Supplemental Materials

Seasonal transcriptional decoupling drives microbial biosecurity risks in a large river

Qi Song, Zhibin He, and Pengfei Liu et al.,

*Corresponding author:

Zhibin He (zhibin.he@pku.edu.cn); Pengfei Liu (liupf@lzu.edu.cn)

This PDF file includes:

Other Supplementary Materials for this manuscript include the following:

The HMM profiles used for RNA virus prediction, the alignment file used for phylogenetic analysis of putative pathogenic microorganisms, and Supplementary Tables S1–S27 are available on Figshare (https://figshare.com/s/a8c530c2eb60d4a9b63a). All raw sequencing data generated in this study have been deposited in the National Omics Data Encyclopedia (NODE) under accession number OEP00007169. All code used in this study is available on GitHub (https://github.com/Songq-20/YellowRiver-Risk/tree/main/workflow).

Supplementary Text

Supplementary Results

Data collected from hydrological stations

Hydrological parameters used in this study, including suspended sediment concentration, runoff and precipitation were compiled from publicly available datasets and platforms. Data sources included the Department of Water Resources of Gansu Province and Qingyue Data (epmap.org). Hydrological monitoring records were obtained from the Department of Water Resources of Gansu Province and Qingyue Data (Supplementary Fig. 3; Supplementary Table S2).

Seasonal dynamics of environmental parameters of the Yellow River water

Environmental parameters of the Yellow River water exhibited pronounced seasonal dynamics, with significant monthly variations (Supplementary Figs. 4, 5; Supplementary Table S2). The pH remained weakly alkaline throughout the year (7.75-8.61), peaking in winter (8.26-8.61). Water temperature (Temp) was lowest in winter (1.80-5.30 °C) and peaked in summer (15.70-17.80 °C). In contrast, electrical conductivity (EC) and dissolved oxygen (DO) displayed inverse seasonal patterns: EC showed the highest values in winter (607-635 μS/cm) and lowest in summer (474-550 μS/cm), whereas DO was elevated in winter (8.92-11.74 mg/L) and reduced in summer (7.37-8.05 mg/L).

Turbidity (Turb) and particulate organic carbon (POC) peaked in summer, ranging from 34.10 to 1387.98 NTU and 45.26 to 1539.30 mg/L, respectively. Turb exhibited particularly large annual fluctuations (0.40-1387.95 NTU). Dissolved organic carbon (DOC), total phosphorus (TP), and nitrite (NO_2^-) were elevated in summer (2.18-31.75 mg/L, 0.02-0.06 mg/L, and 0.86-1.37 mg/L, respectively), whereas nitrate (NO_3^-) was higher in autumn and winter (0.75-3.31 mg/L). Total nitrogen (TN) peaked in winter (1.98-4.77 mg/L). Seasonal fluctuations in chlorophyll-a (Chla, 0.56-5.30 ng/L) and sulfate (SO_4^{2-}, 54.92-132.77 mg/L) were minimal. Dissolved inorganic carbon (DIC) varied little over the year (14.00-34.70 mg/L), with slightly lower values in winter (14.00-23.11 mg/L).

Chloride (Cl^-) was highest in winter (35.09-41.23 mg/L), while sodium (Na^+, 12.88-41.28 mg/L), ammonium (NH_4^+, 0.00-2.95 mg/L), potassium (K^+, 0.69-3.59 mg/L), magnesium (Mg^{2+}, 10.62-27.99 mg/L), and calcium (Ca^{2+}, 23.04-65.41 mg/L) showed negligible seasonal changes. Fluoride (F^-) reached a minimum in winter (0.27-0.74 mg/L).

Viral abundance (VA) and bacterial abundance (BA) both peaked in autumn (19.35-23.12 × 10^6 VLPs/mL and 1.6-2.47 × 10^6 cells/mL, respectively). Algal density (AD) and biomass (AB) were highest in spring (231.34-25,767.83 ind/L and 1.13-20.21 μg/L, respectively). Zooplankton density (ZD) was lowest in summer (0.00-9.39 ind/L), whereas zooplankton biomass (ZB) peaked in spring (0.01-24.57 μg/L).

Sand concentration was highest in summer (0.40-1.96 kg/m^3) and lowest in winter (0.00-0.01 kg/m^3). Water flow (Flow) and suspended sediment concentration (SSC) both peaked in summer (1,140-1,670 m^3/s and 649-2,560 kg/s, respectively), while precipitation (Pre) showed the maximum value in autumn (1.80-78.50 mm) (Supplementary Figs. 4, 5; Supplementary Table S2)

Phylogenetic evidence for putative bacterial and viral pathogens

To assess the pathogenic potential of putative pathogens and bacterial MAGs, we integrated phylogenetic placement with pairwise sequence identity comparison against known pathogenic references (Supplementary Figs. 9-14; Supplementary Table S5).

Among DNA viruses, several *Iridoviridae* sequences clustered with fish viral pathogen clades, including lymphocystis disease virus, grouper iridovirus, and scale drop disease virus. Pairwise sequence identities ranged from 63% to 73%, including 66% and 73% identity to lymphocystis disease virus from *Sparus aurata*, 71% to lymphocystis disease virus 1, 63% to grouper iridovirus, and 65%-66% to scale drop disease virus. Their consistent phylogenetic placement within fish iridovirus lineages suggests a persistent aquatic animal-associated biosecurity risk (Supplementary Figs. 9-10).

Among RNA viruses, the *Astroviridae* sequence clustered with Human astrovirus and shared 96% sequence identity, indicating a potential human-associated risk. Several *Reoviridae* sequences were phylogenetically linked to Banna virus (84% identity) and Rotavirus A (93%-95% identity). In addition, a *Picobirnaviridae* sequence was placed close to Human picobirnavirus, with 76% sequence identity, suggesting a moderate yet notable human-associated viral signal. In contrast, most *Picornaviridae* sequences grouped with environmental or invertebrate-associated picorna-like viruses, indicating limited evidence for direct pathogenic relevance (Supplementary Figs. 11-13).

Bacterial MAG assignments were strongly supported by high genome-wide sequence identity (ANI; ≥ 95%) to reference pathogens. The recovered MAGs exhibited near-complete sequence similarity to known strains, including 99% ANI to *Afipia broomeae*, 95% to *Acinetobacter johnsonii*, 100% to *A. junii*, 99% to *A. haemolyticus*, 99% to *Brevundimonas vesicularis*, and 100% to *Methylorubrum extorquens*. Together with their phylogenetic placement, these high sequence identities support the widespread occurrence of opportunistic pathogen-associated lineages, particularly *Acinetobacter*-related bacteria (Supplementary Fig. 14).

Virulence factor dynamics and their linkage with mobile genetic elements in MAGs

Across 601 non-redundant MAGs (nrMAGs) carrying VFs, adhesive-related VFs exhibited the highest copies/MAG ratios (504 copies/MAG) and highest relative transcriptional activity (TPM = 7.1×10^6), whereas exoenzyme VFs showed a lower copies/MAG ratio (23 copies/MAG), but still relatively high transcriptional activity

(TPM = 1.26×10^5) (Supplementary Figs. 32a-b). Among these, 131 MAGs affiliated with *Burkholderiaceae* were identified as the primary VFs carriers. Of these, 6 MAGs harbored adhesion-related genes and exhibited the highest VF relative transcriptional activity (TPM = 1.52×10^6), with adhesion-related gene transcription reaching 1.5×10^6 TPM (Supplementary Figs. 32e-f). No significant differences were observed between putative pathogenic and non-pathogenic MAGs in either VF diversity per genome or transcriptional activity per MAG (Supplementary Figs. 32c-d). Lastly, VF and MGE abundances were significantly positively correlated across multiple analytical frameworks, including reads-based ($R^2_{DNA} = 0.465$, $P < 0.01$; $R^2_{RNA} = 0.41$, $P < 0.0001$), and contigs-based analyses ($R^2_{DNA} = 0.81$, $P < 0.0001$; $R^2_{RNA} = 0.873$, $P < 0.001$; Supplementary Fig. 33).

Supplementary Discussion

While disease-causing microbial and viral pathogens are ubiquitous across global rivers, their concentrations are highly variable and context-dependent[1]. In heavily sewage-impacted systems, specific opportunistic bacterial pathogens, such as *A. johnsonii* and *Prevotella* spp., can reach concentrations exceeding 5.9×10^6 copies/L, comprising up to 2%-10% of the total microbial community[2]. Similarly, viral pathogens, particularly human enteric viruses derived from wastewater effluent, are frequently detected in polluted rivers at concentrations ranging from 10^2 to 10^6 genome copies/L[3]. RNA viral pathogen families can represent an overwhelming 4% to 22% of total RNA viral reads in severely contaminated waters[4].

In contrast, our analysis of the upper Yellow River basin through monitoring the Zhongshan Bridge cross-section reveals a relatively low overall burden of recognized bacterial and viral pathogens. Both potential bacterial and viral pathogens constitute only a marginal fraction of the total river bacterial and viral communities, respectively (Figs. 1a, c, e, g; Supplementary Figs. 8a, c). Furthermore, the potential bacterial pathogens identified predominantly map to low-risk categories. Among the detected DNA and RNA viruses, except for *Astroviridae* and *Reoviridae*, sequence homology to high-risk human viral pathogens remained low (ranging from 4% to 49%). Interestingly, certain DNA viruses (e.g., 2307KPC2_k141_2506506) exhibited a high degree of similarity (up to 73%) to known piscine viruses (fish-infecting viruses, e.g., lymphocystis disease virus 3 from Sparus aurata; Supplementary Fig. 9), underscoring that the dominant viral pressures in this ecosystem are directed toward aquatic fauna rather than mammalian hosts.

Absolute risk assessment indices confirm that the baseline biosecurity risk of the Yellow River water from the Zhongshan Bridge is comparable to previously assessed urban river ecosystems[5]. Further, our findings also demonstrate that classical pathogenic taxa do not occupy dominant ecological niches in this specific river section; instead, the community architecture is dictated by highly adapted environmental clades.

This pattern aligns with the oligotrophic status and high water quality (Class II) of the Lanzhou section, which is characterized by the low concentrations of total dissolved nitrogen (TDN), total dissolved phosphorus (TDP), and chloride (Cl^-)[6].

Notably, *Picobirnaviridae* emerged as a dominant RNA viral family in the river water, with its relative abundance fluctuating between 0.20% and 8.77%. This substantial presence is consistent with previous observations in other major river ecosystems globally, such as the Red and Assiniboine rivers in Winnipeg, Manitoba, where *Picobirnaviridae* abundances reached 3%-10% of the viral community[7]. Given the broad host range of *Picobirnaviridae* and a high sequence similarity (56% to 76%) to viruses known to infect specific aquatic hosts, e.g. freshwater macrophyte-associated picobirna-like virus (Supplementary Fig. 13; Supplementary Table S5), we infer that these viral populations primarily target endemic aquatic organisms.

Although there are some differences between the two methods, both found that the overall microbial-associated risks in the Yellow River water exhibit distinct seasonal variation characteristics, being higher in summer and autumn. Aside from ARGs, both methods revealed similar patterns for DNA viral and bacterial pathogens. Similar to the seasonal dynamics of community composition, risk index at the RNA level indicated higher levels of ARG risks in autumn (Fig. 5d), but the absolute quantitative analysis at the DNA level showed that ARGs carried by Rank I and II (B1) ESKAPE+EV were highest in summer (B2; Fig. 5f). Because absolute quantification of RNA viruses was not performed in the absolute quantitative framework of this study, and other indicators (A2, C1, and D2) were only considered in the absolute quantitative framework, these indicators cannot be compared between the two methods.

Supplementary Methods

Measurements of environmental parameters for the Yellow River water

All samples were delivered to the laboratory within 2 hours for analysis of environmental parameters. For each sample, 1 L of water was filtered through GF/F glass microfiber filters (Whatman™, Cytiva). The retained filters were immersed in 10 mL of cold anhydrous ethanol. Subsequently, chlorophyll a (Chla, ng/L) was extracted for 24 h at 4°C in the dark, and its concentration was measured using a fluorometer (Trilogy; Turner Designs, Sunnyvale, CA, USA)[8]. Another 1 L of water was filtered through pre-combusted (450°C, 6 hours) GF/F glass microfiber filters[9]. The filter and filtrate were subsequently used to determine the concentrations of particulate organic carbon (POC, mg/L), dissolved organic carbon (DOC, mg/L) and dissolved inorganic carbon (DIC, mg/L). POC, DOC, and DIC were determined using a TOC-L series TOC analyzer (SM5300A, Shimadzu Corporation, Japan). The filtrates were also retained for nutrient analysis, including nitrite nitrogen (NO_2^--N, mg/L), nitrate nitrogen (NO_3^--N, mg/L) and ammonium nitrogen (NH_4^+-N, mg/L), which were quantified with an Automated Discrete Chemistry Analyzer (AMS Alliance, USA). The concentrations of

total dissolved nitrogen (TDN, mg/L), and total dissolved phosphorus (TDP, mg/L) were measured using a GENESYS^{TM} 180 spectrophotometer (Thermo Fisher Scientific, China). Anion and cation concentrations-specifically fluoride (F^-, mg/L), chloride (Cl^-, mg/L), sulfate (SO_4^{2-}, mg/ L), sodium (Na^+, mg/L), potassium (K^+, mg/L), magnesium (Mg^{2+}, mg/L), and calcium (Ca^{2+}, mg/L)-were determined with an ICS-600 ion chromatography system (Thermo Fisher Scientific, China). All environmental parameters are provided in (Supplementary Table S2).

Measurements of the plankton community

At each sampling site, 10 L of raw water was collected and filtered through a 25# plankton net. These samples were preserved in 10 mL of Lugol's iodine solution for later analysis of phytoplankton after a 48 h settling period. Colonies or filamentous algal cells were separated using an ultrasonicator (JY88-II, Scientiz, China). The phytoplankton in residue samples was placed in a counting chamber (Sedgewick-Rafter S51 Counting Cell, USA) with 1 mL of water to count the number of cells under 100–400× magnification using a regular Olympus microscope (Olympus BX53, Japan). Counts were converted to a density and biomass based on size measurements of at least 30 cells of each taxon and using formulae for geometric shapes approximating cell forms. Phytoplankton species were identified and classified in reference to previous study and the standard methods[10].

Zooplankton were also collected by filtering a 10 L sample through a 25# plankton net, after which the retained material was preserved in 10 mL of formaldehyde solution for later analysis of zooplankton after a 48 h settling period. Finally, the sample was placed in a counting chamber (Sedgewick-Rafter S51 Counting Cell, USA) with 1 mL of water to count zooplankton. In these samples, all cladocerans and copepods were counted and identified to species using 10-40× magnification following standard methods. Protozoans and rotifers were also counted to species level, when possible, at 100-200× magnification. At least 30 individuals of each species were measured with an Olympus microscope (Olympus BX53, Japan) and converted to a density and biomass using equations as previously described[11]. Data of the plankton community are provided in Supplementary Table S3.

Measurements of bacterial and viral abundance

Bacterial cells were counted by flow cytometry, whereas viral particles were counted by fluorescence microscopy after SYBR Green I staining (Molecular Probes, 10,000×original dilution, Thermo Fisher Scientific, USA). Two 1.98 mL aliquots of raw Yellow River water were transferred into sterile cryovials and fixed with glutaraldehyde (2% final concentration), followed by incubation in the dark for 15 min and directly stored at -80°C. For viral enumeration, 1-10 mL samples were filtered onto 0.02 μm Anodisc membrane filters (Anodisc Al_2O_3, 25 mm, Whatman, USA), with a 0.8 μm backing membrane filter. The filters were stained with 20× SYBR Green I for 15 min

in the dark. After drying completely, the filters were mounted on a glass slide with about 30 μL Molecular Probe Slow Fade-antifade solution. Viral particles were then observed using a Leica DM3000 fluorescence microscope (Leica Microsystems GmbH, Wetzlar, Germany) under blue-green light excitation (FITC filter, excitation at 480-495 nm)[12]. For bacterial enumeration, the fixed samples were stained with SYBR Green I and counted using a Beckman flow cytometer[13]. Data on bacterial and viral abundance are provided in Supplementary Table S2.

DNA extraction and metagenomic sequencing

Extracted DNA from 15 samples were carried out as previously reported[14]. In brief, ~6 L Yellow River water was filtered through a 0.22 μm pore-size Isopore™ polycarbonate filters (47 mm; Millipore). The filters were stored at -80°C until further processing. Prior to DNA extraction, the filters were thawed at 4°C, cut into small pieces, and incubated in lysis buffer at 70°C for 15 min. DNA was then extracted using the DNeasy PowerSoil Kit (QIAGEN, Hilden, Germany) according to the manufacturer's instructions. DNA concentration was measured using a Qubit fluorometer (Thermo Fisher Scientific, USA). Library preparation and sequencing were performed by Guangdong Magigene Biotechnology Co., Ltd. (Guangzhou, China). Fifteen DNA libraries were sequenced on an Illumina NovaSeq 6000 platform with 2 × 150 bp paired-end chemistry.

Viral particle enrichment, DNA extraction, and viral metagenomic (metaviromic) sequencing

A total of 6 L of Yellow River water were initially filtered through 3 μm pore-size Isopore™ polycarbonate filters (47 mm; Millipore) to eliminate substantial sediment and large suspended particulates. The resulting filtrate was filtered through a 0.22 μm pore-size Isopore™ polycarbonate filters (47 mm; Millipore) to remove cells and large debris. The filtrate of each sample was treated with $FeCl_3$ to precipitate the viral particles by adding 25 mL of concentrated Fe^{3+} stock solutions (10 g/L Fe^{3+}) to every 10 L of filtrate[15]. The mixture was vigorously shaken for at least 1 min and then allowed to stand for 1 h at room temperature. The $FeCl_3$-treated filtrate was filtered using 0.45 μm Millipore Express® PLUS membranes (47 mm; Millipore), which were stored at -80°C until further processing.

DNA was extracted from the filter-retained biomass. Briefly, the membrane filters were cut into small pieces and processed using the DNeasy PowerSoil Pro Kit (QIAGEN, Hilden, Germany) according to the manufacturer's protocol. The extracted DNA was quantified using a Qubit fluorometer (Thermo Fisher Scientific, USA). Library preparation and sequencing were performed by Guangdong Magigene Biotechnology Co., Ltd. (Guangzhou, China). Thirty-six DNA libraries were sequenced on an Illumina NovaSeq 6000 platform with 2 × 150 bp paired-end chemistry.

Assembly of metagenomic and metaviromic reads and binning

Raw reads of each microbial and viral metagenome were quality-controlled using Trimmomatic v.0.39[16] with default parameters to remove sequencing adapters and low-quality sequences. All high-quality sequences were co-assembled using MEGAHIT with default parameters[17] with kmers: 21-141. All assembled scaffolds longer than 1.0 kb were binned using BASALT with default parameters[18]. Genome completeness and contamination were evaluated with CheckM v1.2.2[19], and metagenome-assembled genomes (MAGs) with more than 50% completeness and less than 10% contamination level were retained for further analyses. Bins from different samples were dereplicated using dRep v.2.3.2 with ANI of 95% to generate a non-redundant set of medium-quality genomes[20]. Taxonomic classification was performed based on the Genome Taxonomy Database (GTDB; release R226) using the GTDB-Tk toolkit (v.2.4.0) with the classify workflow[21,22].

Identification, quality checking and clustering of DNA viral contigs

Contigs from each metagenomic and metaviromic datasets with length ≥ 10 kb were kept for virus identification. Viral contigs were retrieved from assemblies by the combined use of VirSorter2 v2.2.1, VirFinder v1.1, geNomad v1.11.2 and VIBRANT v1.2.1[23]. In brief, VirSorter2 (with all group parameters), VirFinder, geNomad and VIBRANT were run with default settings. For VirSorter2, contigs with a score ≥ 0.9 were assigned 1 point, those with scores of 0.5-0.9 were assigned 0.5 points, and those with scores < 0.5 were assigned 0 points. For geNomad, contigs with scores ≥ 0.8 were assigned 1 point, those with scores of 0.7-0.8 were assigned 0.5 points, and those with scores < 0.7 were assigned 0 points. For VirFinder, contigs with a p-value < 0.05 and score ≥ 0.9 were assigned 1 point, those with scores of 0.7-0.9 were assigned 0.5 points, and those with scores < 0.7 were assigned 0 points. Contigs identified by VIBRANT were assigned 1 point. Contigs with a total sum score ≥ 1 across the four tools were retained as candidate viral contigs (vContigs). Candidate vContigs from four pipelines mentioned above were merged and the quality was checked and trimmed by using CheckV v0.7.0[24]. Contigs with host gene number >0 and without any viral hallmark gene detected were discarded as previously recommended by Luo et al.,[25] and the VirSorter2 SOP protocol[26]. Finally, a total of 111,912 CheckV-trimmed clean vContigs were obtained. Then, vContigs were clustered at 95% average nucleotide identity (ANI) across 85% of the shorter contigs (AF) using the CheckV scripts (https://bitbucket.org/berkeleylab/checkv/src/master/) to generate a non-redundant set of species-level of viral populations (vOTUs) as previously described[27], yielding 31,877 vOTUs (Supplementary Table S7).

Taxonomic assignment of DNA vOTUs

To assign taxonomy for obtained vOTUs, five pipelines were applied, including a protein-sharing network approach using vConTACT2 v0.9.19[28], genome-based

taxonomic prediction using geNomad, viral protein family-based classification using VPF-Class[29], host-informed viral classification using PhaGCN3[30], and ViralRecall v2.1[31] to identify Nucleo-cytoplasmic large DNA viruses (NCLDVs). In brief, we first used Prodigal v2.6.3[32] with '-p meta' parameter to predict the open reading frame (ORF), and the resulted protein sequences were used for vConTACT2 with default parameters as previously described [27]. The nucleotide sequences of vOTUs were then subjected to taxonomic prediction using geNomad, VPF-Class, and PhaGCN3. The final taxonomy of each vOTU was assigned based on the available classified results from the four approaches mentioned above, following the order of vConTACT2, geNomad, VPF-Class, and PhaGCN3. If conflicts exist at a lower taxonomic level (e.g., family level) for different pipelines but a common higher level of taxonomy (e.g., order level) exits, then this common higher level of taxonomy was assigned [27]. In addition, signatures of NCLDVs were identified by ViralRecall v2.1 with default parameters and contigs with scores > 0 were considered to be putative NCLDVs (Supplementary Table S7).

Relative abundance of vOTUs and MAGs in metavirome and metagenomes

Bowtie2 (v2.4.5)[33] with the '-sensitive' parameter and coverM (https://github.com/wwood/CoverM; v0.6.1)[34] with the 'genome' and 'contigs' module with default parameters were employed to calculate the relative abundance of MAGs and vOTUs. Briefly, trimmed metaviromic or metagenomic reads were mapped to vOTUs derived from metaviromic or metagenomic samples and to MAG sequences with default parameters. The resulting bam files were sorted using Samtools v1.4[35]. Transcripts per million (TPM) and reads count were generated based on the sorted bam files using coverM with '-m tpm' and '-m count' parameters. The 'coverm genome' and 'coverm contigs' modules were applied to estimate TPM values of MAGs and vOTUs, respectively (Supplementary Table S6-8).

Total RNA extraction, quality checking, sequencing, and reads processing

Extracted RNA from 36 samples were carried out as previously reported[36]. In brief, a total of ~6 L Yellow River water was filtered through the 3.0 μm Isopore™ polycarbonate membranes (47 mm; Millipore). The filters were stored at -80°C in the laboratory until further processing. Total RNA from all samples was extracted according to the manufacturer's instructions using the RNAeasy PowerSoil Total RNA Kit (QIAGEN, Hilden, Germany). Residual DNA was removed, and RNA purification was performed using DNase I and the RNA Clean&Concentrator-5 (Zymo, Irvine, USA). The concentration of the extracted RNA was measured using the RNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA) with a Qubit 3 Fluorometer (Invitrogen, Carlsbad, CA, USA). Subsequently, library preparation and sequencing were performed by Guangdong Magigene Biotechnology Co., Ltd. (Guangzhou, China). Thirty-six RNA libraries were sequenced on an Illumina NovaSeq 6000 platform with

150 bp paired-end reads.

Over 839 GB of raw read data were generated, averaging approximately 155 million reads per sample. The raw reads were processed using Trimmomatic v0.39 with default parameters to remove low-quality adapters and adapter-dimers. Subsequently, rRNA reads were removed from the total RNA using SortMeRNA v4.3.6[37], resulting in 110 GB of non-rRNA reads, averaging approximately 61 million mRNA reads per sample (Supplementary Table S4).

Metatranscriptomic reads assembly, RNA virus identification and verification

The protein sequence containing RdRp domains with an E-value ≤ 0.01 were retained as candidate RdRp-containing ORFs: First, all protein sequences with putative RdRp domains were subjected to the Palmscan algorithm3 screening based on PALMdb[38]. Sequences with a palmprint (a segment of the palm sub-domain robustly delineated by well-conserved catalytic motifs) score ≥ 20 were treated as positive RdRp sequences. Second, we applied a metagenomic reads mapping strategy as described by Liu et al.[36] to remove DNA remnants since most of the metatranscriptomes were accompanied by a metagenome from the same sample. Metagenome reads were mapped against the RdRp nucleotide sequences from prodigal using Bowtie2 with the "end-to-end" setting to check whether there was a DNA counterpart. RdRp sequences with any metagenomic read mapping evidence were excluded. RdRp sequences that passed the two verification processes were confirmed as true positive RdRps and contigs containing these RdRps were considered as *bona fide* RNA viral genomes. In total, we obtained 9,579 RNA viral genomes (vContigs) from all samples, with an average of 266 vContigs per sample (Supplementary Table S9).

Taxonomy classification of RNA viruses

Taxonomic assignment of RNA viruses was performed using a hierarchical framework based on RdRp amino acid similarity and vContig nucleotide similarity against multiple public reference databases, including the global RNA virome dataset (RVMT), Tara Oceans RNA virus datasets (TO), IMG/VR-v4 and LucaProt[39-41]. Taxonomic annotations were assigned according to a predefined confidence priority from Level 1 to Level 11.

