Published online 3 August 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 3 **1**
https://doi.org/10.1093/nargab/lqac057

Mining bacterial NGS data vastly expands the complete genomes of temperate phages

Xianglilan Zhang [1,†], Ruohan Wang [2,†], Xiangcheng Xie [3,†], Yunjia Hu [4,5,†], Jianping Wang [2,†], Qiang Sun [6], Xikang Feng [7], Wei Lin [4], Shanwei Tong [8,9], Wei Yan [16], Huiqi Wen [1,4], Mengyao Wang [2], Shixiang Zhai [10,11,12], Cheng Sun [13], Fangyi Wang [14], Qi Niu [13], Andrew M. Kropinski [15], Yujun Cui [1], Xiaofang Jiang [16], Shaoliang Peng [13,*], Shuaicheng Li [2,*] and Yigang Tong [4,*]

[1]State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing 100071, People's Republic of China, [2]Department of Computer Science, City University of Hong Kong, Hong Kong 999077, People's Republic of China, [3]College of Computer, National University of Defense Technology, Changsha 410073, People's Republic of China, [4]Beijing Advanced Innovation Center for Soft Matter Science and Engineering (BAIC-SM), College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100029, People's Republic of China, [5]School of Medicine, Shanghai University, Shanghai 200444, People's Republic of China, [6]The 964th Hospital, Changchun 130021, People's Republic of China, [7]School of Software, Northwestern Polytechnical University, Xi'an 710072, People's Republic of China, [8]Bioinformatics Graduate Program, University of British Columbia, VancouverBC V6T 1Z4, Canada, [9]Faculty of Health Sciences, Simon Fraser University, Burnaby, BC V5A 1S6, Canada, [10]Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai 264003, People's Republic of China, [11]University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China, [12]Center for Ocean Mega-Science, Chinese Academy of Sciences, Qingdao 266071, People's Republic of China, [13]School of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, People's Republic of China, [14]Department of Statistics, the Ohio State University, Columbus, OH 43210, USA, [15]Departments of Food Science, and Pathobiology, University of Guelph, Guelph, ON N1G 2W1, Canada and [16]National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received August 19, 2021; Revised June 13, 2022; Editorial Decision June 27, 2022; Accepted July 20, 2022

ABSTRACT

Temperate phages (active prophages induced from bacteria) help control pathogenicity, modulate community structure, and maintain gut homeostasis. Complete phage genome sequences are indispensable for understanding phage biology. Traditional plaque techniques are inapplicable to temperate phages due to their lysogenicity, curbing their identification and characterization. Existing bioinformatics tools for prophage prediction usually fail to detect accurate and complete temperate phage genomes. This study proposes a novel computational temperate phage detection method (TemPhD) mining both the integrated active prophages and their spontaneously induced forms (temperate phages) from next-generation sequencing raw data. Applying the method to the available dataset resulted in 192 326 complete temperate phage genomes with different host species, expanding the existing number of complete temperate phage genomes by more than 100-fold. The wet-lab experiments demonstrated that TemPhD can accurately determine the complete genome sequences of the temperate phages, with exact flanking sites, outperforming other state-of-the-art prophage prediction methods. Our analysis indicates that temperate phages are likely to function in the microbial evolution by (i) cross-infecting different bacterial host species; (ii) transferring antibiotic resistance and virulence genes and (iii) interacting with hosts through restriction-modification and CRISPR/anti-CRISPR systems. This work pro-

*To whom correspondence should be addressed. Tel: +86 10 64 45 17 81; Email: tong.yigang@gmail.com
Correspondence may also be addressed to Shuaicheng Li. Tel: +852 34429412; Email: shuaicli@cityu.edu.hk
Correspondence may also be addressed to Shaoliang Peng. Email: slpeng@hnu.edu.cn
†The authors wish it to be known that, in their opinion, the first five authors should be regarded as Joint First Authors.

vides a comprehensively complete temperate phage genome database and relevant information, which can serve as a valuable resource for phage research.

INTRODUCTION

Temperate phage is a vital component of the microbiome. The temperate phagescan undergo both lysogenic and lytic cycles. In the lytic cycle, the temperate phages can kill the infected bacteria to release phage descendants. However, when the temperate phages integrate (as prophages) into bacterial chromosomes, they usually enter the lysogenic cycle to participate in essential bacterial cellular processes (1). In the lysogenic cycle, they become prophages that replicate with the bacterial (host's) genomes while the host cell divides (2). Temperate phages can regulate bacterial gene expression and behavior in pathogenic and environmental bacterial species (3). As the temperate phages have an inherent capacity to mediate the gene transfers between bacteria when the temperate phages integrate into bacterial chromosomes, they may increase bacterial virulence and promote antibiotic resistance (2,4).

Temperate phages widely exist in their hosts and can act as weapons of bacterial competition by encoding defense mechanisms, such as Restriction-Modification (RM) systems, CRISPRs/anti-CRISPRs (2,5,6). Given the long and complex dynamic interaction of bacteria and their phages, temperate phages possibly provide many additional viral defence systems that have yet to be fully identified (7). Even though the RM systems are often found on bacterial chromosomes, temperate phages can also protect the bacteria against virulent phage infection by encoding restriction systems (8). The phages also encode the CRISPR-Cas system to counteract a phage inhibitory chromosomal island of the bacterial host (9). Six types of CRISPR-Cas systems (I, II, III, IV, V, VI) have been defined and updated based on Cas protein content and arrangements in CRISPR-Cas loci (10). On the other hand, anti-CRISPRs provide a powerful defense system that helps phages escape injury from the CRISPR-Cas system (5). For example, anti-CRISPR proteins with anti-I-E and anti-I-F activities have been found in *Pseudomonas aeruginosa* phages (11,12); the anti-CRISPR proteins inhibiting type II-C systems have been found in *Neisseria meningitides* (13).

A comprehensive temperate phage genome database is necessary for phage-related clinical applications. For example, phage therapy uses virulent phages to eliminate pathogenic bacteria. The phage therapy virulent candidate accidentally mixed with temperate phages that may increase pathogenic risks by possibly mediating transfer of antibiotic resistance genes or virulence factors among bacteria (4). Therefore, temperate phage detection mitigates the therapy risks. Also, isolating suitable virulent phages is challenging for anaerobes, such as *Clostridioides difficile* and *Mycobacterium tuberculosis* (14–16). Hence, genetically modifying temperate phages into virulent phages is promising to prevent bacterial infections. In recent years, fecal microbiota transplantation (FMT) has increasingly become a prominent therapy to treat *C. difficile* infection (CDI) by normalizing microbial diversity and community structure in patients (17,18). FMT may also help manage other disorders associated with gut microbiota alteration (17). However, drug-resistant and highly virulent bacteria in the donor's stool pose a potential threat to the patient's life. Temperate phages transfer antibiotic resistance genes and virulence factors among bacteria. Identifying and profiling the temperate phages that carry important antibiotic resistance genes and virulence genes is crucial for FMT success.

Even though temperate phages are highly prevalent in bacterial genomes and play essential roles in medical treatments (4,19,20), a few complete temperate phage genomes are accessible until now, seriously hindering the related research. The temperate phages have been largely ignored mainly due to the challenges in detecting them and achieving meaningful annotations. Traditionally, phage detection relies on culture-based methods to isolate the temperate phage after inducing it from a lysogenic strain, amplify the phage to high titers, and characterize it often by transducing the phage into different strains (21). However, as temperate phages are readily integrated into the host genome and stay silent, forming phage plaques is difficult after transduction, which causes a big challenge for the traditional culture method.

Nowadays, an increasing amount of bacterial genome sequences have become available in databases due to the advances in next-generation sequencing (NGS) technologies. The sequence data harbor many phage sequences, most of which are the remains of temperate phages. Due to the mutation or deletion of some phage genes, these sequences can no longer produce replicable phages. Only a few phage-like sequences in the bacterial genome remain functional and can produce active phages again.

In most cases, the current bioinformatics methods hardly obtain accurate and complete temperate phage genome sequences. Bioinformatics tools have been developed to predict potential prophage sequences within assembled bacterial genomes, including Phage_Finder (22), Prophage finder (23), Prophinder (24), PHAST (25), PHASTER (26), PhiSpy (27), VirSorter (28), Prophage Hunter (29) and VIBRANT (30). Among these tools, PHASTER/PHAST, VirSorter and Prophage Hunter consider the completeness of a predicted prophage sequence. Typically, these tools rely on phage protein clusters to locate possible prophage regions on assembled bacterial genome sequences, yet fail to acquire exact active prophage (temperate phage) sequence boundaries (25,27,29,30). The above bioinformatics methods are suitable for predicting phage remnants while cannot determine whether the predicted phage sequences are inducible.

This study developed a computational temperate phage detection (TemPhD) method to detect complete temperate phage genome sequences using the raw data of bacterial next-generation genome sequencing. Unlike other *in silico* tools adopting machine learning or statistical classifiers to predict possible prophage regions from the bacterial genomes (Table 1), our method identifies the temperate phages by incorporating the biological principle. That is, the temperate phages exhibit spontaneous induction, during which the linear phage genome circularizes and replicates. As long as a circular phage genome sequence is detected, this sequence is recognized as a complete temperate phage genome sequence. In principle and theoretically speaking,

Table 1. General characteristics of our method TemPhD and eight commonly used prophage prediction tools. In the feature of 'Latest Update', we listed the years when the methods were last updated, not the years when the methods were first shown in public. The last updated years can be founded on their websites. In the feature of 'Boundary Identification', the 'repeated sequence' is used to identify the flanking sites (*attP* and *attB* sites) of temperate phage and its integrated host genome, the paired-end reads of 'NGS data' is used to check the completeness of the temperate phage genome sequence

Features		TemPhd	vibrant	Prophage Hunter	PHASTER	virsorter	prophinder	phispy	phage finder	prophage finder
lastEST update		2021	2020	2019	2016	2021	2015	2021	2012	2006
Target		Temperate phage (active prophage)	Virus	Prophage	Prophage	Virus	Prophage	Prophage	Prophage	Prophage
Essential difference		capture replication process of temperate phages	computational prediction	computational prediction	computational prediction	computational prediction	computational prediction	computational prediction	computational prediction	computational prediction
Input		Bacterial NGS data	Assembled bacterial genome sequence	Assembled bacterial genome sequence	Assembled bacterial genome sequence	Assembled bacterial genome sequence	Bacterial genome sequence in GenBank format	Assembled bacterial genome sequence	Assembled bacterial genome sequence	Assembled bacterial genome sequence
METHOD	**Rough region identification**	Alignment-based	Annotation-based	Alignment-based	Alignment-based	Alignment-based	Statistic-based	Machine learning models	Alignment-based	Alignment-based
	Boundary Identification	Repeated sequence and NGS data	—	Repeated sequence	Repeated sequence	—	—	—	Repeated sequence	—
	Activity/ Completeness determination	Detecting with NGS data	Annotation-based	Machine learning models	Scoring	—	—	—	—	—
OUPUT		Temperate phages (real active prophage)	Prophage/negative	active/ambiguous/ inactive	intact/questionable/ incomplete	full/partial	prophage/negative	prophage/negative	prophage/negative	prophage/negative
Available via		https://github.com/NancyZxll/temperate-phage-active-prophage-detection	https://github.com/AnantharamanLab/VIBRANT	https://pro-hunter.genomics.cn/	www.phaster.ca	https://github.com/jiarong/VirSorter2	http://aclame.ulb.ac.be/prophinder (web service only)	https://github.com/linsalrob/Phispy	http://phage-finder.sourceforge.net	http://bioinformatics.uwp.edu/~phage/DOEResults.php (out of maintenance)

TemPhD can detect all the temperate phages when they are spontaneously induced to a particular concentration from their host strains. The amount of temperate phage concentration allows NGS technology to generate the reads of circularized temperate phage genome sequences in the bacterial NGS data.

