frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 19 November 2025
DOI 10.3389/fmicb.2025.1686402

OPEN ACCESS

EDITED BY
Gary Antonio Toranzos,
University of Puerto Rico, Puerto Rico

REVIEWED BY
Clara Rolland,
German Collection of Microorganisms and Cell Cultures GmbH (DSMZ), Germany
Kanchan Bhardwaj,
Manav Rachna International Institute of Research and Studies (MRIIRS), India
Etan Dieppa-Colón,
University of Wisconsin-Madison, United States

*CORRESPONDENCE
Chao Wei
✉ chao2v@163.com
Zhe Chen
✉ 1158059974@qq.com

RECEIVED 15 August 2025
REVISED 30 October 2025
ACCEPTED 04 November 2025
PUBLISHED 19 November 2025

CITATION
Wei C and Chen Z (2025) Comprehensive analysis of phage genomes from diverse environments reveals their diversity, potential applications, and interactions with hosts and other phages.
Front. Microbiol. 16:1686402.
doi: 10.3389/fmicb.2025.1686402

Comprehensive analysis of phage genomes from diverse environments reveals their diversity, potential applications, and interactions with hosts and other phages

Chao Wei* and Zhe Chen*

National Key Laboratory of Pig Genetic Improvement and Germplasm Innovation, Jiangxi Agricultural University, Nanchang, China

Phages are ubiquitous and diverse, playing a key role in maintaining microbial ecosystem balance. However, their diversity, potential applications, and their interactions with hosts and other phages remain largely unexplored. To address this, we collected 59,652,008 putative viral genomes from our laboratory, 45 public viral datasets, and an integrated public viral genome database (IGN), covering seven habitats. We obtained 741,692 phage genomes with completeness ≥50% (PGD50), and most (93.83%, 695,938/741,692) of these phage genomes were classified into the Caudoviricetes class. We found that 158,522 species-level viral clusters that contained 28.96% (214,814/741,692) phage genomes without any known phage genomes in the IGN, indicating substantial novelty. Global phylogenetic trees for five iterations based on complete phage genomes significantly expanded the known diversity of the virosphere. Genome analysis revealed phage potential divergence with habitat types and highlighted the utilization of alternative genetic codes. Furthermore, 3D structural similarity searches demonstrated significant potential for annotating previously uncharacterized viral proteins. Analysis of CRISPR spacer inferred potential hosts of phages and competitive networks among phages, highlighting virulent phages as promising candidates for phage therapy against pathogenic bacteria. Intriguingly, diverse CRISPR-Cas systems were detected within phage genomes themselves, suggesting their enormous potential as novel gene editing tools. Collectively, this study provides a comprehensive phage genome resource, foundational for future research into phage–host and phage–phage interactions, phage therapy development, and the mining of next-generation genetic tools.

KEYWORDS

phage–phage interactions, phage diversity, potential applications, diverse CRISPR-Cas systems, diverse environments

1 Introduction

Phages, ubiquitous, highly diverse viral components, are key regulators of microbial ecosystem balance, primarily through infection and lysis of bacteria and archaea (Clokie et al., 2011). They shape microbial community dynamics, metabolism, and diversity via established interactions (e.g., "kill-the-winner," "piggyback-the-winner," and evolutionary arms races) (Brown et al., 2022; Yan and Yu, 2024). Specifically, phages maintain diversity by lysing

dominant strains, enhance host adaptability through horizontal gene transfer, and drive microbial diversification via adaptive co-evolution (Dion et al., 2020; Mangalea and Duerkop, 2020). Their therapeutic promise is exemplified in combating multidrug-resistant pathogens through phage therapy (Federici et al., 2022). Furthermore, phages engage in complex co-evolutionary dynamics with their hosts and environments. For instance, under heavy metal stress like chromium contamination in soil, phage–host interactions can shift from a predatory relationship to a potentially mutualistic one, with an increase in lysogeny and phage-mediated horizontal gene transfer potentially aiding host adaptation (Touchon et al., 2017). Similarly, in freshwater lake systems subjected to multiple environmental stressors, the complexity and stability of virus-bacteria interaction networks can be significantly reduced, altering the composition of viral auxiliary metabolic genes and consequently impacting ecosystem functions like carbon cycling (Wang T. et al., 2025). These findings underscore the critical role of environmental factors in shaping phage–host interaction networks. Although a number of phage genome databases have been established, the data remain largely fragmented and exhibit significant habitat-specific biases (Resch et al., 2024; Wang et al., 2024). However, a significant research gap persists because two key resources are lacking: a unified, high-quality genome resource for phages from diverse habitats, and a comprehensive understanding of the global-scale architecture of phage–host interaction networks. This gap fundamentally limits systematic ecological and evolutionary insights (Bignaud et al., 2025; Wang B. et al., 2025).

To counter phage predation, prokaryotes have evolved the CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats) system, an adaptive immune mechanism that provides sequence-specific defense against invading nucleic acids (DNA and RNA) through the stages of adaptation, expression, and interference (Burstein et al., 2017; Makarova et al., 2020). Diverse CRISPR-Cas systems have been identified within metagenome-assembled genomes (MAGs) across specific prokaryotic phyla (Shmakov et al., 2015; Yan et al., 2018), and this suggests a rich landscape of interacting phage genomes. Phages themselves have been found to harbor CRISPR-Cas systems, inspiring novel gene-editing tools, a comprehensive overview of CRISPR-Cas systems across entire prokaryotic host and phage populations is still lacking (Pausch et al., 2020; Al-Shayeb et al., 2022). This gap hinders our understanding of the tripartite interactions among phages, other phages, and host bacteria or archaea, and their collective role in maintaining microbial community homeostasis.

Herein, to bridge these knowledge gaps, we present the construction and comprehensive characterization of the PGD50 database, a curated collection of high-quality phage genomes integrated from diverse habitats. The primary objective of this study is to employ this unified resource to systematically evaluate global phage diversity, evolutionary patterns, and ecological interactions, with an emphasis on uncovering novel phages and elucidating their functional traits. To address these objectives, we designed a series of targeted analytical approaches: (1) Taxonomic classification and phylogenetic analysis were applied to delineate evolutionary relationships and quantify phylogenetic novelty. (2) CRISPR spacer matching was leveraged to infer phage–host interaction networks and uncover competitive dynamics among phages. (3) Structure-based functional annotation enabled the prediction of protein functions beyond sequence homology, expanding the functional landscape of phage genomes. (4) Comparative genomics of CRISPR-Cas systems identified their diversity and potential activity within phage genomes. Together, these integrated approaches provide a multidimensional perspective on phage ecology and evolution, while also facilitating the identification of phage-encoded systems with potential biotechnological utility.

2 Methods

2.1 Construction of a phage genome database (PGD50)

To obtain phage genomes from diverse environments, we collected and combined putative viral contigs from our laboratory, 45 public viral datasets across 7 habitats, and an Integrated Genomic Database [IGN: IMG/VR (Camargo et al., 2023a), GenBank (Benson et al., 2013), NT (Sayers et al., 2022); Supplementary Tables S1, S2], yielding 59,652,008 contigs for analysis. Among them, we performed a dereplication step on all viral genomes included in the IGN database using MMseqs2 with the "easy-linclust -c 1.0 --min-seq-id 1.0" options, clustering them at 100% sequence identity to ensure non-redundancy. Viral contig identification utilized a custom pipeline developed by Nayfach et al. (2021b) based on four signatures: presence of viral protein families, absence of microbial protein families, viral nucleotide signatures, and multiple adjacent genes on the same strand. Briefly, to identify the presence of viral protein families, we constructed hidden Markov models (HMMs) for 23,841 viral protein families from the IMG/VR database, after excluding 1,440 families that are commonly found in microbial genomes or plasmids. Conversely, to confirm the absence of prevalent microbial protein families, HMM profiles were constructed for 16,260 families from the Pfam-A database, following the removal of 452 families that are also common in viruses. All protein sequences were searched against these HMMs using hmmsearch [HMMER v3.3.2 (Potter et al., 2018); parameters: -Z 1, *E*-value $<1 \times 10^{-10}$], with the database of the top hit determining the classification. Concurrently, viral nucleotide signatures were identified using VirFinder v.1.1 (Ren et al., 2017), which employs k-mer frequencies and machine learning. Genomic organization was assessed by calculating the strand switch rate (number of strand switches divided by gene count) for contigs with multiple adjacent genes. Finally, 9,607,235 viral contigs with genome size ≥3 kb were obtained for subsequent analysis.

Phage identification employed two complementary methods (Devoto et al., 2019; Al-Shayeb et al., 2020; Shang et al., 2022). First, protein sequences derived from contigs were annotated against Pfam-A (Mistry et al., 2021), TIGRFAM (Haft et al., 2003), and VOGDB[1] databases using HMMER v3.3.2 (Potter et al., 2018) with the "hmmsearch -E 1e-5" parameter. Genomes required two or more genes containing virus-specific keywords ("capsid, phage, terminase, base plate, baseplate, prohead, virion, virus, viral, tape measure, tapemeasure neck, tail, head, bacteriophage, prophage, portal, DNA packaging, T4, p22, and holin"), exclusion of prokaryote-specific terms ("ribosomal protein, preprotein translocase, and DNA gyrase subunit A"), and at least one spacer match from bacterial or archaeal genomes. Second, we used

1 http://vogdb.org

PhaMer v1.0 (Shang et al., 2022; Hu et al., 2024) with default parameters, which applies a Transformer model for metagenomic phage prediction.

Removing false positives involved assessing bacterial universal single-copy orthologs (BUSCOs) ratios (Simao et al., 2015) and curated viral protein family modules (VPFs) ratios (Gregory et al., 2020). Genomes were retained only if they exhibited a BUSCO ratio <0.067, or a BUSCO ratio >0.067 with at least three VPF hits. Subsequent processing detected provirus boundaries, removed host bacterial sequence contamination, and evaluated genome completeness using CheckV v0.8.1 (Nayfach et al., 2021a). The final PGD50 database comprised 741,692 phage genomes with ≥50% completeness.