First, RdRp sequences were compared against the RVMT dataset using CD-HIT[42] clustering at 90% amino acid identity to assign family- to genus-level taxonomy (Level 1). Viral contigs were subsequently compared against RVMT genomes using ANI- and alignment fraction (AF)-based clustering implemented in CheckV scripts (anicalc.py and aniclust.py) for nucleotide-level family/genus classification (Level 1a). RdRp sequences were also searched against RVMT using BLASTp[43] (Identity ≥ 90%, Query Coverage ≥75%) to identify highly similar homologs (Level 2), while viral contigs were searched against RVMT using dc-megablast for nucleotide-level classification (Level 3). To assign broader taxonomic affiliations, RdRp sequences were clustered against

RVMT at 50% amino acid identity using CD-HIT for family-level classification (Level 4).

The same strategy was independently applied to the Tara Oceans RNA virus dataset. RdRp sequences were clustered against TO references using CD-HIT at 90% amino acid identity for family/genus-level assignment (Level 5), searched using BLASTp (Identity ≥ 90%, Query Coverage ≥ 75%) (Level 6), and clustered at 50% amino acid identity for broader family-level classification (Level 8). Viral contigs were additionally compared against TO genomes using dc-megablast (Level 7).

For contigs without confident assignments from RVMT or TO datasets, nucleotide similarity searches against IMG/VR-v4 using dc-megablast were further conducted (Level 9). High-confidence family-level classifications from VITAP[44] were subsequently incorporated (Level 10). Finally, RdRp sequences were clustered against LucaProt references using CD-HIT at 50% amino acid identity to obtain additional family-level annotations (Level 11).

Taxonomic assignments followed a hierarchical confidence scheme, where annotations derived from higher-priority levels superseded those from lower-priority levels. Overall, RdRp-based approaches primarily supported protein-level evolutionary classification, whereas viral contig-based approaches provided complementary nucleotide-level taxonomic evidence (Supplementary Table S9).

Transcriptional profile analysis of MAGs and DNA vOTUs

Bowtie2 with '-sensitive' parameter and coverM with 'genome' and 'contigs' module with default parameters were used to identify active bacterial and viral taxa. Briefly, the total mRNA reads were mapped to vOTUs derived from metagenomic samples and MAG sequences with default parameters. The resulting bam files were sorted using Samtools v1.4. TPM and reads count were generated based on the sorted bam files using coverM with '-m tpm' and '-m count' parameters. The 'coverm genome' and 'coverm contigs' modules were applied to estimate TPM values for transcriptional profile of MAGs and vOTUs, respectively (Supplementary Table S10-11).

Resistome, virulome, and mobilome gene analysis

All open reading frames (ORFs) of MAGs and Contigs (> 2.5 kbp) were aligned to the Structured ARG (SARG v2.2) protein database[45] using BLASTp[43] with an e-value ≤ 1e-10 and a minimum amino acid identity of ≥70%. Similarly, virulence factor genes (VFGs) and mobile genetic elements (MGEs) were annotated by aligning predicted ORFs against the Virulence Factor Database (VFDB)[46] and the MGE marker gene database (MGE_db)[47], respectively.

To identify and quantify ARGs, VFGs, and MGEs at the read level, clean reads from all samples were analyzed using ARGs-OAP v3.2[48]. In addition to the built-in SARG v2.2 database for ARG detection, customized VFDB and MGE_db databases were incorporated into the ARGs-OAP workflow to enable parallel identification of the

virulome and mobilome across all samples.

To construct non-redundant gene catalogs, ARG, VF, and MGE sequences identified across all samples were separately clustered using MMseqs2 (v 13.45111)[49] with the parameters -e 0.001 --min-seq-id 0.95 -c 0.80. This generated three non-redundant databases representing the resistome, virulome, and mobilome, comprising 1,344 ARGs, 13,494 VFs, and 35,964 MGEs, respectively. ARG risk levels were assessed following previously described criteria[50], with ARGs classified into five risk ranks and further grouped as high-risk (Ranks I and II), low-risk (Ranks III and IV), and Others (Supplementary Table S12-14).

Bowtie2 with '-sensitive' parameter and coverM with 'contigs' module with default parameters were used to map reads to the ARGs, VFs, and MGEs databases[45]. Briefly, the clean reads and assembled contigs of metagenome and metatranscriptome samples were mapped to the ARGs, VFs, and MGEs databases with default parameters, and yielded bam files were sorted using Samtools v1.4. TPM were generated based on the sorted bam files using coverM with '-m tpm' and '-m count' parameters. The 'coverm contigs' modules were applied to estimate TPM values of ARGs, VFs, and MGEs in metagenome and metatranscriptome samples.

Taxonomy annotation of unbinned contigs

Taxonomic classification of unbinned ARG-carrying contigs was performed to infer the potential bacterial hosts of ARGs that were not assigned to MAGs. All unbinned contigs longer than 2.5 kb and carrying at least one ARG were extracted and analyzed using the Contig Annotation Tool (CAT)[51]. Briefly, the CAT contigs workflow was applied to classify contigs based on predicted protein-coding genes and their homology to reference protein databases. Taxonomic assignments were made according to the lowest common ancestor strategy implemented in CAT, thereby providing conservative lineage-level annotations for ARG-carrying contigs.

The resulting taxonomic output was further processed using the CAT add_names function to append complete taxonomic lineage names, including phylum, class, order, family, genus, and species where available. To improve confidence in the classification of contigs carrying high-risk ARGs, taxonomic assignments for these contigs were additionally refined using Kraken2[52] against the prebuilt PlusPF database, which includes the standard Kraken2 database supplemented with RefSeq protozoa and fungi genomes. The resulting taxonomic assignments were used for downstream contig-level taxonomic composition analyses. The CAT- and Kraken2-based classifications were then compared to evaluate consistency and to support host assignment of high-risk ARG-carrying contigs.

Taxonomic information for unbinned ARG-carrying contigs, including their assigned lineages and associated ARGs, is provided in Supplementary Table S15.

Functional verification of high risks ARGs

To validate the resistance phenotypes of high-risk ARGs, all 80 high-risk ARGs were first screened using AlphaFold3[53] and Phyre2.2[54] to evaluate the structural integrity of their encoded proteins. From the subset of ARGs with AlphaFold3 pTM values > 0.8, nine representative genes were selected for experimental validation by further considering ARG class, resistance mechanism, predicted protein completeness, and the availability of corresponding antibiotics for phenotypic testing. The selected genes covered four major antibiotic resistance categories, including β-lactamases, tetracycline resistance genes, MLS resistance genes, and aminoglycoside resistance genes. This selection strategy ensured that the tested ARGs represented structurally reliable resistance proteins and diverse resistance mechanisms detected in the high-risk ARG pool (Supplementary Table S16).

Identification of multi-species ARGs as putative mobile ARGs

Potential mobile ARGs were identified by detecting identical or nearly identical assembled nucleotide sequences shared across different species. As vertical inheritance leads to significant sequence divergence over large taxonomic distances, finding such high nucleotide identity (≥ 99%) across distantly related taxa serves as a strong proxy for recent HGT events. We first pooled the nucleotide sequences of ARG ORFs from contigs and MAGs together, and then clustered the sequences at 99% identity and 90% coverage using MMseqs2 (v 13.45111)[49], with options '-min-seq-id 0.99 -c 0.9-cov-mode 1'. ARG ORFs within the same cluster were considered identical or nearly identical genes. For each ARG cluster, we determined the host ranges based on the taxonomic classifications of the contigs or MAGs with the ARG ORFs, as well as the species-level cluster IDs of MAGs. The latter was obtained by the dRep (v2.3.2) dereplication workflow, which was used with options 'dRep dereplicate -comp 50 -sa 0.95' to cluster the MAGs at 95% average nucleotide identity (ANI). The ARG clusters were then classified into one of the three categories: 'multi-species' when the hosts belonged to different species or higher taxonomic levels, or when they were assigned to distinct species-level MAG clusters[55].

High-throughput quantitative PCR of ARGs and MGEs

High-throughput quantitative PCR (HT-qPCR) was performed by MagiGene (Guangzhou, China) using a SmartChip real-time PCR platform. After DNA extraction and quality control, DNA samples and qPCR reagents were loaded into a 384-well sample plate, while gene-specific primers and qPCR reagents were loaded into a separate 384-well primer plate. An automated nanoliter-scale liquid handling system was used to dispense the sample and primer mixtures into the microwells of the HT-qPCR chip. Amplification and fluorescence detection were then performed using a SmartChip Real-Time PCR Cycler, and amplification curves and melting curves were generated automatically. In total, 144 ARG- and MGE-associated target genes were

quantified in triplicate for each sample using standard curves and negative controls (Supplementary Table S17).

The 16S rRNA gene amplicon sequencing

The indexed and purified amplicons were pooled in equal amounts and sequenced using the Illumina NovaSeq 6000 platform in paired-end mode (2 × 250 bp) (MagiGene). Raw 16S rRNA gene amplicon sequences were processed using USEARCH (v11)[56]. Paired-end reads were merged, primer sequences were removed, and low-quality reads were filtered based on an expected error threshold of ≤1.0. Sequences with abnormal lengths were discarded prior to dereplication and denoising with UNOISE3[57], yielding 3,477 zero-radius operational taxonomic units (zOTUs). Quality-filtered reads were mapped to representative zOTUs sequences to construct the abundance table (Supplementary Table S18). Taxonomic assignments were obtained using the SILVA database (v138)[58], retaining only classifications with bootstrap support ≥ 80%.

Identify of risk microbes within DNA, RNA viruses, and pathogens

Biosafety risk information was primarily obtained from CDC and gcPathogen[59] and was standardized into four risk levels, from level 1 representing higher risk to level 4 representing lower risk. For pathogens lacking risk-level annotations, additional information was manually retrieved from international culture collection catalogues, including ATCC and the Leibniz Institute DSMZ–German Collection of Microorganisms and Cell Cultures GmbH, as well as institutional biosafety manuals and relevant literature (Supplementary Table S19).

Representative genomes of pathogenic bacteria were downloaded from GTDB and NCBI. Bacterial 16S rRNA gene sequences were extracted from the GTDB ssu_all_r226.fna.gz file and matched to the curated pathogenic-bacteria list to construct a 16S rRNA reference dataset. The GTDB file description defines ssu_all_*.fna.gz as FASTA files containing 16S rRNA sequences identified across GTDB genomes passing quality control, supporting its use as a GTDB-linked 16S rRNA reference source. For viral pathogens, representative genomes and marker genes, including RdRp genes for RNA viruses and other viral hallmark genes for DNA viruses, were collected from NCBI, ICTV-related resources and curated viral databases.

Potential pathogenic bacteria were first screened from 16S rRNA amplicon data. zOTU sequences were compared against the curated pathogenic-bacteria 16S rRNA reference dataset using MAPseq v1.2.6 with default parameters[60]. MAPseq is a reference-based rRNA sequence classification tool designed for taxonomic assignment of ribosomal RNA sequences. Because the amplicon-derived zOTUs had an average length of approximately 372 bp, only zOTUs with ≥99% nucleotide identity to known pathogenic bacterial 16S rRNA sequences were retained as potential pathogenic bacteria[61]. In parallel, all metagenome-assembled genomes recovered from Yellow

River water were compared with representative pathogenic bacterial genomes, and MAGs sharing ≥ 95% average nucleotide identity (ANI) with known pathogenic bacteria were considered potential pathogenic bacterial MAGs.

Potential pathogenic viruses were identified by searching the viral taxonomic annotation results against the curated viral pathogen list. DNA and RNA viral contigs or vOTUs assigned to known pathogenic viral taxa were retained as potential pathogenic viruses. In addition, predicted hosts of DNA and RNA viruses were compared with the curated pathogenic-bacteria list. Viruses predicted to infect bacterial hosts belonging to the same species as known pathogenic bacteria were considered associated with potential pathogenic bacterial hosts.

Finally, candidate pathogens were further evaluated using phylogenetic and genomic evidence. For bacterial candidates, 16S rRNA gene sequences and/or MAGs were compared with closely related pathogenic and non-pathogenic reference taxa. For viral candidates, RdRp or other viral hallmark genes were used for phylogenetic placement. Candidate genomes were further annotated for pathogenicity-related features, including virulence factors, antimicrobial resistance genes and mobile genetic elements. The combined evidence from taxonomic assignment, sequence similarity, phylogenetic placement, host association and genome annotation was used to assess the potential pathogenic risk of microorganisms and viruses detected in Yellow River water (Supplementary Table S19).

Phylogenetic analysis

To further support the identification of putative pathogenic risk elements, phylogenetic analyses were performed for DNA viruses, RNA viruses and MAGs that were initially annotated as potentially pathogenic.

For DNA and RNA viruses, phylogenetic trees were constructed based on whole-genome nucleotide sequences. Taxonomic information and GenBank accession numbers were obtained from the ICTV Virus Metadata Resource, "VMR MSL40 v1 2025.xlsx"[62]. For each putative pathogenic viral sequence, whole-genome nucleotide sequences of reference viruses within the corresponding family were downloaded using efetch. In addition, two viruses belonging to different families within the same order were selected as outgroups. Sample-derived viral genomes and reference viral genomes were aligned using MAFFT[63]. The resulting alignments were trimmed using trimAl[64] to remove poorly aligned regions, and maximum-likelihood phylogenetic trees were constructed using IQ-TREE[65]. The trees were visualized and annotated using iTOL[66]. Putative pathogenic viral sequences were evaluated based on their phylogenetic placement relative to known pathogenic viral lineages.

For MAGs with potential pathogenic risk, reference genomes were selected from our curated pathogenic microorganism database. Genomes clustered with the target MAGs at 95% ANI using dRep were used as close reference genomes, and two genomes

from other categories in the database were selected as outgroups. The target MAGs, close reference genomes and outgroup genomes were then analysed using GTDB-Tk to generate a bacterial 120-marker-gene alignment (bac120.user_msa.fasta.gz). This alignment contained only the genomes included in this analysis. The alignment was trimmed using trimAl, and maximum-likelihood phylogenetic trees were inferred using IQ-TREE. The resulting trees were visualized and annotated using iTOL. MAGs were considered putative pathogenic MAGs when they clustered with known pathogenic reference genomes in the phylogenetic tree.

Risk index assessment of risk virus, pathogen, and ARGs based on relative abundance

The abundance of each risk virus, pathogen, and ARG (α_i) was calculated using TPM-normalized abundance derived from metagenomic or metatranscriptomic read mapping, normalized by total TPM across samples (n), as previously described [67]. The average abundance ($\bar{\alpha}_\iota$) was calculated as:

$$\bar{\alpha}_\iota = \frac{1}{n}\sum_{i=1}^{n}\alpha_i \quad (1)$$

For each risk virus, pathogen, and ARGs i, risk index (μ_i) represents its ability to transfer from water to others and was calculated as:

$$\mu_i = \bar{\alpha}_\iota \times \lambda_i \times \gamma_i \quad (2)$$

where μ_i is risk index of risk virus, pathogen, and ARGs i, $\bar{\alpha}_\iota$ is its average relative abundance, λ_i is its occupancy (the ratio of samples detected risk virus, pathogen, and ARGs i to total samples), and γ_i is its hazard classification formulated by the National Health Commission of the People's Republic of China (https://zwfw.nhc.gov.cn/kzx/zcfg/gzbxbywswsyhdsp_242/202007/t20200716_1677.html, in Chinese, translated in Supplementary Table S20-24).

For each sample, sample richness (β) was also incorporated, and the sample-level risk index (μ) was calculated as:

$$\mu = \beta \times \sum_{i=1}^{n}(\alpha_i \times \mu_i) \quad (3)$$

where μ is risk index of a sample, β is sample richness (Supplementary Table S25), α_i is abundance of risk virus, pathogen, and ARGs i in the sample, and μ_i is its risk index calculated using Eq. (2)[67].

Risk assessment of microbial risks based on absolute quantification data

The AQM-based framework based on absolute quantification were applied to calculate the risk indices of microbial risks as previously described[5]. Specifically, the risk elements included: (i) putative pathogenic viruses; (ii) putative pathogenic bacteria; (iii) faecal indicators, including Escherichia coli and enterococci; and (iv) high risk

antibiotic resistance genes (ARGs) and ARGs associated with ESKAPE+EV pathogens; and mobile genetic elements (MGEs), including plasmids and other mobilome-associated elements. Here, ESKAPE+EV refers to Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter spp., Escherichia coli and Vibrio spp. These taxa were included because they are frequently associated with multidrug resistance, environmental persistence and hospital-acquired infections.

The absolute abundance of each risk element was calculated as follows. For pathogenic viruses, the absolute abundance was estimated by multiplying the relative abundance of each viral taxon by the total virus-like particle abundance determined by epifluorescence microscopy:

$$N_x = RA_x \times V$$

where N_x is the absolute abundance of viral taxon x, RA_x is its relative abundance, and V is the total virus-like particle abundance.

For pathogenic bacteria, absolute abundance was calculated by multiplying the relative abundance of each bacterial taxon by total bacterial cell abundance measured by flow cytometry, followed by correction using the taxon-specific 16S rRNA gene copy number:

$$N_x = \frac{RA_x \times B}{rrn_x}$$

where B is the total bacterial cell abundance and rrn_x is the estimated 16S rRNA gene copy number of taxon x. For 12 key pathogens within the QMRA framework that were not recovered in metagenome-assembled genomes, their relative abundances were estimated using Kraken2-based taxonomic profiles. Briefly, all clean reads from all samples were taxonomically classified with Kraken2 against the PlusPF database, and the resulting classified reads assigned to each target pathogen were used to calculate its sample-specific abundance.

For ARGs and MGEs, absolute copy numbers were estimated by combining ARGs-OAP-derived gene abundance normalized to 16S rRNA genes with bacterial abundance and mean 16S rRNA gene copy number:

$$N_x = A_x \times B \times \overline{rrn}$$

where A_x represents the abundance of ARG or MGE x expressed as copies per 16S rRNA gene, B is the bacterial cell abundance measured by flow cytometry, and $\overline{rrn}$ is the mean 16S rRNA gene copy number estimated from qPCR. Plasmid copy numbers were directly quantified by qPCR.

All absolute abundances of risk elements were subsequently used to calculate the microbial risk index according to the framework described by Shi et al.[5] (Supplementary Table S26).

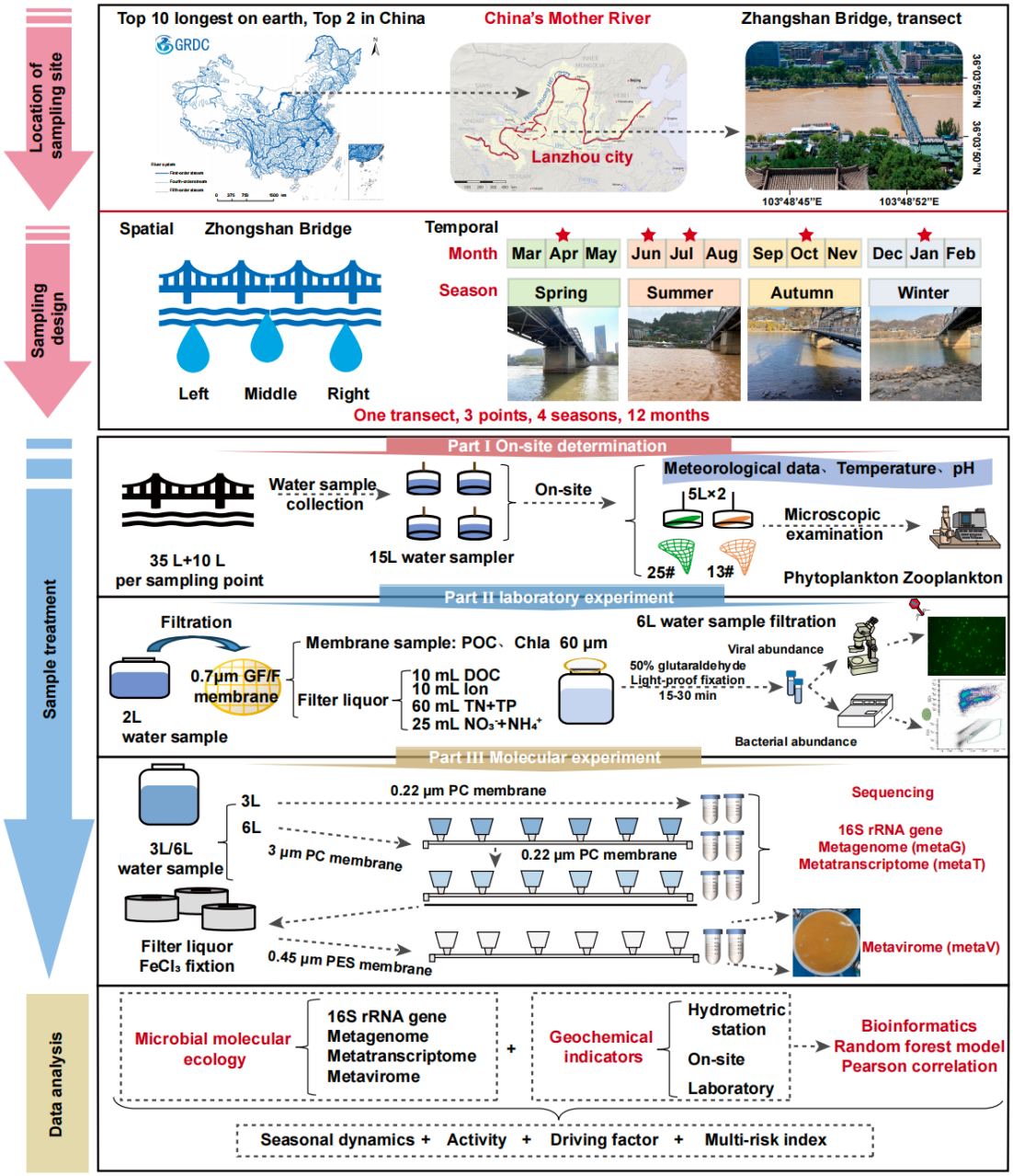
Determination of antibiotic concentrations

For each sampling site, three independent water samples were collected as biological replicates, with 500 mL of water collected for each replicate. After collection, the samples were immediately protected from light, transported to the laboratory under refrigerated conditions, and stored at 4°C until analysis. Antibiotic concentrations in water samples were quantified by the analytical service provider using liquid chromatography–tandem mass spectrometry. Each 500 mL water sample was enriched and concentrated through the established pretreatment procedure and finally reconstituted to 1.5 mL for instrumental analysis. Each biological replicate was analyzed with two technical injections, and the averaged value of the technical replicates was used for downstream statistical analyses.

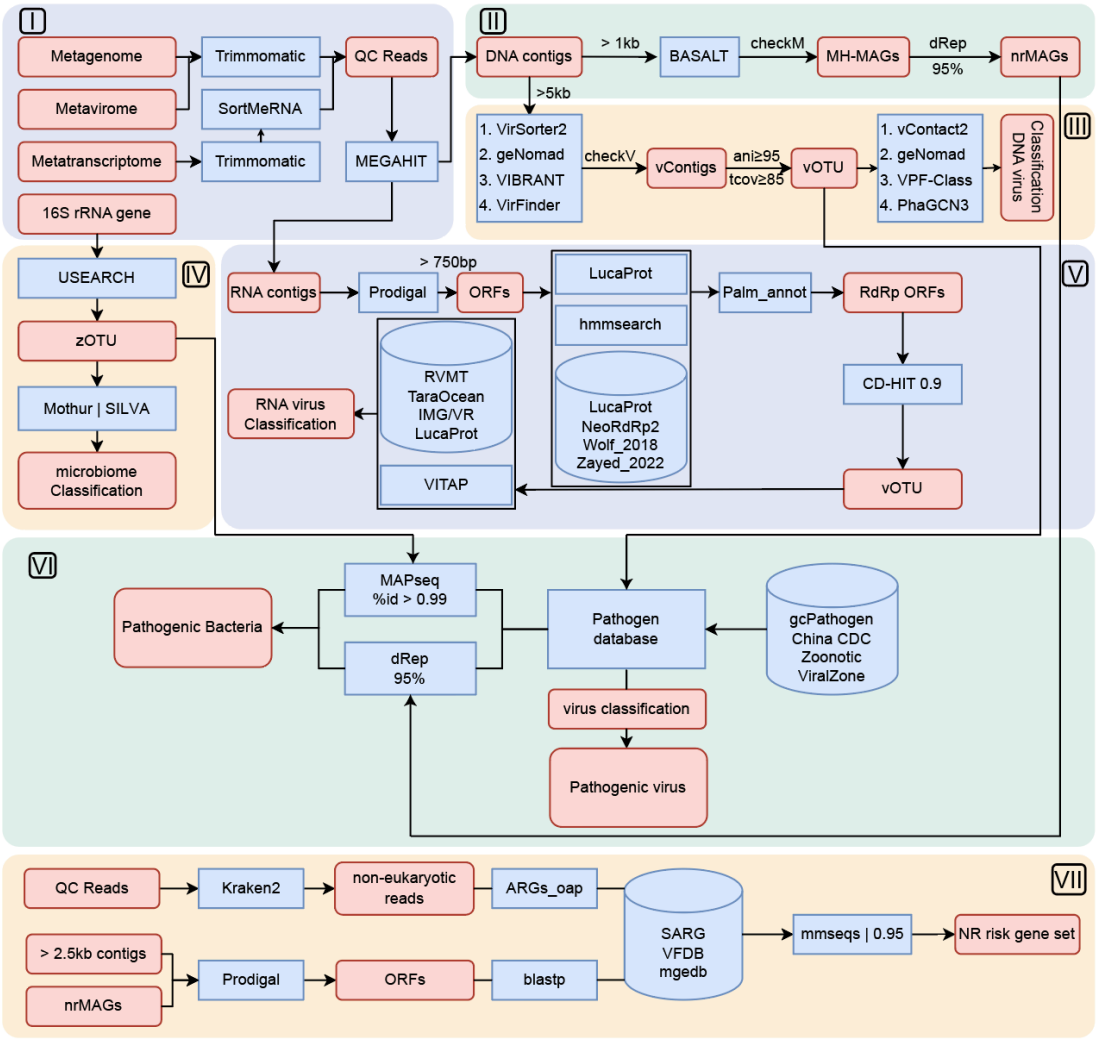
A total of 48 target antibiotics were determined, covering sulfonamides, quinolones, tetracyclines, macrolides, β-lactams, lincosamides, amphenicols, and other antimicrobial agents. Sample pretreatment mainly included filtration, solid-phase extraction, elution, nitrogen evaporation, and reconstitution. Chromatographic separation was performed using a C18 column, and mass spectrometric detection was conducted with an electrospray ionization source in multiple reaction monitoring mode. Individual target compounds were quantified using compound-specific calibration curves. The limits of detection and limits of quantification were provided by the analytical service provider and corresponded to concentrations at signal-to-noise ratios of 3 and 10, respectively.