To validate TemPhD, we sequenced the bacterial strains preserved in our laboratory using NGS technology. TemPhD detected 17 temperate phages from 15 of 148 lab-preserved bacterial strains belonging to seven species (Extended Data Table 1). Subsequently, the wet-lab experiments were then conducted to induce the temperate phages from these bacterial strains. The induced temperate phages were then sequenced using NGS technology, and their genome sequences serve as ground truth.

We then benchmarked TemPhD with the seven state-of-the-art prophage prediction tools. In contrast to Prophage finder (23), Prophinder (24), PHASTER (26), PhiSpy (27), VirSorter (28), Prophage Hunter (29) and VIBRANT (30). TemPhD identified exact boundaries and acquired accurate temperate phage genome sequences (Supplementary Figure S1, Extended Data Table 2 and Supplementary Materials, Verification of our temperate phage detection method). We used TemPhD to analyze a large number of raw NGS host data sets from GenBank and increased the number of complete temperate phage sequences by ∼107-fold.

To further investigate how temperate phages influence their hosts, we examined the phage–host interactions and identified the mediated horizontal gene transfers (HGT), the restriction-modification (RM) systems, and anti-CRISPRs encoded by the temperate phage genome sequences extracted in our study. Our study shed light on the mosaicity and diversification of temperate phages at the genomic level, and illustrates that these temperate phages, along with their hosts, are interconnected through a complex network.

MATERIALS AND METHODS

Details of the temperate phage detection method proposed in this study

Our temperate phage detection method includes three main steps (Figure 1).

1) Next-generation sequencing (NGS) data processing. FastQC (31) and Trimmomatic (32) are first used for quality control and NGS reads filtering. The filtered NGS reads are fed into SPAdes (33) to acquire the assembled scaffolds. The SPAdes was used with default parameters and was designated '–meta' for metagenomic data. Recently, META-SPADES (34) was reported to be the best tool for assembly of individual genomes within a metagenome (35). META-SPADES can be used by simply adding the parameter '–mata' when using the SPAdes program. Here, we also suggested using the meta parameter when assembling metagenomic data.
2) Prophage region detection. Essentially, TemPhD incorporates three biological principles from phage life cycles to detect temperate phages: (i) In the replication process of the lytic cycle, the phage genome typically circularizes (36,37). (ii) During the cycle of lysogenization, a

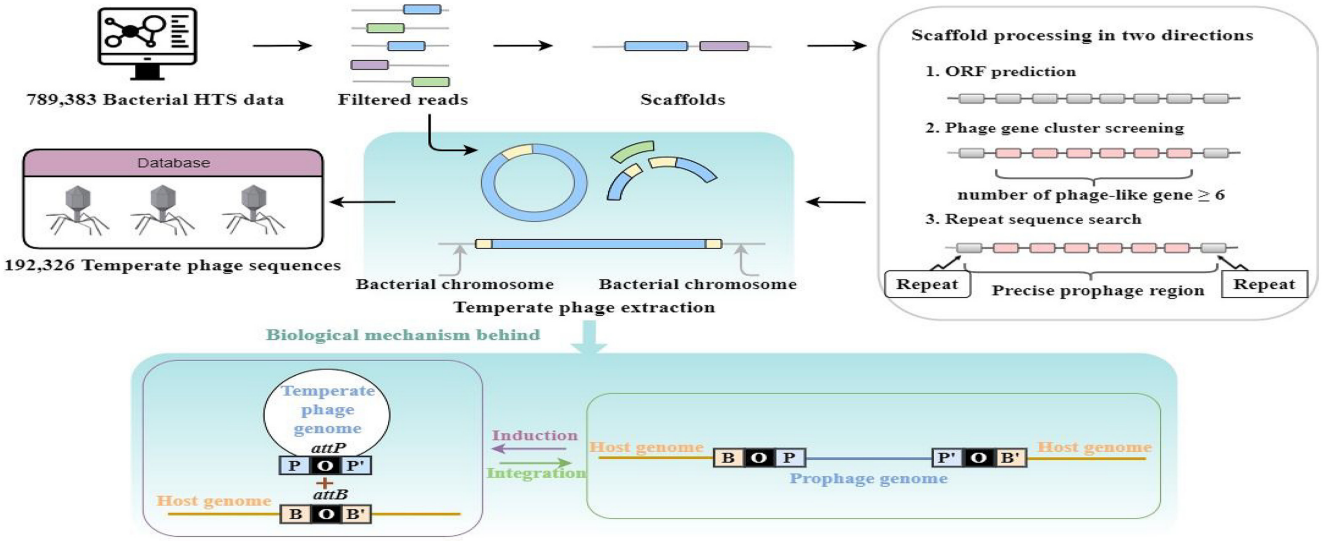

Figure 1. Workflow of our temperate phage detection method and illustration of temperate phage induction and integration processes. The main step of temperate phage detection is based on the temperate phage induction and integration processes, which is illustrated at the bottom of the figure. *attP* is short for attachment site of phage, *attB* represents the attachment site of host strain, *attL* stands for the left attachment site after integration, *attR* is short for the right attachment site after integration, O stands for core region of phage and bacterium, B represents host, while P represents phage. Temperate phage is also called prophage after integrating into a lysogenic host strain.

temperate phage integrates into the host's chromosome and shares a core (repeat) sequence of attachment sites (attP/attB) with its host's genome sequence (Figure 1). (iii) Active prophages usually undergo spontaneous induction at a relatively low frequency, resulting in a small number of temperate phages in the bacterial culture. To better find the genes in the above bacterial scaffolds, we use both Prodigal (38) (a dependency of Prokka (39)) and GLIMMER v3.023 (40) to predict open reading frames (ORFs).The ORF is predicted by either Prodigal or GLIMMER. For the overlapped ORF region predicted by both of them, we keep the longer ORF in our final results. For Prodigal and Prokka, we use the default settings to predict ORFs and annotate the genes, respectively. For GLIMMER, we train it separately on each bacterial scaffold. Specifically, we first use the 'long-orfs' program to identify long, non-overlapping ORFs in each bacterial assembled DNA scaffold. The parameters of 'long-orfs' were: -z 11 –n -t 1.15 –l. That is, (i) use Genbank translation table number 11 to specify stop codons; (ii) do not include the program-settings header information in the output file; (iii) only genes with an entropy distance score <1.15 will be considered; (iv) assume a linear rather than circular genome. Then the 'extract' program was taken to acquire the training sequences as standard output by giving both the bacterial assembled DNA scaffolds and the identified ORFs in the first step. The 'build-icm' program then constructed an interpolated context model (ICM) from the training sequences. Then the glimmer3 program itself was run to analyze the sequences and predict ORFs. Finally, 'multi-extract' program was conducted to acquire the predicted ORFs in the nucleotide FASTA file. BLASTp is then used to align the ORFs with the local phage protein database generated from the GenBank public protein database. To acquire as many prophage regions as possible, hypothetical and functional phage genes, including capsid, terminase, protease, integrase, transposase, lysis, tail, spike, holin, portal, and baseplate, are all taken into consideration. The clustering algorithm DBSCAN can discover arbitrary-shaped clusters and eliminate noisy data. Therefore, DBSCAN (density-based spatial clustering of applications with noise) (41) is used to identify phage gene clusters (Extended Data Table 3). The DBSCAN uses two parameters: the size of the cluster (M) and the radius of the cluster (E). Here, M is the minimum number of phage genes in a possible prophage region, and E is the largest distance between two neighbour genes. The default settings of M and E were based on our daily phage analysis findings. The prophages normally contain more than five phage genes, thus we set M as five. Also, to identify as many rough prophage regions as possible, we set E as five, illustrating that there are at most five other genes between two neighbor phage genes. In our work, we have analyzed many phage genome sequences and found that >90% of them have lengths longer than 30 000 bp. Therefore, in this study, a potential prophage region includes more than 30,000 bases before and after the center on the same scaffold. The center is a hypothetical phage gene cluster or a functional phage gene cluster. The integrated overlap of any two afore-defined prophage regions by DBSCAN is treated as a rough prophage region. TemPhD defines a precise prophage region based on the integration and induction mechanism of a temperate phage (Figure 1). The *attL* and *attR* in Figure 1 are treated as the termini of a temperate phage when the phage inserts into a bacterial chromosome. Considering that *attL* and *attR* have a 14–50 bp core region, TemPhD defines two sliding windows with the same sizes to detect the core region. The

distance between the two sliding windows is defined as $d \in$[10 000, length (rough prophage region)]. Considering that most prophage genomes are larger than 10 000 bp, the initial value of the distance is set as 10 000 bp. For each d, the sequences in the two sliding windows are compared base by base (Supplementary Figure S2). The comparison stops once the two bases in the two windows are different or reach the end of the window. The repeats with sequence between them are recorded as the terminal candidates of the precise prophage region.

3) Temperate phage detection. In our method TemPhD, a prophage is treated as a temperate phage if the prophage can circulate itself. As shown in Supplementary Figure S3, the 1000 bp sequences at both ends (A and B) of the precise prophage region are extracted, with their positions being reverted to form a new sequence. Currently, most NGS platforms use paired-end sequencing technology. If this prophage can circulate itself, there must be paired reads that match A and B concurrently, which is evidence that NGS captures the replication status in the lytic cycle of the temperate phage. Finally, a complete temperate phage sequence is acquired after using an in-house sequence-end-extend script to fill the gap between the paired reads matching A and B.

In TemPhD, the scaffolds with character N inside are processed in two directions (red rectangle in Figure 1). The reason for doing this is that when 'N' appears in the scaffold, these Ns cause GLIMMER v 3.023 to predict ORFs differently in the original and reverse orders of the given sequence. That is, we may miss some temperate phages if searching from only one direction. Therefore, reverse complementary forms of these sequences containing character N are considered so that more temperate phages (within the reverse complementary host genome) are identified. All the identical and reverse complementary phage sequences in one host sample were treated as one phage and hence only kept one sequence in our analysis. To balance between quality and variety, we excluded a temperate phage sequence if the number of character N in the sequence takes up more than 5%.

Biological verification of temperate phages detected in this study

As TemPhD is host species-free, we randomly chose 148 lab-preserved bacterial strains of 21 species and sequenced them (Extended Data Table 1). Mitomycin C was used to induce the potential temperate phages in the 148 bacteria. Mitomycin C is reported as a standard chemical treatment to induce temperate phages (42). Considering the efficacy and safety of mitomycin C in temperate phage induction, we used it as our inducer. The nucleotides in bacterial culture supernatant were extracted to do NGS to identify the temperate phages inside.