2.2 Lifestyle prediction and taxonomy assignment of phage genomes (PGD50)

We predicted phage lifestyles using BACPHLIP v0.9.3 (Hockenberry and Wilke, 2021), which classifies genomes as virulent (score <0.5), uncertain (score 0.5–0.9), or temperate (score >0.9). Since temperate phages exhibit both lytic and lysogenic states, we integrated prophages identified by CheckV with BACPHLIP-predicted temperate phages to define the final temperate category. Taxonomic assignment was performed using geNomad v1.7.4 (Camargo et al., 2023b), which leverages viral taxon markers covering most ICTV-recognized lineages.

2.3 Clustering phage genomes to species-level viral clusters and identification of potential novel phage genome clusters

We clustered 741,692 PGD50 genomes into species-level viral clusters using a greedy centroid-based algorithm (Roux et al., 2019; Nayfach et al., 2021b; Tomofuji et al., 2022; Zeng et al., 2024; Wei et al., 2025) with threshold criteria of 95% average nucleotide identity (ANI) and ≥85% genome coverage, as recommended by Roux et al. (2019). Clusters lacking any phage genomes from the IGN database were subsequently classified as novel phage genome clusters. Furthermore, we obtained all phage genomes from the PhageScope (Wang et al., 2024) database (total 873,718 genomes). To ensure a fair comparison with our PGD50 dataset (completeness ≥50%), we first processed the PhageScope genomes through CheckV, retaining only those with ≥50% completeness (446,062 genomes). These were then dereplicated at 100% average nucleotide identity using MMseqs2 (--min-seq-id 1.0 -c 1.0), resulting in a high-quality, non-redundant PhageScope reference set of 334,616 genomes. A comparative analysis at the species-level viral cluster was performed based on the PGD50 and PhageScope reference sets.

2.4 Performing global phylogenetic analysis for phage genomes based on five iterations

To evaluate the phylogenetic novelty and contribution of our obtained phage genomes within the global context of phage diversity, we conducted a large-scale phylogenetic analysis. This approach allowed us to quantify the phylogenetic distance (PD) between our genomes and established reference sequences, thereby assessing the expansion of the known evolutionary landscape.

Specifically, we combined 44,311 complete phage genomes from PGD50 with 5,658 reference complete phage genomes from the VMR database.[2] The combined dataset was processed through a five-iteration phylogenetic workflow adapted (Low et al., 2019). First, duplicate genomes were removed using MMseqs2 v2.0 ("--min-seq-id 1.0 -c 1.0") (Steinegger and Soding, 2017). Protein coding sequences were then predicted using Prodigal v2.50 (Hyatt et al., 2010). The resulting protein sequences were clustered with MMseqs2 ("--min-seq-id 0.3 -c 0.7"), yielding 353,315 protein clusters. Clusters containing ≥3 proteins were used to build HMM profiles with MUSCLE v3.8.1551[3] and HMMER. These were supplemented with 77 existing Caudoviricetes HMM profiles from single-copy protein markers, generating a total of 219,604 phage-associated HMM profiles.

In each iteration, core HMM profiles were identified by scanning progressively refined genome subsets against all profiles using HMMER (*E*-value $\leq 1 \times 10^{-3}$; coverage ≥50%). A profile was considered core if it was present in ≥10% of genomes, had an average copy number ≤1.2, and an average protein length >100 residues. For phylogenetic tree construction, gene markers in retained genomes were identified via HMMER searches (*E*-value $\leq 1 \times 10^{-3}$) against the core HMM profiles (Nayfach et al., 2021b). Multiple sequence alignments of these markers were trimmed with trimAl v1.4.rev22 (Capella-Gutierrez et al., 2009), retaining fragments with <50% gaps. Genomes needed to possess ≥3 markers present in >5% of alignment columns to be included. The final phylogeny was reconstructed using IQ-TREE2 v2.1.3 (Nguyen et al., 2015) under the LG + F + G4 model with 1,000 ultrafast bootstraps, and visualized in iTOL.[4] Finally, phylogenetic distances between genomes were computed from the resulting tree branch lengths using the ape v5.7-1 (Paradis and Schliep, 2019) and picante v1.8.2 (Kembel et al., 2010) packages in R, enabling quantitative assessment of the novel diversity introduced by our dataset.

2.5 Potential divergence analysis of complete phage genomes with habitat types

To minimize confounding effects from genomic fragmentation and unannotated habitats, we analyzed 26,439 complete phage genomes with verified habitat origins. These genomes were clustered into genus-level viral clusters based on average amino acid identity (AAI) and gene sharing (Nayfach et al., 2021b; Tomofuji et al., 2022; Zeng et al., 2024; Wei et al., 2025). Protein sequences were first predicted using Prodigal, followed by all-vs-all BLASTP alignments in DIAMOND v2.1.9.163 (Buchfink et al., 2015). For each phage pair, we calculated AAI percentages and shared gene proportions. Genome pairs exhibiting <50% AAI or <20% shared genes were clustered using MCL v14-137 (van Dongen, 2008) with an inflation factor of 2.0.

2 https://ictv.global/vmr
3 http://www.drive5.com/muscle/
4 https://itol.embl.de/

For clusters containing ≥4 genomes distributed across ≥2 habitats, we constructed phylogenetic trees (Wu et al., 2024). Core genes were identified using Roary v1.7.8 (Page et al., 2015) (-i 50 option) and aligned to create multi-FASTA files. Phylogenies were reconstructed with FastTree v2.1.10 (Price et al., 2010) and visualized in iTOL. Branch lengths between all genome pairs were systematically measured within each cluster. We performed two one-tailed Wilcoxon rank sum tests to compare branch lengths: (1) between genomes from identical habitats versus (2) between genomes from different habitats. Clusters where cross-habitat branch lengths significantly exceeded same-habitat distances ($p < 0.05$) were designated as exhibiting potential habitat-specific divergence.

2.6 Identifying alternative genetic codes in phage genomes (PGD50)

We employed a custom Prodigal v2.50 to identify open reading frames in all 741,692 PGD50 genomes using four genetic coding schemes: the standard genetic code (11) and three alternative codes—TAG recoding (15), TAA recoding (90), and TGA recoding (91) (Ivanova et al., 2014; Nayfach et al., 2021b; Lou et al., 2024). Briefly, for a phage with a genome size <100 kb, if its protein-coding density with the genetic codes 15, 90, or 91 increased >10% compared to that with the standard genetic code 11, we considered that this phage genome tended to use the corresponding alternative genetic code. For those phages with a genome size ≥100 kb, the threshold for considering the utilization of an alternative genetic code was the increase of protein-coding density >5%.

2.7 Functional annotation of phage genomes (PGD50)

We predicted proteins from all 741,692 PGD50 genomes using their corresponding alternative genetic codes and clustered them into 4,372,210 protein clusters via MMseqs2 ("--min-seq-id 3.0 -c 0.7"). To account for the mixed genetic repertoire of phages, which often includes genes of bacterial origin acquired via horizontal gene transfer, we utilized a multi-database approach including Pfam-A (Mistry et al., 2021), TIGRFAM (Haft et al., 2003), and VOGDB (see text footnote 1) for functional annotation to ensure broad coverage of both viral and bacterial protein domains. Representative sequences from each cluster were then functionally annotated against the three databases using HMMER (hmmsearch) (Potter et al., 2018) with an *E*-value threshold of 1×10^{-5}.

To address unannotated proteins, we employed an approach of structural similarity searches developed by Nomburg et al. (2024) leveraging conserved structural domains from horizontal gene transfer events between viruses and cells. Given the substantial computational demands of structure prediction, our structural analysis was limited to representative sequences from the top 100 largest no-hit clusters. These structures were generated using ColabFold (Tunyasuvunakool et al., 2021), which leverages the AlphaFold2 algorithm. To infer functional insights, we performed structural alignments of our predicted models against the AlphaFold database using Foldseek (v1.0) (van Kempen et al., 2024). A TM-score threshold of ≥ 0.4 was employed to filter the alignments, retaining only those pairs with a statistically significant topological similarity for functional inference.

2.8 Host prediction for phage genomes (PGD50) and identification of phage–phage interactions

We predicted hosts for 741,692 PGD50 genomes through CRISPR spacer matching. CRISPR spacers were identified from microbial genomes and MAGs in GTDB (Genome Taxonomy Database) (Chaumeil et al., 2022), UHGG (Unified Human Gastrointestinal Genome) (Almeida et al., 2021), and pig gut (Chen et al., 2021) databases using MinCED v0.4.2[5] with default parameters. Taxonomic classification of MAGs employed GTDB-tk v2.0.0 (classify_wf mode) (Chaumeil et al., 2022). Spacer-phage mapping used BLAST v2.12.0 + (BLASTn, -max_target_seqs 10,000,000 -dust no -word_size 8 -evalue 10) (Camacho et al., 2009), with matches requiring ≤1 mismatch and 100% alignment. Successful mappings indicated host-phage relationships. For pathogenic targeting analysis, we downloaded complete *Escherichia coli* and *Klebsiella pneumoniae* genomes from GenBank,[6] removed duplicates using dRep v3.2.2 (-pa 0.9 -sa 1) (Olm et al., 2017), and annotated virulence factors (VFDB, http://www.mgc.ac.cn/VFs/), antibiotic resistance genes (CARD, https://card.mcmaster.ca/), and pathogenic bacterial proteins (PHI database, http://www.phi-base.org/).

We identified CRISPR spacers within PGD50 genomes using MinCED with default parameters and performed reciprocal BLASTn searches against all phage CRISPR spacers. Interactions were confirmed when spacers mapped to other phage genomes with ≤1 mismatch and 100% alignment. CRISPR-Cas systems in both phages and hosts were predicted using CRISPRCasFinder v 4.3.2 (Couvin et al., 2018) with default parameters. Furthermore, Cas12 proteins were obtained and the Cas12 phylogeny was reconstructed using IQ-TREE2 v2.1.3 under the LG + F + G4 model with 1,000 ultrafast bootstraps, and visualized in iTOL (see text footnote 4).

2.9 Statistical analysis

All statistical analyses were performed using R packages (v4.2.1).