Human activity pressure mapping

Human activity pressure in the upstream Yellow River basin above the Zhongshan Bridge control section in Lanzhou was assessed using the global annual terrestrial Human Footprint dataset of Mu et al.[68]. The Zhongshan Bridge section was used as the outlet control point to delineate the upstream contributing basin. Human Footprint raster data within the basin boundary were extracted and classified following the thresholds proposed by Venter et al.[69], generating four mapped pressure categories: low, moderate, high, and very high human activity pressure. The area and proportion of each pressure category were calculated within the basin. The classified Human Footprint layer was then overlaid with the Yellow River mainstream, major rivers, cities, and sub-basin boundaries, and the final map was produced in ArcGIS.

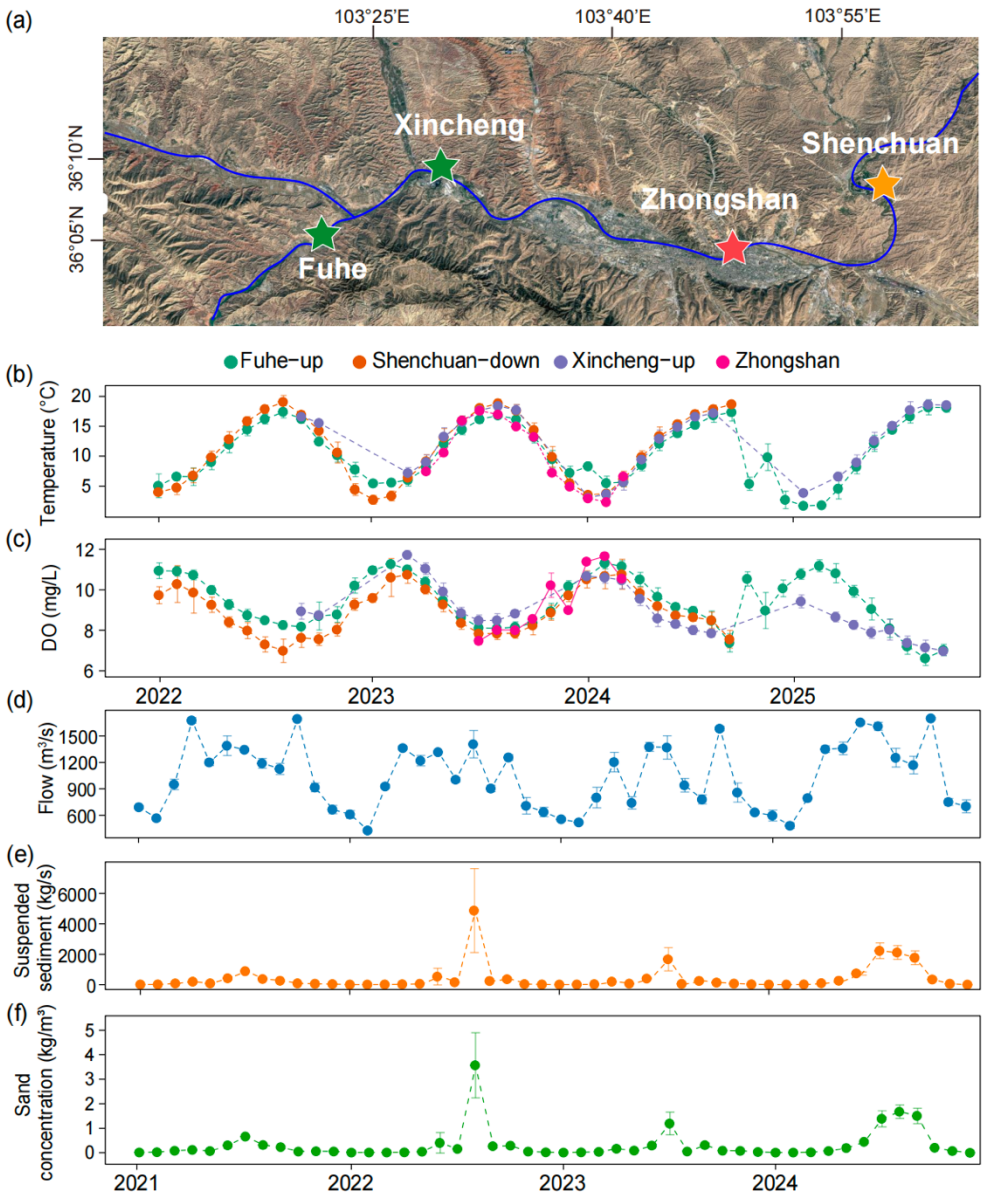


Supplementary Figure 1. River water sampling and sample preparation workflows. This figure outlines the key steps, including sampling point selection, seasonal sampling design, sample pretreatment, physicochemical parameter measurement, microscopic classification and counting of zooplankton and phytoplankton, microscopic enumeration of viral-like particles, nucleic acid extraction and sequencing, and data analysis. Red stars on top of the Month indicate representative samples for metagenome sequencing and high-throughput qPCR of ARGs and MGEs.

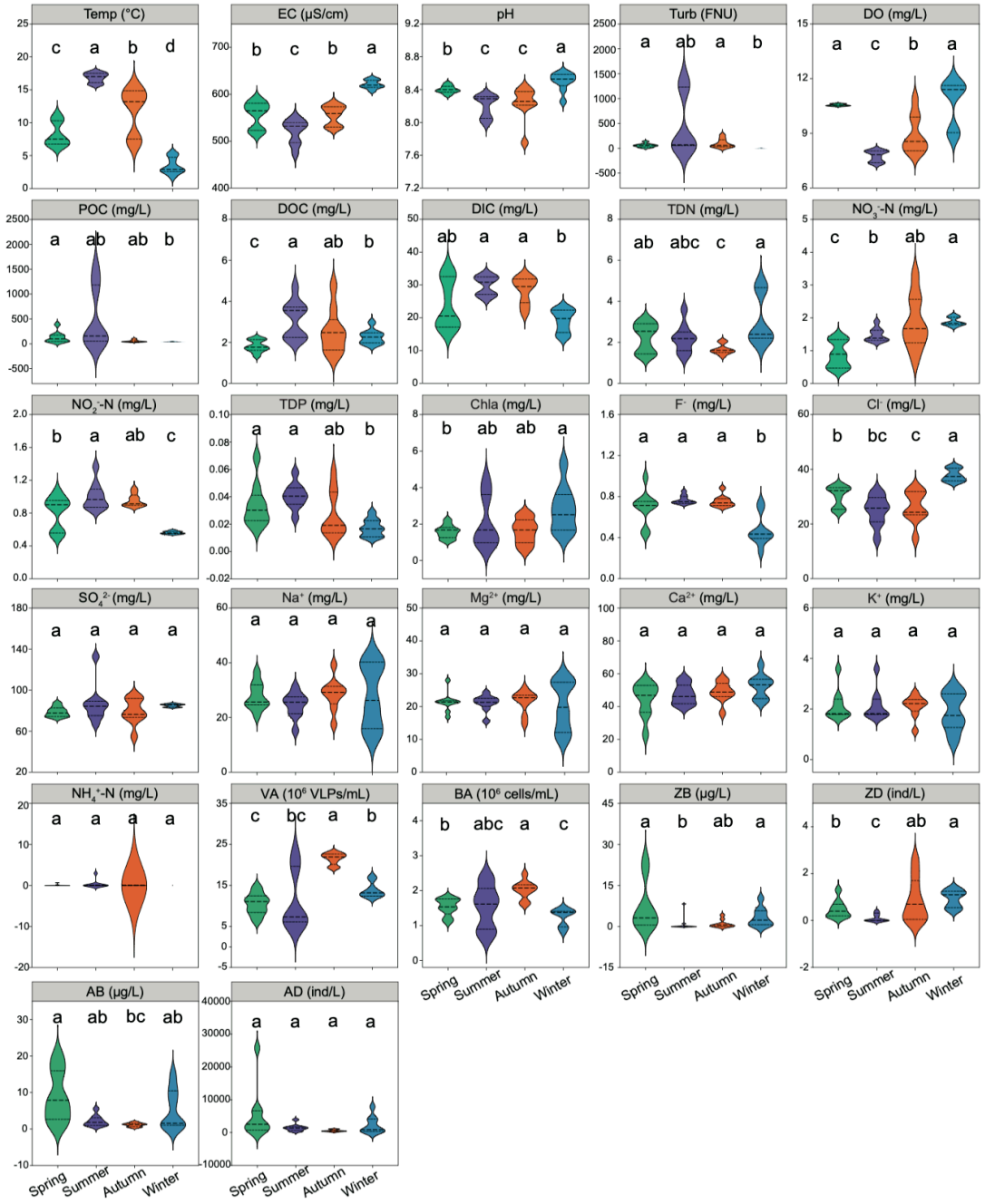


688 **Supplementary Figure 2. Bioinformatic workflow.**

689 This workflow illustrates the main steps (I-VII) of bioinformatics pipelines, mainly including:
690 processing sequencing reads; assembly, genome binning, and MAGs classification; DNA and
691 RNA virus identification and classification; pathogen database construction and identification
692 of potential pathogens, including bacterial, DNA virus, and RNA virus pathogens; and risk-
693 associated genetic elements.

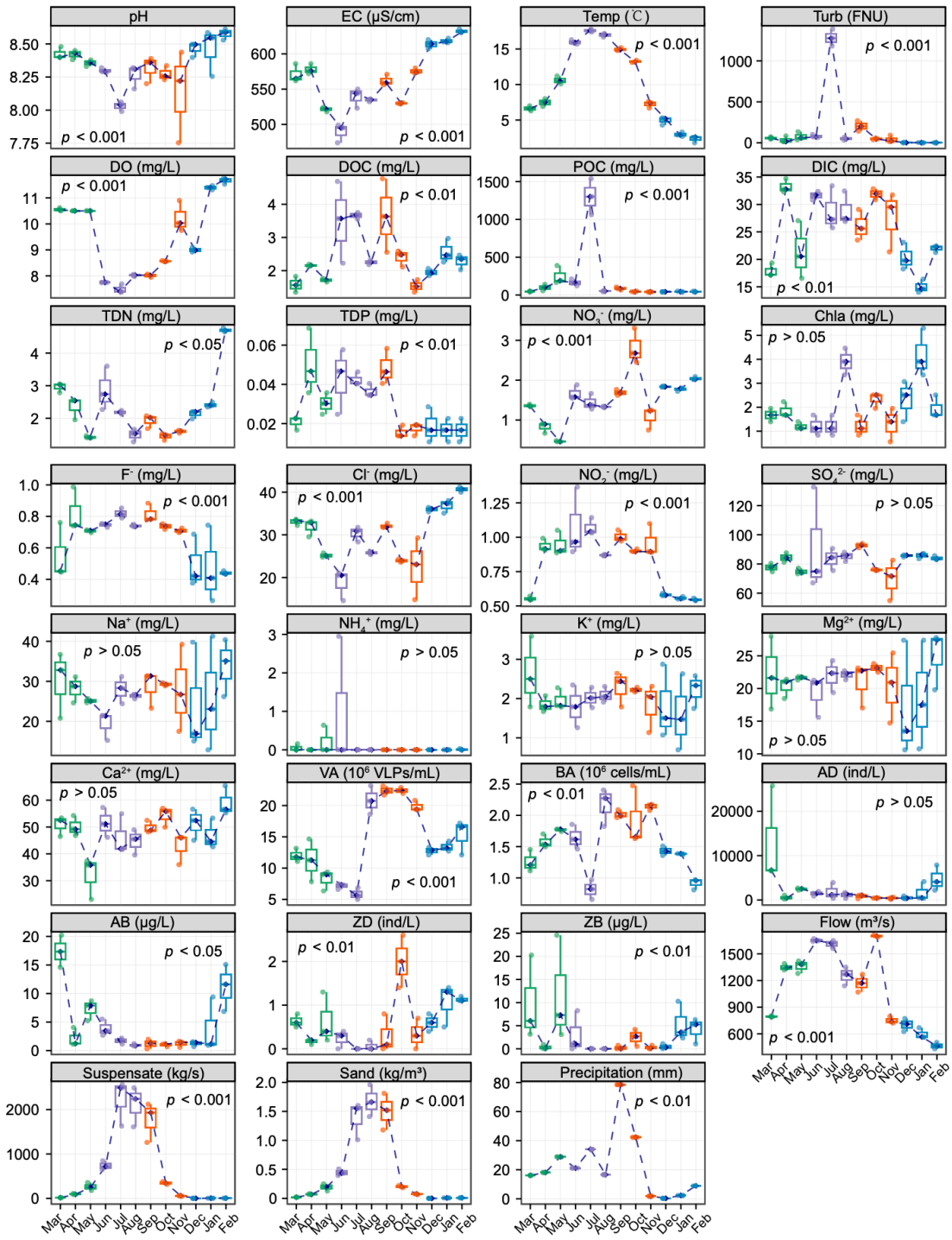


694 **Supplementary Figure 3. Comparison of river water parameters from different hydrology**
695 **stations.**
696 **a**, location of hydrological stations; **b-c**, comparison of temperature and DO data from different
697 stations with our own data; **d-f**, flow, suspended sediment concentration and sand concentration
698 data from Zhongshan hydrological stations. To evaluate whether the Zhongshan Bridge site
699 accurately represents hydro-ecological conditions in the Lanzhou reach of the Yellow River,
700 we compiled publicly available physicochemical and hydrological data from monitoring
701 stations within a 42.5-km buffer around the site. The resulting dataset includes DO and water
702 temperature parameters from Shenchuan, Fuhe, and Xincheng monitoring stations, together
703 with precipitation, flow, suspended sediment and sand concentration measurements from the
704 Zhongshan Bridge hydrological station. The Shenchuan, Fuhe, and Xincheng stations provided
705 monthly observations spanning 2022 to 2025, whereas the Zhongshan Bridge station recorded
706 daily hydrological dynamics over the same period. To ensure consistency across datasets with
707 differing temporal resolutions, the daily observations were summarized as monthly-scale values
708 by averaging, to characterize monthly-scale variability. Error bars denote standard deviation (n
709 = 3) (see Supplementary Results and Table S2).



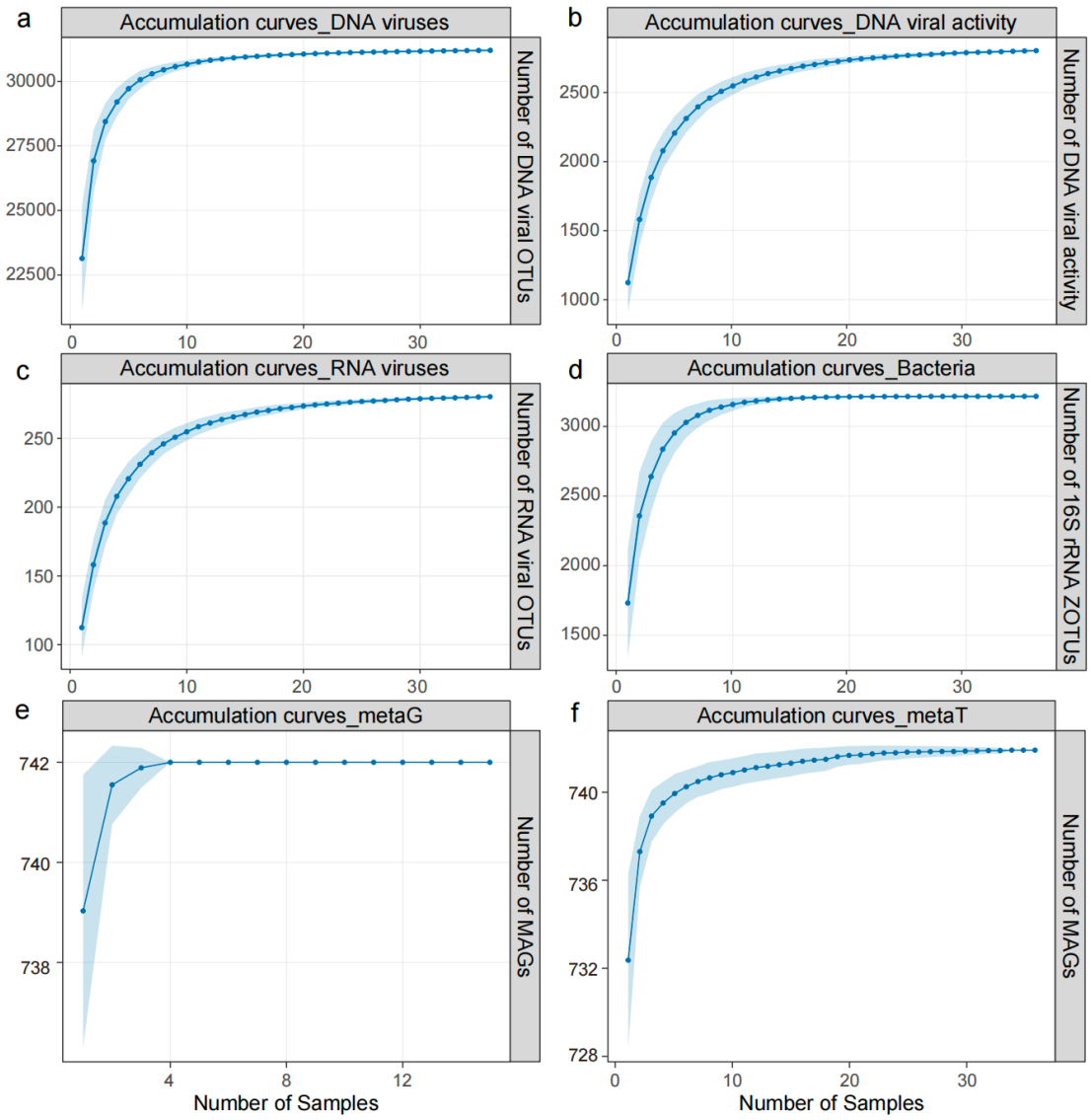
Supplementary Figure 4. Seasonal variation of biotic and abiotic factors in the Yellow River water.

Data that do not share a common letter are significantly different between seasons ($P < 0.05$; multiple comparisons examined with Kruskal-Wallis tests). The unit was put on top of each panel. Notes: EC: electrical conductivity; Temp: temperature; Turb: Turbidity; DO: Dissolved oxygen; DOC: Dissolved Organic Carbon; POC: Particulate Organic Carbon; DIC: Dissolved Inorganic Carbon; TDN: Total Dissolved Nitrogen; TDP: Total Dissolved Phosphorus; NO_3^-:Nitrate Nitrogen; Chla: Chlorophyll a; F^-: Fluoride Ion; Cl^-: Chloride Ion; NO_2^-: Nitrous Oxide Nitrogen; SO_4^{2-}: Sulfate Ion; Na^+: Sodium Ion; NH_4^+: Ammonium Nitrogen; K^+: Potassium Ion; Mg_2^+: Magnesium Ion; Ca_2^+: Calcium Ion; VA: Virus-like Particles abundance; BA: Bacterial Abundance; AD: Algal Density; AB: Algal Biomass; ZD: Zooplankton Density; ZB: Zooplankton Biomass.



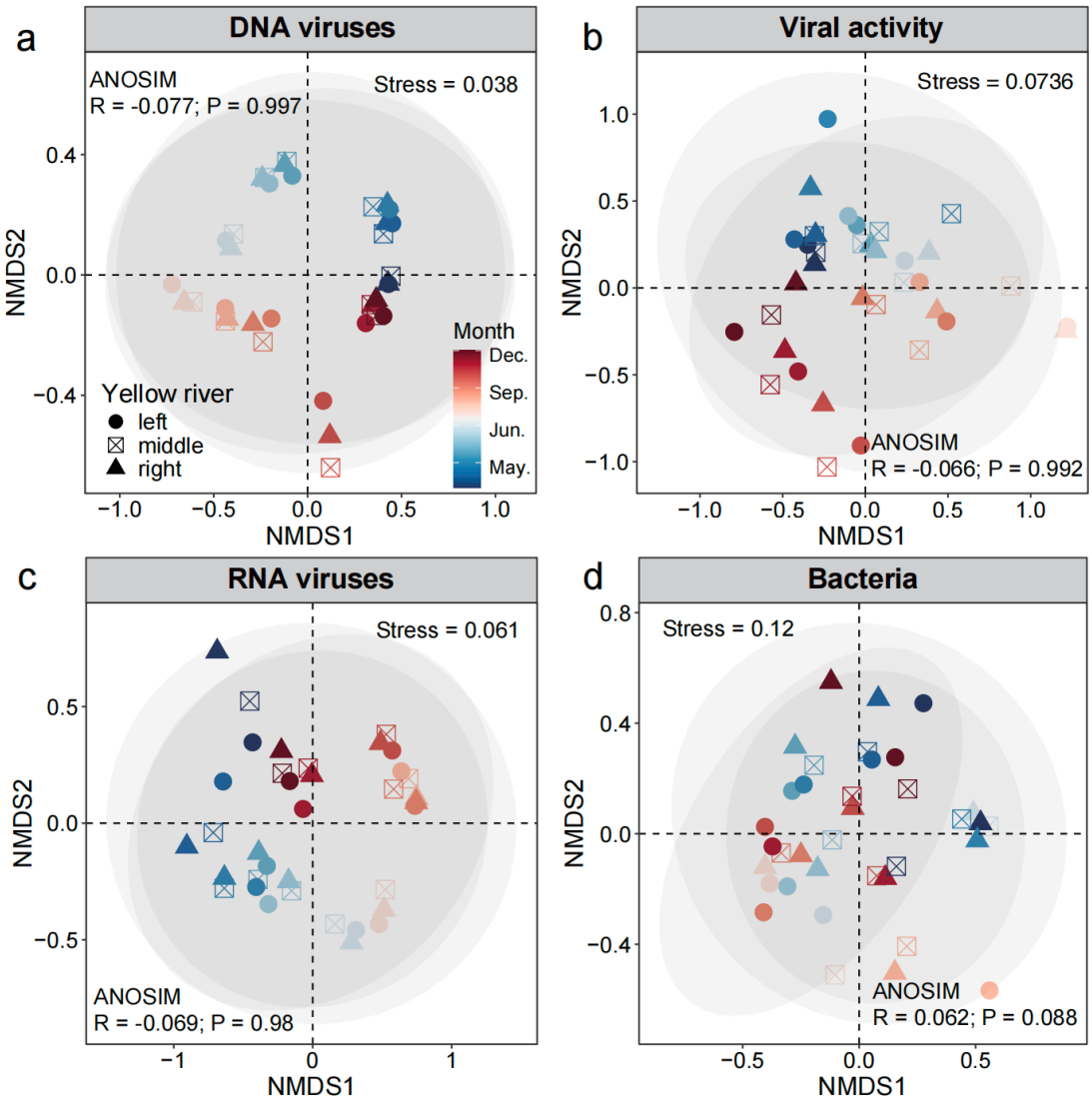
Supplementary Figure 5. Dynamics of biotic and abiotic factors across months in the Yellow River water.

Kruskal-Wallis test were used to examine the significant difference between seasons, with $P < 0.05$ indicating a significant difference between seasons. Notes: Flow, runoff.



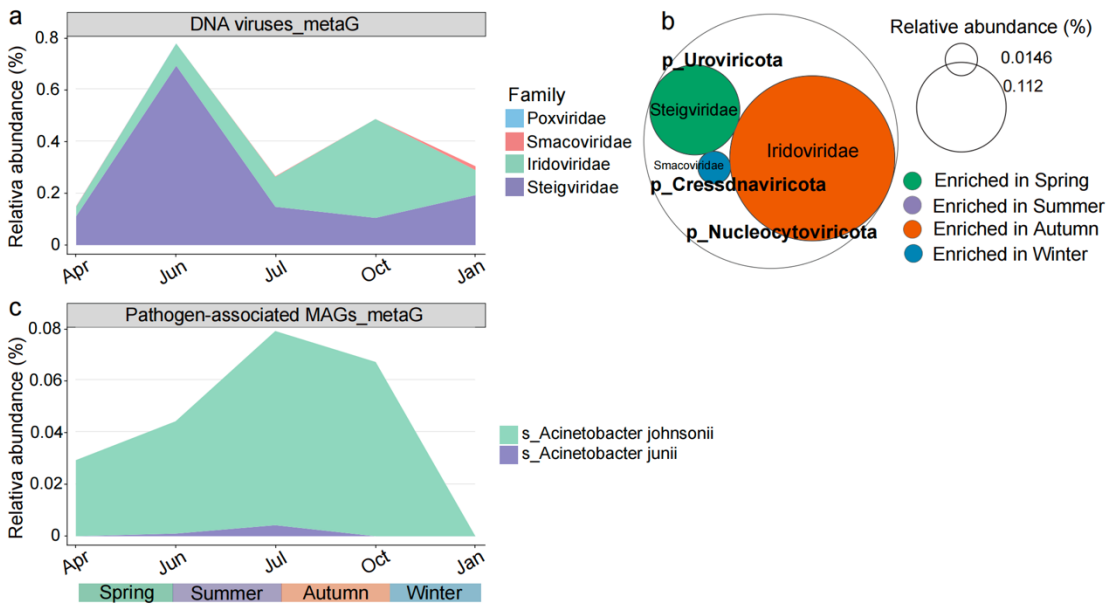
Supplementary Figure 6. Rarefaction curves.

Species accumulation curves are shown for DNA viral OTUs (vOTUs) based on metavirome (metaV) sequences (**a**) and metatranscriptome (metaT) sequences (**b**), RNA viral OTUs (vOTUs) based on metaT sequences (**c**), 16S rRNA zero-radius OTUs (ZOTUs) based on 16S rRNA amplicon sequences (**d**), mOTUs for MAGs based on metagenome (metaG) (**e**), metaT sequences data (**f**). Dots indicate the mean number of OTUs across 36 samples (except for panel **e**, which includes 15 samples) in Yellow River water, with error ribbons representing 95% confidence intervals.

