Long-Read Sequencing Reveals Extensive DNA Methylations in Human Gut Phagenome Contributed by Prevalently Phage-Encoded Methyltransferases

Chuqing Sun, Jingchao Chen, Menglu Jin, Xueyang Zhao, Yun Li, Yanqi Dong, Na Gao, Zhi Liu,* Peer Bork,* Xing-Ming Zhao,* and Wei-Hua Chen*

DNA methylation plays a crucial role in the survival of bacteriophages (phages), yet the understanding of their genome methylation remains limited. In this study, DNA methylation patterns are analyzed in 8848 metagenome-assembled high-quality phages from 104 fecal samples using single-molecule real-time sequencing. The results demonstrate that 97.60% of gut phages exhibit methylation, with certain factors correlating with methylation densities. Phages with higher methylation densities appear to have potential viability advantages. Strikingly, more than one-third of the phages possess their own DNA methyltransferases (MTases). Increased MTase copies are associated with higher genome methylation densities, specific methylation motifs, and elevated prevalence of certain phage groups. Notably, the majority of these MTases share close homology with those encoded by gut bacteria, suggesting their exchange during phage–bacterium interactions. Furthermore, these MTases can be employed to accurately predict phage–host relationships. Overall, the findings indicate the widespread utilization of DNA methylation by gut DNA phages as an evasion mechanism against host defense systems, with a substantial contribution from phage-encoded MTases.

1. Introduction

Bacteriophages (phages) are viruses that infect bacteria and archaea,[1] and modulate many ecological and evolutionary processes in complex microbial communities, including the human gut microbiome.[2] Phages often have narrow host ranges and are thus ideal tools for precision manipulation of the gut microbiota.[3] To defend the phages, prokaryotic organisms adopt a variety of defense mechanisms that are deployed against xenogeneic DNAs, among which the restriction-modification (RM) systems are ubiquitous and have been found in ≈90% of sequenced bacterial genomes.[4] RM systems often consist of a restriction endonuclease (REase) that recognizes a highly specific target DNA sequence (i.e., a distinctive, usually recurrent, molecular sequence, or motif) and degrades the unmethylated ones, and a corresponding DNA methyltransferase (MTase) that protects the same

C. Sun, J. Chen, M. Jin, Y. Li, N. Gao, W.-H. Chen
Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular Imaging, Center for Artificial Intelligence Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology
Huazhong University of Science and Technology
Wuhan, Hubei 430074, P. R. China
E-mail: weihuachen@hust.edu.cn

X. Zhao, W.-H. Chen
College of Life Science
Henan Normal University
Xinxiang, Henan 453007, P. R. China

Y. Dong, X.-M. Zhao
Institute of Science and Technology for Brain-Inspired Intelligence
Fudan University
Shanghai 200433, P. R. China
E-mail: xmzhao@fudan.edu.cn

N. Gao
Department of Laboratory Medicine, Zhongnan Hospital of Wuhan University
Wuhan University
Wuhan 430071, P. R. China

Z. Liu
Department of Biotechnology, College of Life Science and Technology
Huazhong University of Science and Technology
Wuhan 430074, P. R. China
E-mail: zhiliu@hust.edu.cn

P. Bork
European Molecular Biology Laboratory
Structural and Computational Biology Unit
69117 Heidelberg, Germany
E-mail: peerbork@embl.org

P. Bork
Max Delbrück Centre for Molecular Medicine
13125 Berlin, Germany

DNA sequence via the DNA methylation of the bacterial genome.[4] RM systems thus enable bacteria to distinguish self-genome from invading phage DNAs that are either unmethylated or not properly methylated. To escape from the host immunity, phages have developed strategies to overcome the RM systems, which can involve the methylation of their genomes through the hijacking of the host MTases, packing host MTase proteins into their virions, or incorporating host MTase genes into their genomes.[5]

Recent exploration of large human virome/phageome datasets[6] has identified a substantial amount of novel gut phage genomes and revealed their diversity in human gut.[2] Furthermore, we have considerable anecdotal knowledge of the vital roles of phages in shaping the microbial community structure,[5a,7] mediating horizontal gene transfers among bacteria,[8] and modulating host metabolic capacities.[9] In contrast, we know little about the epigenome landscape of entire phage microbial communities and the abundance of particular survival tactics, such as the encoding of their own MTases and the potential impact on the survival of phages.