Specifically, this biological verification includes five steps:

1) Inducing temperate phages. A single bacterial colony was extracted and put in a 5 ml LB liquid medium. The cultures were incubated at 37°C overnight and then added to 400 ml of fresh LB medium to grow to the stationary phase ($OD_{600} = 0.5$). Mitomycin C was added (1 g/ml) into the medium and then incubated at 37°C for 12 h until the medium became clear. The bacterial culture supernatant was then collected.
2) Extracting temperate phages. The phage was precipitated using a modified PEG 8000 and subsequent chloroform extractions protocol (43). The above extracted bacterial culture supernatant was then added 23.4 g NaCl to acquire a final concentration of 1 mol/L, then stirred and cooled for 1 h in an ice bath. The cooled mixture was then centrifuged at 1000 × g at 4°C for 10 min. The supernatant in the mixture was then collected and transferred into a clean 500 ml flask. PEG8000 was then added to the supernatant at 10 mg/100 ml, followed by being stirred and cooled for 3 h in an ice bath. The culture was then centrifuged at 1000 × g at 4°C for 10 min. After centrifugation, the phage precipitation was then collected and resuspended in SM buffer. The same volume of chloroform was added to the phage precipitation. The phage was then collected at its hydrophilic phase. This phage precipitation was then filtered using a 0.45 μm filter and stored at 4°C.
3) Purifying temperate phages. The phage particles were purified by isopycnic centrifugation through CsCl gradients. The high quality of solid CsCl was added into the SM buffer and made three kinds of CsCl solutions with different densities (P: 1.45 g/ml, P: 1.50 g/ml, P: 1.70 g/ml). Each CsCl solution was added into different Beckman Ultra-Clear centrifuge tubes, which were further added 10 ml phage precipitation prepared at step 2. The tubes were then centrifuged at 25 000 r/min at 4°C for 3 h. The phage layer was then transferred to a new centrifuge tube to remove CsCl using SM buffer in a 100 kDa dialysis bag for 10 h. Finally, the pure phage suspension was acquired.
4) Extracting temperate phage genomes. According to a modified standard phenol-chloroform extraction protocol (43), DNase I and RNase A (Thermo Fisher Scientific, MA, USA) with a final concentration of 1 μg/ml were added to the above-purified phage suspension and incubated at 37°C for 10 h. After inactivation at 80°C for 15 min, the lysis buffer with a final concentration of 0.5% SDS (Sigma-Aldrich, Darmstadt, Germany), 50 μg/ml protease K (Thermo Fisher Scientific, MA, USA), and 20 mM EDTA (Solarbio, Beijing, China) was added and incubated at 56°C for 1 h. An equal volume of phenol–chloroform–isoamyl alcohol (25:24:1) (Solarbio, Beijing, China) was added to the mixture, then centrifugated at 12 000 × g for 10 min. The aqueous phase was collected. An equal volume of isopropanol (Macklin, Shanghai, China) was added to the aqueous and incubated at –20°C for at least 1 hr. After 10 000 × g centrifugation at 4°C for 20 min to precipitate phage genomes DNA, the precipitation was collected and washed twice with 1 ml of 75% cold ethanol and then resuspended in 30 μl of deionized water.

5) High-throughput sequencing for temperate phage genomes. The preparation of paired-end libraries and whole-genome sequencing by generating 2 × 150 bp paired-end reads were performed using the Illumina MiSeq sequencing platform by Annoroad Genomics Co., Ltd (China).

Method comparison of temperate phages detected in this study

The bacterial scaffolds assembled in the first step of our temperate phage detection method were used as the input for each prophage prediction method, including Phage_Finder (22), Prophinder (24), PHASTER (26), PhiSpy (27), VirSorter (28), Prophage Hunter (29) and VIBRANT (30). As Prophage finder (23) has been out of maintenance, we are unable to run it in our study. CLC Workbench v3 was used to map the NSG reads of the wet-lab induced temperate phages to their hosts' assembled scaffolds, with parameter settings length as 1.0 and sequence fraction as 1.0.

Lineage assignment to the hosts of the temperate phages

After applying five filtration criteria about LibraryLayout, LibraryStrategy, LibrarySelection, LibrarySource, and Platform on NGS data (Extended Data Table 4), we downloaded 789 383 raw NGS host data sets from NCBI sequence read archive (SRA) (March 2020). We performed TemPhD analysis on the downloaded NGS, obtained the temperate phages, and assembled host scaffolds for each SRA data. To ensure that the taxonomy information associated with the SRA data is correct, we blasted the assembled host scaffolds against the NCBI nucleotide (nt) database (downloaded on 16 May 2020). The species of the top blast hits (with the threshold of e-value $< 1 \times 10^{-5}$ and identity $> 90\%$) were assigned as the species of the assembled host scaffold. If the species of the top blast hits of the assembled scaffolds is inconsistent with that associated with the SRA data, we would assign the species of the top blast hits as the host species to the temperate phage. Otherwise, the taxonomy associated with the SRA data was used to assign the host species of the phage.

Two .dmp format files named names and nodes in the folder of taxdump were download from https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/. An in-house script in Python 3.6 was used to acquire the complete taxonomic lineages of the host based on National Center for Biotechnology Information (NCBI) Taxonomy. The lineage information of each phage is listed in Extended Data Table 5.

Analysis of temperate phage genomes

All the primary analysis in this study, including the distributions of sequence length and GC content, and genome classification, was conducted using the Python programming language (version 3.6.0) and text processing tool csvtk 0.19.1 (44). The related images were generated using the ggplot2 package (45). Genome annotations were conducted using Prokka (39), while the genomic mapping was generated using an in-house Python script. We grouped the same/reverse complementary temperate phage genomes as one phage entry using the in-house script. After excluding the data whose species listed on NCBI are inconsistent with their BLAST results, the host species listed in the NCBI record for the sequence is treated as the host species for each phage entry in our study.

The connected network was implemented by an in-house developed script in Python 3.6 and used to display the connections between phages and their multi-hosts, where nodes represent hosts while edges (lines) are phages. Within a temperate phage-host connection network, only hosts with identical/reverse complementary temperate phage will be connected. We blasted temperate phage genomes in gene sharing networks against the ARG/VFDB gene databases. A gene was reported if its coverage >90%. If this gene is identified in different temperate phages, these phages were connected. A cluster is formed as long as the hosts inside can be infected by at least one identical or reverse complementary temperate phage. In Figure 2, one independent temperate-phage-sharing network is defined as a cluster. We first detect the temperate phage from bacterial assembled sequences, and then group the identical temperate phage sequence as one entry. All the bacterial hosts where the identical temperate phage was detected are connected through this temperate phage sequence. Therefore, we labelled the bacterial hosts as nodes and the identical temperate phage infecting them as an edge. A single temperate-phage-sharing network (cluster) is formed when no other phage infects any other hosts outside of the cluster. Therefore, the cluster illustrates the taxonomic classification of the hosts of the temperate phages. With the development of whole-genome sequencing (WGS), the phylogenetic tree built on single nucleotide polymorphisms (SNPs) of core genomes is widely used in bacterial genome analysis. Better than using single or multiple conserved genes, the core SNPs on the whole genome capture all the genetic signals on the genome sequence. Therefore, we used core SNPs from whole-genome nucleotide alignments to build the phylogenetic tree of the bacterial host genome sequences. The core SNPs are defined as the SNPs that all the genome sequences have. As we aimed to acquire bacterial sequence clusters rather than evolution rates, we used the neighbor-joining (NJ) method to construct the phylogenetic tree.

An in-house script was used to build the phylogenetic tree of bacterial hosts. Specifically, we first set the reference genome sequence using the standard bacterial strain from the RefSeq database on Genbank (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/assembly_summary.txt). Our in-house script has integrated the assembly-based SNVcalling pipeline called core.ope (46). To be specific, the processes include nine steps: (i) use MUMMER to do pair-wise alignments between any bacterial strain and the reference bacterial genome sequence, with the alignment threshold set to 0.1; (ii) merge all the MUMmer alignment to a multiple fasta file; (iii) call relaxed core genome; (iv) call reference genome repeat region; (v) remove the repeat regions in the core genome; (vi) call SNPs from the core genome; (vii) remove the SNPs in the repeat regions; (viii) generate an SNP-only fasta file and (ix) build NJ tree.

When analyzing the phylogenetic relationships among temperate phages, we took the idea of the phylogenetic anal-

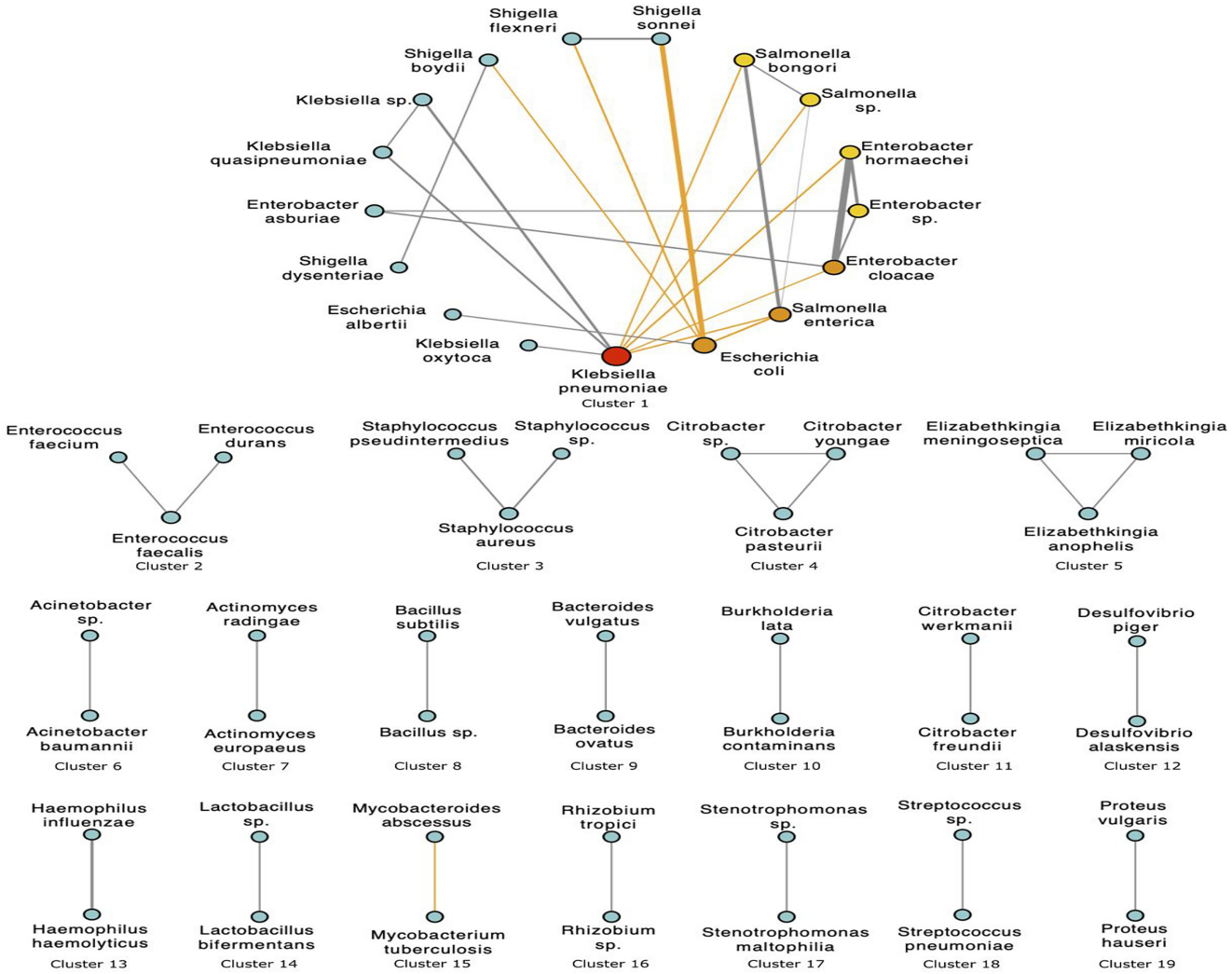

Figure 2. The connected network of temperate phages (edge) with their hosts (nodes). All data labeled as metagenome or unclassified by GenBank are not included. The more temperate phages shared, the thicker the line (the node) is. The gray line represents the same phages identified within the same host genus, while the yellow line represents the phages identified across the host genera.