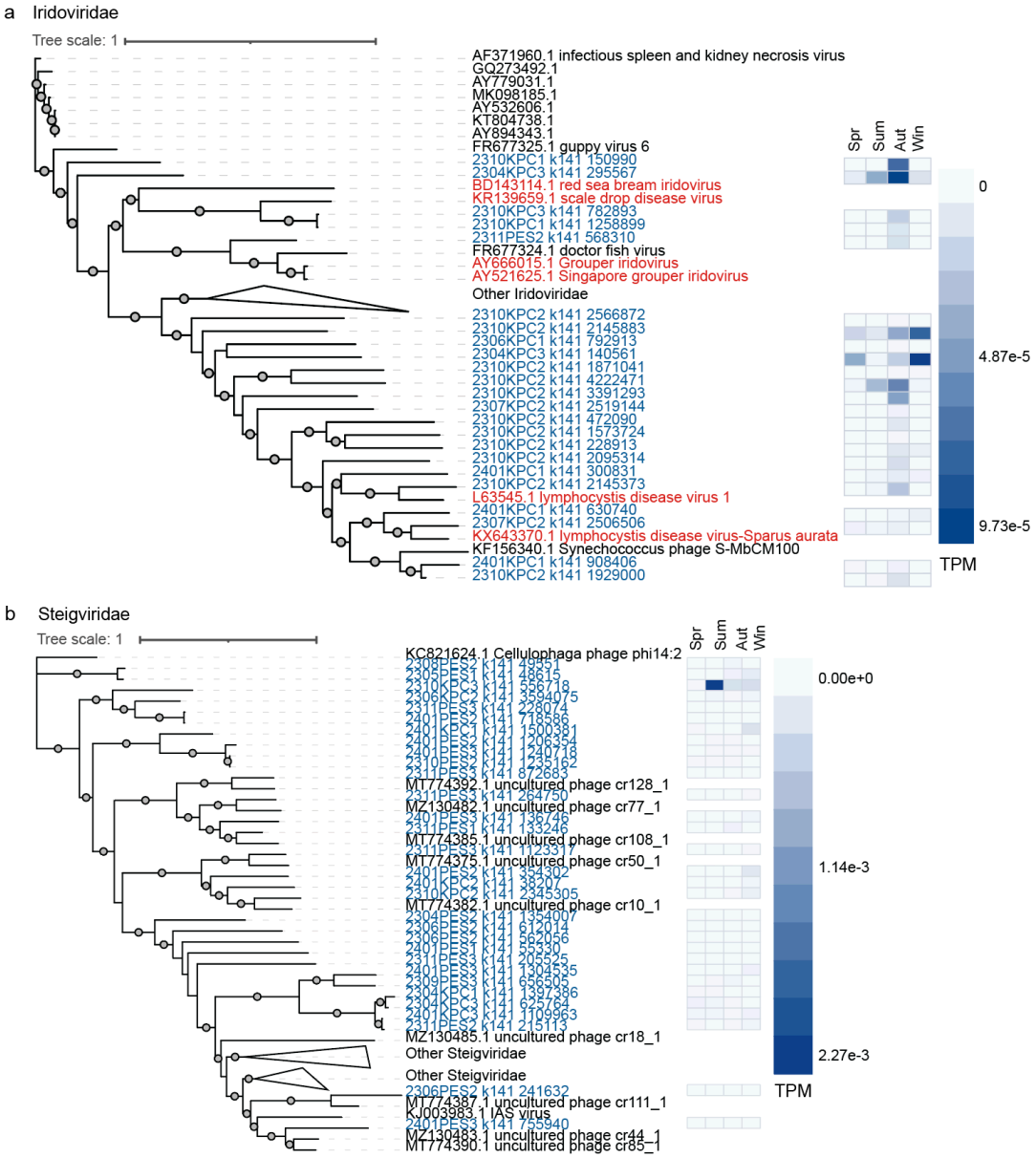


Supplementary Figure 7. Dissimilarity of samples from the left, middle and the right sampling sites.

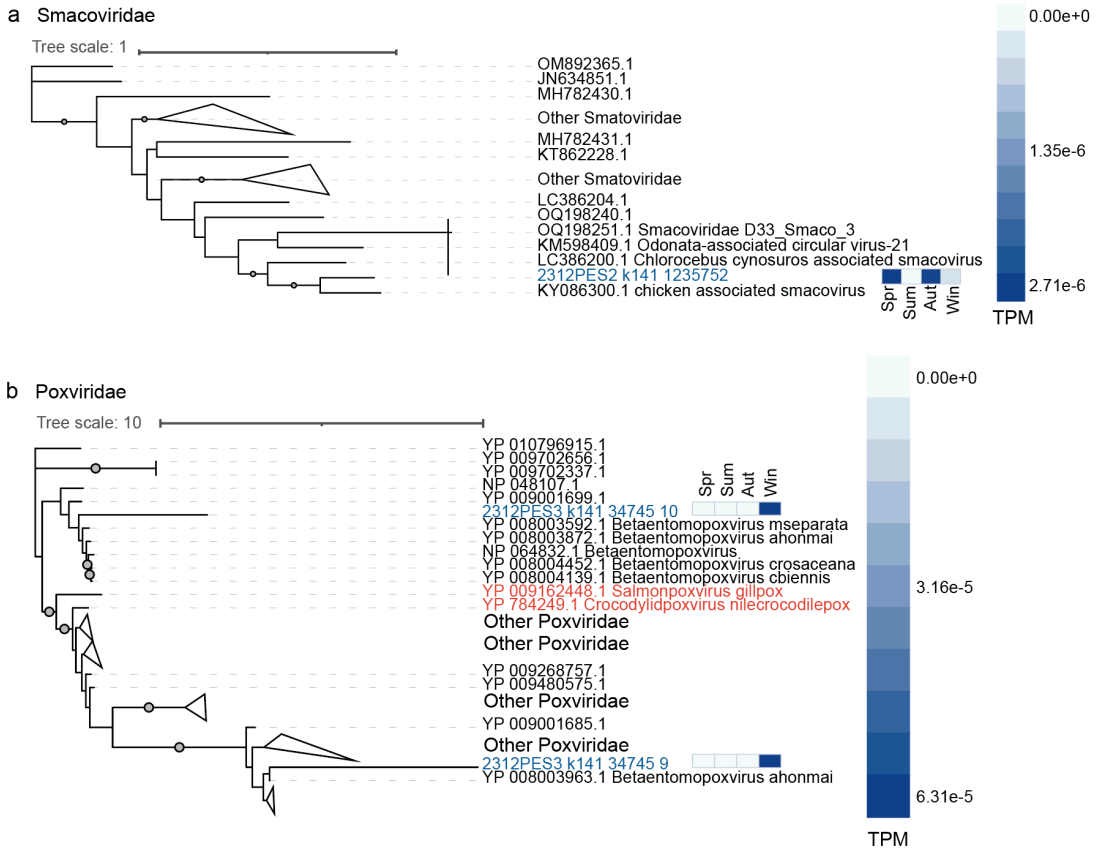
NMDS shows the clustering of samples based on the Bray-Curtis distance matrix of: relative abundance of DNA viral pathogens (**a**), relative transcripts abundance DNA viral pathogens (**b**), relative transcripts abundance of RNA viral pathogens (**c**), and relative abundance of potential bacterial pathogens (**d**), respectively. Panel a, b-c, d are based on metavirome, metatranscriptome and 16S rRNA gene amplicon sequences data, respectively. 95% confidence ellipses were shown around the samples grouped based on the left, middle, and right sampling sites. Statistical significance was examined using the ANOSIM test.



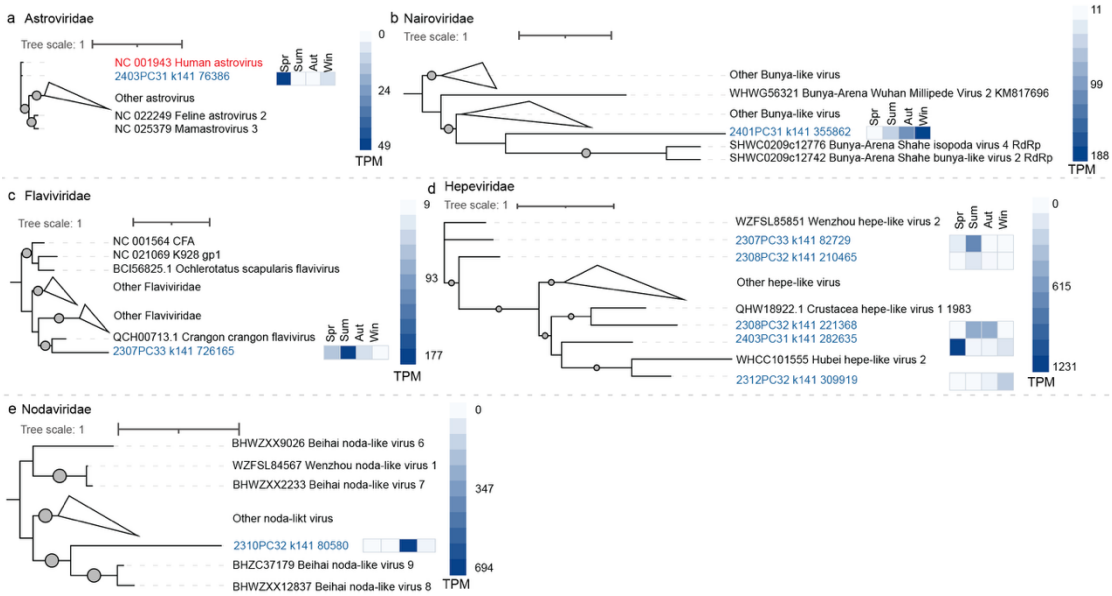
Supplementary Figure 8. Seasonal dynamics and biomarkers of potential DNA viral and bacterial pathogens in the Yellow River water as revealed by metagenomic sequence data. **a–b**, Seasonal dynamics and biomarkers of potential pathogenic DNA viruses based on metagenomic reads mapped to viral operational taxonomic units (vOTUs). **c**, Seasonal dynamics of potential bacterial pathogens based on metagenome reads mapped to metagenome operational taxonomic units (mOTUs). Seasonal biomarkers were identified using LEfSe analysis. Circle size is proportional to the relative abundance of each taxon. Note, the data is not enough for LEfSe analysis of bacterial pathogens' biomarkers.



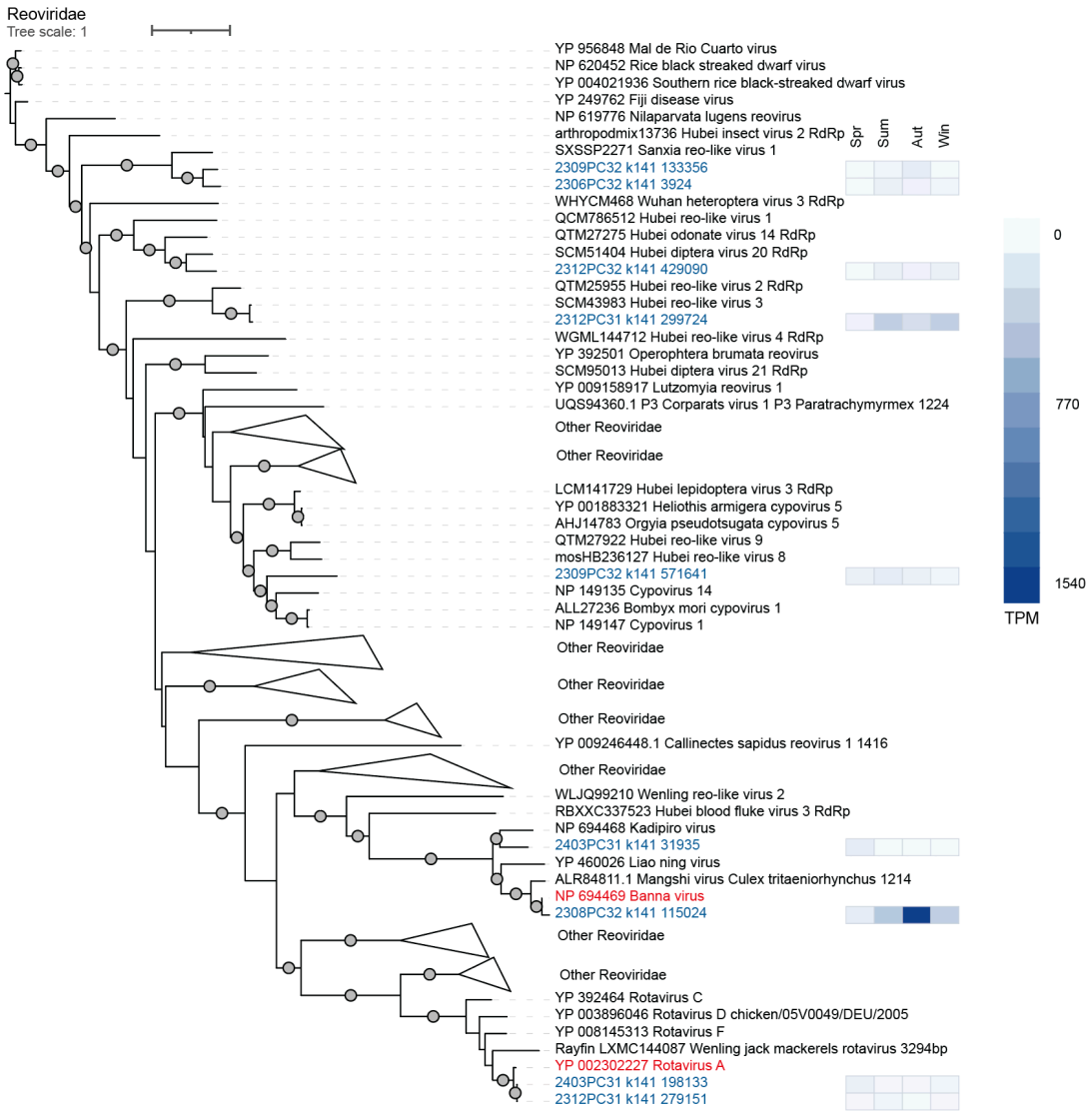
Supplementary Figure 9. Phylogenetic trees of potential DNA viral pathogens in Yellow River water: *Iridoviridae* and *Steigviridae*. a-b: Evolutionary relationships of *Iridoviridae* and *Steigviridae* with potential pathogenic risks. Each phylogenetic tree was constructed using a maximum likelihood method. Within each phylogeny, viral sequences recovered from the samples are shown in blue, and reference sequences with potential pathogenic risks are shown in red. Grey circles indicate branches with bootstrap support > 0.95. The relative activity of sample-derived viruses is displayed as a heatmap to the right of each tree.



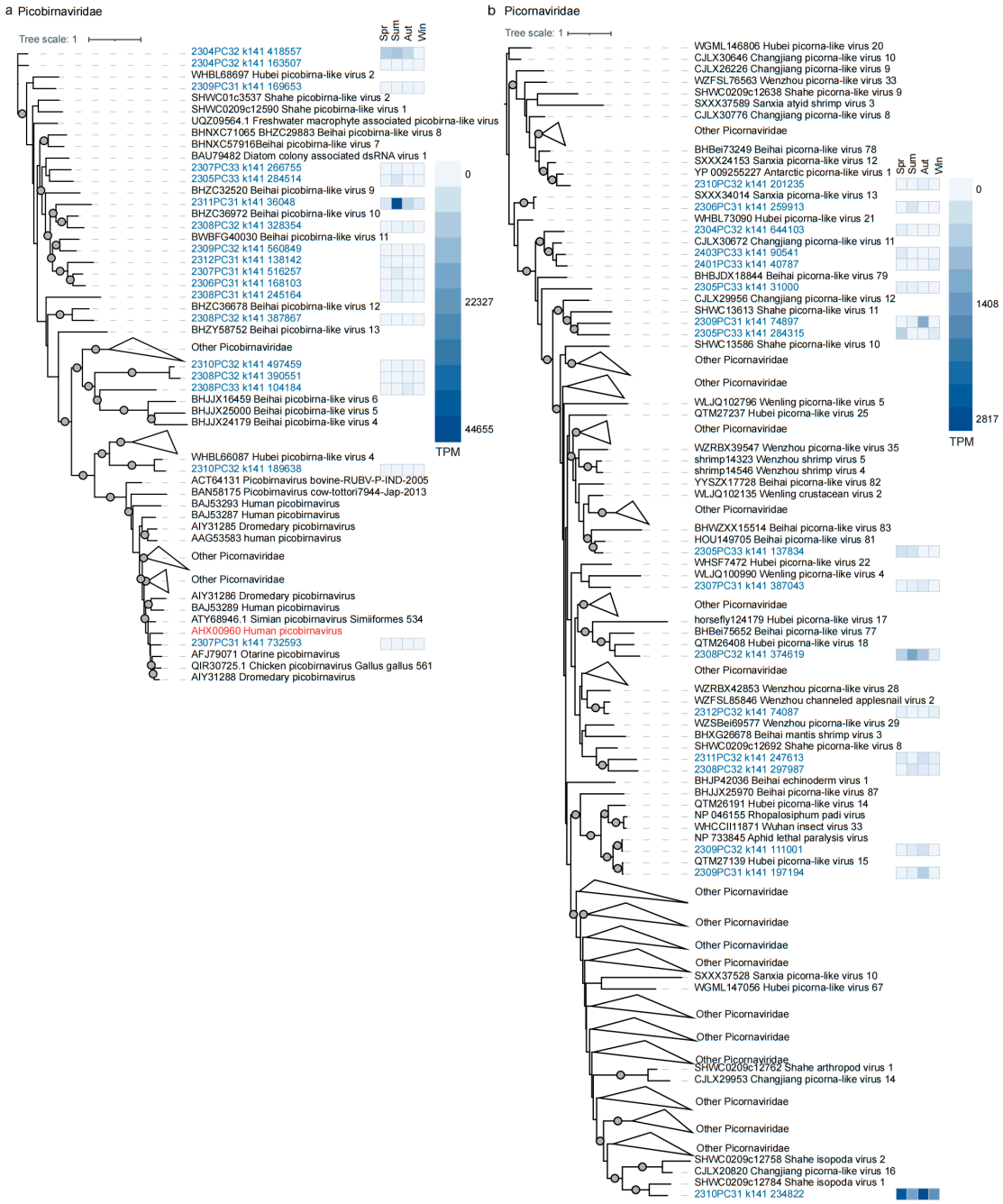
Supplementary Figure 10. Phylogenetic trees of potential DNA viral pathogens in Yellow river water: *Smacovirdae* and *Poxviridae*. a–b: Maximum-likelihood phylogenetic trees of *Smacoviridae* and *Poxviridae*. Sample-derived viral contigs are shown in blue, and reference sequences associated with potential pathogenic risks are shown in red. Grey circles indicate branches with bootstrap support > 0.95. Heatmaps to the right of each tree show seasonal relative activity of sample-derived viral contigs.



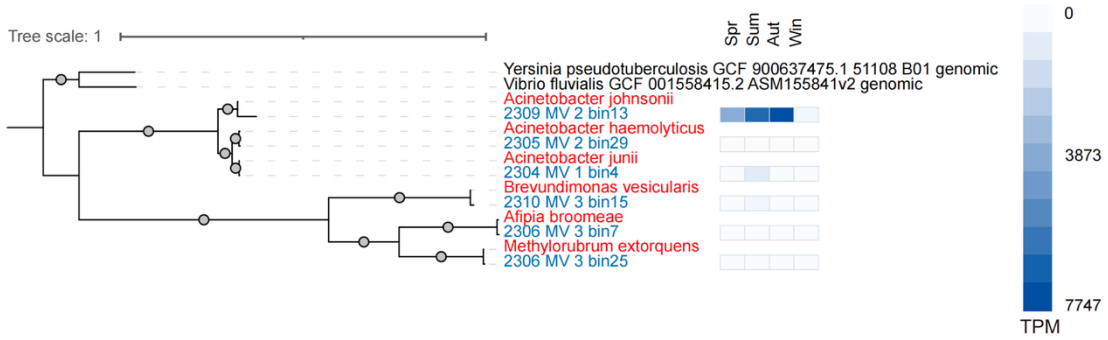
Supplementary Figure 11. Phylogenetic trees of potential RNA viral pathogens in Yellow river water: *Astroviridae*, *Nairoviridae*, *Flaviviridae*, *Hepeviridae*, and *Nodaviridae*.
a–e: Maximum-likelihood phylogenetic trees of five RNA virus families. Sample-derived viral contigs are shown in blue, and reference sequences associated with potential pathogenic risks are shown in red. Grey circles indicate branches with bootstrap support > 0.95. Heatmaps to the right of each tree show seasonal relative activity of sample-derived viral contigs.



Supplementary Figure 12. Phylogenetic tree of potential RNA viral pathogens in Yellow river water: *Reoviridae*. Maximum-likelihood phylogenetic trees of *Reoviridae*. Sample-derived viral contigs are shown in blue, and reference sequences associated with potential pathogenic risks are shown in red. Grey circles indicate branches with bootstrap support > 0.95. Heatmaps to the right of each tree show seasonal relative activity of sample-derived viral contigs.

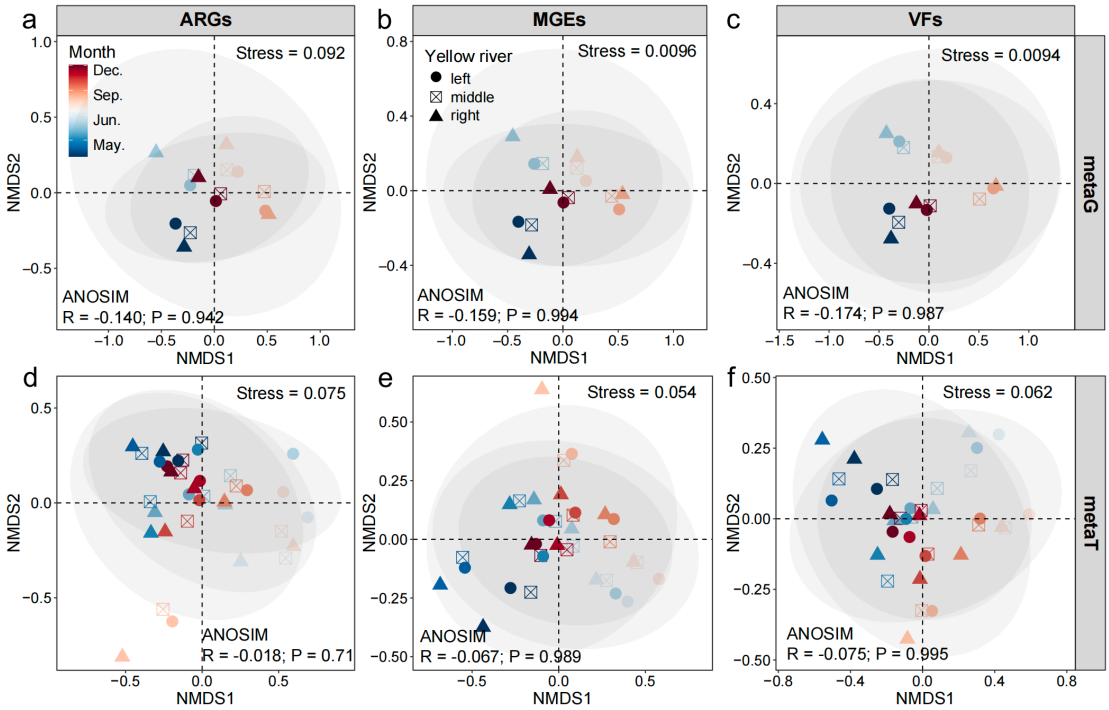


Supplementary Figure 13. Phylogenetic tree of potential RNA viral pathogens in Yellow river water: *Picobirnaviridae* and *Picornaviridae*. a–b: Maximum-likelihood phylogenetic trees of *Picobirnaviridae* and *Picornaviridae*. Sample-derived viral contigs are shown in blue, and reference sequences associated with potential pathogenic risks are shown in red. Grey circles indicate branches with bootstrap support > 0.95. Heatmaps to the right of each tree show seasonal relative activity of sample-derived viral contigs.

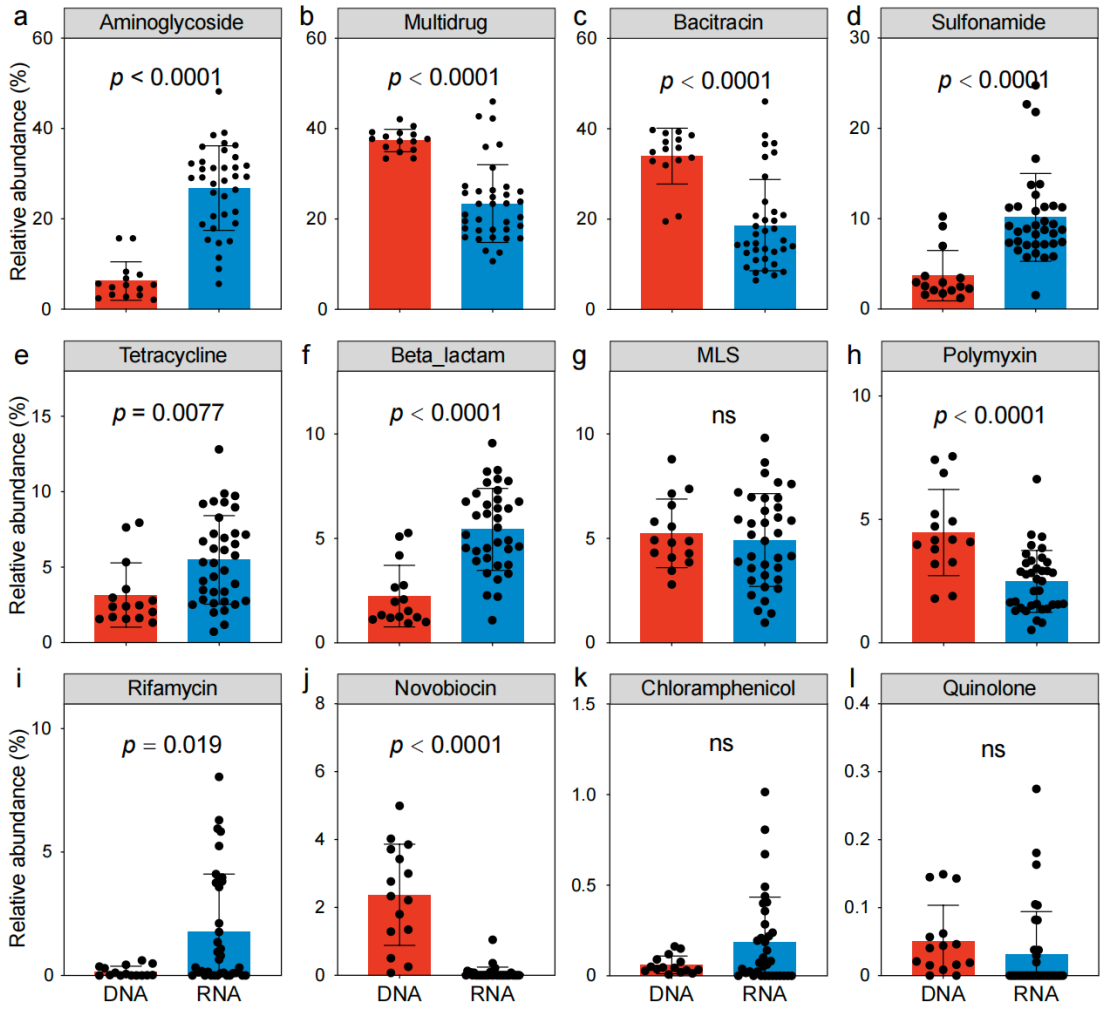


Supplementary Figure 14. Phylogenetic trees of metagenome-assembled genomes (MAGs) belonging to potential bacterial pathogens in Yellow river water.