3 Results

3.1 Characterization of phage genomes with completeness ≥50% (PGD50) from diverse environments

To expand phage genome recovery across habitats, we collected putative viral genomes from our laboratory, 45 public viral datasets, and an integrated public viral genome database (IGN). Using a custom pipeline, we identified 5,893,090 phage contigs from 59,652,008 total

5 https://github.com/ctSkennerton/minced

6 https://www.ncbi.nlm.nih.gov/genbank/

contigs based on: (1) using a custom viral pipeline, (2) removal of contigs with a genome size <3 kb, (3) retention of contigs encoding ≥2 virus-specific hallmark genes, (4) retention of contigs with ≥1 CRISPR spacer match, (5) using the PhaMer tool, and (6) confirmation using BUSCO and VPFs. Following validation using multiple methods, we retained 741,692 high-confidence genomes with completeness ≥50% (termed PGD50; Figure 1a).

We estimated the source distribution of phage genomes from PGD50, and found that 230,600 and 88,706 phage genomes were recovered from the human gut and pig gut, respectively (Figure 1b). Completeness assessment of phage genomes in the PGD50 identified 44,311 complete genomes, which represented a valuable resource for the known virosphere diversity (Figure 1c). We further focused on the lifestyle of phages in the PGD50 and found 55.35% (410,503/741,692) phage genomes were predicted as virulent phages, highlighting therapeutic potential against pathogenic infections (Figure 1d). Taxonomic analysis of phages in the PGD50 using geNomad assigned 93.83% (695,938/741,692) to the Caudoviricetes class, yet only 7.94% (58,902/741,692) achieved family-level resolution, demonstrating both substantial novelty and persistent classification challenges (Figure 1e and Supplementary Table S3).

3.2 Assessing novelty of PGD50 and global phylogenetic analysis of complete phage genomes

To evaluate the novelty of phage genomes in the PGD50, we clustered 741,692 phage genomes into 420,230 species-level viral clusters at the threshold of 95% average nucleotide identity (ANI) and 85% coverage. We found 69,198, 45,937, and 23,611 species-level viral clusters were specifically identified in the human gut, pig gut, and rumen, respectively. Analysis of the species-level viral clusters confirmed the substantial novelty of our dataset. Specifically, 37.72% of the clusters themselves were novel, as they contained no sequences from the IGN database including IMG/VR, GenBank, and NT. These novel clusters comprised 28.96% of all the phage genomes analyzed (Figure 2a). Furthermore, a comparative analysis at the species-level viral cluster revealed the distinct contribution of our resource: 331,784 (44.73%) of the 741,692 genomes in our PGD50 dataset are not present in the PhageScope database. In contrast, only 76,410 (22.84%) of the 334,616 PhageScope genomes with completeness ≥50% are absent from our dataset. It demonstrated that our study has contributed a massive number of novel phage genomes that were absent from a leading, recently published database.

To resolve global evolutionary relationships of phages, we constructed phylogenetic trees for five iterations using 44,311 complete phages from this study and 5,658 reference genomes from the Virus Metadata Resource (VMR) from the International Committee on Taxonomy of Viruses (VMR_MSL40.v1). Briefly, we first obtained 353,315 protein clusters and 219,604 HMM profiles based on protein clustering and HMM profile generating using MMseqs2 and HMMER. Notably, we filtered and generated core HMM profiles for these complete phage genomes and performed five iterations to construct global phylogenetic trees (Figure 2b). Briefly, this iterative process was essential due to the vast diversity of our dataset. In each iteration, phage genomes not placed in the phage phylogenetic tree were identified, their specific marker genes were inferred, and these new markers were added to a composite set. This strategy progressively captured a broader spectrum of phage diversity, enabling a more inclusive and robust global phylogeny than would be possible with a single, static marker set. Interestingly, core HMM profiles of five iterations showed low inter-iteration similarity (Figure 2c), confirming representation of distinct phage diversity subsets. Phylogenetic distance (PD) metrics from all iterations (Figure 2d) collectively demonstrate significant expansion of known virosphere diversity.

3.3 Potential divergence analysis with habitat types using complete phage genomes

To minimize impacts of the genomic fragmentation and unknown habitat, we analyzed 26,439 complete phage genomes from seven known habitats. These were clustered into genus-level viral clusters at the threshold of <50% average amino acid identity (AAI) or <20% of shared genes and an inflation factor of 2.0, yielding 2,517 genus-level viral clusters. Our analysis revealed a substantial number of habitat-specific genus-level viral clusters, with 687 uniquely identified in the pig gut, 525 in the human gut, 458 in the rumen, 138 in soil and lake sediments, 60 in the ocean, and 38 in the non-human primate gut (Figure 3a). These clusters contained no complete phage genomes from any other habitat, highlighting the distinct viral populations endemic to each environment.

To investigate potential divergence with habitat types, we analyzed 327 genus-level clusters containing ≥4 genomes distributed across ≥2 habitats (Supplementary Table S4). For each genus-level viral cluster, phylogenetic trees were constructed to test whether genomes from the same habitat exhibited closer evolutionary distances than cross-habitat counterparts. We observed that 62.69% (205/327) of clusters showed significantly closer phylogenetic distances among same-habitat genomes ($p < 0.05$), supporting potential habitat-phage divergence (Figure 3b). As the examples, genus-level viral clusters 1 and 2 demonstrated clear habitat-based phylogenetic clustering. No divergent pattern was detected in genus-level viral clusters 3–5, indicating taxon-specific variation in evolutionary dynamics (Figure 3c).

3.4 Functional potentials of phage genomes in the PGD50

We investigated whether phages utilize alternative genetic codes to maintain low coding density and prevent protein fragmentation. Using custom Prodigal (v2.50), we evaluated four genetic codes (11, 15, 90, 91) based on total potential coding scores. The standard genetic code (11) dominated (97.97%, 726,601/741,692), while a small proportion (2.03%, 15,091/741,692) recoded stop codons as glutamine (Q, genetic codes 15) and Glycine (G, genetic codes 90). Notably, no genomes recoded TAA as glutamine (Q, genetic code 91) (Figure 4a). To identify phages employing alternative genetic codes, we applied a specific threshold during prediction. The use of the correct, corresponding genetic code for these identified phages then led to a significant improvement in functional annotation, as evidenced by a higher match rate against the Pfam-A database.

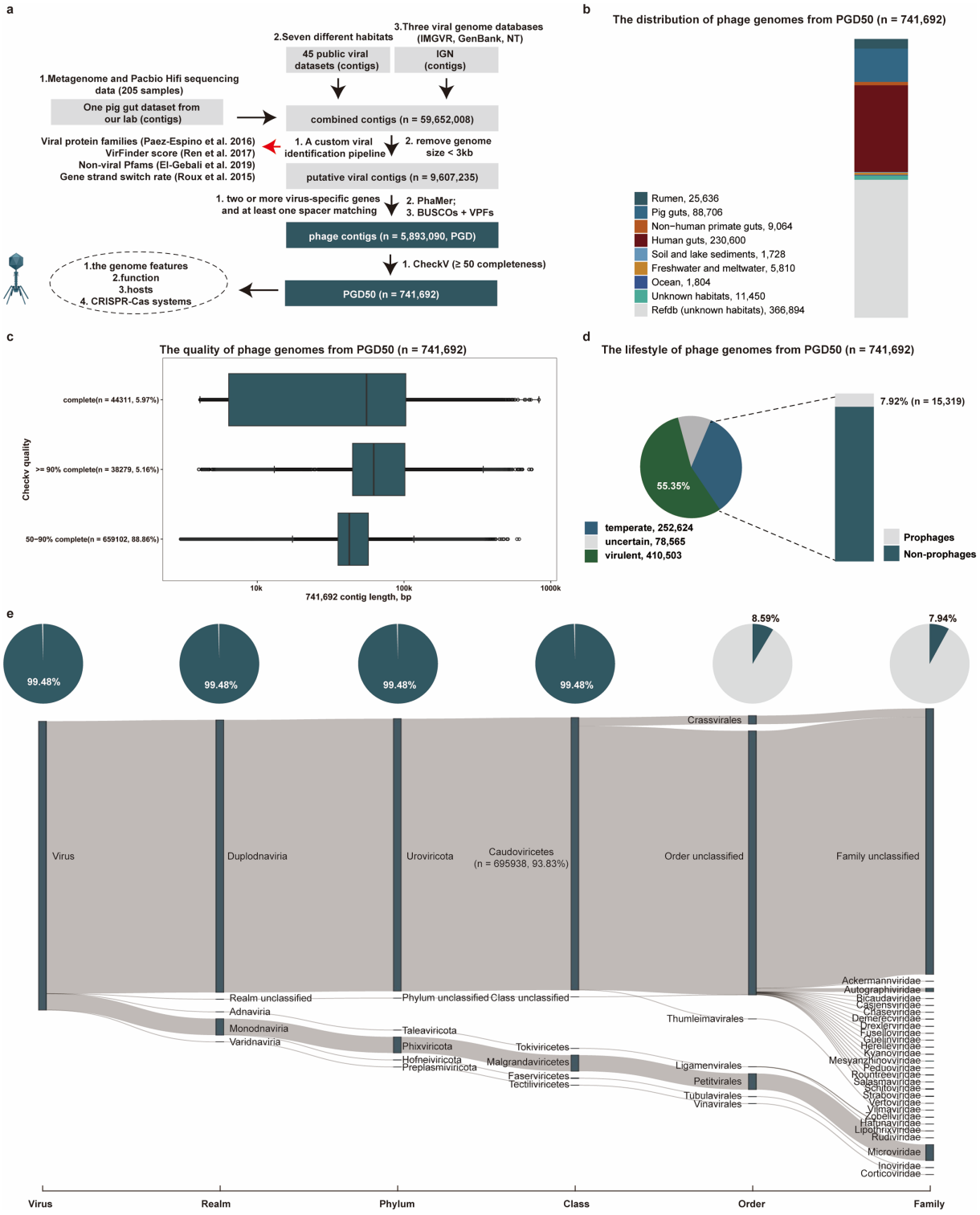

FIGURE 1
Identification and characterization of phage genomes from diverse environments. **(a)** The pipeline for identification of phage genomes (PGD50) from diverse environments. **(b)** The distribution of phage genomes (PGD50) in different habitats. Different colors represent phage genomes from different habitats. **(c)** The detailed distribution of genome length and quality for phage genomes (PGD50). **(d)** The lifestyle prediction of phage genomes (PGD50) and the proportion of prophages in temperate phages. The pie chart (left) represents different lifestyles of phage genomes with different colors, and the bar chart (right) shows the proportion of prophages in temperate phages. **(e)** The detailed taxonomy of phage genomes and proportion of known taxonomy for phages genomes at each taxonomic level. The Sankey diagram represents the detailed taxonomy of phages genomes and these pie charts show proportion of known taxonomy for phages genomes at each taxonomic level.

