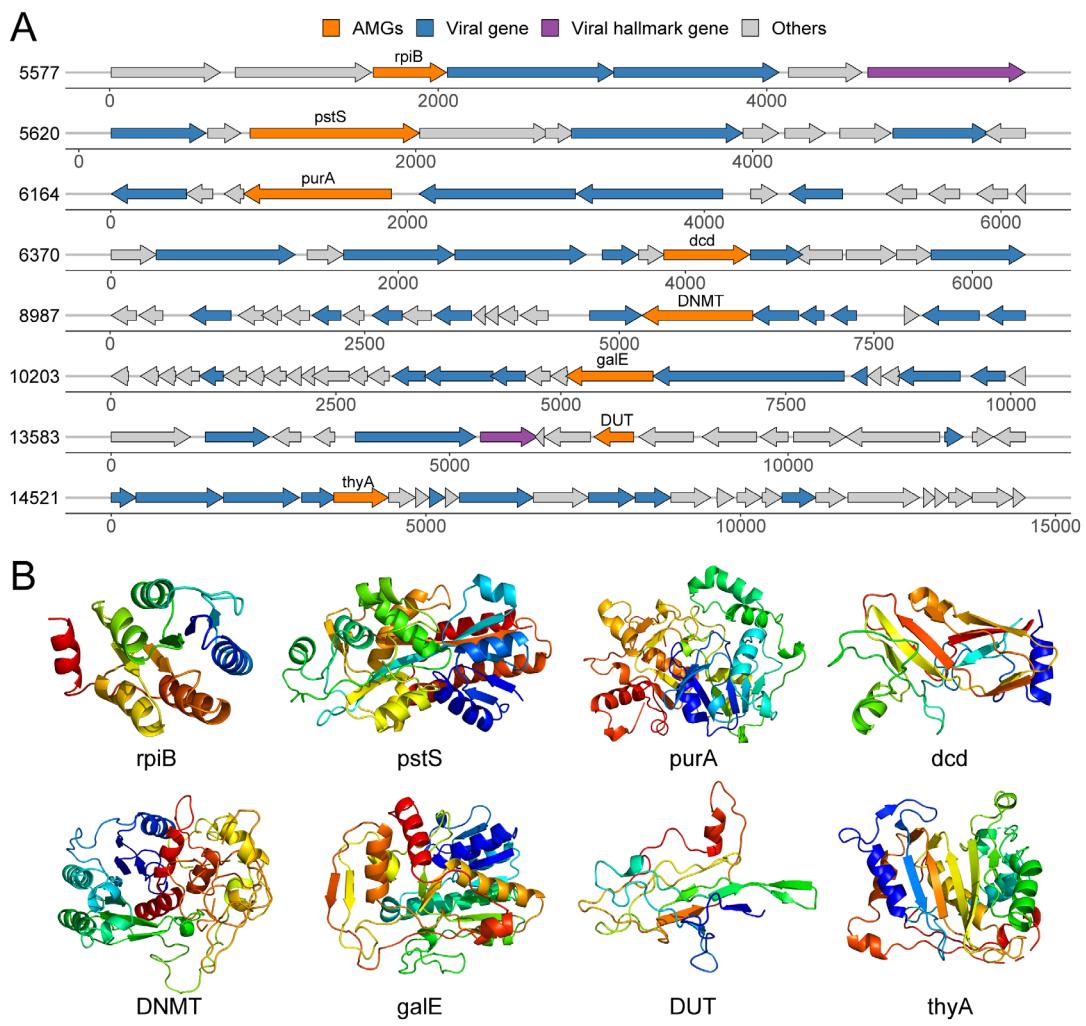
Five largest modules
for example:
● ● ● ● ●