Long-read sequencing techniques such as the single-molecule real-time (SMRT) and Nanopore have allowed us to explore the large-scale, genome-wide DNA methylations; they were successfully applied to genomes of eukaryotes[10] and prokaryotes,[11] and recently also were used to characterize the methylation patterns in individual viral genomes[12] and those in environmental samples.[13] Although the Nanopore sequencing could identify more types of DNA methylations including the N^4-methyl-cytosine (m4C), N^6-methyl-adenine (m6A), and N^5-methyl-cytosine (m5C),[14] the SMRT sequencing, especially in its circular consensus sequencing (CCS) mode could provide better resolution and higher accuracy (≈85%) in DNA methyla-

P. Bork
Yonsei Frontier Lab (YFL)
Yonsei University
Seoul 03722, South Korea
P. Bork
Department of Bioinformatics, Biocenter
University of Würzburg
97070 Würzburg, Germany
X.-M. Zhao
MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence and MOE Frontiers Center for Brain Science
Fudan University
Shanghai 200433, P. R. China
X.-M. Zhao
State Key Laboratory of Medical Neurobiology, Institute of Brain Science
Fudan University
Shanghai 200433, P. R. China
X.-M. Zhao
Department of Neurology, Zhongshan Hospital
Fudan University
Shanghai 200032, P. R. China
X.-M. Zhao
International Human Phenome Institutes (Shanghai)
Shanghai 200433, P. R. China
W.-H. Chen
Institution of Medical Artificial Intelligence
Binzhou Medical University
Yantai 264003, P. R. China

tion detection,[15] and has been successfully used to characterize viral DNA methylations in marine samples.[13a] Hence, we applied SMRT sequencing to 104 viral-like particles (VLPs) enriched human fecal samples and conducted a comprehensive survey of the DNA methylation landscape of human gut DNA phages, which allowed us to identify and quantify DNA methylations and their contributing factors at a large scale.

2. Results

2.1. A Set of High-Quality Phage Genomes Representing the Human Gut Phageome

To obtain high-quality phage genomes for subsequent DNA methylation analysis, we first established a Chinese Human Gut Virome (CHGV) catalog consisting of 21 646 non-redundant phage genomes, via the combined assembly of short-(Illumina) and long-(PacBio in CCS mode) reads (Experimental Section; see also ref. [16]) Briefly, we enriched double-stranded DNA phages from fecal samples of 135 individuals of a Chinese population, subjected them to short-read Illumina sequencing, and selected 104 samples with sufficient amounts of high-molecular weight DNA for PacBio SMRT long-read sequencing (Experimental Section). We reconstructed the phage genomes through a hybrid assembly pipeline using both the short- and long-reads, followed by de-replication, and viral genome recognition to generate a non-redundant set of phage genomes (**Figure 1**A).[16] The viral recognition procedure included viral prediction by VirSorter,[17] VirFinder,[18] PPR-Meta,[19] protein annotation against the NCBI POG (Phage Orthologous Groups) database[20] and comparing the assembled contigs to NCBI Viral RefSeq using BLASTn,[21] followed by evaluation of the outputs based on six commonly used criteria (Experimental Section).

To avoid biases brought by the fragmented phage genomes, we selected a subset of 8848 high-quality phages with ≥ 90% completeness (CHGV-HQ hereafter; Table S1, Supporting Information) according to CheckV;[22] in addition to the CheckV recognition, 68.80% of the genomes were recognized by at least two viral recognition criteria (Figure 1B).