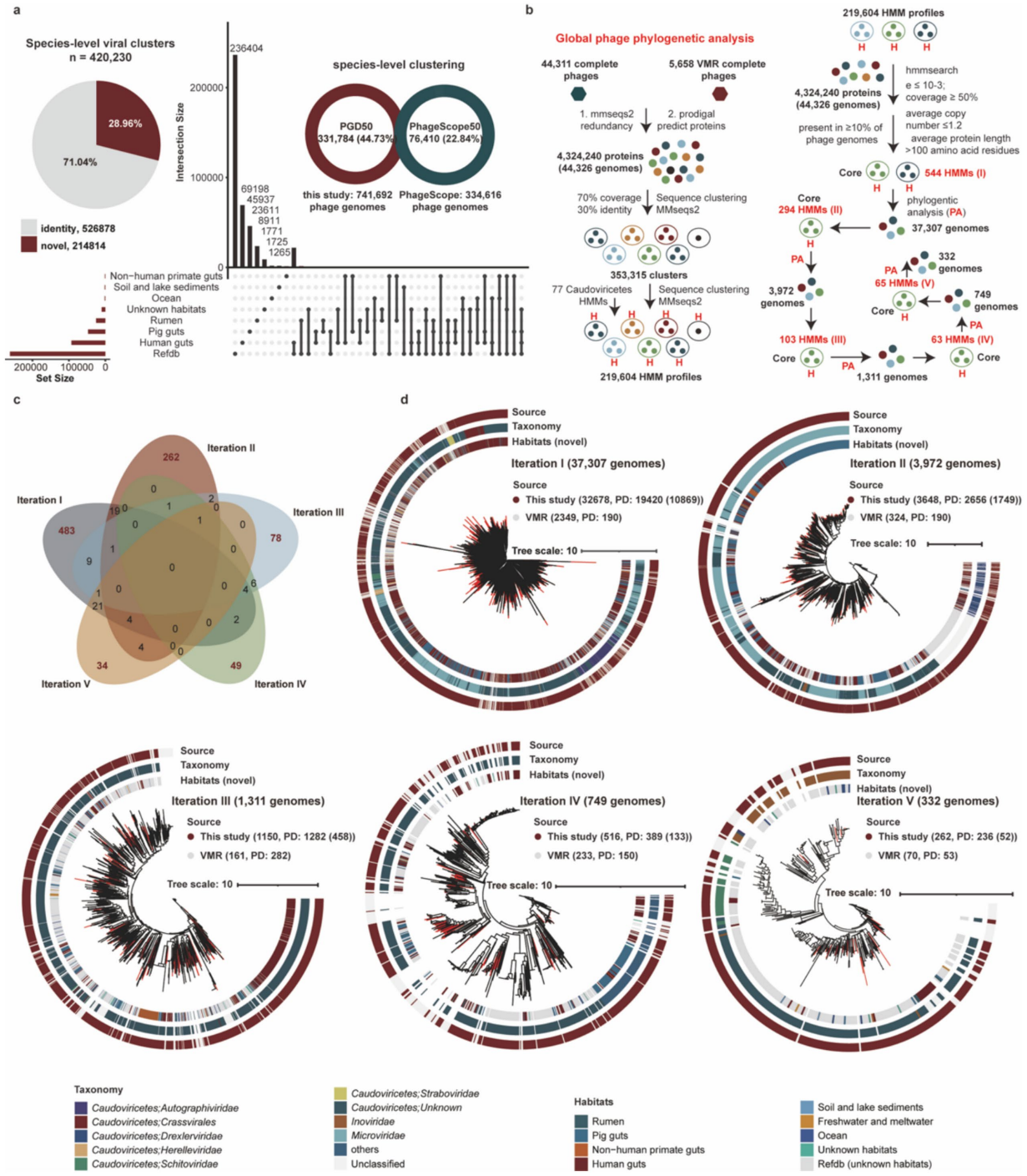

FIGURE 2
Assessing novelty of PGD50 and construction of global phylogenetic trees based on complete phage genomes. **(a)** The assessing novelty of PGD50 at species-level viral clusters. The pie represents the proportion of novel phages from PGD50 and the UpSet plot compares phage populations among different habitats. **(b)** The pipeline of global phage phylogenetic analysis of complete phages and core HMM profiles and number of phage genomes for five iterations. The pipeline (left) is performed to build all 219,604 HMM profiles based on the MMseqs2 and HMMER software. The pipeline (right) is performed to generate core HMM profiles and phylogenetic trees for five iterations, and the core HMM profiles for each iteration are generated based on all 219,604 HMM profiles. The parameters for generating core HMM profiles are shown in first iteration, and the parameters for five iterations are consistent. Phylogenetic trees for five iterations are constructed based on the method developed by Low et al. (2019) and corresponding core HMM profiles. **(c)** The sharing and unique core HMM profiles for five iterations to global phage phylogenetic trees. The distribution of sharing and unique core HMM profiles are described by the Venn diagram. The red numbers represent the unique core HMM profiles and the black number represent the sharing core HMM profiles for five iterations. **(d)** The global phage phylogenetic trees for five iterations and the source distribution of phage genomes from this study and the VMR database. The different colors OD outer circle for the phylogenetic trees represent phage genomes from different sources and the red clades represent novel complete phage genomes from this study.

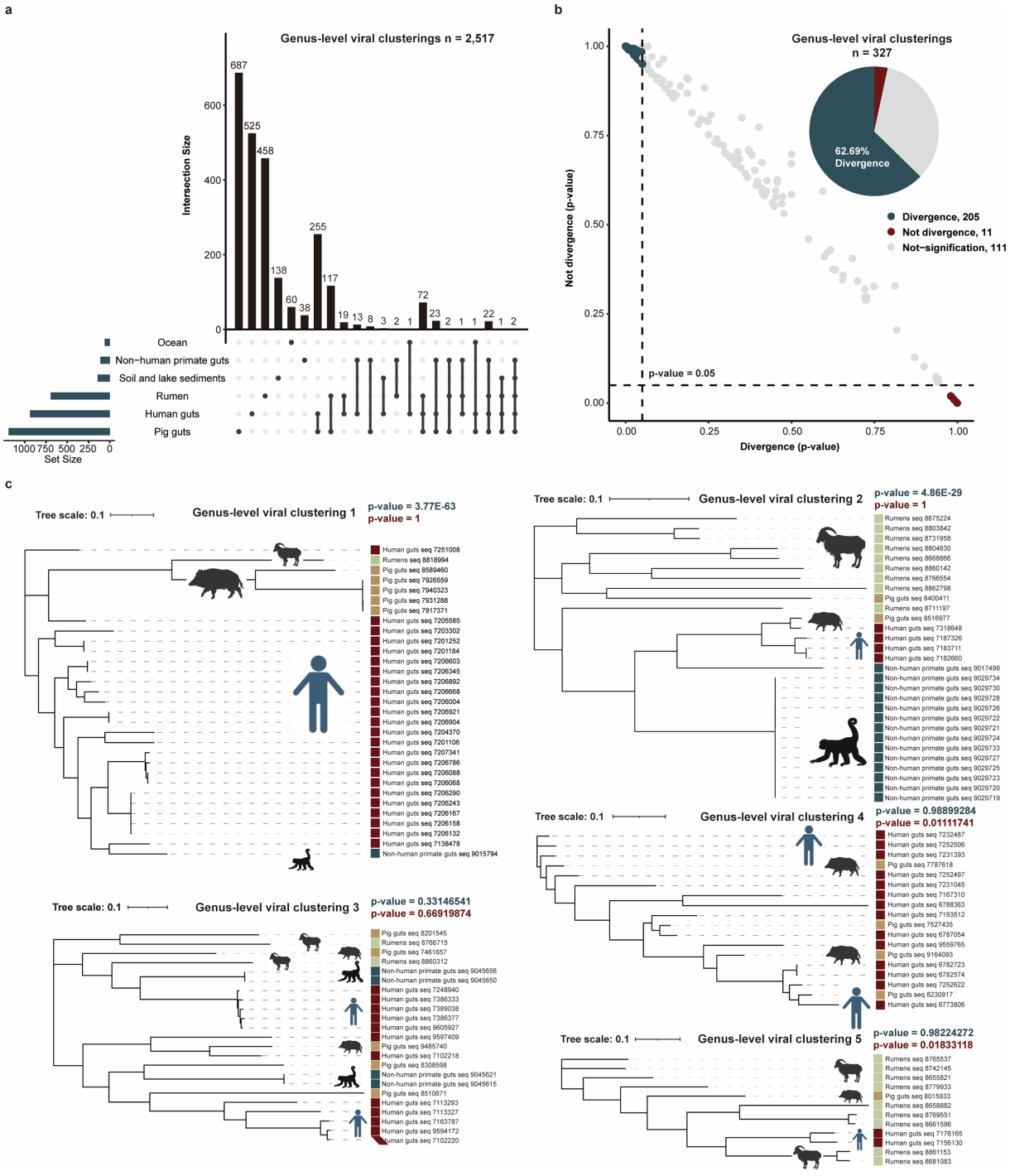

FIGURE 3
Potential divergence analysis for complete phage genomes with habitat types. **(a)** Comparing genus-level viral clusters among six habitats. **(b)** Potential divergence analysis of genus-level viral clusters with different habitat types. The dot plot shows the likelihood (*p*-values) distributions of 327 genus-level viral clusters that were diverged (black green dots) or not diverged (red dots) with habitat types. The pie chart shows the proportion of divergent and not-divergent genus-level viral clusters. **(c)** Five examples of genus-level viral clusters potentially diverged or did not diverge with habitat types. The blue *p*-value ($p < 0.05$) and the red *p*-value ($p < 0.05$) represent the significant divergence and significant not divergence for phage genomes with habitats.