Number of vOTUs: 4,659
Number of MAGs: 3,930
Number of linkages: 33,391

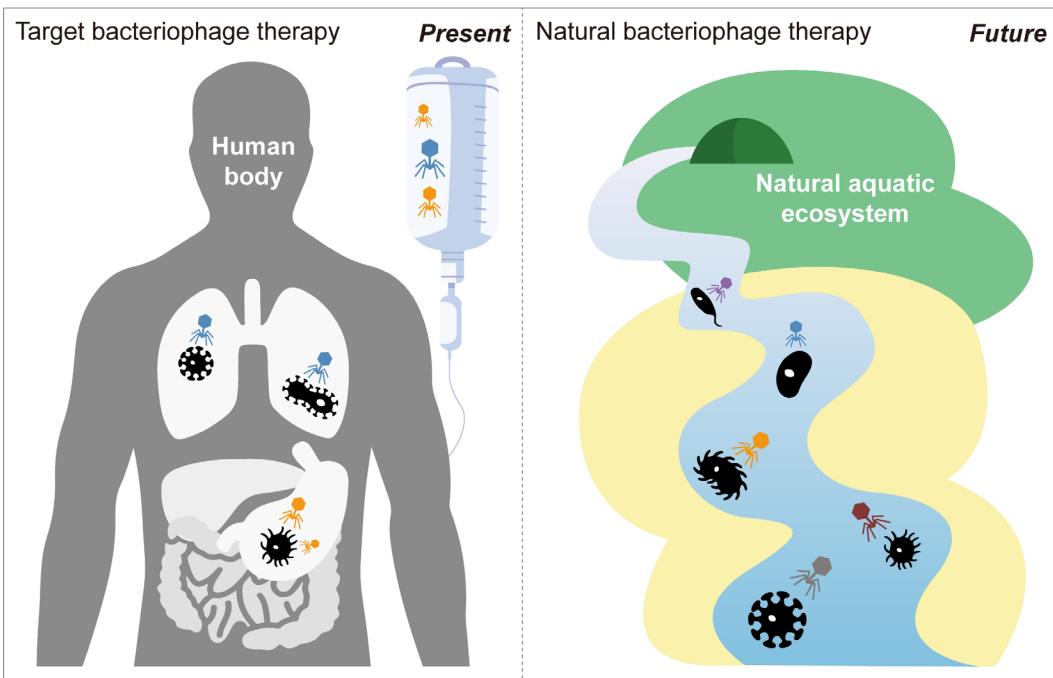
Averaged degree: 3.888
Number of modules: 153
Modularity: 0.881

177

178 **Fig. S16 Co-occurrence network of virus–host interactions.** Nodes with or
179 without black outlines represent MAGs or vOTUs, respectively. Each edge
180 marks a specific virus–host linkage. The modularity of the network is calculated
181 using community detection algorithm built in Gephi. Top five modules are shown
182 in different colors. Source data are provided in the Source Data file.



183
184 **Fig. S17 Genomic context and protein structure of selected virus-encoded**
185 **AMGs.** **A** Genome map of representative AMG-encoding viral contigs. Each
186 contig is marked by its genome length. **B** Tertiary structures of selected AMGs
187 based on structural modelling using Phyre2. rpiB: ribose 5-phosphate
188 isomerase B; pstS: phosphate transport system substrate-binding protein; purA:
189 adenylosuccinate synthase; dcd: dCTP deaminase; DNMT: DNA (cytosine-5)-
190 methyltransferase 1; galE: UDP-glucose 4-epimerase; DUT: dUTP
191 pyrophosphatase; thyA: thymidylate synthase. Source data are provided in the
192 Source Data file.



193

194 **Fig. S18 Outlook of practical utility of natural bacteriophage therapy in**
 195 **natural aquatic ecosystems compared to the target bacteriophage**
 196 **therapy used in human body.** Highly specific infections of viruses have helped
 197 to facilitate precise treatments of pathogen-induced human diseases in clinical
 198 practice. Future studies would be expected to widen the application prospects
 199 of natural bacteriophage therapy to eliminate the waterborne pathogens.

200 **Supplemental Tables**

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

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

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

207 **Table S4.** Annual average TP concentrations (mg/L) measured in the MR-
208 SNWDC (present study) and those observed in other representative river/lake
209 ecosystems (from the Global Freshwater Quality Database). The records are
210 sorted based on the average TP concentration in all years.