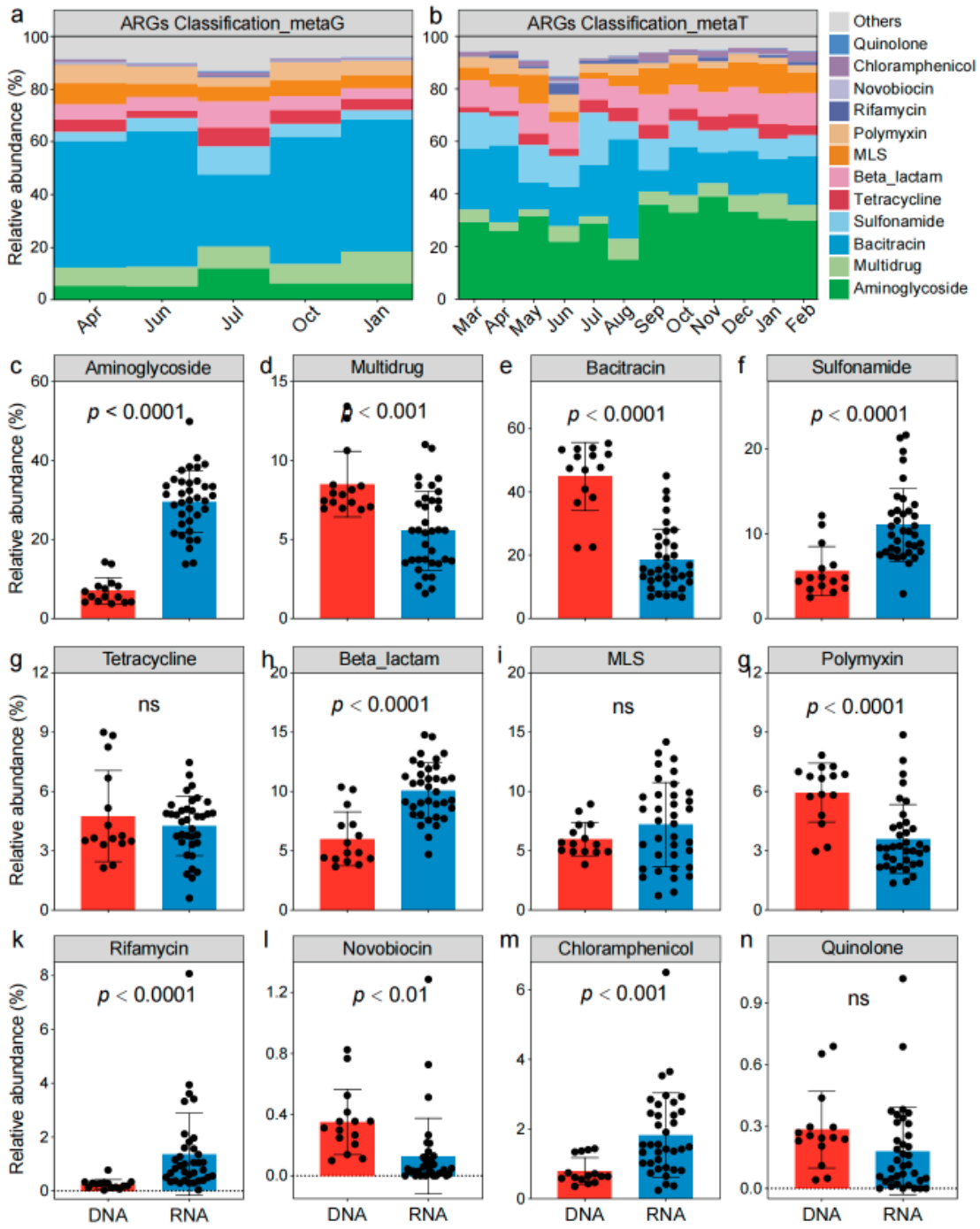
Phylogenetic relationships of six potentially pathogenic MAGs and related reference genomes. The phylogenomic tree was inferred using a maximum likelihood method. The tree includes six MAGs from the samples, the reference genomes clustered with them, and two outgroup genomes. MAGs from the samples are shown in blue, and potentially pathogenic reference genomes are shown in red. Grey circles indicate branches with bootstrap support > 0.95. The relative activity of sample-derived MAGs is displayed as a heatmap to the right of each tree.



795 **Supplementary Figure 15. Dissimilarity of risk-associated genetic element communities**
796 **from the left, middle and the right sampling sites.**
797 NMDS shows the structure of microbial community for ARGs, MGEs and VFs based on
798 metagenome (metaG) reads mapping (**a-c**) and metatranscriptome (metaT) reads mapping (**d-**
799 **f**). 95% confidence ellipses were shown around the samples grouped based on the left, middle,
800 and right sampling sites. Statistical significance was examined using the ANOSIM test.

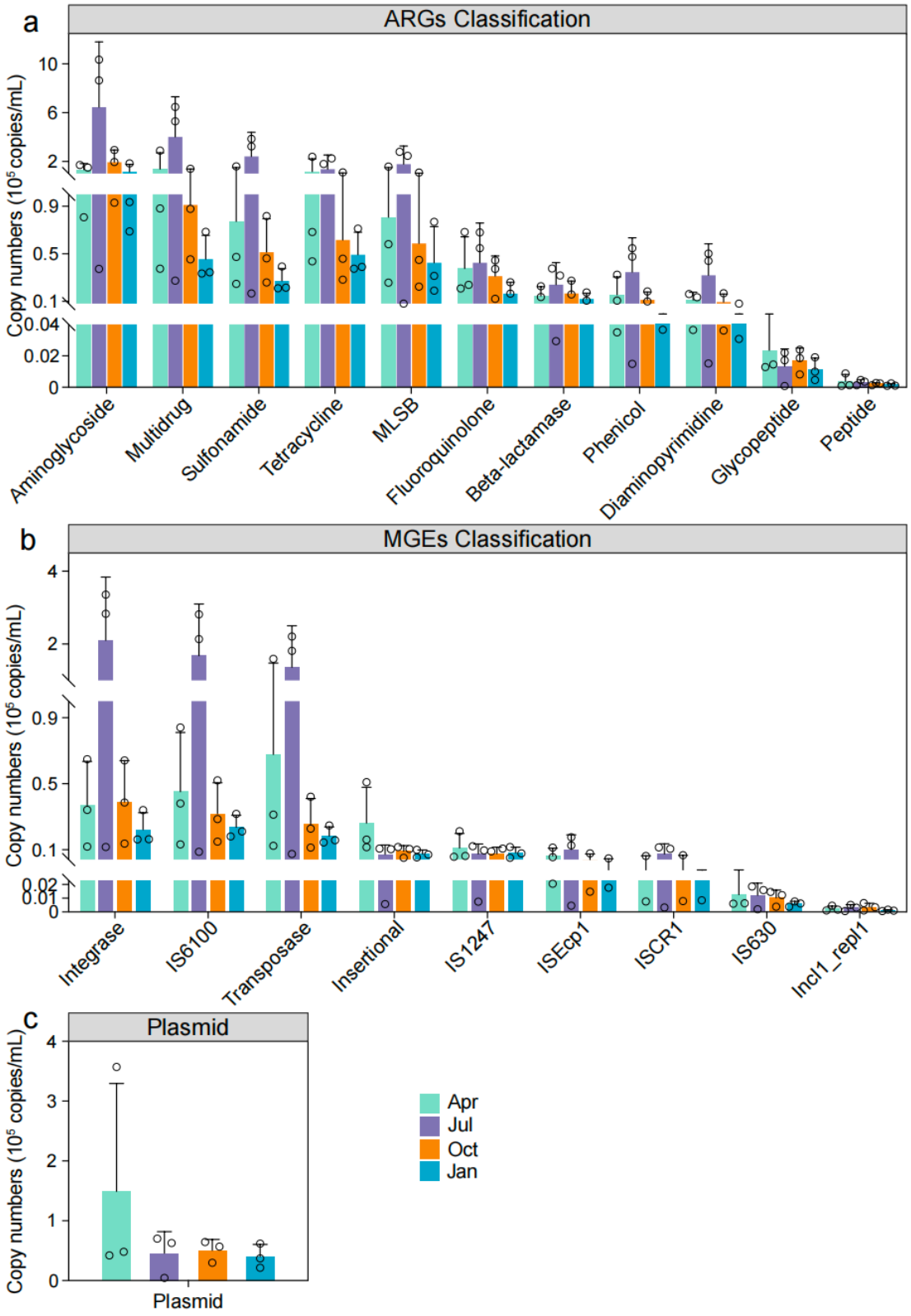


Supplementary Figure 16. Relative abundance of ARGs in the Yellow River water. **a–l**, Bar plots show the differences in the relative abundance of each ARG type between the DNA (metagenome-based) and RNA (metatranscriptome-based) levels. The data represent reads mapped to nucleotide sequences of ARGs identified from all contigs. "DNA" and "RNA" indicate that relative abundances were calculated based on metagenome and metatranscriptome reads mapped to genes, respectively. Comparisons between DNA and RNA samples were performed using the Wilcoxon rank-sum test (n = 15 for DNA, n = 36 for RNA). $P < 0.05$ indicate a significant difference; ns, not significant. MLS, Macrolide–Lincosamide–Streptogramin.

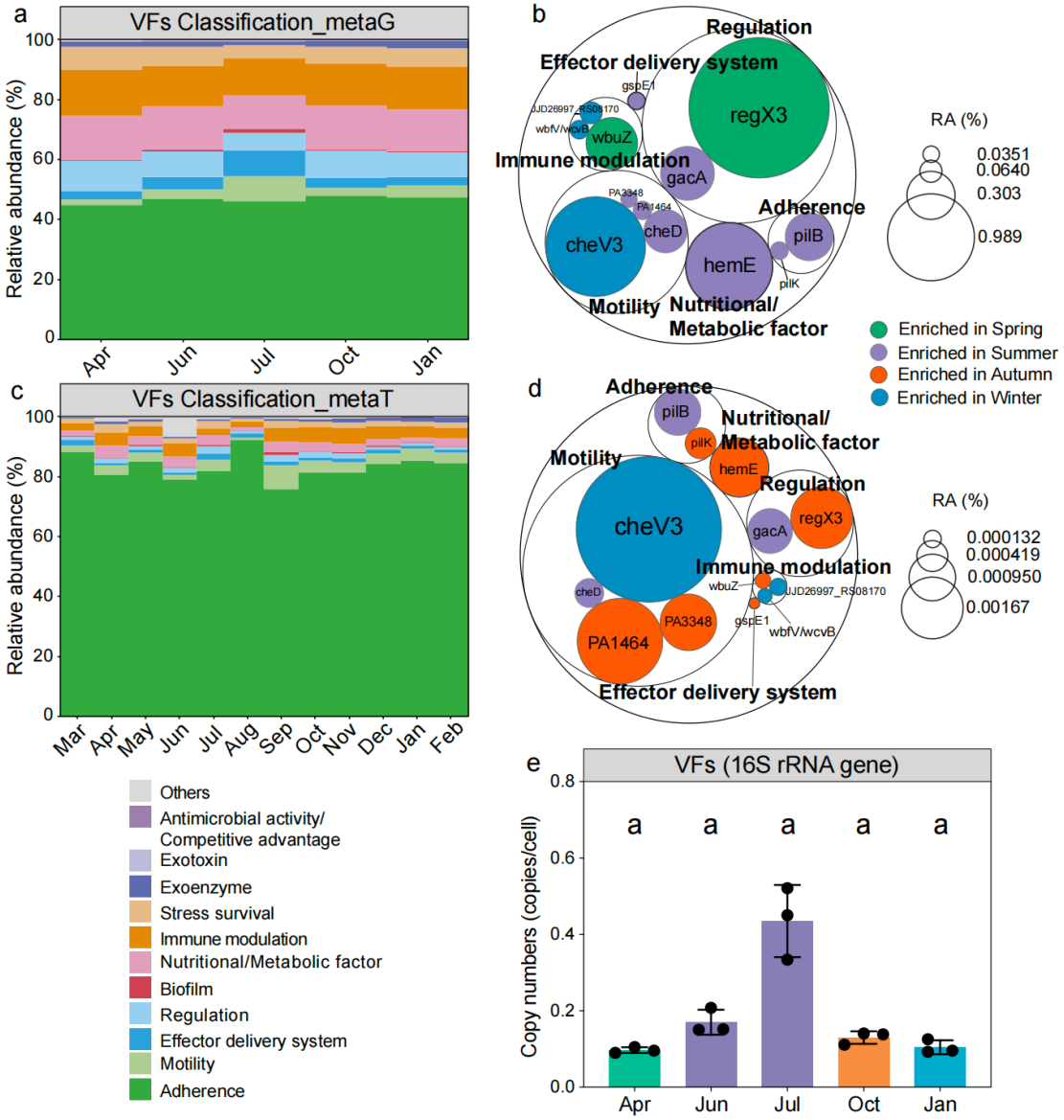


Supplementary Figure 17. Cmposition of ARGs in Yellow River water based on read-based ARG profiling using the ARGs-OAP v3 pipeline.

a, Composition of ARG types inferred from quality-filtered metagenomic reads using ARGs-OAP v3. **b**, Composition of ARG types inferred from quality-filtered metatranscriptomic reads using ARGs-OAP v3. **c–n**, Comparisons of the relative abundance of each ARG type between DNA-level and RNA-level ARG profiles. "DNA" and "RNA" indicate ARG profiles generated from metagenomic and metatranscriptomic reads, respectively, using ARGs-OAP v3. Comparisons between DNA and RNA samples were performed using the Wilcoxon rank-sum test (n = 15 for DNA and n = 36 for RNA). $P < 0.05$ indicates a significant difference; ns, not significant. MLS, macrolide–lincosamide–streptogramin.

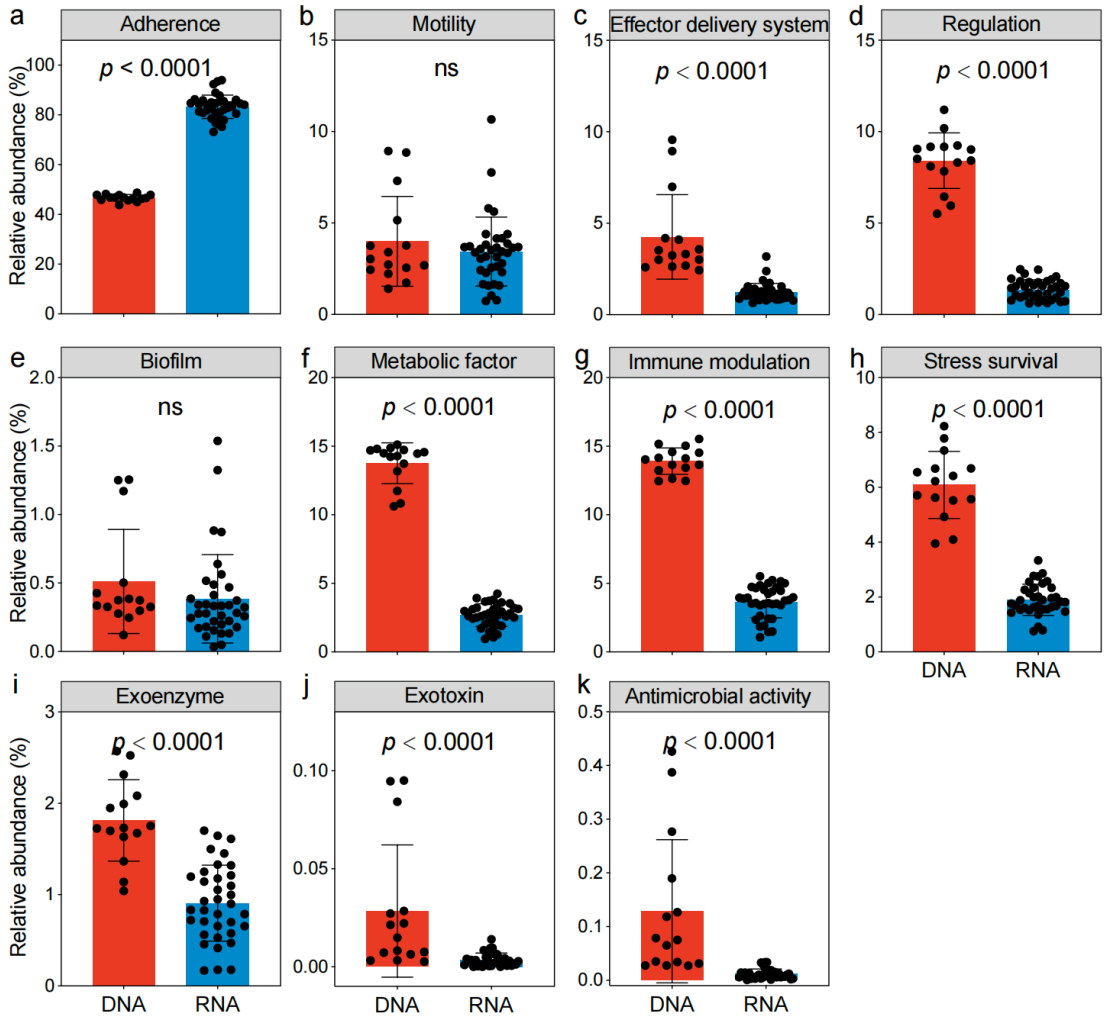


820 **Supplementary Figure 18. Absolute copy number of ARGs, MGEs and Plasmids across**
821 **four representative months based on high-throughput qPCR.**
822 Absolute copy number of different types of ARGs (**a**), MGEs (**b**), and Plasmid (**c**).

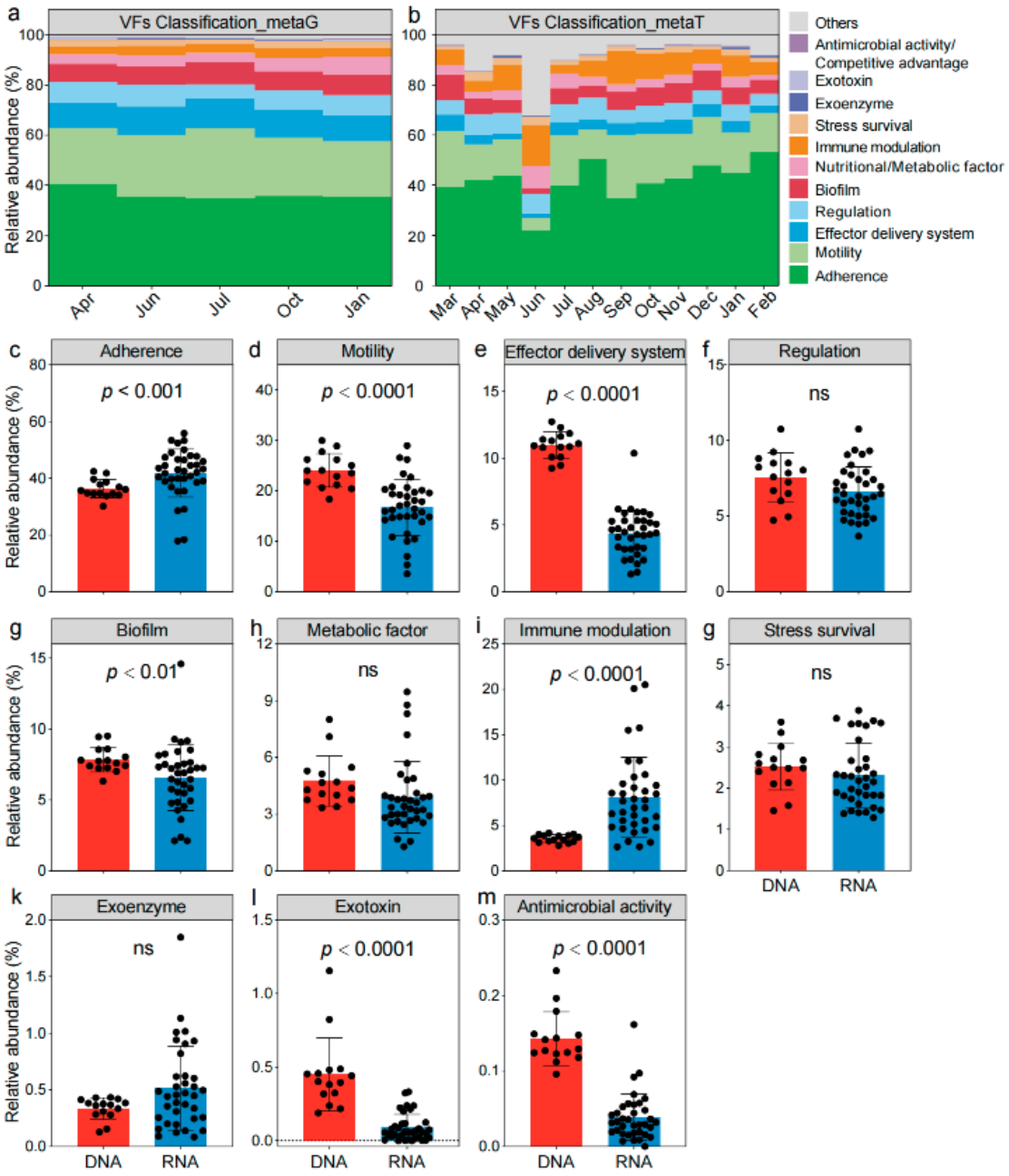


Supplementary Figure 19. Seasonal dynamics of Virulence factors (VFs) in the Yellow River water.

a, Relative abundance of different types of VFs based on metagenome (metaG) sequences data; **b**, Biomarkers of VFs based on metaG sequences data; **c**, relative transcripts abundance of different types of VFs based on metatranscriptome (metaT) sequences data; **d**, Biomarkers of VFs based on metaT sequences data; **e**, Copy number of VFs normalized by 16S rRNA genes from representative months of four seasons (with June for the summer) based on metaG sequence data. Seasonal biomarkers in the Yellow River water were identified through LEfSe analysis. Circle size is proportional to the relative gene (**b**) or transcripts (**d**) abundance of each type. Bold and light font styles represent VF types and subtypes, respectively. In panel **e**, Data that do not share a common letter are significantly different between months ($P < 0.05$; multiple comparison with Kruskal-Wallis tests).

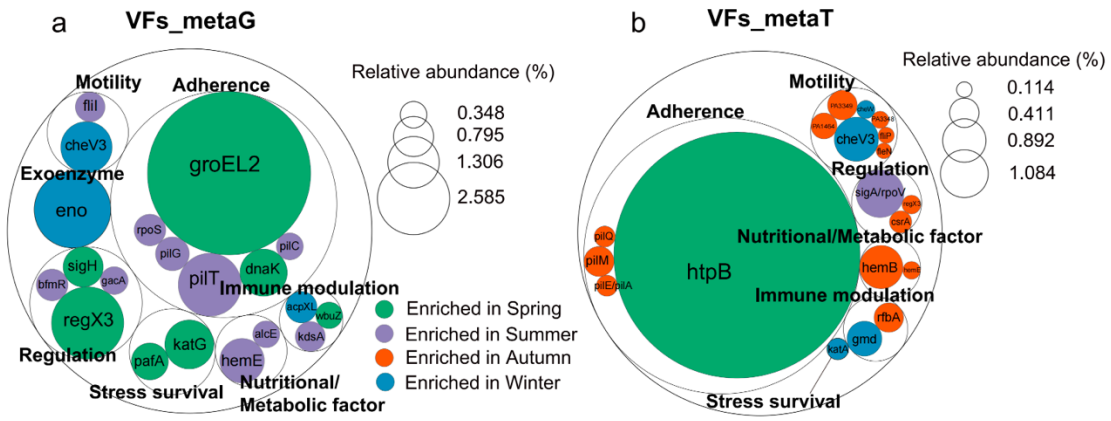


Supplementary Figure 20. Relative abundance of VFs in the Yellow River water. **a–k**, Bar plots showing differences in the relative abundance of each VF type between the DNA (metagenome-based) and RNA (metatranscriptome-based) levels. The data represent reads mapped to nucleotide sequences of VF genes identified from all contigs. “DNA” and “RNA” indicate that relative abundances were calculated based on metagenomic (metaG) and metatranscriptomic (metaT) reads mapped to genes, respectively. Comparisons between DNA and RNA samples were performed using the Wilcoxon rank-sum test (n = 15 for DNA, n = 36 for RNA). $P < 0.05$ indicate a significant difference; ns, not significant.