Viral proteins were highly divergent even within the same virus family, limiting the utility of sequence-based similarity searches when amino acid identity fell below 30%. To overcome the limitations of sequence-based annotation (e.g., for hits with <30% AA identity), we performed structural similarity searches. This approach leverages the fact that protein structural domains were often more conserved than their amino acid sequences, allowing for the detection of distant evolutionary relationships that were otherwise missed (Figure 4b).

FIGURE 4
Functional annotations of phage genomes in the PGD50. **(a)** The proportion of using alternative genetic codes for phages genomes (pie chart), the number of proteins annotated by the Pfam-A database for phages using five alternative genetic codes (bar chart). **(b)** The pipeline of functional annotations for 55,294,737 phage genes. **(c)** The proportion of annotated phage genes for 55,294,737 phage genes. **(d)** Functional items with the numbers of phage genes in the top 50. Different colors represent different functional categories. **(e)** The proportion of further annotated phage genes using structure searching method for genes of top 100 clusters from no hit. The pie chart represents the proportion of annotated phage genes using structure searching method, and the 3D structures of phage genes were compared with Alphafold2 and Foldseek.

Interestingly, 87.42% (48,339,065/55,294,737) proteins were annotated (Figure 4c) and we further classified these genes into functional items, and the items with the number of annotated genes in the top 50 were listed. Among them, functional items related to the structure, assembly and packaging, DNA replication and transcription, and lysis, all of which were typical functional capacities of phages were enriched by annotated genes (Figure 4d). Critically, structural searches resolved 53% (53/100) of previously unannotated proteins (top 100 clusters with no sequence hits), demonstrating its power for annotating divergent viral proteins (Figure 4e and Supplementary Table S5).

3.5 Revealing phage–host relationships and pathogen targeting potential via CRISPR spacer matching

The distribution of host bacteria or archaea is a strong determinant for the distribution of phages, and the indigenous phage community also greatly affects the structure and function of the host bacterial or archaeal community. To establish phage–host linkages, we leveraged CRISPR spacer similarity, a key determinant linking phage distribution to their bacterial or archaeal hosts. Analysis of 741,692 phage genomes identified putative hosts for 56.75% (420,907/741,692) phages through spacer matches (Supplementary Table S5). Our analysis revealed that 35.38% (262,423/741,692) of phage genomes were linked via CRISPR spacers to multiple bacterial genera, with some connections spanning different phyla (Figure 5a). There were 21.37% (158,484/741,692) phage genomes only targeting one host genus, and host for these specialist host viruses mainly belonged to the keystone genera *Bacteroides* and *Prevotella*, critical in gut or hypersaline ecosystems.

Our analysis focused on *Escherichia coli* and *Klebsiella pneumoniae* given their predominant role in the global burden of antimicrobial resistance. This focused approach allows for a deeper investigation into phage solutions for these clinically paramount threats (Murray et al., 2022). We analyzed virulent phages targeting pathogens

including *Escherichia coli* and *Klebsiella pneumoniae*, and first estimated the distribution of PBP proteins, VFs, and ARGs in collected pathogenic bacteria genomes from the GenBank database. Interestingly, we found virulent phage genomes in the PGD50 could target 67.83% (4,295/6,332) *Escherichia Coli* and 31.08% (1,288/4,050) *Klebsiella pneumoniae* based on CRISPR spacer matching (Figure 5b), suggesting that these virulent phages in the PGD50 might be an ideal tool for phage therapy via targeted lysis of pathogenic bacteria.

3.6 Competitive phage networks and CRISPR-Cas system distribution

We identified 37,708 CRISPR spacers within 4,430 phage genomes in the PGD50. Among these, 8.35% (3,149/37,708) targeted 52,909 phage genomes, establishing extensive phage–phage interaction networks (Figure 6a). Target pair analysis revealed single-directed relationships (where one phage targets another without reciprocal targeting) dominated these interactions at 89.83% (237,936/264,882), while double-directed pairs (reciprocal targeting) constituted the remaining 10.17%. Critically, 94.60% (250,585/264,882) of targeted pairs consisted of phages infecting the same host, revealing a high prevalence of potential competitive relationships (Supplementary Table S6).

CRISPR-Cas systems are adaptive immune systems widespread in hosts but rarely found in phage genomes. We identified 243 CRISPR-Cas systems within phage genomes that specifically target other phages, with the most prevalent subtypes being I-C, I-F, and II-C (Figure 6b and Supplementary Table S7). Among these, 37 systems (15.23%) were complete. More broadly, a total of 299 CRISPR-Cas systems were identified across all phage genomes. In contrast, we found 30,222 CRISPR-Cas systems encoded by host genomes, which were predominantly subtypes I-C, II-A, and I-E (Figure 6c and Supplementary Table S8). Notably, phage-encoded CRISPR-Cas systems (83.61%, 250/299) frequently lacked spacer acquisition proteins (Cas1, Cas2, and Cas4), suggesting partial horizontal gene transfer (HGT) during acquisition. Furthermore, focusing on the two most prevalent subtypes, we found that 96.97% (116/120) of the I-C systems and 17.86% (5/28) of the II-C systems were missing these Cas proteins. Given the biotechnological significance of Cas12 proteins, particularly their compact size for gene-editing applications, we performed a phylogenetic analysis to explore their diversity in phages. Our analysis incorporated a total of 1,311 Cas12 sequences, comprising 1,310 derived from host genomes (spanning subtypes Cas12a, Cas12b, Cas12f, and Cas12k) and a single, notable Cas12a sequence identified from a phage genome in our study. The resulting phylogenetic tree (Figure 6d) revealed distinct clustering of Cas12 subtypes, with the phage-encoded Cas12a nesting within the diversity of host-encoded Cas12a sequences. This placement suggested a potential evolutionary history of horizontal gene transfer between phages and their bacterial hosts for this particular system. While the single phage sequence precluded a conclusion on phage-specific diversity, its presence alone was significant. The Cas12 protein identified in the phage contained the conserved RuvC nuclease domain. This domain was critically important as it was responsible for the DNA cleavage activity that formed the foundation for all DNA-targeting applications of Cas12 in biotechnology (Makarova et al., 2020). The preservation of this functional domain in the phage protein underscored its potential as a functional nuclease and a valuable resource for mining novel gene-editing tools.

4 Discussion

Phages play a key role in maintaining the balance of microbial ecosystems (Rascovan et al., 2016; Emerson et al., 2018), but their interactions with hosts and other phages are largely unknown. This study presents a comprehensive and expansive resource of phage genomes, significantly augmenting our understanding of global phage diversity, evolutionary dynamics, and functional potential. By integrating massive datasets from diverse habitats, we have assembled a collection of 741,692 medium-to-high-quality phage genomes, vastly exceeding the scale of most previous individual studies and significantly enriching existing public databases like the IMG/VR (Roux et al., 2021), GenBank (Benson et al., 2013), and NT (Sayers et al., 2022) (IGN). The sheer number of phage genomes analyzed here provides unprecedented resolution for exploring the virosphere.

The most striking finding is the immense proportion (28.96%) of phage genomes clustering into 158,522 species-level viral clusters that lack any representatives in the IGN. This underscores a profound gap in our current cataloging of viral diversity. These novel species-level viral clusters likely represent phages endemic to understudied environments, highly divergent lineages, or those infecting uncultured hosts. Their discovery dramatically reshapes our perception of the virosphere's true breadth and complexity, suggesting that known phages represent merely a fraction of the total diversity. The construction of global phylogenetic trees reveals that our dataset substantially expands the known diversity of the *Caudoviricetes*, filling critical phylogenetic gaps and introducing novel, deep-branching lineages. This expansion was particularly pronounced in previously underexplored habitats such as the pig gut and rumen. However, we must acknowledge that this apparent "expansion" is partially shaped by the inherent unevenness of existing genomic databases. The deeper sequencing of certain environments like the human gut naturally allows for the resolution of finer-scale genetic diversity, while the true evolutionary breadth of under-sampled habitats likely remains underestimated (Nayfach et al., 2021b). Consequently, the present evolutionary map should be viewed as a robust yet interim framework, one that will be refined as future metagenomic surveys encompass a broader spectrum of global ecosystems.