ysis of their bacterial hosts. Considering the variety of temperate phage sequences, we calculated their phylogenetic distances based on their whole genomes. We first aligned all the 196 nonredundant temperate phage sequences (phage entries), which have multiple hosts (Supplementary Figure S4) and 98 phage entries that share the most number of hosts (Figure 2, Cluster 1). Then we built two neighbor-joining (NJ) trees based on the SNPs of the two sets of aligned sequences, 222 782 SNPs and 193 830 SNPs, separately. In particular, multiple alignments of the temperate phage sequences were implemented by MAFFT v5 (47). We used the interactive MAFFT program by choosing '3. Sorted fasta format' as our output format, '1. –auto' as strategy, no additional arguments as our parameters. The phylogenetic relationship was then analyzed by the neighbor-joining method implemented in MEGA X (48) with default settings and then displayed using iToL (49).

All the 147 phage entries formed host-sharing clusters (Fiugre 3A) were taxonomically clustered together using vContact2 and the ProkaryoticViralRefSeq94-Merged database with default parameters (50). The input of vContact2 was the phage entry ORFs annotated by Prokka (39).

Detection of antibiotic resistance genes, virulence genes, genomic islands, restriction-modification, CRISPR and anti-CRISPR systems

The large expansion of the temperate phage genomes and host species enabled us to examine the genomic diversity, perform a preliminary investigation of phage-host interactions, explore mediated horizontal gene transfers (HGT), and examine the encoded restriction-modification (RM) and anti-CRISPRs. Specifically, the antibiotic resistance genes were identified using the database in ResFinder (51), and virulence gene identification was conducted by VFDB (52). Genomic islands (GIs) are gene clusters in prokaryotic genomes of probable horizontal origin (53). IslandPath-DIMOB (54) was used to detect genomic islands with the input of phage genomes annotated by Prokka (39). We used IslandPath-DIMOB and Prokka by default settings.

The output of IslandPath-DIMOB is the start and end positions of the identified genomic island. The restriction-modification (RM) systems were searched against REBASE (55). BLAST was then used to align all the temperate phage sequences with the above databases by setting e-value as 1×10^{-5} and choosing 90% as the lower bound for sequence coverage. Moreover, the restriction (R) and the modification (M) enzyme genes are often tightly linked to form a restriction-modification (RM) gene complex (56). The type I systems also have sequence recognition (S) subunit genes to form multi-subunit enzymes for modification (SM) or restriction (SMR) (56,57). These genes are found to be separated by an intergenic region, such as 56 bp (58), 76 bp (59), 109 bp (60), 110 bp (61), 330 bp (8), 365 bp (62). Therefore, we set the upper bound of the distance between such genes as 400 bp. The CRISPR-Cas systems were searched using CRISPRCasFinder (63). Only the temperate phages containing both the CRISPR array and the Cas proteins are counted in our study. The anti-CRISPR systems were searched against the database Anti-CRISPRdb (5) by BLAST to align all the temperate phage sequences with setting e-value as 1×10^{-5} and choosing 90% as lower bound for sequence coverage.

RESULTS

Biological verification of our method TemPhD

To validate TemPhD, we sequenced the bacterial strains preserved in our laboratory using NGS technology. As TemPhD is host species-free, we randomly chose 148 lab-preserved bacterial strains of 21 species and sequenced them (Extended Data Table 1). 148 bacterial strains include *Staphylococcus aureus*, *Enterococcus faecalis*, *Gemella morbillorum*, *Alloiococcus otitis*, *Escherichia coli*, *Klebsiella pneumoniae*, *Acinetobacter baumannii* etc. In total, our computational method detected 17 temperate phages. The size ranges of the 17 phages are from 15 707 bp to 73 289 bp, with an average length of 42 969 bp, while 15 of them are from 32 427 bp to 55 642 bp. These 17 temperate phages are from 15 of these bacterial strains belonging to seven species, including *Klebsiella pneumoniae* (five temperate phages from five host strains), *Staphylococcus wokerii* (three temperate phages from two host strains), *Staphylococcus aureus* (three temperate phages from two host strains), *Serratia Marcescens* (two temperate phages from two host strains), *Staphylococcus haemolyticus* (one temperate phage from one host strain), *Escherichia coli* (one temperate phage from one host strain), *Citrobacter freundii* (two temperate phages from two host strains). Subsequently, the wet-lab experiments were then conducted to induce the temperate phages from these bacterial strains. It is noted that using mitomycin C does not always successfully induce all the temperate phages. No chemical or physical inducing treatment can induce all the temperate phages in nature. But fortunately, it does not impact our conclusion. The main point of our wet-lab experiments is to verify our method. That is, all the temperate phages induced by mitomycin C are real temperate phages. As long as the wet-lab induced temperate phages are also detected by our method, our method is verified—the temperate phages detected by our method are real temperate phages. Next, these wet-lab induced temperate phages were sequenced using NGS technology, and their genome sequences were treated as ground truth.

In contrast to Prophage finder (23), Prophinder (24), PHASTER (26), PhiSpy (27), VirSorter (28), Prophage Hunter (29) and VIBRANT (30), only our method TemPhD acquired accurate temperate phage genome sequences (Supplementary Figure S1, Extended Data Table 2 and Supplementary Materials, Verification of our temperate phage detection method). In our study, we compared the performance of prophage predictions by whether identifying the exact boundaries of the phage genome sequences. We used CLC Workbench v3 to map the NSG reads of the wet-lab induced temperate phages to their hosts' assembled scaffolds, with parameter settings length as 1.0 and sequence fraction as 1.0. For all the 17 bacterial assembled sequences, TemPhD detects the exactly same regions (100% nucleotide identity and coverage) of the temperate phage genome sequences as the wet-lab experiments. Phage_Finder does not report any phages in 12 of 17 bacteria, followed by Prophage Hunter (missing four temperate phages), and PhiSpy (missing two temperate phages). PHASTER acquires two temperate phage genome sequences with the exactly same regions (100% nucleotide identity and coverage) as the wet-lab experiments (the temperate phages in Bac1320Scaffold5 and Bac1325Scaffold5), PhiSpy obtains the accurate start position of the temperate phage in Bac2747Scaffold6, Prophage Hunter acquires the accurate start position of the temperate phage in Bac2756Scaffold3 (Supplementary Figure S1).

We also compared the calculation time and memory usage of the offline tools, including VirSorter, VIBRANT, PhiSpy, Phage_finder and TemPhD (Supplementary Figure S5, details in Extended Data Table 6). Due to the consideration of sequence boundaries (Table 1), VirSorter and TemPhD take more calculation time than the other three tools. Notably, TemPhD takes much less time than VirSorter and is comparable to VIBRANT,which does not consider sequence boundaries. Because TemPhD needs bacterial NGS reads to detect sequence circulation, it requires the most memory space (∼700MB) among all the offline tools.

Expansion of the temperate phages

The expansion of the temperate phages includes the temperate phage genome sequences and their host species. We discovered 192,326 complete temperate phage sequences within 2717 host species of 710 host genera (Extended Data Table 7, Supplementary Materials, Expansion of temperate phages). Using integrase/transposase as a marker, we identified 1800 complete temperate phage genomes with 1790 nonredundant sequences on GenBank (December 2021, Extended Data Table 8). Compared with all the 1,800 GenBank public complete temperate phage genomes of 186 host species and 93 host genera, our result represents an ∼107-fold (192 326/1800) increase in the number of complete temperate phage genomes, with an ∼15-fold (2717/186) increase in the number of host species, and a 8-fold (710/93) rise in the number of host genera. Our work represents an

~37-fold (66,823/1,790) increase in the number of nonredundant temperate phage genome sequences. All the temperate phage genome sequences are freely available at https://phage.deepomics.org/. Among the host species of all the detected 192 326 temperate phages, *Salmonella enterica* contains the largest number of temperate phages, followed by *Escherichia coli*, *Listeria monocytogenes*, *Klebsiella pneumoniae*, *Staphylococcus aureus* etc. (Supplementary Figure S6C). The temperate phages in *L. monocytogenes* have the widest genome size ranges (from 40 407 bp to 77,097 bp), while these in the human gut metagenome have the broadest range of GC content, from 37.1 to 52.3 (Figure 3A and B, Extended Data Table 9). Considering that metagenome contains many bacteria, it is reasonable that the temperate phages in metagenome have the broadest range of GC content. Despite metagenomes, the temperate phages in *Neisseria gonorrhoeae* have the broadest range of GC content (from 44.2 to 54.7) (Figure 3B). Compared with temperate phages, their host species have a relatively narrow range of GC content, with most of them having relatively consistent GC content (Figure 3C). We also compared the core regions of the temperate phages and their hosts, where the sizes range from 13 bp to 248 bp (Extended Data Table 10). Within the top 20 most host species, the core regions of temperate phages in *S. enterica* and *E. coli* have the broadest size range, followed by *S. sonnei*, *L. monocytogenes*, *Haemophilus influenzae* (Supplementary Figure S7).

It is noted that identical temperate phage genome sequences can be identified in different bacterial genomes. This is the basic fact to analyze phage-host interactions. Furthermore, by aligning the 192 326 temperate phage genome sequences with each other, the identical/reverse complementary sequences were treated as one phage entry. Thus we harvest 66 823 nonredundant complete temperate phage genomes (phage entries).

To compare the 192 326 temperate phage genome sequences with the phage genome sequences in Genbank, we downloaded all the 5907 complete phage genome sequences in Genbank (December 2021), including all the virulent phages and temperate phages. We then aligned all the 192 326 temperate phages to the 5907 public phage genome sequences. The 5907 sequences have already included the previously downloaded 1800 complete temperate phages. Totally, 10 909 of 192 326 temperate phage genome sequences, which include 604 nonredundant genomes/phage entries, are fully aligned to the phage genome sequences in Genbank (Extended Data Table 11).