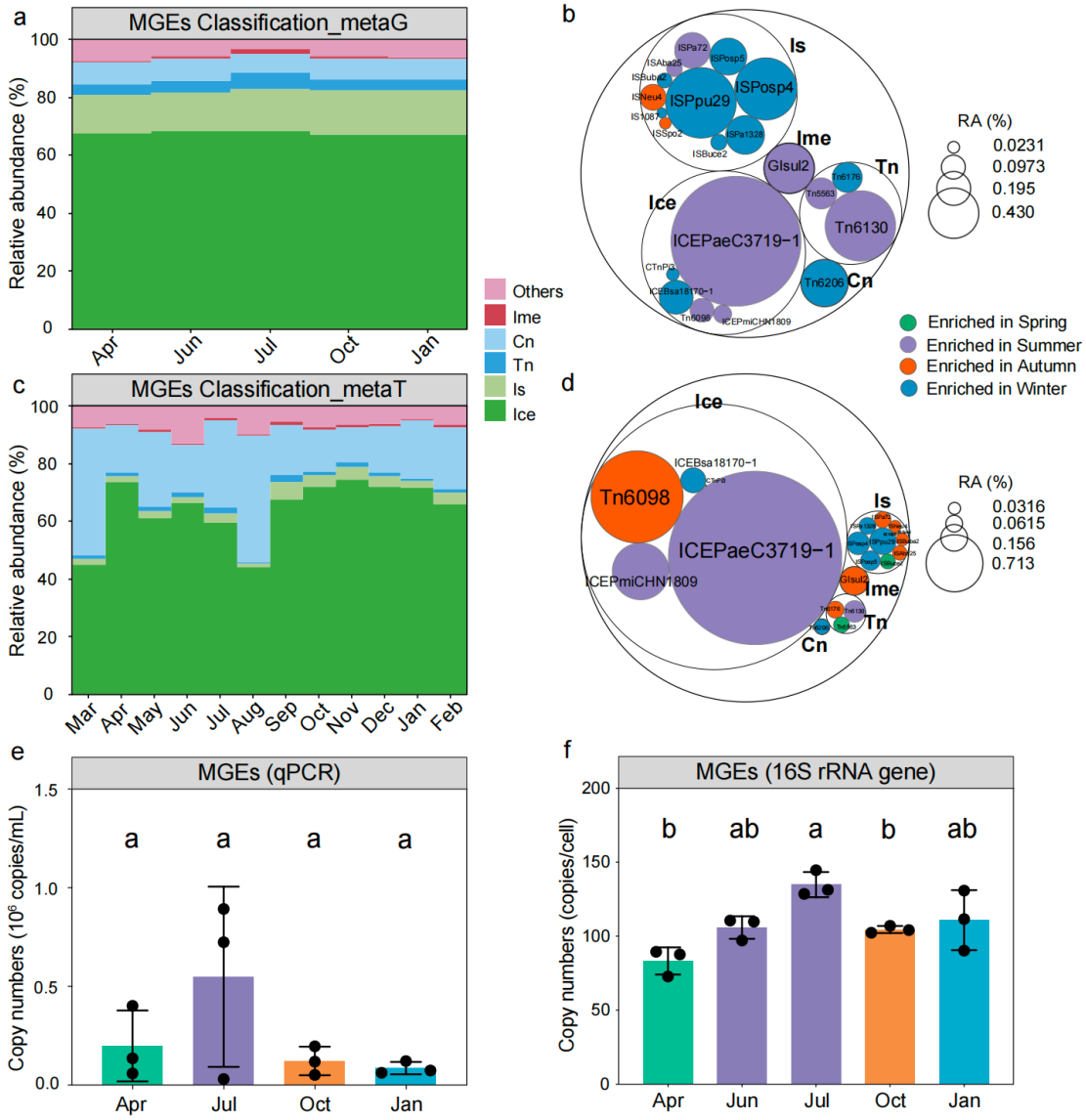
Supplementary Figure 21. Compositions of VFs in the Yellow River water across months as revealed reads classification using ARGs-OAP v3 pipeline.

a, Composition of VF types inferred from quality-filtered metagenomic reads using ARGs-OAP v3 pipeline. **b,** Composition of VF types inferred from quality-filtered metatranscriptomic reads using ARGs-OAP v3 pipeline. **c–m,** Comparisons of the relative abundance of each VF type between DNA-level and RNA-level VF profiles. "DNA" and "RNA" indicate VF profiles generated from metagenomic and metatranscriptomic reads, respectively, using ARGs-OAP v3 pipeline. Comparisons between DNA and RNA samples were performed using the Wilcoxon rank-sum test (n = 15 for DNA and n = 36 for RNA). **P** < 0.05 indicates a significant difference; ns, not significant.



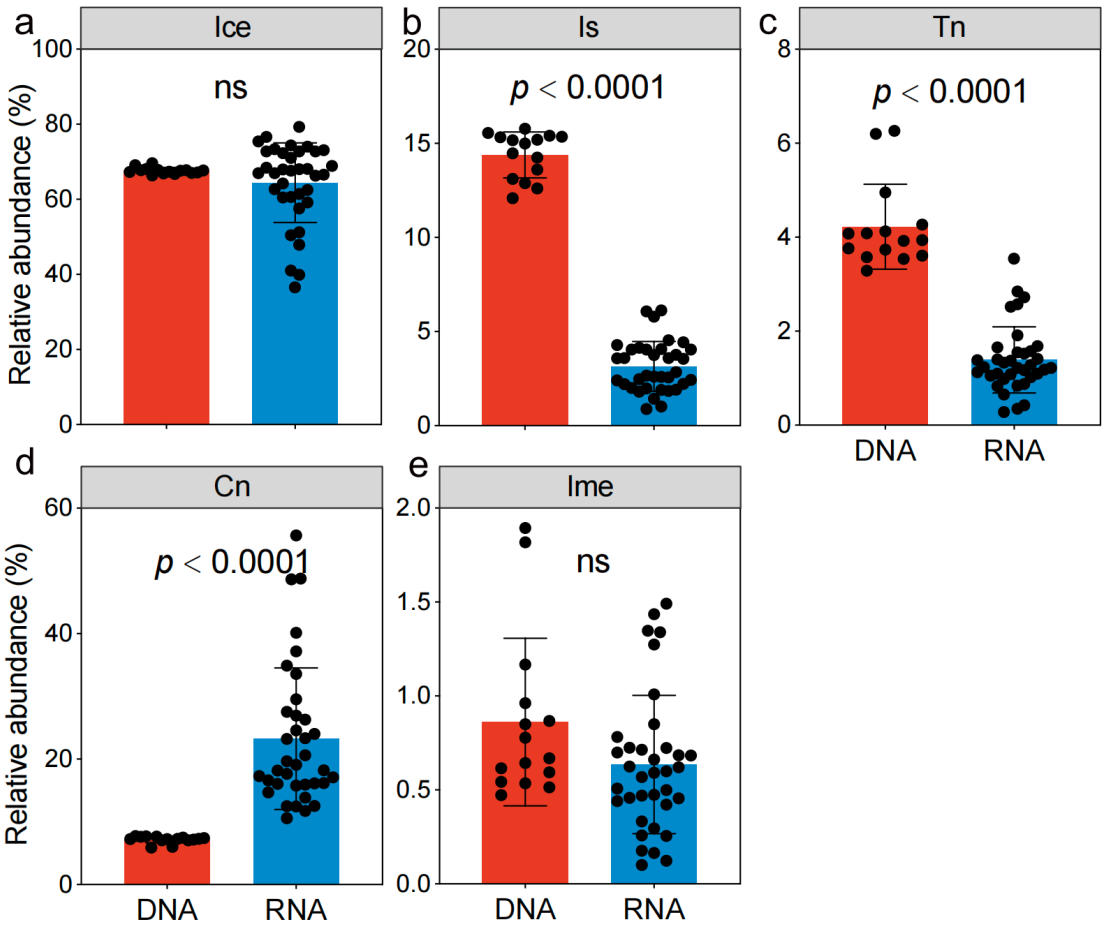
Supplementary Figure 22. Biomarkers of VFs.

a, Biomarkers of VFs based on metagenome (metaG) sequences data; **b**, Biomarkers of VFs based on metatranscriptome (metaT) sequences data; Only VFs with the relative abundance of the TOP 20 are shown. Seasonal biomarkers in the Yellow River Water were identified through LEfSe analysis. Circle size is proportional to the relative abundance of each taxon. Bold and light font styles represent VF types and subtypes, respectively.

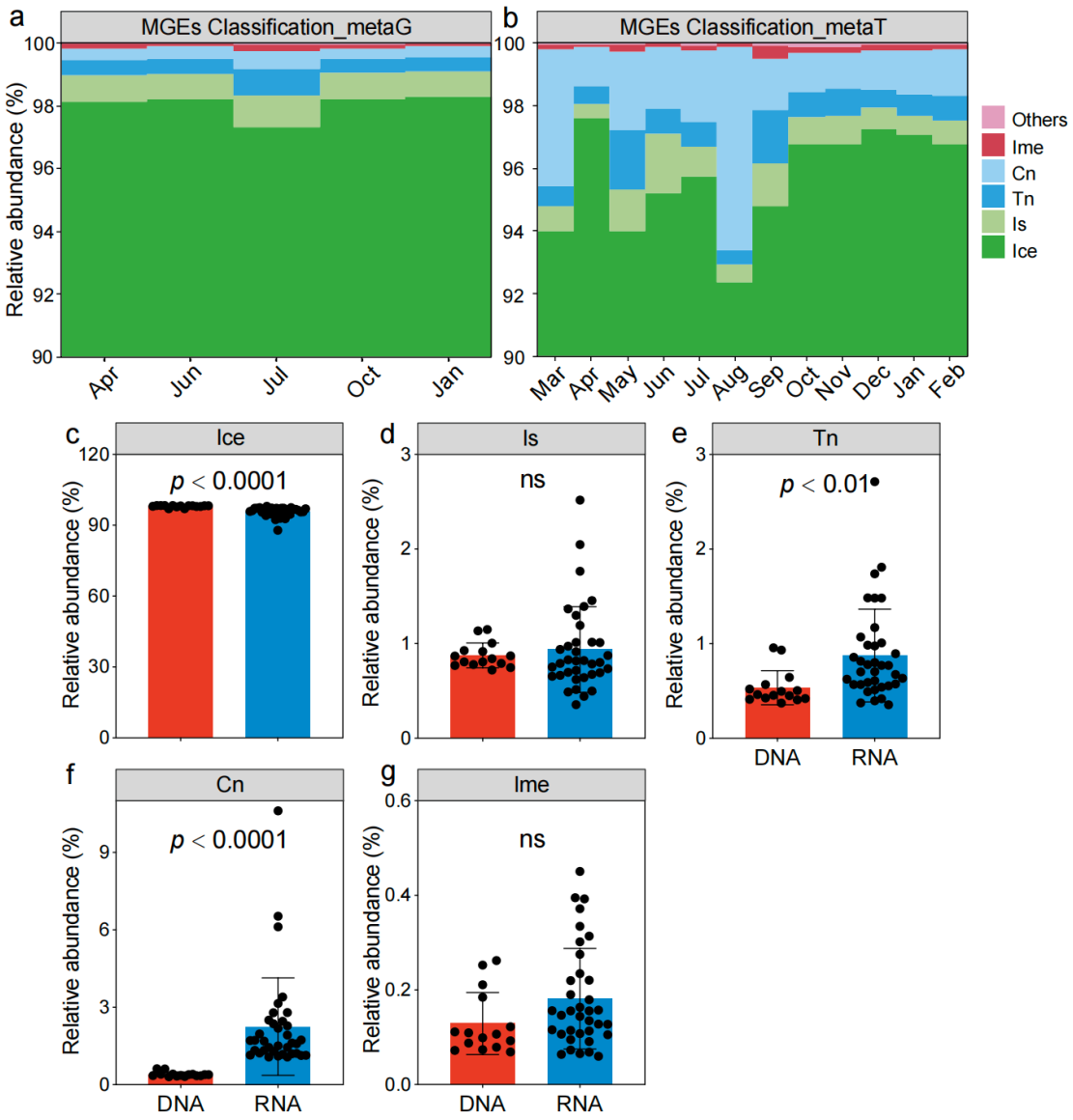


Supplementary Figure 23. Seasonal dynamics of Mobile genetic elements (MGEs) in the Yellow River water.

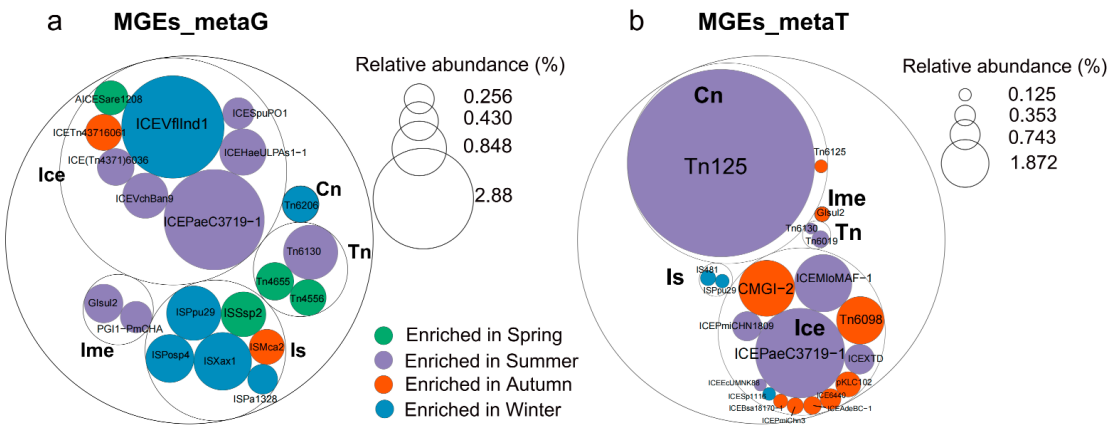
a, Relative abundance of different types of MGEs based on metagenome (metaG) sequences data; **b**, Biomarkers of MGEs based on metaG sequence data; **c**, relative transcripts abundance of different types of MGEs based on metatranscriptome (metaT) sequences data; **d**, Biomarkers of MGEs based on metaT sequences data; **e**, The absolute abundance of different types of MGEs in samples from representative months of four seasons. **f**, Copy number of MGEs normalized by 16S rRNA genes from representative months of four seasons (with June for the summer) based on metaG sequence data. Seasonal biomarkers in the Yellow River water were identified through LEfSe analysis. Circle size is proportional to the relative gene (**b**) or transcripts (**d**) abundance of each type. Bold and light font styles represent MGE types and subtypes, respectively. In panels **e** and **f**, Data that do not share a common letter are significantly different between months ($P < 0.05$; multiple comparison with Kruskal-Wallis tests). The data represent reads mapped to nucleotide sequences of MGE genes identified from all contigs.



874 **Supplementary Figure 24. Relative abundance of MGEs in the Yellow River water.**
875 **a–e**, Bar plots showing differences in the relative abundance of each MGE type between the
876 DNA (metagenome-based) and RNA (metatranscriptome-based) levels. The data represent
877 reads mapped to nucleotide sequences of MGEs identified from all contigs. “DNA” and “RNA”
878 indicate that relative abundances were calculated based on metagenomic (metaG) and
879 metatranscriptome (metaT) reads mapped to genes, respectively. Comparisons between DNA
880 and RNA samples were performed using the Wilcoxon rank-sum test (n = 15 for DNA, n = 36
881 for RNA). $P < 0.05$ indicate a significant difference; ns, not significant.
882

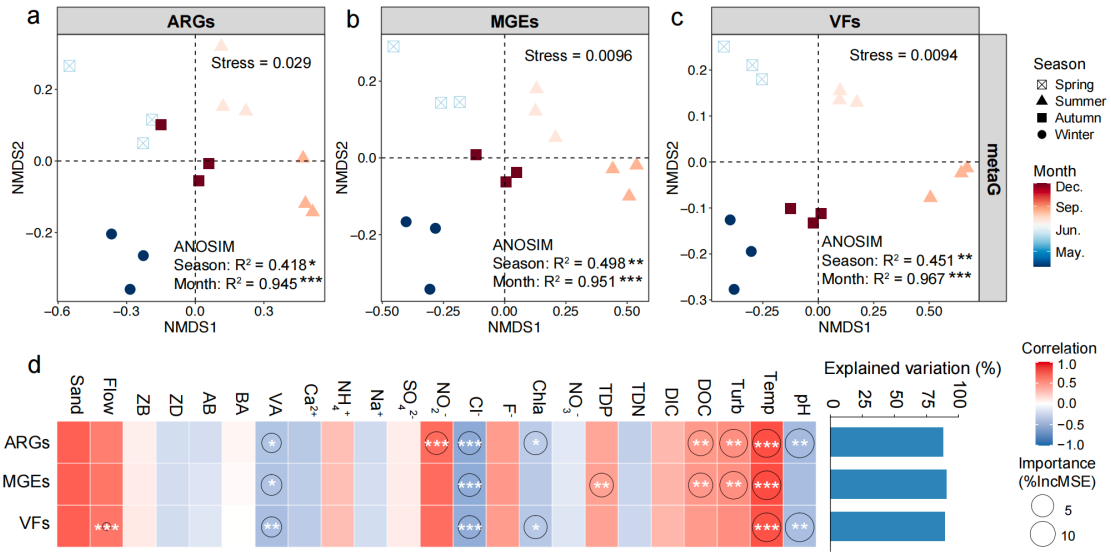


883 **Supplementary Figure 25. Compositions of MGEs in the Yellow River water across**
884 **months as revealed reads classification using ARGs-OAP v3 pipeline.**
885 **a**, Composition of MGE types inferred from quality-filtered metagenomic reads using ARGs-
886 OAP v3 pipeline. **b**, Composition of MGE types inferred from quality-filtered
887 metatranscriptomic reads using ARGs-OAP v3 pipeline. **c–g**, Comparisons of the relative
888 abundance of each MGE type between DNA-level and RNA-level MGE profiles. "DNA" and
889 "RNA" indicate MGE profiles generated from metagenomic and metatranscriptomic reads,
890 respectively, using ARGs-OAP v3 pipeline. Comparisons between DNA and RNA samples
891 were performed using the Wilcoxon rank-sum test (n = 15 for DNA and n = 36 for RNA). P <
892 0.05 indicates a significant difference; ns, not significant.
893



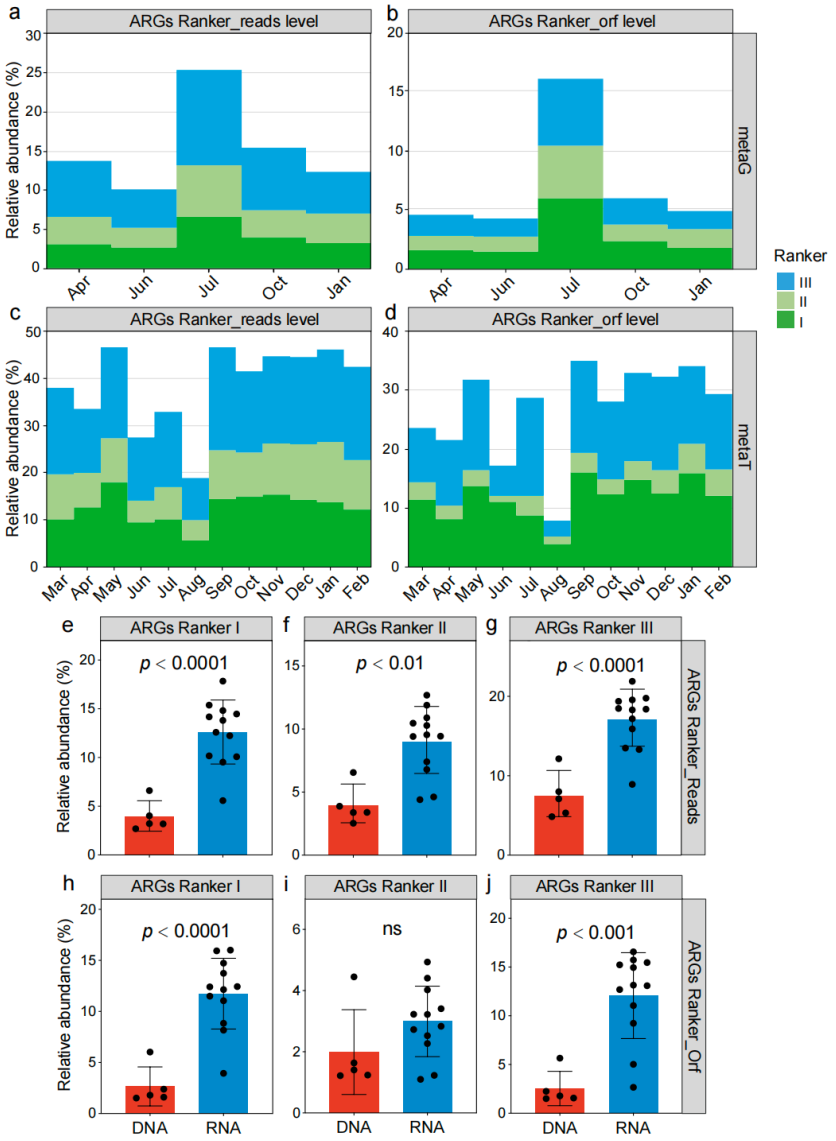
Supplementary Figure 26. Biomarkers of MGEs.

a, Biomarkers of MGEs based on metagenome (metaG) sequences data; **b**, Biomarkers of MGEs based on metatranscriptome (metaT) sequences data; Only MGEs with the relative abundance of the TOP 20 are shown. Seasonal biomarkers in the Yellow River Water were identified through LEfSe analysis. Circle size is proportional to the relative abundance of each taxon. Bold and light font styles represent MGE types and subtypes, respectively.

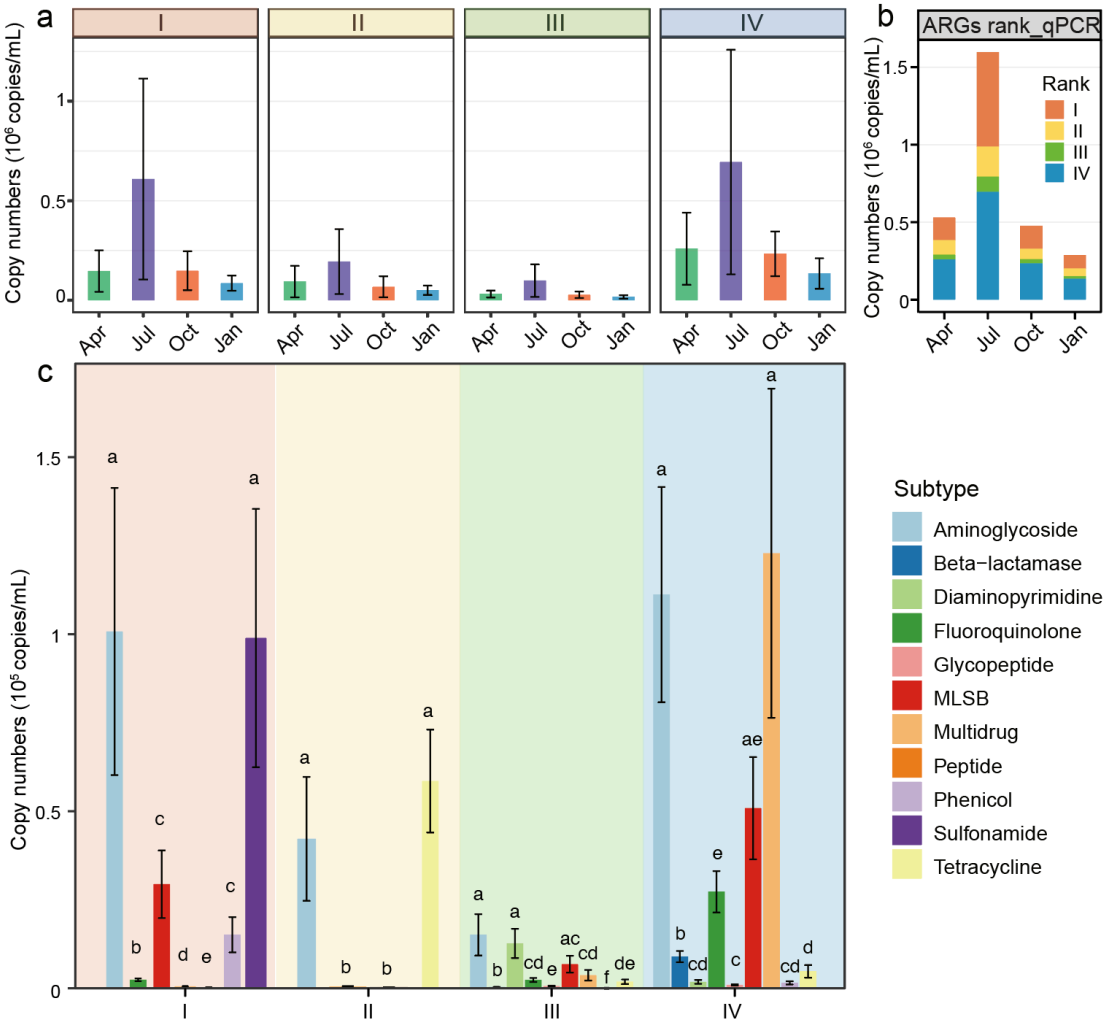


Supplementary Figure 27. Seasonal variation and driving factors risk-associated genetic elements in the Yellow River water.

NMDS shows the structure for the ARGs (**a**), MGEs (**b**), VFs activities (**c**), respectively, based metagenome (metaG) reads mapped to corresponding genes detected from all contigs (see Methods). 95% confidence ellipses were shown for samples grouped by seasons. Statistical significance was examined using the ANOSIM test (**P < 0.01, ***P < 0.001). d, Contributions of different factors to the dissimilarities of risk-associated genetic elements. Circle size indicates variable importance measured by percentage increase in mean squared error (%IncMSE) from random forest; only circles with permutation-based $P < 0.05$ are shown. The right bar plot shows the explained variation (%) by random forest for each response variable. Colors represent Pearson correlations, with red for Positive and blue for negative correlation. *, ** and *** indicate significant Pearson correlation with $P < 0.05$ and $P < 0.01$, and $P < 0.001$, respectively. Notes: Pre, precipitation; Flow, runoff.



Supplementary Figure 28. Seasonal dynamics of different ranks of ARGs across months. **a**, Relative abundance of different ARGs ranks based on clean metagenomic (metaG) reads classification using ARGs-OAP v3 pipeline (see Methods). **b**, Relative abundance of different ARGs ranks based on clean metaG reads mapped to nucleotide sequences of ARGs identified from all contigs (see Methods). **c**, Relative abundance of different ARGs ranks based on clean metatranscriptomic (metaT) reads classification using ARGs-OAP v3 pipeline (see Methods). **d**, Relative abundance of different ARGs Ranks based on clean metaT reads mapped to nucleotide sequences of ARGs identified from all contigs (see Methods). **e–g**, Bar plots show the differences in the relative abundance of each ARG Rank between the DNA (metagenome-based) and RNA (metatranscriptome-based) levels. Data are based on ARGs-OAP v3 read-based ARG profiling, as in panels a and c. **h-j**, Bar plots show the differences in the relative abundance of each ARG Rank between the DNA (metagenome-based) and RNA (metatranscriptome-based) levels. The data represent reads mapped to nucleotide sequences of ARGs identified from all contigs, as in panels b and d. "DNA" and "RNA" indicate ARG profiles generated from metagenomic and metatranscriptomic reads, respectively. Comparisons between DNA and RNA samples were performed using the Wilcoxon rank-sum test (n = 15 for DNA, n = 36 for RNA). $P < 0.05$ indicates a significant difference; ns, not significant.