Beyond cataloging diversity, our genomic analyses revealed distinct evolutionary patterns reflected in the association between phages and their habitats. While the observed signals are consistent with potential divergence with habitats, we interpret these patterns as strong evidence of environmental filtering and habitat adaptation. Phages are likely finely tuned to the physicochemical and biological conditions of their respective niches, a phenomenon driven by factors such as host availability, nutrient constraints, and inter-phage competition (Koskella and Brockhurst, 2014; Sharma et al., 2018). However, caution is warranted in ascribing these distribution patterns solely to strict co-evolution. Habitat filtering, where environmental conditions selectively favor both compatible hosts and their phages represents a powerful, alternative mechanism shaping these ecological relationships (Lennon and Martiny, 2008). In other words, the signal we detect may reflect phage adaptation to their host's ecological niche, rather than direct, synchronous genome evolution between phage and

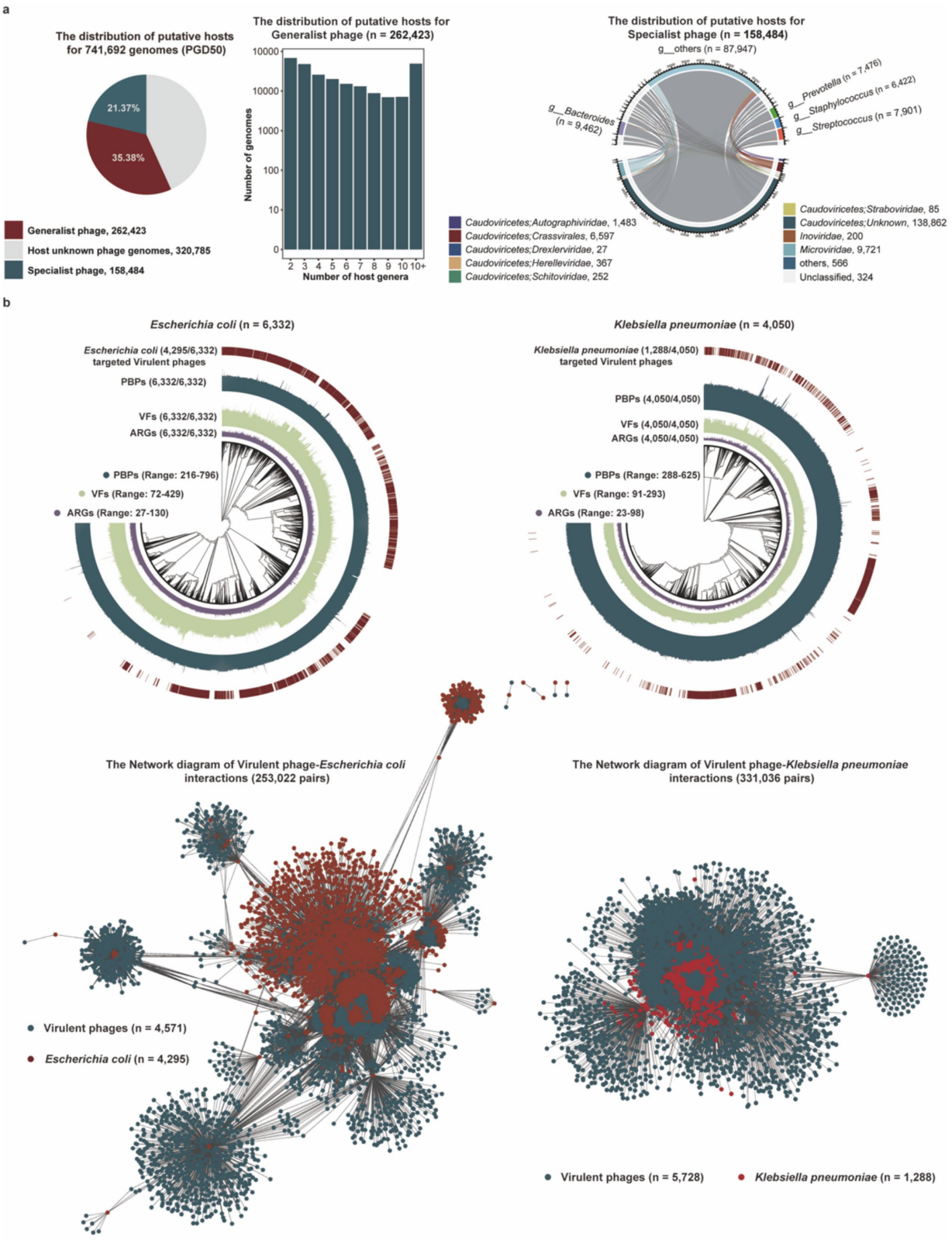

FIGURE 5
Potential hosts and targeting pathogen for phage genomes in the PGD50. **(a)** The proportion of generalist and specialist phages in the PGD50 (pie), the distribution of putative host numbers for 262,423 generalist phages at the genus level (box plot), and the distribution of putative prokaryotic hosts for 158,484 specialist phages (cycle diagram). **(b)** The phylogenetic trees of pathogenetic bacteria including *Escherichia coli* and *Klebsiella pneumoniae* and the network diagram of virulent phages targeting *Escherichia coli* and *Klebsiella pneumoniae*. The outer circle in phylogenetic trees represents *Escherichia coli* and *Klebsiella pneumoniae* genomes targeted by virulent phages, and the inner three circles represent *Escherichia coli* and *Klebsiella pneumoniae* genomes with VFs, ARGs, and PBP proteins, respectively.

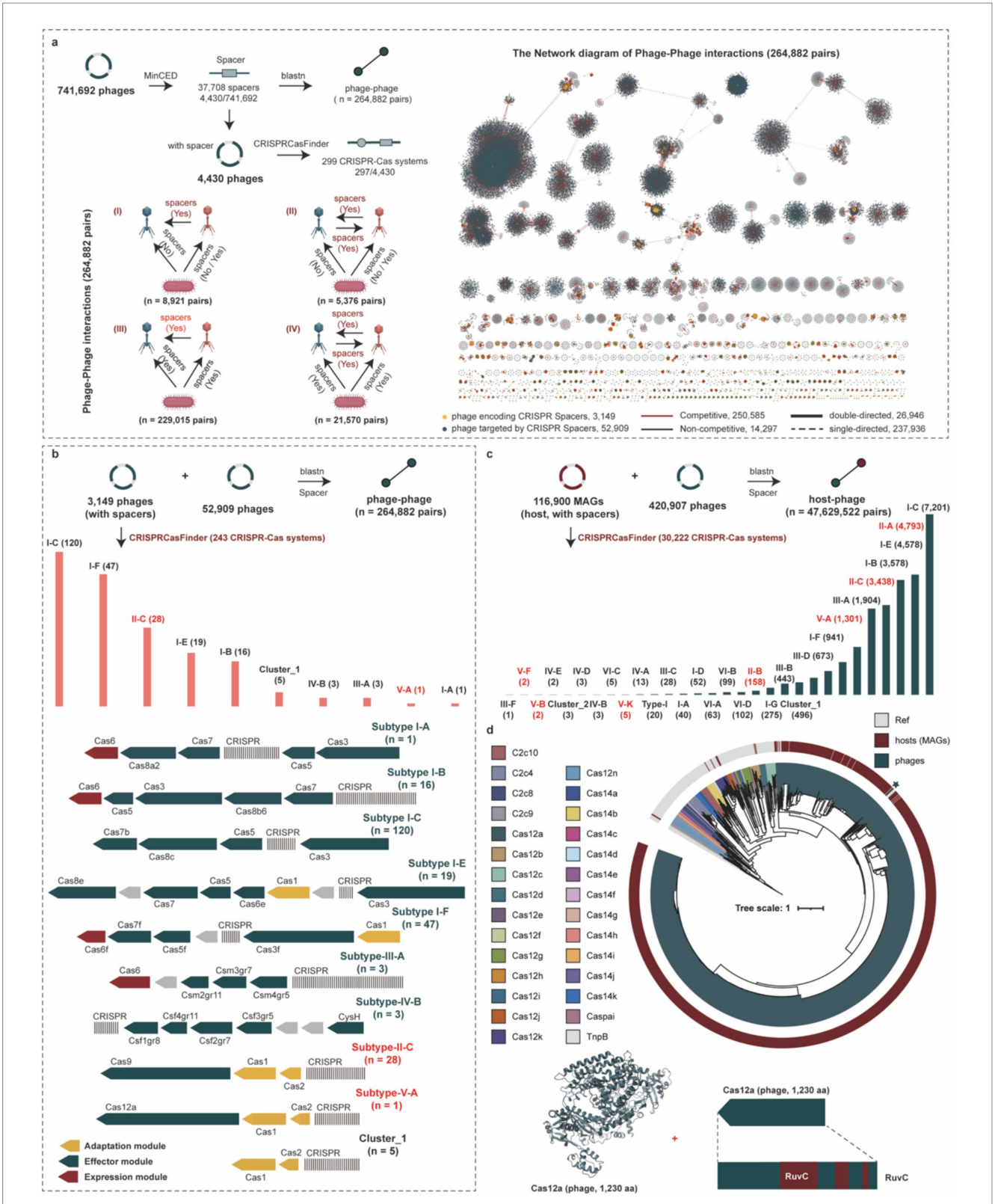

FIGURE 6
Phage–phage interaction and revealing diverse CRISPR-Cas systems for hosts and phages. **(a)** The pipeline of identification for CRISPR-Cas systems and CRISPR spacers for phage genomes, and the four specific phage–phage interactions (left). The network diagram of phage–phage interactions for phage genomes (right). **(b)** The detailed distribution and structural diagram of CRISPR-Cas systems for phage genomes. **(c)** Distribution of CRISPR-Cas system types across host genomes involved in predicted phage-host interactions. **(d)** The phylogenetic tree of Cas12 proteins from phages, hosts, and reference proteins, and the 3D structure and RUVC domain of the Cas12a protein from phage genomes. The different colors of outer circle in the phylogenetic tree represent the sources of Cas12 proteins, and the colors of clades represent different types of Cas12 proteins.

host. For instance, a longitudinal study of *Aeromonas* and its phages demonstrated that their interaction dynamics oscillated over time between "arms race" and "fluctuating selection" modes (Marston et al., 2012). Our cross-sectional study may have captured only a single snapshot of this complex, dynamic process. Disentangling these possibilities will require future longitudinal time-series sampling of the same habitats, combined with experimental validation through controlled co-evolution experiments of specific host-phage pairs (McGrath, 2024). Furthermore, the detection of alternative genetic code usage in some phages highlights an intriguing evolutionary strategy (Devoto et al., 2019), possibly conferring advantages like evasion of host defenses or optimization of replication efficiency under specific conditions, warranting deeper investigation.

The functional annotation of phage genomes, particularly for proteins with no known homologs, remains a major challenge (Nomburg et al., 2024). Our application of 3D structural similarity searches represents a significant methodological advance. By moving beyond sequence-based homology, this approach provided functional predictions for 53% of the top 100 previously unannotated viral proteins based on 3D structural resolution. This not only enhances our understanding of the functional repertoire encoded within this novel phage diversity but also provides a powerful strategy for future viral metagenomic studies.