Multiple host infection of temperate phages

By analyzing temperate phage-host interactions (Supplementary Materials, Multiple host infection of temperate phages), we found many temperate phages in common pathogens, such as *Klebsiella pneumoniae*, *Salmonella enterica*, *Enterobacterclo*acae, *E. coli* (Extended Data Table 12). The host species sharing network shows that these temperate phages in common pathogens could also infect other species and genera. In particular, *K. pneumoniae* has the most temperate phages in common with other species, including *Klebsiella sp.*, *Klebsiella quasipneumoniae*, *Klebsiella oxytoca*, *S. enterica*, *Salmonella bongori*, *Salmonella* spp., *Enterobacter hormaechei* and *E.cloacae* (Figure 2, Cluster 1).

At the nucleotide sequence level of the hosts of these temperate phages, *K. pneumoniae* strains are most similar, the strains from the genera *Salmonella* group together, and the strains from *Enterobacter* have high similarity with each other (Figure 4A). On the other hand, *E. coli* shares the most temperate phages with *Shigella sonnei*, *Shigella flexneri*, *Shigella boydii* and *S. enterica*.

In the aspect of sequence similarity of these hosts, the *E. coli* strains are most similar with each other, *S. sonnei* strains have the highest similarity with each other, and the only two strains from *S. enterica* are most similar with the two strains from *E. coli* (Figure 4B). The phylogeny network shows that the temperate phages, constituting the largest cluster in the host species sharing network (Figure 2, cluster 1), are similar at the sequence level (Figure 4C, cluster 1).

In the aspect of taxonomic classification of the temperate phage themselves, the temperate phages in the biggest cluster (Cluster 1 in Figure 2) connected to those temperate phages in the other seven clusters, including cluster 4, 6, 10, 11, 13, 17 and 19 (Figure 5, VC1). The temperate phages infecting the bacterial genera of *Enterobacter*, *Salmonealla*, *Escherichia*, *Klebsiella*, *Burkholderia*, *Shigella*, *Stenotrphomonas* and *Acinetobacter*, also show high homology with the phages from other 26 genera, including *Pseudomonas*, *Yersinia*, *Vibrio*, *Mannheimia*, *Ralstonia* etc. (Figure 5, VC1). The temperate phages infecting *Enterococcus* (Figure 3, Cluster 2), *Staphylococcus* (Figure 3, Cluster 3), *Actinomyces* (Figure 3, Cluster 7), *Bacillus* (Figure 3, Cluster 8), and *Lactobacillus* (Figure 2, Cluster 14), also have high similarity with the phages from Deep-sea, *Listeria*, *Lactococcus*, *Weissella*, *Brochothrix*, *Croceibacter*, *Clostridium*, *Thermus*, and *Brevibacillus* (Figure 5, VC2). The temperate phages infecting *Mycobacterium* (Figure 2 Cluster 15) reveal high homology to the phages from *Gordonia* and *Tsukamurella* (Figure 5 VC4). Four independent VCs (VC3, VC5, VC6 and VC7) were formed by the temperate phages in Cluster 1 shown in Figure 2, displaying high similarities. Our results show high complexity in the similarity of temperate phages from different hosts, thus suggesting that the temperate phages have the potential to cross host hierarchies barriers.

Horizontal gene transfers mediated by temperate phages

The phage/gene sharing networks illustrate that temperate phages play essential roles in HGT among host species (Figure 2, Supplementary Figure S8, Supplementary Figure S9). In total, 31,932 (16.6%, 31 932/192 326) temperate phages were identified to encode antibiotic resistance genes and 12 089 (6.3%, 12 089/192 326) temperate phages encoded virulence factors. In particular, tetracycline, macrolide and aminoglycoside resistance genes were shared widely among temperate phages of different host species, as well as the virulence factors of adherence and invasion and secretion systems (Supplementary Figure S10A, B). Regarding the host species, the temperate phages in *E. coli* and *S. enterica* contain the most antibiotic resistance gene types (eight gene types for each). The temperate phages infecting *S. enterica* take up more than 50% of the total number of virulence fac-

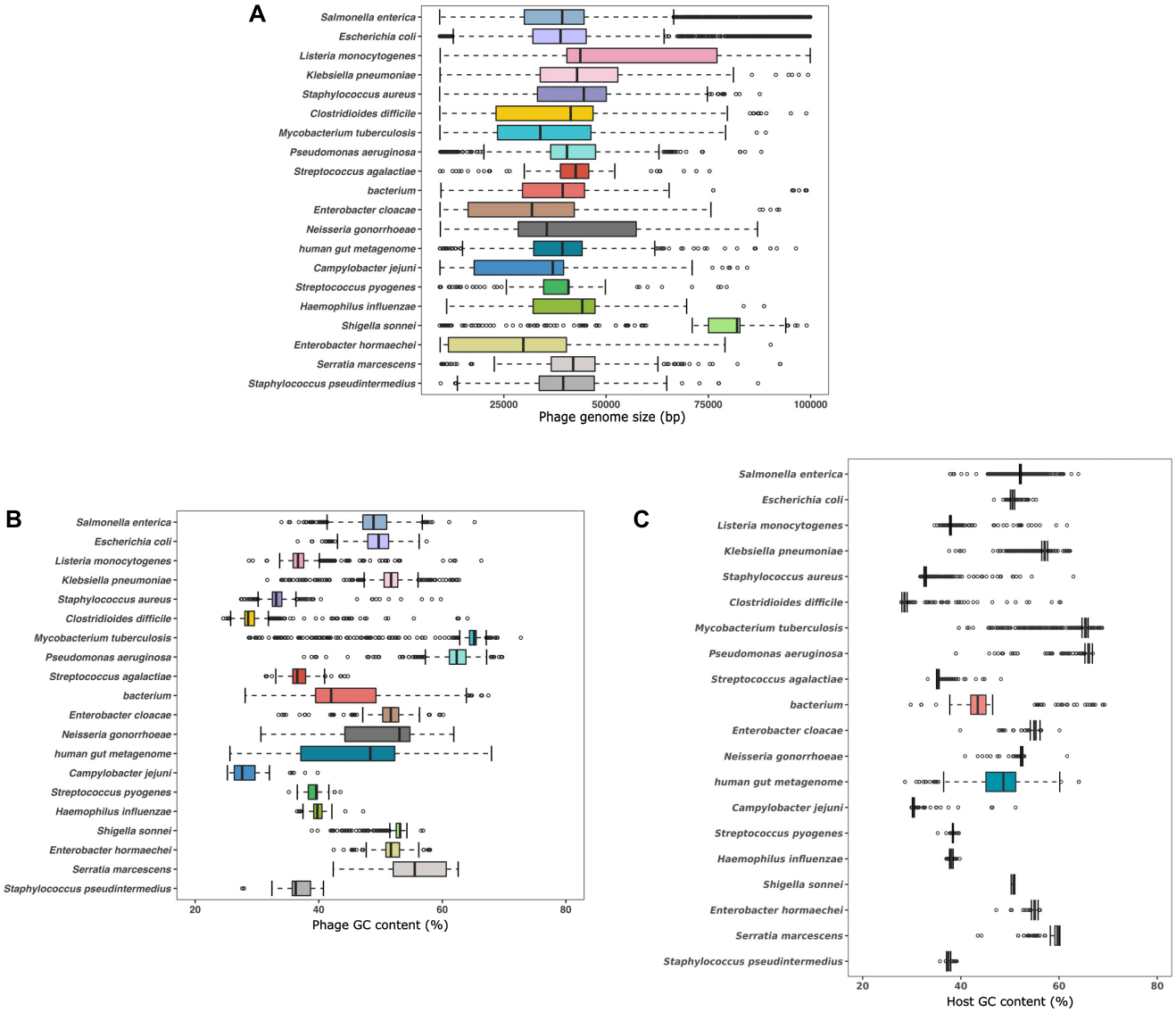

Figure 3. Genome size and GC content distribution of temperate phages in the top 20 host species listed on NCBI. We also display the GC content distribution of the top 20 host species here. We keep the original names in the item of NCBI host species, including bacterium and human gut metagenome. The 'bacterium' relates to NCBI taxonomy ID 1869227 which includes all the unclassified bacteria that have not been separated into NCBI taxonomic hierarchies.

tors (Extended Data Table 13). For the size distributions of the top 20 host species containing the most GI entries, the GI entries in phages of the species *S. sonnei* have the broadest size range (from 4657 bp to 16 631 bp) while with a relatively consistent GC content (50.3–56.0%) (Supplementary Figure S11A, B). Even though the phages in *S. enterica* contain the most GI entries, most of them have a relatively narrow size range (5845–7927 bp) with GC content from 45.1% to 54.0% (Supplementary Figure S11A, B, Extended Data Table 14).

Examination of restriction-modification systems encoded by temperate phages

Restriction-modification (RM) systems have been classified into three classes designated I, II and III (64,65). By aligning with REBASE (55), this study preliminarily identified the three RM systems encoded by 2626 (1.4%, 2626/192 326) temperate phages, containing 76 host species (Supplementary Materials, Examination of restriction-modification systems encoded by temperate phages). Our research also showed that type II RM systems widely existed in the most temperate phages with different host species (Supplementary Figure S10C, Extended Data Table 15). We have also researched the host species of these temperate phages, to identify if the host species contain known RM systems (Extended Data Table 16). In the previously published findings, *C. jejuni*, *E. coli*, *P. sanguinis*, *S. enterica*, *S. aureus*, *S. suis* and *L. monocytogenes* encode all the three RM systems (I, II and III). On the other hand, the RM system II has been found in most host species. Although RM systems are the weapons for the hosts to fight against phages, our