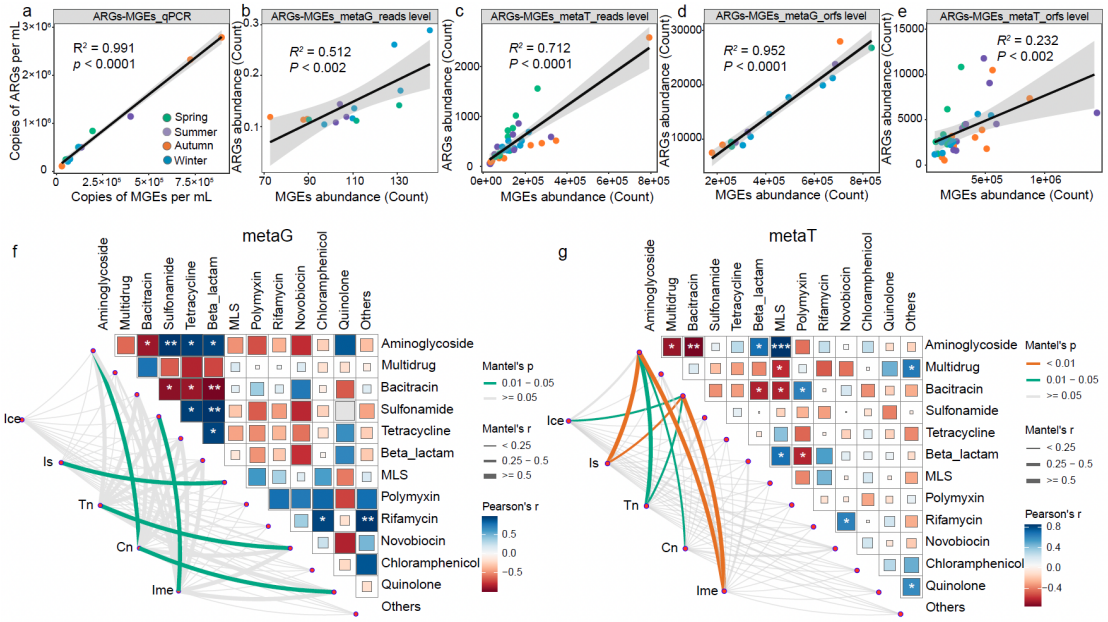


Supplementary Figure 29. Copy number of ARGs belonging to different risk ranks across seasons.

a, Total number ARGs copies. **b**, Copies of ARGs types belonging to different ranks. **c**, Relative abundance of different ranks of ARGs based on high-throughput qPCR. In panel c, data that do not share a common letter are significantly among ARGs subtype of different risk levels ($P < 0.05$; multiple comparisons using Kruskal-Wallis tests). Data shown are means ± SD, n = 3.

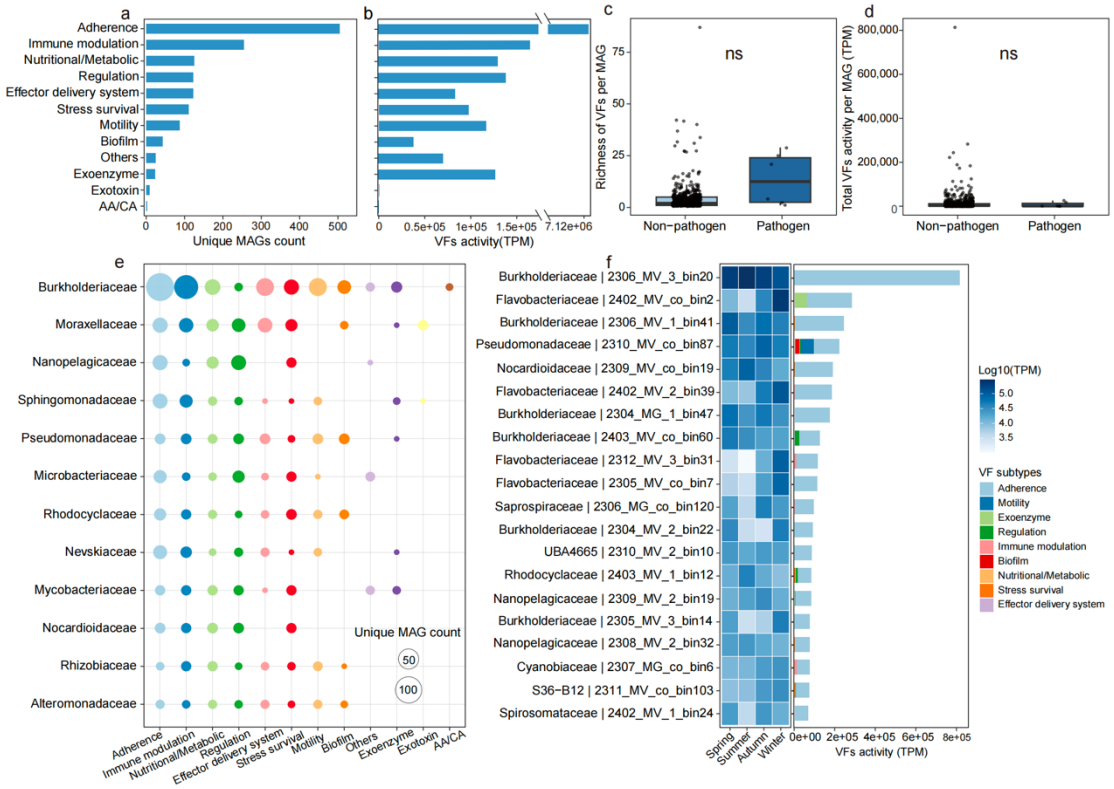
Supplementary Figure 30. Genetic contexts of contigs carrying Rank I or Rank II ARGs.

Gene arrangements of contigs carrying Rank I or Rank II ARGs. Arrows represent predicted ORFs and their transcriptional orientations. Gene names are shown above arrows, and nucleotide positions are indicated below each contig. Notes: Src, source of the contig; Tax, taxonomic assignment.

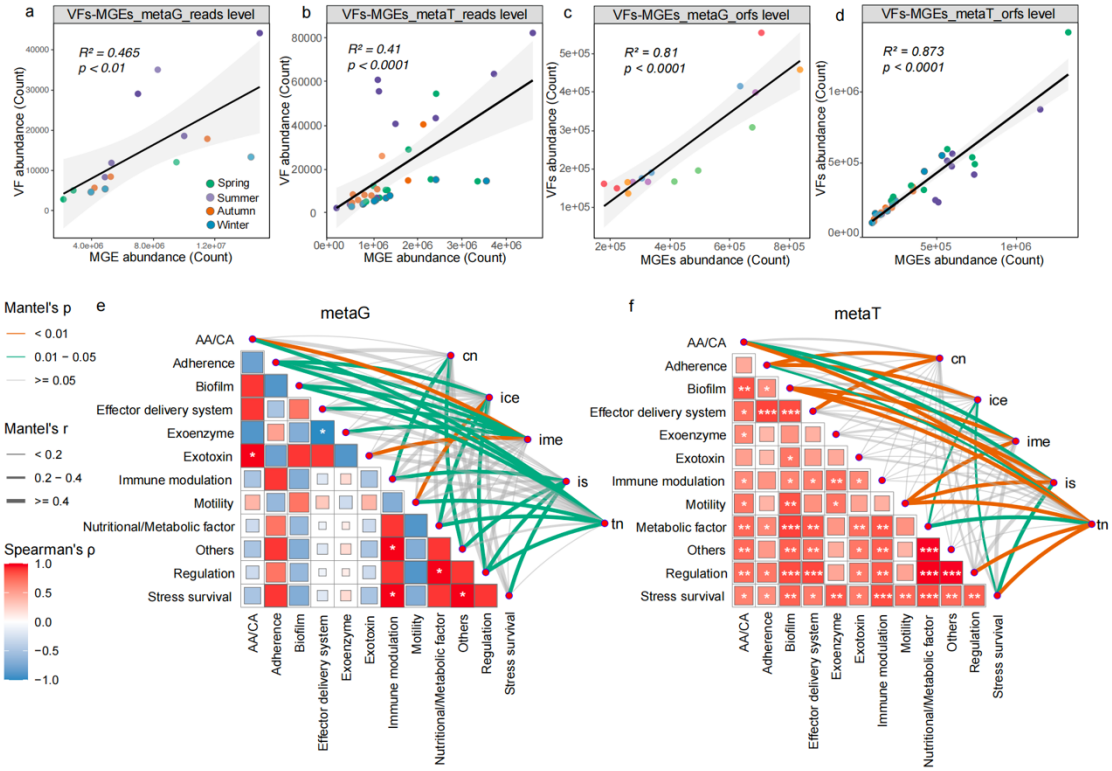


944 **Supplementary Figure 31. Relation between ARGs and MGEs.**

945 **a-e,** Pearson correlations between the abundance of ARGs and MGEs measured by different
946 methods. **a,** copy number of ARGs vs. MEGs quantified by qPCR; **b**, normalized ARGs vs
947 MGEs abundances (copy/16S rRNA gene) based on metagenome (metaG) reads classification
948 ARGs-OAP v3 pipeline. **c**, ARGs vs MGEs counts based on metatranscriptome (metaT) reads
949 classification ARGs-OAP v3 pipeline. **d**, ARGs vs MGEs counts based on metaG reads mapped
950 to all ARGs and MGEs genes from all contigs. **e**, ARGs vs MGEs counts based on metaT reads
951 mapped to all ARGs and MGEs genes from all contigs. The linear regression trend was shown
952 with a 95% confidence interval. **f-g**, Pearson's correlation coefficients of ARGs and MGEs
953 based on Mantel tests by metaG and metaT samples. Edge width corresponds to Mantel's r
954 value, and the color of the edge indicates the statistical significance based on 999 permutations.
955 Pairwise correlations of these variables are displayed with square size, while the color gradient
956 indicates Pearson's correlation coefficients. Square with asterisks represent the significant
957 levels at $P < 0.05$ (*), $P < 0.01$ (**), and $P < 0.001$ (***). The data is based on metaG and
958 metaT contigs mapped to genes, respectively.
959

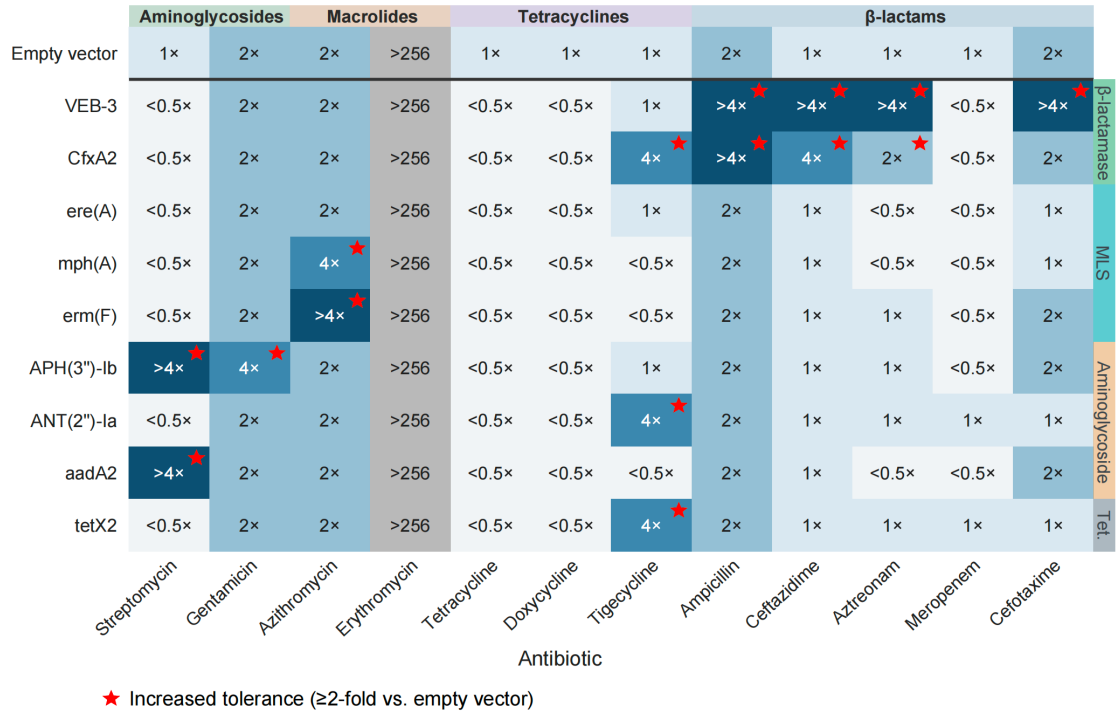


960 **Supplementary Figure 32. Distribution and activity of virulence factors (VFs) carried by**
961 **MAGs.**
962 **a,** number of MAGs carrying different types of VFs. **b**, cumulative TPM of different types of
963 virulence factors (VFs). **c**, Comparison of the number of types of virulence factor genes (VF
964 genes) carried by each MAG in pathogenic and non-pathogenic bacteria. **d**, Comparison of the
965 total relative activity of virulence factor genes (VF genes) carried by each MAG in pathogenic
966 and non-pathogenic bacteria. In panels c and d, differences between the two groups were
967 assessed using the Wilcoxon rank-sum test. ns, not significant. **e**, Number of MAGs carrying
968 different types of VFs among the top 12 families with the most MAGs carrying VFs. **f**, Seasonal
969 dynamics and compositional patterns of VFs carried by the top 20 MAGs with the highest total
970 relative activity.



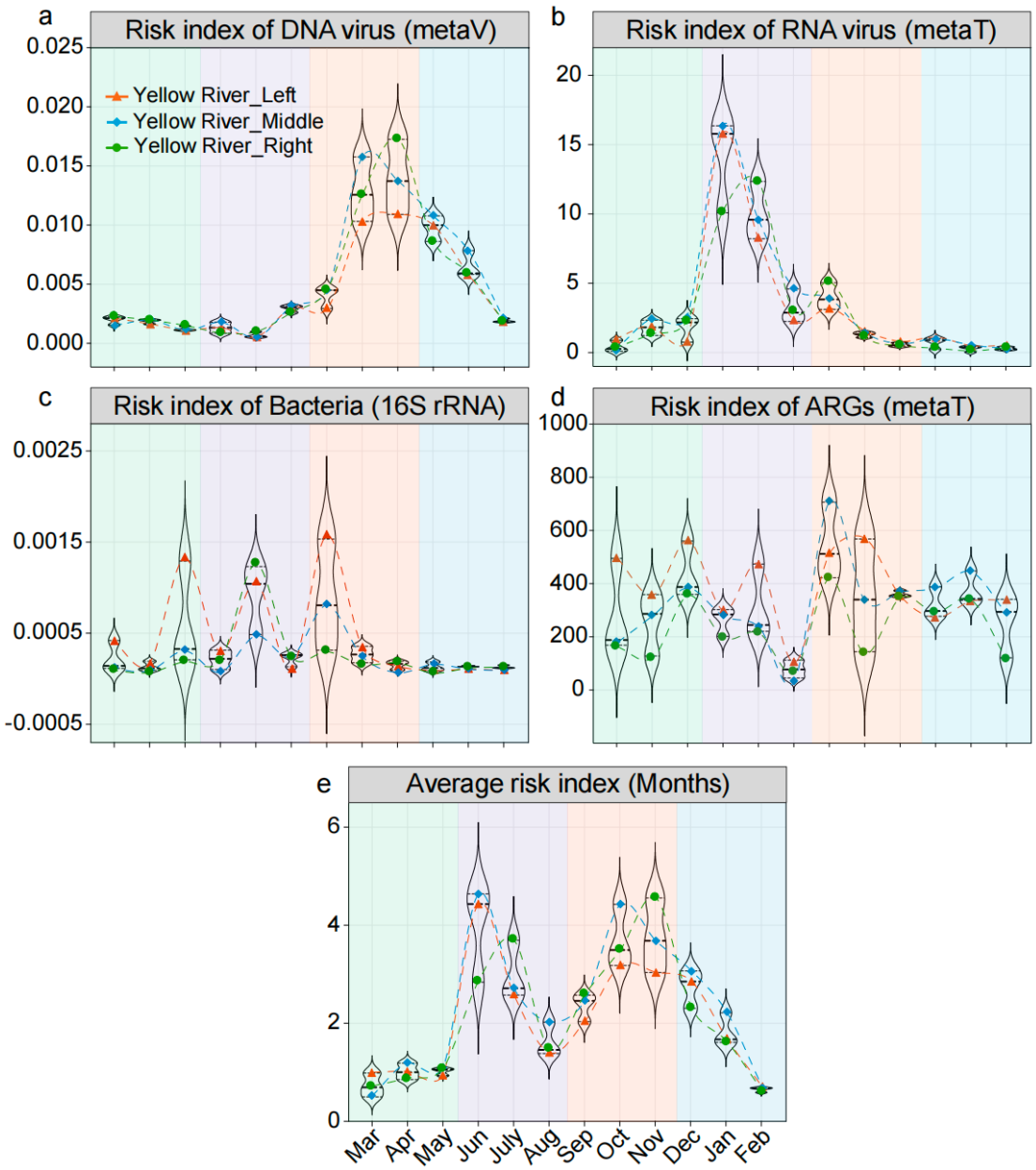
971 **Supplementary Figure 33. Relation between VFs and MGEs.**

972 **a-e,** Pearson correlations between the abundance of VFs and MGEs measured by different
973 methods. **a**, ARGs vs MGEs counts based on m metagenome (metaG) reads classification
974 ARGs-OAP v3 pipeline. **b**, VFs vs MGEs counts based on metatranscriptome (metaT) reads
975 classification ARGs-OAP v3 pipeline. **c**, ARGs vs MGEs counts based on metaG reads mapped
976 to all VFs and MGEs genes from all contigs. **d**, ARGs vs MGEs counts based on metaT reads
977 mapped to all ARGs and MGEs genes from all contigs. The linear regression trend was shown
978 with a 95% confidence interval. **e-f**, Pearson's correlation coefficients of VFs and MGEs
979 based on Mantel tests by metaG and metaT samples. Edge width corresponds to Mantel's r
980 value, and the color of the edge indicates the statistical significance based on 999 permutations.
981 Pairwise correlations of these variables are displayed with square size, while the color gradient
982 indicates Pearson's correlation coefficients. Square with asterisks represent the significant
983 levels at $P < 0.05$ (*), $P < 0.01$ (**), and $P < 0.001$ (***). The data is based on metaG and
984 metaT contigs mapped to genes, respectively.
985

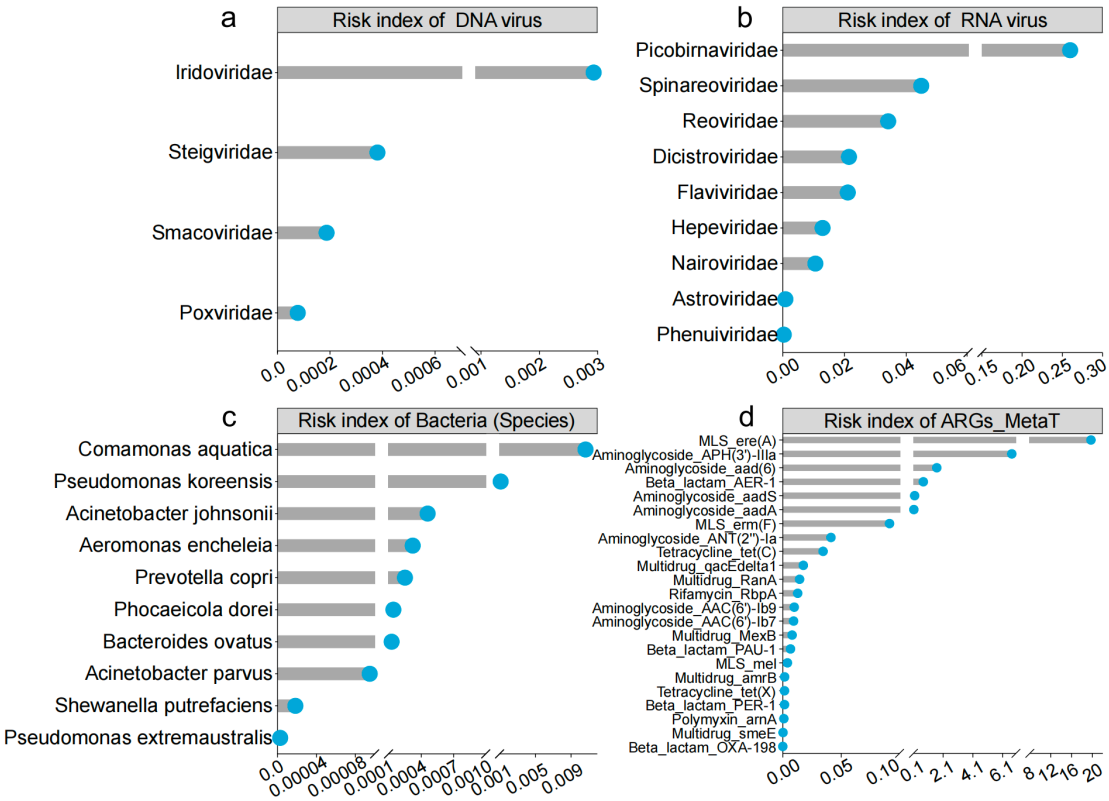


Supplementary Figure 34. Functional validation of high-risk ARGs by MIC-based growth assays.

The heatmap shows the highest tested antibiotic level supporting growth of *E. coli* MG1655 transformants carrying nine representative Rank I/II ARGs across 12 antibiotics. Tile labels indicate tolerance levels relative to the minimum inhibitory concentration (MIC) of the empty-vector control, with darker shading representing higher tolerance. Red stars indicate ARG–antibiotic combinations showing a ≥2-fold increase in tolerance relative to the empty-vector control. >4× denotes growth beyond the highest MIC-fold concentration tested, whereas >256 μg mL^{-1} for erythromycin represents the upper assay boundary. Notes: Tet., tetracycline.

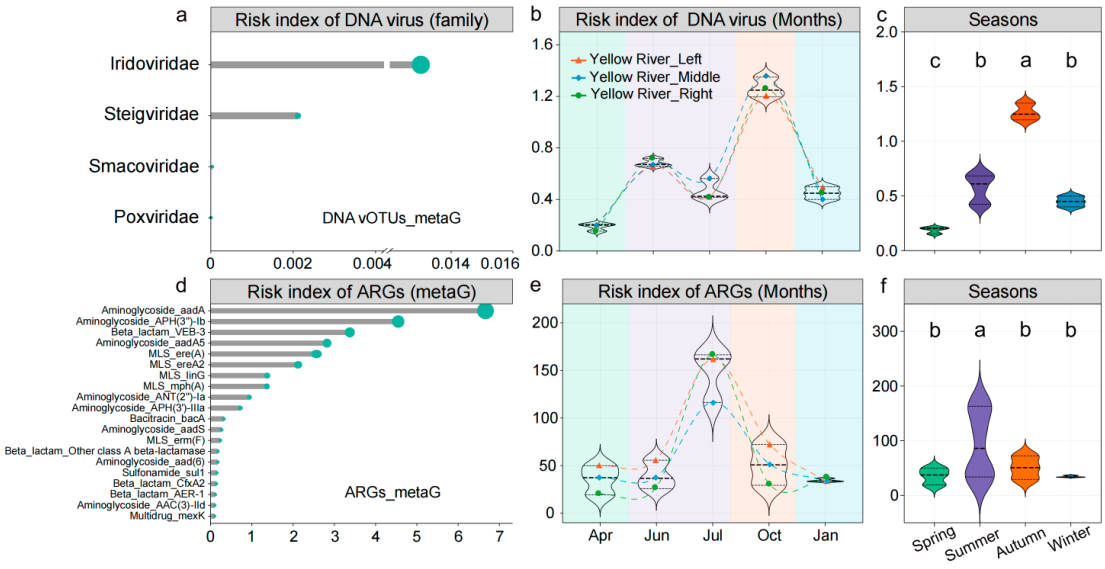


997 **Supplementary Figure 35. Risk assessment of the Yellow River water across months.**
998 Risk index of water sample for each month derived from detected DNA viral pathogens (**a**),
999 risk RNA virus (**b**). bacterial pathogen (**c**), and ARGs (**d**). Risk index shown in a, b, c, and d
1000 are based on metavirome (metaV) reads mapping to DNA vOTUs, metatranscriptome (metaT)
1001 reads mapping to RNA vOTUs, 16S rRNA gene amplicon sequences data and metaT mapping
1002 to ARGs detected in all contigs, respectively (see Mothods).