Our analysis of phage–host interactions, leveraging CRISPR spacer matching, provided crucial insights into the ecological networks connecting phages and their potential hosts (Tisza and Buck, 2021; Johansen et al., 2023). A key finding was that a substantial proportion (35.38%) of phage genomes were linked via spacer matches to hosts spanning multiple genera or even phyla, suggesting the potential for broad host ranges (Nishijima et al., 2022; Bignaud et al., 2025). However, these *in silico* predictions warrant a nuanced interpretation. True ecological generalists capable of productively infecting distantly related hosts are considered rare in nature. The observed patterns may therefore stem from several alternative factors: the presence of common integrative genetic elements shared across diverse hosts, the inherent limitations of predictive bioinformatics, or the fact that spacer matches can reflect past, non-productive infection events rather than active, concurrent replication. Despite these important caveats, this spacer-based approach proved highly valuable for generating specific, testable hypotheses, robustly predicting potential hosts for numerous phages, including those with links to clinically relevant pathogenic bacteria (López-Beltrán et al., 2024) and thereby highlighting promising candidates for further therapeutic exploration. This approach not only revealed complex ecological networks of phage competition and co-existence mediated through shared CRISPR targets (Tisza and Buck, 2021) but also proved particularly valuable for identifying strictly lytic (virulent) phages with therapeutic potential. The strictly lytic life cycle of these virulent phages makes them ideal therapeutic candidates, as it enables the direct and rapid eradication of target pathogens. However, translating these foundational discoveries into clinical practice faces significant challenges. Two of the most prominent hurdles are the typically narrow host range of phages, which can limit their applicability against diverse bacterial strains, and concerns regarding the potential transduction of bacterial virulence factors (Petrovic Fabijan et al., 2023). The vast and diverse reservoir of virulent phages uncovered in our study provides a unique resource to address these challenges. The many genes of unknown function within these genomes may encode novel proteins capable of modulating or evading host immune responses. Through rational genetic engineering, such as modifying phage tail fibers to broaden host range or knocking out highly immunogenic, non-essential genes, we can leverage these natural blueprints to develop next-generation phage-based therapeutics that are safer, more effective, and better suited to clinical application (Meile et al., 2022).

A particularly exciting and unexpected finding was the detection of diverse CRISPR-Cas systems within the phage genomes themselves (Al-Shayeb et al., 2020; Al-Shayeb et al., 2022). Phage-encoded CRISPR-Cas systems open fascinating new avenues for research into phage–host arms races, where phages may utilize these systems to compete against other mobile genetic elements (including other phages) or even manipulate host defenses (Al-Shayeb et al., 2020). Beyond their biological significance, these phage-borne systems represent a vast, largely untapped reservoir of novel CRISPR-Cas variants with potentially unique properties (e.g., smaller size, different PAM specificities) (Pausch et al., 2020; Carabias et al., 2021; Al-Shayeb et al., 2022). This positions our phage genome collection as an extraordinarily rich source for mining the next generation of gene editing tools with enhanced capabilities for biotechnology and medicine.

While this study provides a landmark resource for viral ecology, several limitations inherent to metagenomic analysis must be acknowledged. First, our reliance on a genome completeness threshold (PGD50) ensured high-quality analysis but may have systematically excluded abundant, fragmented viral sequences, leading to an underestimation of the diversity of certain phage groups. Second, although our functional inference was augmented by structural similarity searches to reveal distant homologies (van Kempen et al., 2024), it remains constrained by homology-based methods; proteins with truly novel folds represent a fundamental blind spot, and all predictions require biochemical confirmation. Finally, our host prediction strategy relies solely on CRISPR spacer matches. While this method provides high-specificity links, it is inherently limited by the incompleteness of microbial genome catalogs and reflects historical infection events rather than active replication. Furthermore, this singular approach leaves phage interactions with many uncultured or un-sequenced hosts undetected; future work incorporating complementary methods, such as k-mer composition analysis, would be essential to systematically expand host assignment coverage and obtain a more comprehensive view of phage–host interaction networks. Future efforts combining more permissive assembly strategies, multi-faceted host prediction, and experimental validation will be crucial to overcome these biases. In total, this study delivers an unparalleled genomic resource that fundamentally expands our knowledge of phage diversity on Earth. We have uncovered a vast reservoir of novel phages, revealed intricate patterns of potential divergence and adaptation, developed innovative methods for functional annotation, and uncovered critical insights into phage–host interactions and competitive networks. Most significantly, we have demonstrated the immense, dual application potential of this resource: firstly, as a targeted library for discovering potent phage therapy agents against pathogenic bacteria, and secondly, as a treasure trove for mining the next generation of innovative

CRISPR-Cas-based gene editing technologies. This dataset provides an essential foundation for future research aimed at understanding the intricate roles of phages in global ecosystems, combating antibiotic resistance, and advancing genetic engineering.

5 Conclusion

In conclusion, this study constructed the PGD50 database, a unified resource of 741,692 high-quality phage genomes, which enabled a systematic reassessment of global phage diversity and ecology. Our key advance lies not in the initial reporting of phage diversity or phage-encoded CRISPR-Cas systems, but in the substantial expansion of their documented scale and diversity. Specifically, we identified a significant number of novel, deep-branching lineages, represented by 158,522 species-level viral clusters that were absent from existing references. Furthermore, our analysis reframed the observed ecological patterns not as definitive "co-evolution," but as a distinct "habitat divergence." This pervasive phylogeographic signal indicates a strong environmental imprint on phage evolution, which may arise from co-evolutionary dynamics, environmental filtering, or a combination of both. Beyond diversity, our integrated approach including combining structural annotation, CRISPR spacer analysis, and comparative genomics, provided foundational insights into phage function, host interaction networks, and the expanded distribution of CRISPR-Cas subtypes within phages, underscoring their potential as future therapeutic and biotechnological tools. Collectively, this work provides a refined framework and resource for future research into phage biology and application.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Author contributions

CW: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft. ZC: Formal analysis, Validation, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

The authors appreciate the colleagues in the National Key Laboratory of Pig Genetic Improvement and Germplasm Innovation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2025.1686402/full#supplementary-material

References

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., et al. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* 39, 105–114. doi: 10.1038/s41587-020-0603-3

Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., et al. (2020). Clades of huge phages from across Earth's ecosystems. *Nature* 578, 425–431. doi: 10.1038/s41586-020-2007-4

Al-Shayeb, B., Skopintsev, P., Soczek, K. M., Stahl, E. C., Li, Z., Groover, E., et al. (2022). Diverse virus-encoded CRISPR-Cas systems include streamlined genome editors. *Cell* 185, 4574–4586.e16. doi: 10.1016/j.cell.2022.10.020

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2013). GenBank. *Nucleic Acids Res.* 41, D36–D42. doi: 10.1093/nar/gks1195

Bignaud, A., Conti, D. E., Thierry, A., Serizay, J., Labadie, K., Poulain, J., et al. (2025). Phages with a broad host range are common across ecosystems. *Nat. Microbiol.* 10, 2537–2549. doi: 10.1038/s41564-025-02108-2

Brown, T. L., Charity, O. J., and Adriaenssens, E. M. (2022). Ecological and functional roles of bacteriophages in contrasting environments: marine, terrestrial and human gut. *Curr. Opin. Microbiol.* 70:102229. doi: 10.1016/j.mib.2022.102229

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Burstein, D., Harrington, L. B., Strutt, S. C., Probst, A. J., Anantharaman, K., Thomas, B. C., et al. (2017). New CRISPR-Cas systems from uncultivated microbes. *Nature* 542, 237–241. doi: 10.1038/nature21059

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Camargo, A. P., Nayfach, S., Chen, I. A., Palaniappan, K., Ratner, A., Chu, K., et al. (2023a). 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. doi: 10.1093/nar/gkac1037

Camargo, A. P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., et al. (2023b). Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* 42, 1303–1312. doi: 10.1038/s41587-023-01953-y

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Carabias, A., Fuglsang, A., Temperini, P., Pape, T., Sofos, N., Stella, S., et al. (2021). Structure of the mini-RNA-guided endonuclease CRISPR-Cas12j3. *Nat. Commun.* 12:4476. doi: 10.1038/s41467-021-24707-3

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 38, 5315–5316. doi: 10.1093/bioinformatics/btac672

Chen, C., Zhou, Y., Fu, H., Xiong, X., Fang, S., Jiang, H., et al. (2021). Expanded catalog of microbial genes and metagenome-assembled genomes from the pig gut microbiome. *Nat. Commun.* 12:1106. doi: 10.1038/s41467-021-21295-0

Clokie, M. R., Millard, A. D., Letarov, A. V., and Heaphy, S. (2011). Phages in nature. *Bacteriophage* 1, 31–45. doi: 10.4161/bact.1.1.14942

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Neron, B., et al. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* 46, W246–W251. doi: 10.1093/nar/gky425

Devoto, A. E., Santini, J. M., Olm, M. R., Anantharaman, K., Munk, P., Tung, J., et al. (2019). Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nat. Microbiol.* 4, 693–700. doi: 10.1038/s41564-018-0338-9

Dion, M. B., Oechslin, F., and Moineau, S. (2020). Phage diversity, genomics and phylogeny. *Nat. Rev. Microbiol.* 18, 125–138. doi: 10.1038/s41579-019-0311-5

Emerson, J. B., Roux, S., Brum, J. R., Bolduc, B., Woodcroft, B. J., Jang, H. B., et al. (2018). Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* 3, 870–880. doi: 10.1038/s41564-018-0190-y

Federici, S., Kredo-Russo, S., Valdes-Mas, R., Kviatcovsky, D., Weinstock, E., Matiuhin, Y., et al. (2022). Targeted suppression of human IBD-associated gut microbiota commensals by phage consortia for treatment of intestinal inflammation. *Cell* 185, 2879–2898.e24. doi: 10.1016/j.cell.2022.07.003

Gregory, A. C., Zablocki, O., Zayed, A. A., Howell, A., Bolduc, B., and Sullivan, M. B. (2020). The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe* 28, 724–740.e8. doi: 10.1016/j.chom.2020.08.003

Haft, D. H., Selengut, J. D., and White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res.* 31, 371–373. doi: 10.1093/nar/gkg128

Hockenberry, A. J., and Wilke, C. O. (2021). BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* 9:e11396. doi: 10.7717/peerj.11396