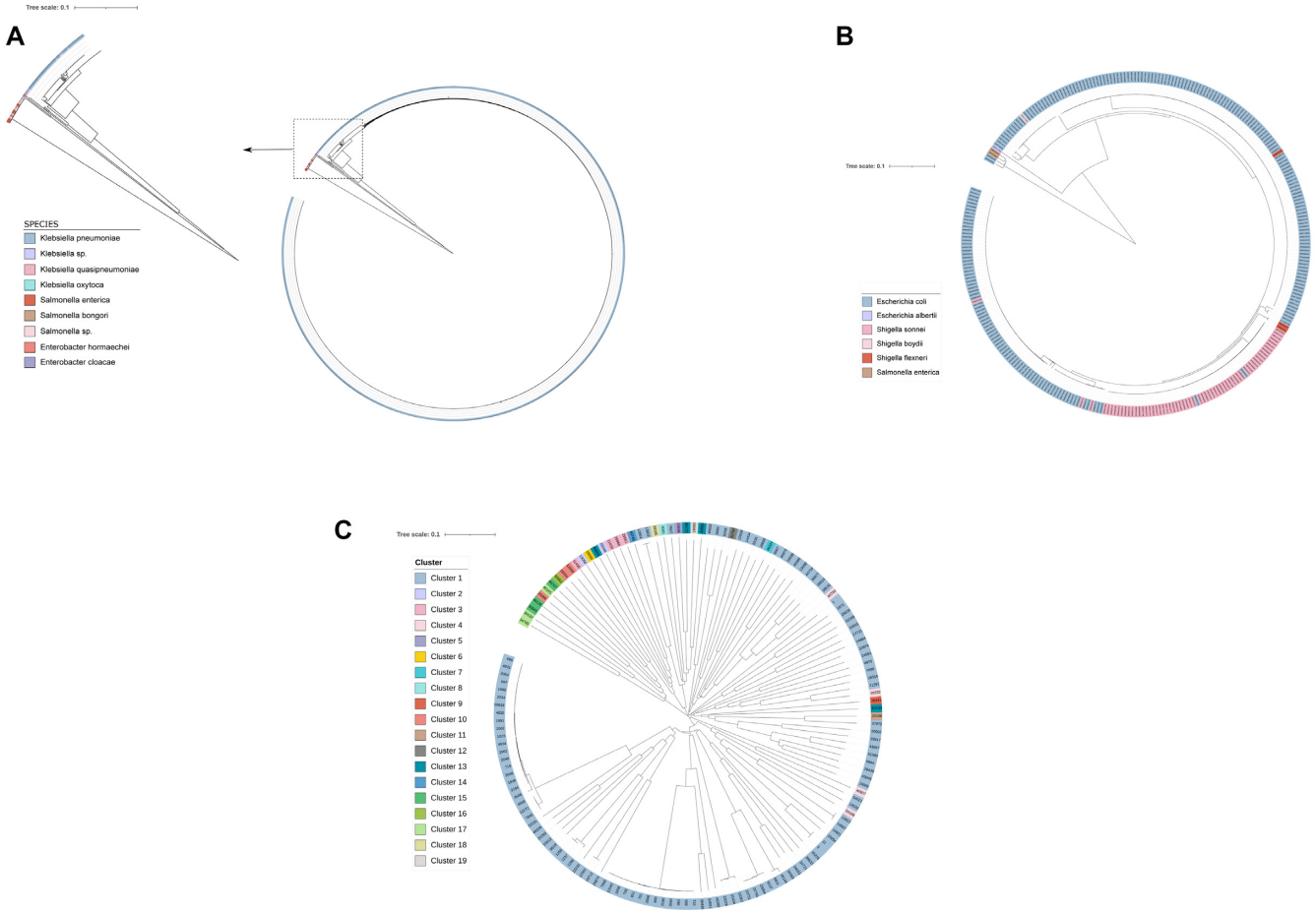

Figure 4. Phylogenetic relationships of the host species shown in Figure 2. (**A**) Phylogenetic relationships of the host species that have identical temperate phages with *Klebsiella pneumonia*. (**B**) Phylogenetic relationships of the host species that have identical temperate phages with *Escherichia coli*. (**C**) Phylogenetic relationships of the phage entries that constitute the phage clusters in Figure 3. The numbers at the tips of the branches represent the phage entries. Our study used phage entry as a short form for nonredundant complete temperate phage genome sequence.

study indicates that temperate phages can protect the hosts with RM systems against other foreign DNA invasions. It is also a self-protection mechanism for the temperate phages to restrict other competing phages.

Examination of CRISPRs and anti-CRISPRs encoded by temperate phages

By searching the CRISPR-Cas systems, our study found 466 (0.2%, 466/192 326) temperate phages in 17 host species. These temperate phages encoded both CRISPR array and Cas proteins, constituting ten subtypes within three main types of CRISPR-Cas systems. The types of CRISPR-Cas systems include types I (I-A∼I-C, I-E∼I-F), II (II-A, II-U) and III (III-A∼III-B, III-U) (Supplementary Materials, Examination of CRISPRs and anti-CRISPRs encoded by temperate phages). The II-U was the most commonly occurring type found in the temperate phages, which infected five host species (Supplementary Figure S10D1, Extended Data Table 17).

By searching the anti-CRISPRdb (5), 1,108 temperate phages in seven host species encode anti-CRISPR proteins inhibiting four important CRISPR systems, including I-E, I-F, II-A and II-C (Supplementary Materials, Examination of CRISPRs and anti-CRISPRs encoded by temperate phages). The type anti-I-F is the most commonly occurring type found in the temperate phages that infect four host species (Supplementary Figure S10D2, Extended Data Table 18). Also, the temperate phages in *P. aeruginosa* encode anti-CRISPR proteins inhibiting I-E and I-F CRISPR systems, and those in *N. meningitidis* and *Neisseria lactamica* encode anti-CRISPR proteins inhibiting type II-C CRISPR system. Our findings are consistent with the previously published findings (11–13). Moreover, it illustrates that not only virulent phages encode the anti-CRISPR proteins but also temperate phages.

DISCUSSION

Temperate phages are an essential portion of the microbial community and play critical roles in microbial ecology, bacterial evolution, bacterial genome diversity, and bacterial pathogenicity. However, due to the lysogenic properties of the temperate phages, it is difficult to observe the activity of the temperate phages by traditional phage techniques based on plaque assay, which limits the study of temperate phages.

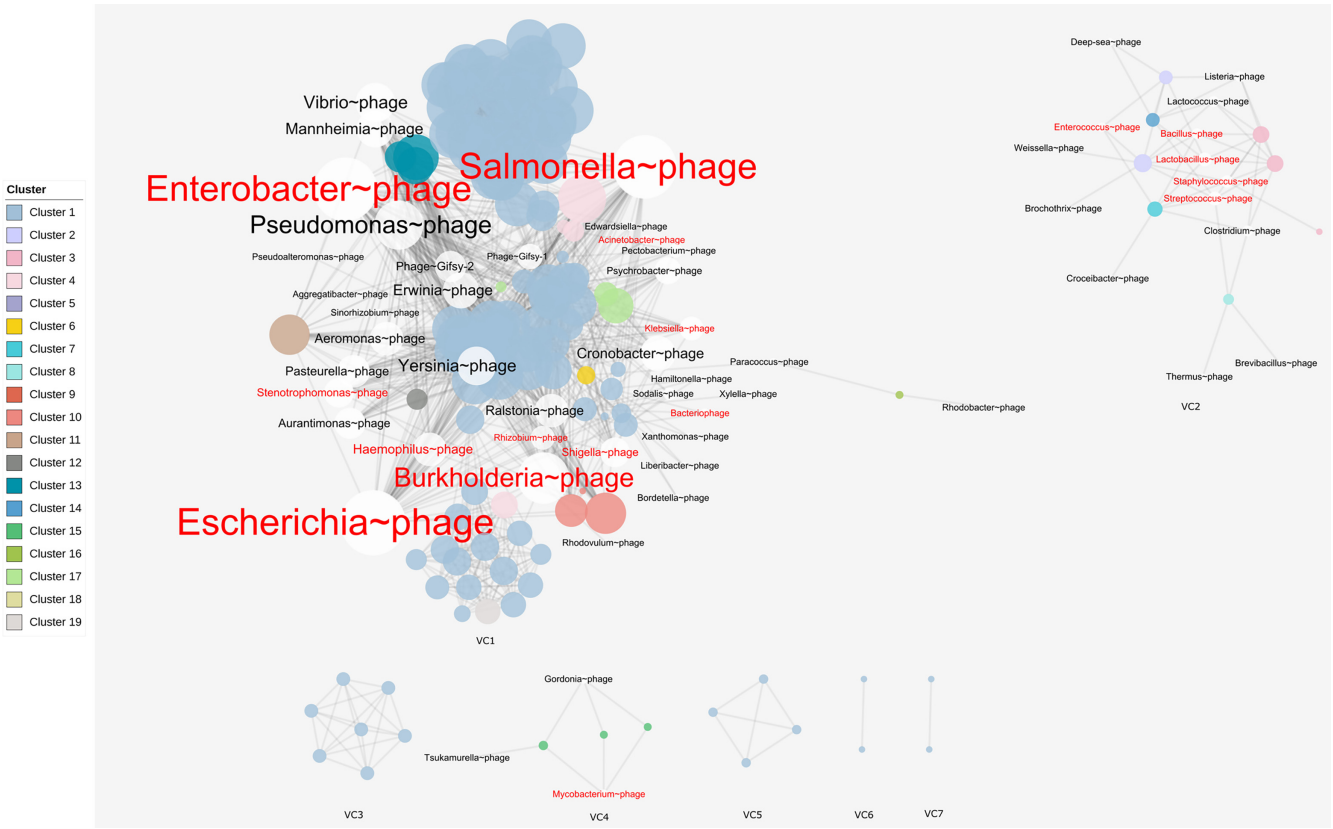

Figure 5. Gene-sharing networks were built using all the 147 phage entries formed host-sharing clusters in Figure 2 and bacterial virus genomes retrieved from Viral RefSeq v.94. VCs were obtained by vConTACT2.

As a result, the number of complete genome sequences of temperate phages in public databases is very small compared with that of virulent phages. The complete temperate phage genome sequences are significant for understanding the genomic characteristics and biological characteristics of temperate phages and their hosts. Also, the complete temperate phage genome sequences can provide vital information for basic and applied research of phages.

This study developed a computational method to detect temperate phages in the raw data of bacterial next-generation genome sequencing. Various bioinformatics tools can predict prophage sequences on bacterial genomes. However, these methods predict rough prophage regions rather than acquire accurate and complete temperate phage genome sequences. Our method TemPhD is based on the biological principle that temperate phages exhibit spontaneous induction, during which the linear phage genome circularizes and replicates.

We then designed a series of experiments to verify TemPhD. Compared with currently available prophage prediction methods, such as Phage_Finder (22), Prophage Finder (23), Prophyinder (24), PHASTER (26), PhiSpy (27), VirSorter (28), Prophage Hunter (29), only TemPhD can accurately determine the exact boundaries of the temperate phage genome and obtain reliable and complete phage genome sequences. Some other tools also predict other prophages different from the wet-lab experiments, such as the prophages predicted by VirSorter in Bac1369Scaffold1 and the prophages predicted by PHASTER, PhiSpy, VirSorter, Prophage Hunter and VIBRANT in Bac2853Scaffold2 (Supplementary Figure S1). Considering these tools do not report the consistent start and end positions of these prophages, it is impossible to tell whether these prophages actually exist. To compare the performance of bioinformatics tools, we conducted wet-lab experiments to induce the temperate phages in bacterial strains and treated them as ground truth. In this study, we only consider the wet-lab induced temperate phages. As long as our method can detect the identical temperate phages as the wet-lab experiments, it is verified as a tool that can detect real temperate phage correctly.

TemPhD can detect a temperate phage when it is in replication process: some are still on bacterial chromosomes, while some are excised from its host. It is noted that most prophages in some bacterial species, such as *Salmonella*, are able to induce without complete excision (66). Therefore, TemPhD reports a temperate phage which satisfies two conditions: (i) the genome is circular to represent its lytic cycle that the temperate phage has excised from its host chromosome; (ii) the genome is also on the host chromosome to distinguish it from other kinds of mobile elements which do not integrate on host chromosomes. The temperate phages which appear to be capable of induction but do not go through the lytic cycle cannot be reported by TemPhD.Generally, TemPhD cannot guarantee to acquire all the real temperate phages, such as the prophages that do

not yet go through induction process but may potentially induce from their host. However, it is theoretically guaranteed that the phage genomes detected by TemPhD are real temperate phages.