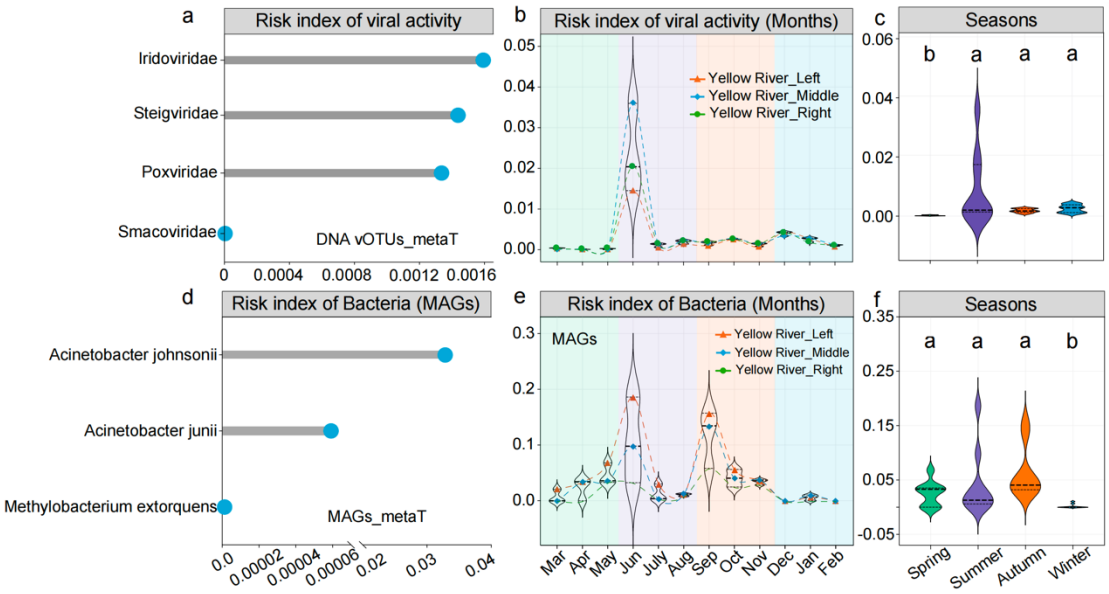
Supplementary Figure 36. Risk assessment for potential pathogens and ARGs in the Yellow River water.

Risk index for detected DNA viral pathogens (**a**), risk RNA virus (**b**). bacterial pathogen (**c**), and ARGs (**d**). Risk index shown in a, b, c, and d are based on metavirome (metaV) reads mapping to DNA vOTUs, metatranscriptome (metaT) reads mapping to RNA vOTUs, 16S rRNA gene amplicon sequences data and metaT mapping to ARGs detected in all contigs, respectively (see Mothods).



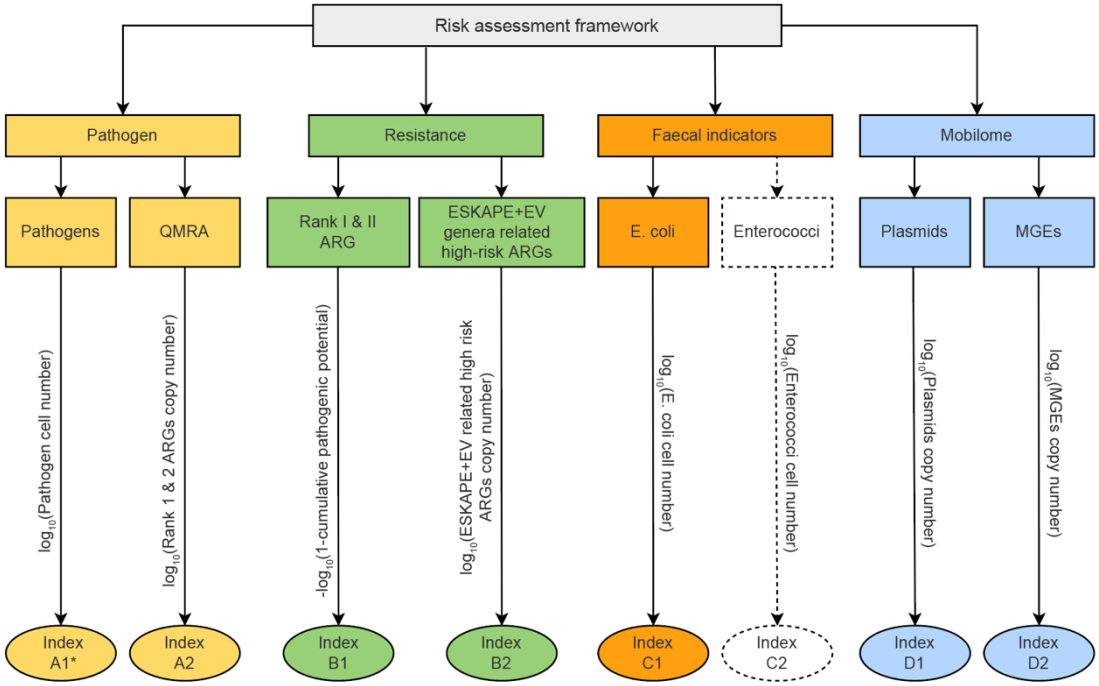
Supplementary Figure 37. Risk assessment of pathogens, ARGs and water samples across months and seasons based on metagenome (metaG) dataset.

Risk index of detected DNA viral pathogens (**a**), water sample for each month (**b**) and seasons (**c**) derived from detected DNA viral pathogens. Risk index of ARGs (**d**), water sample for each month (**e**) and seasons (**f**) derived from detected ARGs. Different letters in panels **c** and **f** indicate significant differences across seasons ($P < 0.05$). Risk indices shown in panels **a-c** and **d-f** are based on metaG reads mapping to DNA vOTUs and to ARGs detected in all contigs, respectively (see Mothods).



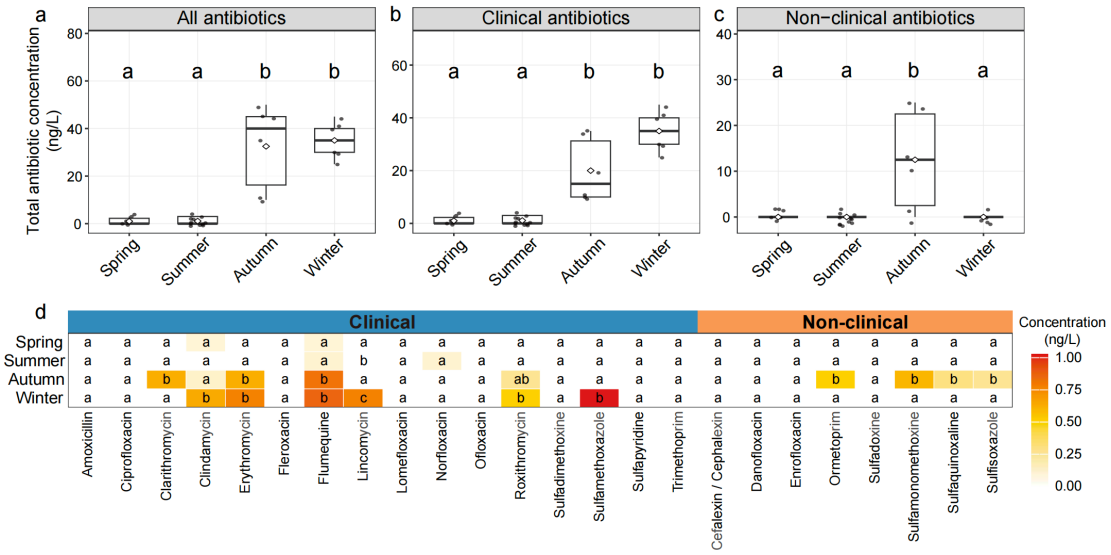
Supplementary Figure 38. Risk assessment of DNA viral, bacterial pathogens, and water samples across months and seasons based on metatranscriptome (metaT) dataset.

Risk index of detected DNA viral pathogens (**a**), water sample for each month (**b**) and season (**c**) derived from detected DNA viral pathogens. Risk index of bacterial pathogens (**d**), water sample for each month (**e**) and season (**f**) derived from detected ARGs. Different letters in panels **c** and **f** indicate significant differences across seasons ($P < 0.05$). Risk indices shown in panels **a-c** and **d-f** are based on metaT reads mapping to DNA vOTUs and to MAGs of pathogens detected, respectively (see Mothods).



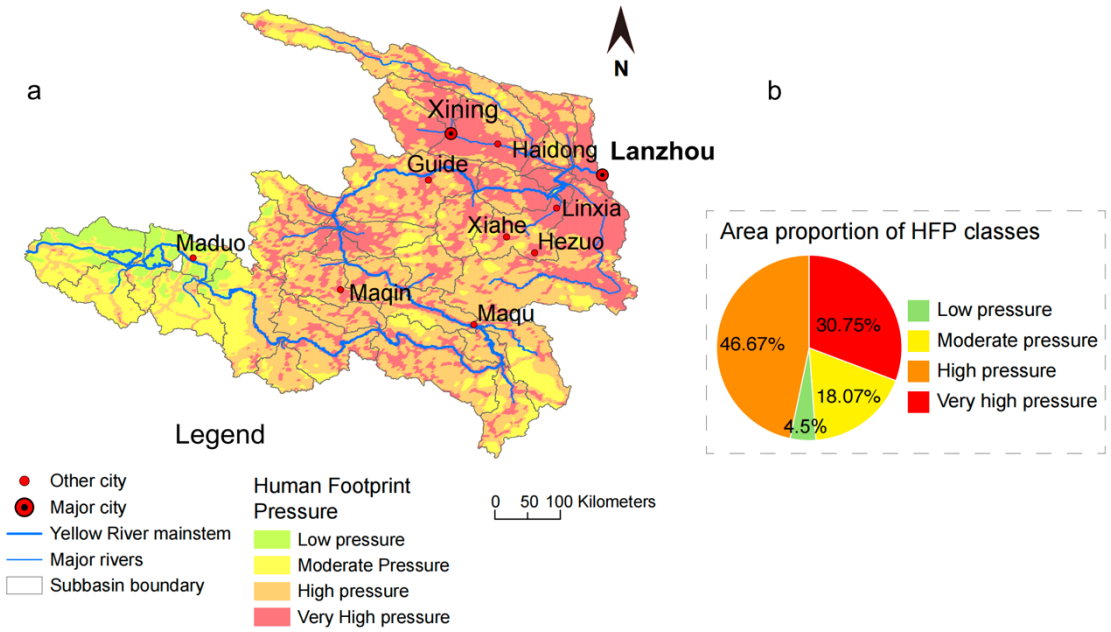
Supplementary Figure 39. Risk assessment of Yellow River water samples using absolute quantification data.

The risk assessment framework was adopted from Shi et al., (2025, Nature Water 3: 473-85. https://dx.doi.org/10.1038/s44221-025-00421-y), which includes factors of risk assessment and scoring methods (cell or gene copy number = cell or gene copy concentration × volume (100 ml)). See Supplementary Methods for more details. A1*: potentially pathogenic DNA viruses are included.



Supplementary Figure 40. Seasonal variations of antibiotic concentrations in the Yellow River water.

a–c, Box plots showing seasonal variations in the total concentrations of all antibiotics, clinical antibiotics, and non-clinical antibiotics. Concentrations were summed within each antibiotic group for each sample. Boxes indicate the 25th–75th percentiles, horizontal lines indicate medians, white diamonds indicate means, and points represent individual samples. **d**, Heatmap showing seasonal mean concentrations of individual antibiotics grouped as clinical or non-clinical antibiotics according to the WHO ATC J01 classification (https://atcddd.fhi.no/atc_ddd_index/?code=J&showdescription=yes). Seasonal differences were evaluated using linear models followed by pairwise comparisons of estimated marginal means, with Benjamini–Hochberg correction. Different letters indicate significant seasonal differences within each group or antibiotic ($P < 0.05$).



Supplementary Figure 41. Spatial distribution of human activity pressure in the upstream Yellow River basin above the Zhongshan Bridge control section, Lanzhou City.

(a) Human activity pressure map. **(b)** Area proportion of Human Footprint (HFP) classes. The map's legend and pie chart illustrate four primary categories: low (1–2), moderate (3–5), high (6–11), and very high pressure (12–50). These levels are categorized based on the global terrestrial HFP classification criteria proposed by Venter et al. (2016, Nature Communications, 7, 12558. DOI: 10.1038/ncomms12558). Areas with no pressure (HFP=0) are also accounted for in the mapping statistics, bringing the combined proportion to approximately 100%. See Supplementary Methods for more details.

References

1. Su, Y. et al. Occurrence, transmission and risks assessment of pathogens in aquatic environments accessible to humans. *J. Environ. Manage*. **354**, 120331 (2024).
2. Shrestha, R.G. et al. Prevalence of arcobacter and other pathogenic bacteria in river water in nepal. *Water.* **11**, 1416 (2019).
3. Wang, H. et al. Variations among viruses in influent water and effluent water at a wastewater plant over one year as assessed by quantitative PCR and metagenomics. *Appl. Environ. Microbiol.* **86**, (2020).
4. Francis, J.D., Yanaç, K. & Uyaguari-Diaz, M.I. Seasonal patterns of viromes in urban aquatic environments of Manitoba. *Appl. Environ. Microbiol.* **91**, e00408-00424 (2025).
5. Shi, X. et al. Microbial risk assessment across multiple environments based on metagenomic absolute quantification with cellular internal standards. *Nat Water.* **3**, 473-485 (2025).
6. Ministry of ecology and environment of the People's Republic of China. Report on the State of the Ecology and Environment in China 2025. (2026).
7. Francis, J.D., Yanaç, K. & Uyaguari-Diaz, M.I. Seasonal patterns of viromes in urban aquatic environments of Manitoba. *Appl. Environ. Microbiol.* **91**, e0040824 (2025).
8. Jespersen, A.-M. & Christoffersen, K. Measurements of chlorophyll-a from phytoplankton using ethanol as extraction solvent. *Arch. Hydrobiol.* **109**, 445-454 (1987).
9. Novak, M.G., Cetinic, I., Chaves, J.E. & Mannino, A. The adsorption of dissolved organic carbon onto glass fiber filters and its effect on the measurement of particulate organic carbon: A laboratory and modeling exercise. *Limnol. Oceanogr. Meth.* **16**, 356-366 (2018).
10. Ministry of ecology and environment of the People's Republic of China. Water quality-determination of phytoplankton-membrane filtration-microscopic counting method. *HJ1215-2021* (2021).
11. China national environmental monitoring centre technical requirements for freshwater zooplankton monitoring. (2022).
12. Noble, R.T. & Fuhrman, J.A. Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquat. Microb. Ecol.* **14**, 113-118 (1998).
13. Marie, D., Brussaard, C.P.D., Thyrhaug, R., Bratbak, G. & Vaulot, D. Enumeration of marine viruses in culture and natural samples by flow cytometry. *Appl. Environ. Microbiol.* **65**, 45-52 (1999).
14. Acharya, K. et al. Metagenomic water quality monitoring with a portable laboratory. *Water Res.* **184**, 116112 (2020).
15. John, S.G. et al. A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep*. **3**, 195-202 (2011).
16. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
17. Read, D.S. et al. Catchment-scale biogeography of riverine bacterioplankton. *ISME J.* **9**, 516-526 (2015).
18. Qiu, Z. et al. BASALT refines binning from metagenomic data and increases resolution of genome-resolved metagenomic analysis. *Nat. Commun.* **15**, 2179 (2024).
19. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P. & Tyson, G.W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043-1055 (2015).
20. Olm, M.R., Brown, C.T., Brooks, B. & Banfield, J.F. dRep: a tool for fast and accurate genomic

comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864-2868 (2017).
21. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P. & Parks, D.H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925-1927 (2020).
22. Parks, D.H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785-D794 (2022).
23. Camargo, A.P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303-1312 (2024).
24. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578-585 (2021).
25. Luo, X.-Q. et al. Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts. *Microbiome* **10**, 190 (2022).
26. Jiarong, G., Dean, V., Akbar Adjie, P., Simon, R. & Matthew, S. Viral sequence identification SOP with VirSorter2. *protocols.io* (2021).
27. Liu, Y. et al. Diversity and function of mountain and polar supraglacial DNA viruses. *Sci. Bull.* **68**, 2418-2433 (2023).
28. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632-639 (2019).
29. Pons, J.C. et al. VPF-Class: taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. *Bioinformatics* **37**, 1805-1813 (2021).
30. Shang, J., Jiang, J. & Sun, Y. Bacteriophage classification for assembled contigs using graph convolutional network. *Bioinformatics* **37**, i25-i33 (2021).
31. Aylward, F.O. & Moniruzzaman, M. ViralRecall-a flexible command-line tool for the detection of giant virus signatures in 'omic data. *Viruses* **13**, 150 (2021).
32. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
33. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359 (2012).
34. Aroney, S.T.N. et al. CoverM: read alignment statistics for metagenomics. *Bioinformatics* **41**, (2025).
35. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
36. Liu, Y. et al. Diverse novel RNA viruses in the cryosphere of the Tibetan Plateau. *Natl. Sci. Rev.* **12**, nwaf361 (2025).
37. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211-3217 (2012).
38. Edgar, R.C. et al. Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142-147 (2022).
39. Neri, U. et al. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023-4037.e4018 (2022).
40. Dominguez-Huerta, G. et al. Diversity and ecological footprint of global ocean RNA viruses. *Science* **376**, 120-1208 (2022).
41. Hou, X. et al. Using artificial intelligence to document the hidden RNA virosphere. *Cell* **187**, 6929-6942.e6916 (2024).
42. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or

nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
43. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
44. Zheng, K. et al. VITAP: a high precision tool for DNA and RNA viral classification based on meta-omic data. *Nat. Commun.* **16**, 2226 (2025).
45. Yin, X. et al. ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics* **34**, 2263-2270 (2018).
46. Chen, L. et al. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* **33**, D325-328 (2005).
47. Johansson, M.H.K. et al. Detection of mobile genetic elements associated with antibiotic resistance in Salmonella enterica using a newly developed web tool: MobileElementFinder. *J. Antimicrob. Chemoth.* **76**, 101-109 (2021).
48. Yin, X. et al. ARGs-OAP v3. 0: antibiotic-resistance gene database curation and analysis pipeline optimization. *Engineering* **27**, 234-241 (2023).
49. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026-1028 (2017).
50. Zhang, A.-N. et al. An omics-based framework for assessing the health risk of antimicrobial resistance genes. *Nat. Commun.* **12**, 4765 (2021).
51. von Meijenfeldt, F.A.B., Arkhipova, K., Cambuy, D.D., Coutinho, F.H. & Dutilh, B.E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 217 (2019).
52. Wood, D.E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
53. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).
54. Powell, H.R., Islam, S.A., David, A. & Sternberg, M.J.E. Phyre2. 2: a community resource for template-based protein structure prediction. *J. Mol. Biol.***437**, 168960 (2025).
55. Wu, L. et al. Decade-long warming accelerates antibiotic resistance in grassland soils. *Nature* 1-10 (2026).
56. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461 (2010).
57. Edgar, R.C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*, 081257 (2016).
58. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-596 (2013).
59. Chen, L. et al. Permafrost carbon cycle and its dynamics on the Tibetan Plateau. *Sci. China Life Sci.* 67, 1833-1848 (2024).
60. Matias Rodrigues, J.F., Schmidt, T.S.B., Tackmann, J. & von Mering, C. MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics* **33**, 3808-3810 (2017).
61. Geng, Y. et al. Anthropogenic activity and climate change exacerbate the spread of pathogenic bacteria in the environment. *Sci. Adv.* **11**, eads4355 (2025).
62. International Committee on Taxonomy of Viruses ICTV 2024 Virus Metadata Resource (VMR). *Zenodo*, VMR_MSL40.v41.20250307 (2025).

63. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).
64. Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
65. Minh, B.Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530-1534 (2020).
66. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293-W296 (2021).
67. Zhao, K. et al. Agriculture increases potential health risks of vertebrate viruses in soils. *iMeta* **4**, e70034 (2025).
68. Mu, H. et al. A global record of annual terrestrial human footprint dataset from 2000 to 2018. *Sci. Data* **9**, 176 (2022).
69. Venter, O. et al. Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation. *Nat. Commun.* **7**, 12558 (2016).

Supplementary Tables S1-S27 (in a separate Excel file, available in figshare: https://doi.org/10.6084/m9.figshare.c.7390684).

Table S1 Sample collection location and basic information

Table S2 Monthly and seasonal biotic and physicochemical parameters of the Yellow River water

Table S3 Abundance and biomass of Zooplankton and Phytoplankton across samples.

Table S4 Metadata of metatranscriptomic, metavirome, and metagenomic samples and libraries.

Table S5 Pairwise sequence similarity matrices of pathogenic microbiome-associated sequences

Table S6 Characterization and Metagenomic (metaG) TPM values of 742 MAGs recovered from metavirome (n = 417) and metagenome (n = 325) samples

Table S7 Taxonomic classification and Metaviromic (metaV) TPM values of DNA viruses.

Table S8 Taxonomic classification and Metagenomic (metaG) TPM values of DNA viruses detected in this study

Table S9 Taxonomic classification and Metatranscriptomic (metaT) TPM values of RNA viruses detected in this study

Table S10 Taxonomic classification and Metatranscriptomic (metaT) TPM values of DNA viruses detected in this study

Table S11 Characterization and Metatranscriptomic (metaT) TPM values of 742 MAGs recovered from metavirome (n = 417) and metagenome (n = 325) samples

Table S12 Metadata of Antibiotic Resistance Genes (ARGs) identified in this study

Table S13 Metadata of virulence factors (VFs) identified in this study

Table S14 Metadata of mobile genetic elements (MGEs) identified in this study

Table S15 Taxonomic classification of unbinned contigs.

Table S16 Structural prediction results of high-risk antibiotic resistance genes (ARGs) using AlphaFold 3 and Phyre2.2.

Table S17 qPCR-based quantification of antibiotic resistance genes (ARGs) and mobile genetic elements (MGEs)

Table S18 MAPseq-based alignment results of putative pathogenic bacterial 16S rRNA gene sequences

Table S19 Overview and metadata of the curated pathogenic bacterial genome database

Table S20 Biosafety levels and metaviromic (metaV) TPM values of DNA viral pathogens detected in this study

Table S21 Biosafety levels and metatranscriptomic (metaT) TPM values of RNA viral pathogens detected in this study

Table S22 Biosafety levels and 16S abundance of bacterial pathogens detected in this study

Table S23 Biosafety levels and metatranscriptomic (metaT) TPM values of risk ARGs detected in this study

Table S24 Biosafety levels and metagenomic (metaG) TPM values of risk ARGs detected in this study

Table S25 Richness diversity of potential pathogens and antibiotic resistance genes (ARGs) in the Yellow River water

Table S26 Risk assessment of microbial risks based on absolute quantification data

Table S27 Pearson correlation coefficients among environmental variables

References in the main text—continuous

81. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
82. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
83. Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
84. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
85. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
86. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2024).
87. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
88. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
89. Pons, J. C. et al. VPF-Class: taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. *Bioinformatics* **37**, 1805–1813 (2021).
90. Shang, J., Jiang, J. & Sun, Y. Bacteriophage classification for assembled contigs using graph convolutional network. *Bioinformatics* **37**, i25–i33 (2021).
91. Aylward, F. O. & Moniruzzaman, M. ViralRecall-A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in 'Omic Data. *Viruses* **13**, 150 (2021).
92. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
93. Aroney, S. T. N. et al. CoverM: read alignment statistics for metagenomics. *Bioinformatics* **41**, btaf147 (2025).
94. Liu, Y. et al. Diverse novel RNA viruses in the cryosphere of the Tibetan Plateau. *Natl Sci. Rev.* **12**, nwaf361 (2025).
95. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
96. Zayed, A. A. et al. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156–162 (2022).
97. Koonin, E. V. et al. Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol. Mol. Biol. Rev.* **84**, e00061–19 (2020).
98. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
99. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).
100. Hou, X. et al. Using artificial intelligence to document the hidden RNA virosphere. *Cell* **187**, 6929–

6942.e16 (2024).
101. Sakaguchi, S. et al. NeoRdRp: A Comprehensive Dataset for Identifying RNA-dependent RNA Polymerases of Various RNA Viruses from Metatranscriptomic Data. *Microbes Environ.* **37**, ME22001 (2022).
102. Neri, U. et al. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023–4037.e18 (2022).
103. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
104. Dominguez-Huerta, G. et al. Diversity and ecological footprint of Global Ocean RNA viruses. *Science* **376**, 1202–1208 (2022).
105. Camargo, A. P. et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.* **51**, D733–D743 (2023).
106. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
107. Roux, S. et al. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol.* **21**, e3002083 (2023).
108. Yin, X. et al. ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics* **34**, 2263–2270 (2018).
109. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
110. Yin, X. et al. ARGs-OAP v3.0: Antibiotic-Resistance Gene Database Curation and Analysis Pipeline Optimization. *Engineering* **27**, 234–241 (2023).
111. Chen, L. et al. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* **33**, D325–D328 (2005).
112. Johansson, M. H. K. et al. Detection of mobile genetic elements associated with antibiotic resistance in Salmonella enterica using a newly developed web tool: MobileElementFinder. *J. Antimicrob. Chemother.* **76**, 101–109 (2021).
113. von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 217 (2019).
114. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
115. Zhao, Y. et al. Global soil antibiotic resistance genes are associated with increasing risk and connectivity to human resistome. *Nat. Commun.* **16**, 7141 (2025).
116. Pedersen, T. L. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. R package version 2.2.2 (CRAN, 2025). doi: 10.32614/cran.package.ggraph.
117. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
118. Powell, H. R., Islam, S. A., David, A. & Sternberg, M. J. E. Phyre2.2: A Community Resource for Template-based Protein Structure Prediction. *J. Mol. Biol.* **437**, 168960 (2025).
119. Li, H.-Q. et al. High-risk ARGs (HRA) Chip: A high-throughput qPCR-based array for assessment of high-risk ARGs from the environment. *Water Res.* **262**, 122106 (2024).
120. Liu, K. et al. Effects of land-use patterns on the biogeography of the sediment bacteria in the Yarlung Tsangpo River. *FEMS Microbiol. Ecol.* **98**, fiac125 (2022).

121. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
122. Ma, W. et al. Exploring methanogenic archaea and their thermal responses in the glacier-fed stream sediments of Rongbuk River basin, Mt. Everest. *FEMS Microbiol. Ecol.* **101**, fiaf044 (2025).
123. Chen, L. et al. Permafrost carbon cycle and its dynamics on the Tibetan Plateau. *Sci. China Life Sci.* **67**, 1833–1848 (2024).
124. Hulo, C. et al. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res.* **39**, D576–D582 (2011).
125. Lefkowitz, E., Smith, D. B. & Hendrickson, R. ICTV 2024 Virus Metadata Resource (VMR). Zenodo, VMR_MSL40.v1.20250307, https://doi.org/10.5281/zenodo.15042310 (2025).
126. R Core Team. R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing, Vienna, Austria, 2025).
127. Oksanen, J. et al. vegan: Community Ecology Package. R package version 2.7-3 (CRAN, 2026). doi: 10.32614/cran.package.vegan.
128. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).
129. Archer, E. rfPermute: Estimate Permutation p-Values for Random Forest Importance Metrics. R package version 2.5.5 (CRAN, 2025). doi: 10.32614/CRAN.package.rfPermute.