Hu, J., Chen, J., Ma, L., Hou, Q., Zhang, Y., Kong, X., et al. (2024). Characterizing core microbiota and regulatory functions of the pig gut microbiome. *ISME J.* 18:wrad037. doi: 10.1093/ismejo/wrad037

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 Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119

Ivanova, N. N., Schwientek, P., Tripp, H. J., Rinke, C., Pati, A., Huntemann, M., et al. (2014). Stop codon reassignments in the wild. *Science* 344, 909–913. doi: 10.1126/science.1250691

Johansen, J., Atarashi, K., Arai, Y., Hirose, N., Sorensen, S. J., Vatanen, T., et al. (2023). Centenarians have a diverse gut virome with the potential to modulate metabolism and promote healthy lifespan. *Nat. Microbiol.* 8, 1064–1078. doi: 10.1038/s41564-023-01370-6

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., et al. (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26, 1463–1464. doi: 10.1093/bioinformatics/btq166

Koskella, B., and Brockhurst, M. A. (2014). Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol. Rev.* 38, 916–931. doi: 10.1111/1574-6976.12072

Lennon, J. T., and Martiny, J. B. H. (2008). Rapid evolution buffers ecosystem impacts of viruses in a microbial food web. *Ecol. Lett.* 11, 1178–1188. doi: 10.1111/j.1461-0248.2008.01225.x

López-Beltrán, A., Botelho, J., and Iranzo, J. (2024). Dynamics of CRISPR-mediated virus–host interactions in the human gut microbiome. *ISME J.* 18:wrae134. doi: 10.1093/ismejo/wrae134

Lou, Y. C., Chen, L., Borges, A. L., West-Roberts, J., Firek, B. A., Morowitz, M. J., et al. (2024). Infant gut DNA bacteriophage strain persistence during the first 3 years of life. *Cell Host Microbe* 32, 35–47.e6. doi: 10.1016/j.chom.2023.11.015

Low, S. J., Dzunkova, M., Chaumeil, P. A., Parks, D. H., and Hugenholtz, P. (2019). Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nat. Microbiol.* 4, 1306–1315. doi: 10.1038/s41564-019-0448-z

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J. J., et al. (2020). Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* 18, 67–83. doi: 10.1038/s41579-019-0299-x

Mangalea, M. R., and Duerkop, B. A. (2020). Fitness trade-offs resulting from bacteriophage resistance potentiate synergistic antibacterial strategies. *Infect. Immun.* 88:e00926-19. doi: 10.1128/IAI.00926-19

Marston, M. F., Pierciey, F. J., Shepard, A., Gearin, G., Qi, J., Yandava, C., et al. (2012). Rapid diversification of coevolving marine *Synechococcus* and a virus. *Proc. Natl. Acad. Sci. U.S.A.* 109, 4544–4549. doi: 10.1073/pnas.1120310109

McGrath, C. (2024). Highlight: forging cheaters in iron-limited microbial communities. *Mol. Biol. Evol.* 41:msae072. doi: 10.1093/molbev/msae072

Meile, S., Du, J., Dunne, M., Kilcher, S., and Loessner, M. J. (2022). Engineering therapeutic phages for enhanced antibacterial efficacy. *Curr. Opin. Virol.* 52, 182–191. doi: 10.1016/j.coviro.2021.12.003

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913

Murray, C. J. L., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., et al. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 399, 629–655. doi: 10.1016/S0140-6736(21)02724-0

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N. C. (2021a). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* 39, 578–585. doi: 10.1038/s41587-020-00774-7

Nayfach, S., Paez-Espino, D., Call, L., Low, S. J., Sberro, H., Ivanova, N. N., et al. (2021b). Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* 6, 960–970. doi: 10.1038/s41564-021-00928-6

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Nishijima, S., Nagata, N., Kiguchi, Y., Kojima, Y., Miyoshi-Akiyama, T., Kimura, M., et al. (2022). Extensive gut virome variation and its associations with host and environmental factors in a population-level cohort. *Nat. Commun.* 13:5252. doi: 10.1038/s41467-022-32832-w

Nomburg, J., Doherty, E. E., Price, N., Bellieny-Rabelo, D., Zhu, Y. K., and Doudna, J. A. (2024). Birth of protein folds and functions in the virome. *Nature* 633, 710–717. doi: 10.1038/s41586-024-07809-y

Olm, M. R., Brown, C. T., Brooks, B., and Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* 11, 2864–2868. doi: 10.1038/ismej.2017.126

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Pausch, P., Al-Shayeb, B., Bisom-Rapp, E., Tsuchida, C. A., Li, Z., Cress, B. F., et al. (2020). CRISPR-CasF from huge phages is a hypercompact genome editor. *Science* 369, 333–337. doi: 10.1126/science.abb1400

Petrovic Fabijan, A., Iredell, J., Danis-Wlodarczyk, K., Kebriaei, R., and Abedon, S. T. (2023). Translating phage therapy into the clinic: recent accomplishments but continuing challenges. *PLoS Biol.* 21:e3002119. doi: 10.1371/journal.pbio.3002119

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46, W200–W204. doi: 10.1093/nar/gky448

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Rascovan, N., Duraisamy, R., and Desnues, C. (2016). Metagenomics and the human virome in asymptomatic individuals. *Ann. Rev. Microbiol.* 70, 125–141. doi: 10.1146/annurev-micro-102215-095431

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5:69. doi: 10.1186/s40168-017-0283-5

Resch, G., Brives, C., Debarbieux, L., Hodges, F. E., Kirchhelle, C., Laurent, F., et al. (2024). Between centralization and fragmentation: the past, present, and future of phage collections. *PHAGE* 5, 22–29. doi: 10.1089/phage.2023.0043

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., et al. (2019). Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.* 37, 29–37. doi: 10.1038/nbt.4306

Roux, S., Paez-Espino, D., Chen, I. A., Palaniappan, K., Ratner, A., Chu, K., et al. (2021). IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* 49, D764–D775. doi: 10.1093/nar/gkaa946

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., et al. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 50, D20–D26. doi: 10.1093/nar/gkab1112

Shang, J., Tang, X., Guo, R., and Sun, Y. (2022). Accurate identification of bacteriophages from metagenomic data using transformer. *Brief. Bioinform.* 23:bbac258. doi: 10.1093/bib/bbac258

Sharma, U., Vipra, A., and Channabasappa, S. (2018). Phage-derived lysins as potential agents for eradicating biofilms and persisters. *Drug Discov. Today* 23, 848–856. doi: 10.1016/j.drudis.2018.01.026

Shmakov, S., Abudayyeh, O. O., Makarova, K. S., Wolf, Y. I., Gootenberg, J. S., Semenova, E., et al. (2015). Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol. Cell* 60, 385–397. doi: 10.1016/j.molcel.2015.10.008

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Steinegger, M., and Soding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028. doi: 10.1038/nbt.3988

Tisza, M. J., and Buck, C. B. (2021). A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2023202118. doi: 10.1073/pnas.2023202118

Tomofuji, Y., Kishikawa, T., Maeda, Y., Ogawa, K., Otake-Kasamoto, Y., Kawabata, S., et al. (2022). Prokaryotic and viral genomes recovered from 787 Japanese gut metagenomes revealed microbial features linked to diets, populations, and diseases. *Cell Genom.* 2:100219. doi: 10.1016/j.xgen.2022.100219

Touchon, M., Moura de Sousa, J. A., and Rocha, E. P. C. (2017). Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Curr. Opin. Microbiol.* 38, 66–73. doi: 10.1016/j.mib.2017.04.010

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Zidek, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. *Nature* 596, 590–596. doi: 10.1038/s41586-021-03828-1

van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.* 30, 121–141. doi: 10.1137/040608635

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., et al. (2024). Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* 42, 243–246. doi: 10.1038/s41587-023-01773-0

Wang, B., Liang, Y., Lian, K., Zhang, C., Han, M., Wang, M., et al. (2025). Correlation with viruses enhances network complexity and stability of co-occurrence prokaryotes across the oceans. *mSystems* 10:e0053925. doi: 10.1128/msystems.00539-25

Wang, R. H., Yang, S., Liu, Z., Zhang, Y., Wang, X., Xu, Z., et al. (2024). PhageScope: a well-annotated bacteriophage database with automatic analyses and visualizations. *Nucleic Acids Res.* 52, D756–D761. doi: 10.1093/nar/gkad979

Wang, T., Zhang, P., Anantharaman, K., Wang, H., Zhang, H., Zhang, M., et al. (2025). Metagenomic analysis reveals how multiple stressors disrupt virus–host interactions in multi-trophic freshwater mesocosms. *Nat. Commun.* 16:7806. doi: 10.1038/s41467-025-63162-2

Wei, C., Wang, Y., and Chen, Z. (2025). Comprehensive discovery and functional characterization of diverse prophages in the pig gut microbiome. *Front. Microbiol.* 16:1662087. doi: 10.3389/fmicb.2025.1662087

Wu, Y., Gao, N., Sun, C., Feng, T., Liu, Q., and Chen, W. H. (2024). A compendium of ruminant gastrointestinal phage genomes revealed a higher proportion of lytic phages than in any other environments. *Microbiome* 12:69. doi: 10.1186/s40168-024-01784-2

Yan, W. X., Chong, S., Zhang, H., Makarova, K. S., Koonin, E. V., Cheng, D. R., et al. (2018). Cas13d is a compact RNA-targeting type VI CRISPR effector positively modulated by a WYL-domain-containing accessory protein. *Mol. Cell* 70, 327–339.e5. doi: 10.1016/j.molcel.2018.02.028

Yan, M., and Yu, Z. (2024). Viruses contribute to microbial diversification in the rumen ecosystem and are associated with certain animal production traits. *Microbiome* 12:82. doi: 10.1186/s40168-024-01791-3

Zeng, S., Almeida, A., Li, S., Ying, J., Wang, H., Qu, Y., et al. (2024). A metagenomic catalog of the early-life human gut virome. *Nat. Commun.* 15:1864. doi: 10.1038/s41467-024-45793-z