Subsequently, we used TemPhD to batch analyze the bacterial genome sequencing raw data from the public database. As it is impossible to do a confidence measure for all the bacterial data on the public database and for strictly speaking, we herein use 'prediction' to describe the temperate phage genomes detected by TemPhD. A total of 192 326 complete temperate phage genome sequences were obtained by analyzing the NCBI NGS data of 789 383 bacteria, which belonged to 2717 species and 710 genera. Our results expand the current number of complete temperate phage genome sequences by >100 times, providing valuable resources for basic and applied research on both bacteria and phages. All of these complete temperate phage genome sequences are available at https://phage.deepomics.org/. Based on the large data set, we also did preliminary research on phage genomic diversity, phage-host interactions, mediated HGTs, the encoded RM and CRISPR/anti-CRISPR systems.

Our work may facilitate phage-related clinical applications:

1) Temperate phage benchmark dataset. Such an expansion of temperate phage genomes includes a large variety of previously unidentified temperate phages. As temperate phages will possibly transfer antibiotic resistance genes or virulence factors among bacteria, we should exclude any temperate phages mixed in the virulent phage candidate for phage therapy to prevent pathogenic risks (4). These temperate phages detected in our study can be used as a benchmark to separate safe and therapy-effective phages from potentially risky temperate phages and thus ensure phage therapy's safety.
2) Sequence references of biological conversion from temperate phages to virulent phages. As a promising alternative for antibiotics, phage therapy is to isolate virulent phages to treat bacterial infection. However, virulent phages for some host bacteria, such as *C. difficile* and *M. tuberculosis*, are difficult to isolate. Temperate phages against such host bacteria, in contrast, can be biologically converted to virulent phages to infect those host species. Our study revealed 1,847 temperate phages in *C. difficile* and 1391 in *M. tuberculosis* (Extended Data Table 5). Their genomes can be further used to transform temperate phages into virulent phages for phage therapy applications.
3) Sequence references of temperate phages infecting uncommon bacteria. By referring to the multiple host connection network provided in this study, researchers can use the temperate phages in these common bacteria to infect the uncommon bacteria for a particular research purpose. The multiple host connection network illustrates the different host species that the same/identical temperate phage can infect. Acquiring the virulent phage for uncommon pathogenic bacterial species is a challenge. Firstly, it is difficult to obtain the uncommon host species, not to mention to culture their virulent phages. Fortunately, our study provides an optional way for obtaining the virulent phages of the uncommon host species. Particularly, from the multiple host connection network, we can find the temperate phage of common hosts, which also infect uncommon/less common hosts. Therefore, we could easily acquire the common host and then induce its temperate phages to co-infect the common and uncommon hosts. For example, the two common species, *E. faecalis* and *Enterococcus faecium*, share identical temperate phages with the less common species Enterococcus durans (Figure 2, Cluster 2). These temperate phages can be further biologically synthesized to become virulent phages that can be used to kill the uncommon host species (4).
4) Warning of potential FMT risks caused by temperate phages. FMT has been successfully applied to treat CDI patients. However, FMT treatment becomes risky if multidrug-resistant or highly virulent bacteria exist in the transplant's microbiota. This situation worsens if temperate phages carry important antibiotic resistance and (or) virulence genes. In this study, we found that 34.9% (912/2613) of human gut metagenome samples contained 912 temperate phage genomes (Extended Data Table 5, Extended Data Table 12). Ten of the 912 temperate phage genomes encode fusidic acid resistance and Macrolide resistance genes; three of them encode virulence factors with the functions of Adherence and invasion, and Serum resistance, immune evasion, and colonization. As all the 912 temperate phage genomes were identified in human gut metagenomes, the above antibiotic resistance genes and virulence genes identified in these temperate phages are also detected in gut metagenome data. The antibiotic resistance (fusidic acid-resistance and macrolide-resistance, Extended Data Table 19) and virulence (adherence and invasion; serum resistance, immune evasion and colonization, Extended Data Table 13) genes in temperate phages may potentially increase the risk of FMT.

Some findings from our study may raise concerns. Explanations to these concerns are as follows:

1) Our study identified three efflux pump genes by VFDB in eighteen temperate phages (1 from *S. enterica*, 1 from *K. pneumonia*, and the other 16 from *Neisseria lactamica*), including VFG049144(gb|YP_002918165.1, acriflavin resistance protein B), VFG036938(gb|NP_273367, fatty acid efflux system protein), and VFG036956(gb|NP_273368, fatty acid efflux system protein). These genes have been proved to relate to bacterial virulence. For example, the acriflavin resistance protein B in *Fusobacterium nucleatum* was virulent (67), while atty acids are functional constituents of several known virulence factors (68).

These virulence factors may not play a role in virulence but are used to maintain the temperate phage's functions for the temperate phage itself. For example, the temperate phage uses endolysins to break open the bacterial cell wall. However, for the host of a temperate phage, the gene may be treated as 'virulence' to the bacterium. The definition of 'virulence' changes according to the research objects. When

referring to the genes identified by VFDB in our study, the researchers can change the 'virulence factor' to any other proper name according to their research purposes.

1) This study identified antibiotic resistance genes (ARG) in 16.6% of total temperate phage genomes. It is reported that virulent phages rarely encode ARG (69). The virulent phages do not integrate into their host genomes but transduce among hosts. Related research also shows that ARG transfer is much less common through phage transduction than through conjugative elements (70). In this study, we research active prophages (temperate phages). In the lysogenic cycles of temperate phages, the active prophages would possibly take ARG (if having any) to integrate into hosts' genomes as 'conjugative elements', not 'transduction' as virulent phages. Considering that conjugative elements systematically transfer ARG among bacteria (69), our results are consistent with the previous findings.
2) Moreover, it is interesting to note that 4604 species do not carry any temperate phages. Such as 43 species in genus *Bacillus*, 15 species in genus *Neisseria*, 83 species in genus *Lactobacillus*, and many uncommon host species (Extended Data Table 20). The no carriage of temperate phages may be because these species are difficult for temperate phages to integrate into or induce from.

CONCLUSION

In conclusion, by developing a novel temperate phage detection method using bacterial NGS data, we essentially filled the temperate phage database with hundreds of thousand complete temperate phage genome sequences. Our work paves the way for a better understanding of interactions between temperate phages and their hosts, and for further explanation of the roles that temperate phages play in bacterial evolvability.

DATA AVAILABILITY

The source code implementing our method TemPhD is available at https://github.com/NancyZxll/temperate-phage-active-prophage-detection. All the raw data for method validation and output of each tool have been made available at https://phage.deepomics.org/files/. The temperate phage genome sequences detected in our study are freely available via https://phage.deepomics.org/. The batched 192 326 temperate phage genome sequences are available at https://phage.deepomics.org/media/data/data.zip. The data generated in our study was also deposited in GenBank with the Bioproject PRJNA848208. According to the NCBI guidelines, the data will be released upon publication of this manuscript. Also, we have provided additional public websites to share the temperate phage genomes: (i) OFS: osf.io/t5a9m/; (ii) on the NCBI ftp site: https://ftp.ncbi.nlm.nih.gov/pub/mgx/phage/data.zip. All additional in-house scripts are available from https://github.com/NancyZxll/temperate-phage-analysis.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We would like to express sincere gratitude to Dr Zhiyi Zhang for providing efficient Python scripts for analyzing the Genbank Taxonomy database and Dr Yarong Wu for providing the Python scripts for bacterial phylogenetic analysis.

Author contributions: X.Z., R.W. designed and performed bioinformatics analyses. X.J. and W.Y. collected all datasets. M.W. curated data for the study. X.X. wrote the essential phage detection program, X.Z., W.L., Q.S., and Q.N. contributed to revising the program. X.Z., S.Z., S.T., H.W., C.S. and F.W. contributed to testing the program and made the tool comparisons. X.F. developed the website. Y.H. conducted the wet-lab experiments. X.Z. wrote the manuscript. Y.T, J.W., A.M.K., S.L., Y.C. and X.J. revised the manuscript. Y.T., S.L. and S.P. conceived and supervised the project. All authors approved the final manuscript.

FUNDING

National Natural Science Foundation of China [31900489, 32001834]; National Key Research and Development Program of China [2018YFA0903000]; Key Research and Development Program of Hebei Province [22322908D]; the work of X.J. was supported by the Intramural Research Program of the National Library of Medicine,National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Golding,I., Coleman,S., Nguyen,T.V. and Yao,T. (2019) In: *Decision Making by Temperate Phages. Reference Module in Life Sciences*. Elsevier, NY.
2. Harrison,E. and Brockhurst,M.A. (2017) Ecological and evolutionary benefits of temperate phage: what does or doesn't kill you makes you stronger. *BioEssays*, **39**, 1700112.
3. Argov,T., Azulay,G., Pasechnek,A., Stadnyuk,O., Ran-Sapir,S., Borovok,I., Sigal,N. and Herskovits,A.A. (2017) Temperate bacteriophages as regulators of host behavior. *Curr. Opin. Microbiol.*, **38**, 81–87.
4. Monteiro,R., Pires,D.P., Costa,A.R. and Azeredo,J. (2019) Phage therapy: going temperate?*Trends Microbiol.*, **27**, 368–378.
5. Dong,C., Hao,G.-F., Hua,H.-L., Liu,S., Labena,A.A., Chai,G., Huang,J., Rao,N. and Guo,F.-B. (2018) Anti-CRISPRdb: a comprehensive online resource for anti-CRISPR proteins. *Nucleic Acids Res.*, **46**, D393–D398.
6. Dempsey,R.M., Carroll,D., Kong,H., Higgins,L., Keane,C.T. and Coleman,D.C. (2005) Sau42I, a bcgi-like restriction–modification system encoded by the staphylococcus aureus quadruple-converting phage π42. *Microbiology*, **151**, 1301–1311.
7. Dedrick,R.M., Jacobs-Sera,D., Bustamante,C.A.G., Garlena,R.A., Mavrich,T.N., Pope,W.H., Reyes,J.C.C., Russell,D.A., Adair,T. and Alvey,R. (2017) Prophage-mediated defence against viral attack and viral counter-defence. *Nat. Microbiol.*, **2**, 16251.
8. Kita,K., Kawakami,H. and Tanaka,H. (2003) Evidence for horizontal transfer of the EcoT38I restriction-modification gene to chromosomal DNA by the P2 phage and diversity of defective P2 prophages in *Escherichia coli* TH38 strains. *J. Bacteriol.*, **185**, 2296–2305.
9. Seed,K.D., Lazinski,D.W., Calderwood,S.B. and Camilli,A. (2013) A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature*, **494**, 489–491.