211 **Table S5.** Virus–host linkages and taxonomic classification of viruses and hosts.

212 **Table S6.** Summary of virus-encoded auxiliary metabolism genes (AMGs)
213 identified in the MR-SNWDC.

214

215 **Supplemental References**

- 216 1. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for
217 exploring and manipulating networks. ICWSM. 2009;8:361-362.
- 218 2. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of
219 communities in large networks. J Stat Mech-Theory E. 2008;2008:P10008.
- 220 3. Chen W, Wang J, Chen X, Meng Z, Xu R, Duoji D, et al. Soil microbial
221 network complexity predicts ecosystem function along elevation gradients
222 on the Tibetan Plateau. Soil Biol Biochem. 2022;172:108766.
- 223 4. Liu B, Arlotti D, Huyghebaert B, Tebbe CC. Disentangling the impact of

- 224 contrasting agricultural management practices on soil microbial
225 communities – Importance of rare bacterial community members. *Soil Biol*
226 *Biochem.* 2022;166:108573.
- 227 5. Nayfach S, Camargo AP, Schulz F, Eloë-Fadrosh E, Roux S, Kyrpides NC.
228 CheckV assesses the quality and completeness of metagenome-
229 assembled viral genomes. *Nat Biotechnol.* 2021;39:578-585.
- 230 6. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ.
231 Prodigal: prokaryotic gene recognition and translation initiation site
232 identification. *BMC Bioinformatics.* 2010;11:119.
- 233 7. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK,
234 Cook H, et al. eggNOG 5.0: a hierarchical, functionally and
235 phylogenetically annotated orthology resource based on 5090 organisms
236 and 2502 viruses. *Nucleic Acids Res.* 2019;47:D309-D314.
- 237 8. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von
238 Mering C, et al. Fast genome-wide functional annotation through orthology
239 assignment by eggNOG-mapper. *Mol Biol Evol.* 2017;34:2115-2122.
- 240 9. Kieft K, Zhou ZC, Anantharaman K. VIBRANT: automated recovery,
241 annotation and curation of microbial viruses, and evaluation of viral
242 community function from genomic sequences. *Microbiome.* 2020;8:90.
- 243 10. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO,
244 et al. VirSorter2: a multi-classifier, expert-guided approach to detect
245 diverse DNA and RNA viruses. *Microbiome.* 2021;9:37.
- 246 11. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa Sabina L, Solden
247 LM, et al. DRAM for distilling microbial metabolism to automate the curation
248 of microbiome function. *Nucleic Acids Res.* 2020;48:8883-8900.
- 249 12. Pratama AA, Bolduc B, Zayed AA, Zhong ZP, Guo JR, Vik DR, et al.
250 Expanding standards in viromics: in silico evaluation of dsDNA viral
251 genome identification, classification, and auxiliary metabolic gene curation.
252 *PeerJ.* 2021;9:e11447.

- 253 13. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2
254 web portal for protein modeling, prediction and analysis. Nat Protoc.
255 2015;10:845-858.
- 256 14. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New
257 and continuing developments at PROSITE. Nucleic Acids Res.
258 2012;41:D344-D347.
- 259 15. Roux S, Páez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, et
260 al. IMG/VR v3: an integrated ecological and evolutionary framework for
261 interrogating genomes of uncultivated viruses. Nucleic Acids Res.
262 2020;49:D764-D775.
- 263 16. Jang HB, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al.
264 Taxonomic assignment of uncultivated prokaryotic virus genomes is
265 enabled by gene-sharing networks. Nat Biotechnol. 2019;37:632-639.
- 266 17. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using
267 DIAMOND. Nat Methods. 2015;12:59-60.
- 268 18. Nepusz T, Yu HY, Paccanaro A. Detecting overlapping protein complexes
269 in protein-protein interaction networks. Nat Methods. 2012;9:471-472.
- 270