10. McDonald,N.D., Regmi,A., Morreale,D.P., Borowski,J.D. and Boyd,E.F. (2019) CRISPR-Cas systems are present predominantly on mobile genetic elements in vibrio species. *BMC Genomic.*, **20**, 105.
11. Bondy-Denomy,J., Pawluk,A., Maxwell,K.L. and Davidson,A.R. (2013) Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*, **493**, 429–432.
12. Pawluk,A., Bondy-Denomy,J., Cheung,V.H., Maxwell,K.L. and Davidson,A.R. (2014) A new group of phage anti-CRISPR genes inhibits the type IE CRISPR-Cas system of pseudomonas aeruginosa. *MBio*, **5**, e00896–00814.
13. Harrington,L.B., Doxzen,K.W., Ma,E., Liu,J.-J., Knott,G.J., Edraki,A., Garcia,B., Amrani,N., Chen,J.S. and Cofsky,J.C. (2017) A broad-spectrum inhibitor of CRISPR-Cas9. *Cell*, **170**, 1224–1233.
14. Hargreaves,K.R. and Clokie,M.R. (2014) *Clostridium difficile* phages: still difficult?*Front. Microbiol.*, **5**, 184.
15. Xiong,X., Zhang,H.-M., Wu,T.-T., Xu,L., Gan,Y.-L., Jiang,L.-S., Zhang,L. and Guo,S.-L. (2014) Titer dynamic analysis of D29 within MTB-infected macrophages and effect on immune function of macrophages. *Exp. Lung Res.*, **40**, 86–98.
16. Carrigy,N.B., Larsen,S.E., Reese,V., Pecor,T., Harrison,M., Kuehl,P.J., Hatfull,G.F., Sauvageau,D., Baldwin,S.L. and Finlay,W.H. (2019) Prophylaxis of mycobacterium tuberculosis H37Rv infection in a preclinical mouse model via inhalation of nebulized bacteriophage D29. *Antimicrob.Agents Chemother.*, **63**, 12.
17. Cammarota,G., Ianiro,G., Tilg,H., Rajilić-Stojanović,M., Kump,P., Satokari,R., Sokol,H., Arkkila,P., Pintus,C. and Hart,A. (2017) European consensus conference on faecal microbiota transplantation in clinical practice. *Gut*, **66**, 569–580.
18. Khoruts,A. and Sadowsky,M.J. (2016) Understanding the mechanisms of faecal microbiota transplantation. *Nat. Rev. Gastroenterol. Hepatol.*, **13**, 508.
19. Davies,E.V., James,C.E., Kukavica-Ibrulj,I., Levesque,R.C., Brockhurst,M.A. and Winstanley,C. (2016) Temperate phages enhance pathogen fitness in chronic lung infection. *ISME J.*, **10**, 2553–2555.
20. Davies,E.V., James,C.E., Williams,D., O'Brien,S., Fothergill,J.L., Haldenby,S., Paterson,S., Winstanley,C. and Brockhurst,M.A. (2016) Temperate phages both mediate and drive adaptive evolution in pathogen biofilms. *Proc. Nat. Acad. Sci. U.S.A.*, **113**, 8266–8271.
21. Sekulović,O. and Fortier,L.-C. (2016) *Clostridium difficile*. Springer, pp. 143–165.
22. Fouts,D.E. (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.*, **34**, 5839–5851.
23. Bose,M. and Barber,R.D. (2006) Prophage finder: a prophage loci prediction tool for prokaryotic genome sequences. *Silico Biol.*, **6**, 223–227.
24. Lima-Mendez,G., Van Helden,J., Toussaint,A. and Leplae,R. (2008) Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*, **24**, 863–865.
25. Zhou,Y., Liang,Y., Lynch,K.H., Dennis,J.J. and Wishart,D.S. (2011) PHAST: a fast phage search tool. *Nucleic Acids Res.*, **39**, W347–W352.
26. Arndt,D., Grant,J.R., Marcu,A., Sajed,T., Pon,A., Liang,Y. and Wishart,D.S. (2016) PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.*, **44**, W16–W21.
27. Akhter,S., Aziz,R.K. and Edwards,R.A. (2012) PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Res.*, **40**, e126.
28. Roux,S., Enault,F., Hurwitz,B.L. and Sullivan,M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ*, **3**, e985.
29. Wenchen,S., Hai-Xi,S., Carolyn,Z., Li,C., Ye,P., Ziqing,D., Dan,W., Yun,W., Ming,H. and Wenen,L. (2019) Prophage hunter: an integrative hunting tool for active prophages. *Nucleic Acids Res.*, **47**, W1.
30. Kieft,K., Zhou,Z. and Anantharaman,K. (2020) VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, **8**, 90.
31. Andrews,S. (2010) Babraham Bioinformatics. Babraham Institute, Cambridge, United Kingdom.
32. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
33. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S. and Prjibelski,A.D. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.
34. Nurk,S., Meleshko,D., Korobeynikov,A. and Pevzner,P.A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res.*, **27**, 824–834.
35. Sutton,T.D., Clooney,A.G., Ryan,F.J., Ross,R.P. and Hill,C. (2019) Choice of assembly software has a critical impact on virome characterisation. *Microbiome*, **7**, 12.
36. Wang,Q., Guan,Z., Pei,K., Wang,J., Liu,Z., Yin,P., Peng,D. and Zou,T. (2018) Structural basis of the arbitrium peptide–AimR communication system in the phage lysis–lysogeny decision. *Nat. Microbiol.*, **3**, 1266–1273.
37. Golding,I. (2016) Single-cell studies of phage λ: hidden treasures under Occam's rug. *Annu. Rev. Virol.*, **3**, 453–472.
38. Hyatt,D., Chen,G.-L., LoCascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf.*, **11**, 119.
39. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
40. Delcher,A.L., Bratke,K.A., Powers,E.C. and Salzberg,S.L. (2007) Identifying bacterial genes and endosymbiont DNA with glimmer. *Bioinformatics*, **23**, 673–679.
41. Ester,M., Kriegel,H.-P., Sander,J. and Xu,X. (1996) Hybrid Genetic Algorithm with K-Means for Clustering Problems. *Int. Conf. Knowl. Discov. Data Mining*, **240**, 6.
42. Oliveira,J., Mahony,J., Hanemaaijer,L., Kouwen,T.R., Neve,H., MacSharry,J. and van Sinderen,D. (2017) Detecting lactococcus lactis prophages by mitomycin C-mediated induction coupled to flow cytometry analysis. *Front. Microbiol.*, **8**, 1343.
43. Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) In: *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press.
44. Shen,W. (2021) csvtk - across-platform, efficient and practical CSV/TSV toolkit. https://github.com/shenwei356/csvtk.
45. Wickham,H. (2009) In: *Ggplot2: Elegant Graphics for Data Analysis*. Springer Publishing Company, Incorporated.
46. Zhou,Z., McCann,A., Litrup,E., Murphy,R., Cormican,M., Fanning,S., Brown,D., Guttman,D.S., Brisse,S. and Achtman,M. (2013) Neutral genomic microevolution of a recently emerged pathogen, salmonella enterica serovar agona. *PLoS Genet.*, **9**, e1003471.
47. Kazutaka,K., Kei-ichi,K., Hiroyuki,T. and Takashi,M. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
48. Sudhir,K., Glen,S., Li,M., Christina,K. and Koichiro,T. (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.*, **35**, 1547–1549.
49. Ivica,Letunic and Peer and BorkPeer and Bork. (2019) Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.*, **47**, W256–W259.
50. Jang,H.B., Bolduc,B., Zablocki,O., Kuhn,J.H., Roux,S., Adriaenssens,E.M., Brister,J.R., Kropinski,A.M., Krupovic,M. and Lavigne,R. (2019) Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.*, **37**, 632–639.
51. Zankari,E., Hasman,H., Cosentino,S., Vestergaard,M., Rasmussen,S., Lund,O., Aarestrup,F.M. and Larsen,M.V. (2012) Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.*, **67**, 2640–2644.
52. Liu,B., Zheng,D., Jin,Q., Chen,L. and Yang,J. (2019) VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.*, **47**, D687–D692.
53. Langille,M.G., Hsiao,W.W. and Brinkman,F.S. (2008) Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinf.*, **9**, 329.
54. Bertelli,Claire and Brinkman,F.S.L. (2018) Improved genomic island predictions with IslandPath-DIMOB. *Bioinformatics*, **34**, 2161–2167.
55. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2015) REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res*, **43**, D298–D299.

56. Furuta,Y., Abe,K. and Kobayashi,I. (2010) Genome comparison and context analysis reveals putative mobile forms of restriction–modification systems and related rearrangements. *Nucleic Acids Res.*, **38**, 2428–2443.
57. Murray,N.E. (2000) Type i restriction systems: sophisticated molecular machines (a legacy of bertani and weigle). *Microbiol. Mol. Biol. Rev.*, **64**, 412–434.
58. Jurenaite-Urbanaviciene,S., Kazlauskiene,R., Urbelyte,V., Maneliene,Z., Petrusyte,M., Lubys,A. and Janulaitis,A. (2001) Characterization of bse MII, a new type IV restriction–modification system, which recognizes the pentanucleotide sequence 5′-CTCAG (N) 10/8↓. *Nucleic Acids Res.*, **29**, 895–903.
59. Beletskaya,I.V., Zakharova,M.V., Shlyapnikov,M.G., Semenova,L.M. and Solonin,A.S. (2000) DNA methylation at the cfr BI site is involved in expression control in the cfr BI restriction–modification system. *Nucleic Acids Res.*, **28**, 3817–3822.
60. Protsenko,A., Zakharova,M., Nagornykh,M., Solonin,A. and Severinov,K. (2009) Transcription regulation of restriction-modification system ecl18kI. *Nucleic Acids Res.*, **37**, 5322–5330.
61. Som,S. and Friedman,S. (1997) Characterization of the intergenic region which regulates the MspI restriction-modification system. *J. Bacteriol.*, **179**, 964–967.
62. Xu,Q., Stickel,S., Roberts,R.J., Blaser,M.J. and Morgan,R.D. (2000) Purification of the novel endonuclease, Hpy188I, and cloning of its restriction-modification genes reveal evidence of its horizontal transfer to the helicobacter pylori genome. *J. Biol. Chem.*, **275**, 17086–17093.
63. Couvin,D., Bernheim,A., Toffano-Nioche,C., Touchon,M., Michalik,J., Néron,B., Rocha,E.P., Vergnaud,G., Gautheret,D. and Pourcel,C. (2018) CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for cas proteins. *Nucleic Acids Res.*, **46**, W246–W251.
64. Yuan,R. (1981) Structure and mechanism of multifunctional restriction endonucleases. *Annu. Rev. Biochem.*, **50**, 285–315.
65. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2010) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **38**, D234–D236.
66. Frye,J.G., Porwollik,S., Blackmer,F., Cheng,P. and McClelland,M. (2005) Host gene expression changes and DNA amplification during temperate phage induction. *J. Bacteriol.*, **187**, 1485–1492.
67. Kumar,A., Thotakura,P.L., Tiwary,B.K. and Krishna,R. (2016) Target identification in fusobacterium nucleatum by subtractive genomics approach and enrichment analysis of host-pathogen protein-protein interactions. *BMC Microbiol.*, **16**, 84.
68. Kenny,J.G., Ward,D., Josefsson,E., Jonsson,I.-M., Hinds,J., Rees,H.H., Lindsay,J.A., Tarkowski,A. and Horsburgh,M.J. (2009) The staphylococcus aureus response to unsaturated long chain free fatty acids: survival mechanisms and virulence implications. *PLoS one*, **4**, e4344.
69. Enault,F., Briet,A., Bouteille,L., Roux,S., Sullivan,M.B. and Petit,M.-A. (2017) Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J.*, **11**, 237–247.
70. Volkova,V.V., Lu,Z., Besser,T. and Gröhn,Y.T. (2014) Modeling the infection dynamics of bacteriophages in enteric escherichia coli: estimating the contribution of transduction to antimicrobial gene spread. *Appl. Environ. Microbiol.*, **80**, 4350–4362.