We next examined whether the CHGV-HQ genomes could allow us to obtain an unbiased view on the DNA methylation landscape of the human gut phageome. Because the human gut phageome was both diverse and individual-specific,[2] rather than directly comparing the phage sequence to those in public human virome databases, we instead checked if our CHGV-HQ phage genomes represented most of the sequence signatures of known human gut phages. Briefly, we trained a virus detection machine learning (ML) model using the 8848 genomes as the true positives and a subset of the Unified Human Gastrointestinal Genome (UHGG)[23] genomes as true negatives (the CHGV-HQ model; Experimental Section). We first tested the model on an independent dataset consisting of viral genomes from the IMG/VR database[24] and the left-out bacterial sequences in the UHGG. The model could accurately distinguish the viral from bacterial sequences with an overall area under the receiver operating characteristic curve (AUC) value of 93% (94% precision and 92% recall rates; Figure 1C). We then applied the model to the public virome databases including the gut virome database (GVD),[6a]
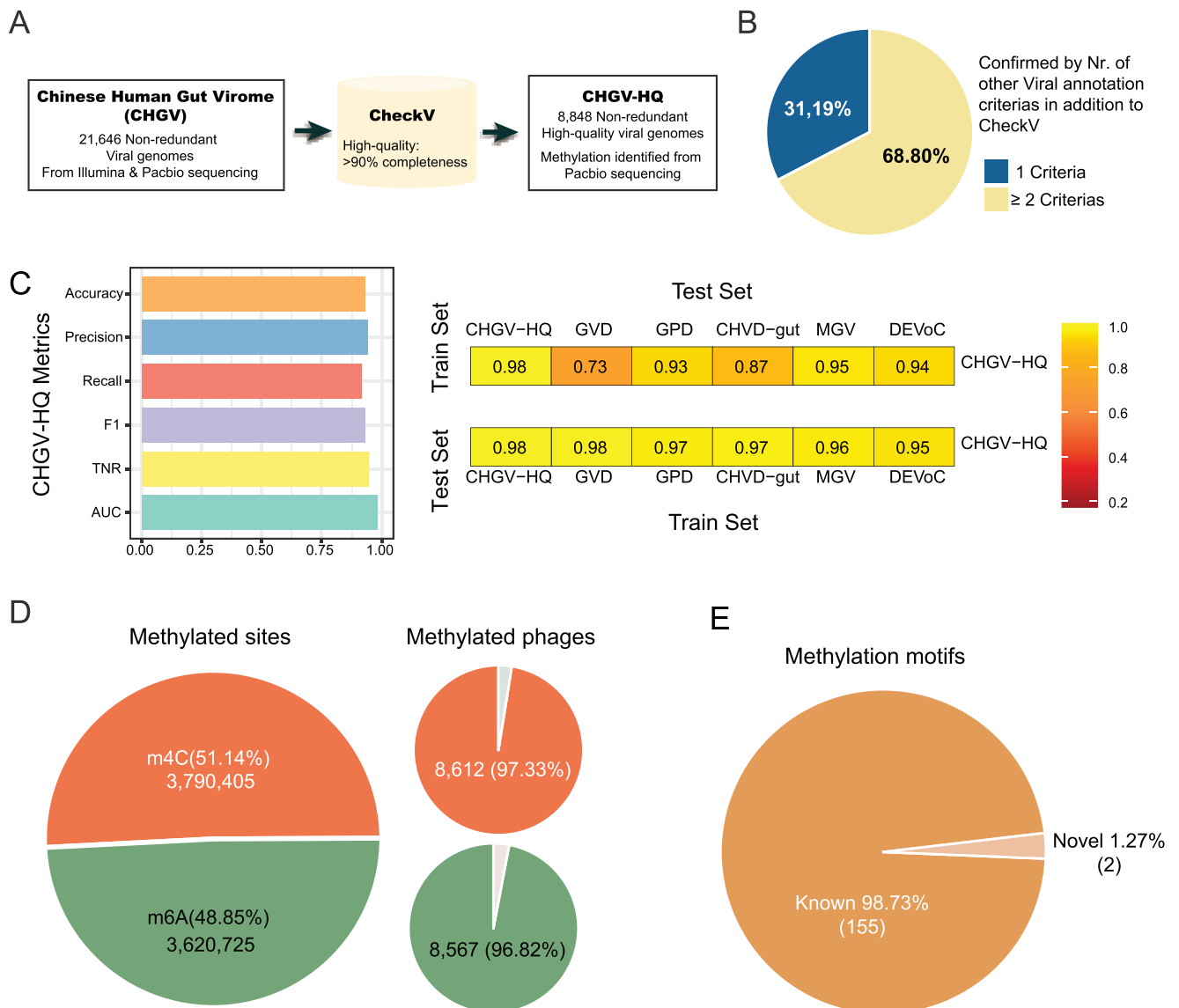
Figure 1. The DNA methylation landscape of the human gut phageome identified using a representative set of 8848 high-quality gut phages. A) Generation of 8848 high-quality non-redundant double-stranded DNA phages from human fecal samples (CHGV-HQ) using combined assembly of short-(Illumina) and long-(PacBio)read sequencing. B) Percentages of CHGV-HQ genomes annotated by the numbers of viral-recognition criteria in addition to CheckV (Experimental Section). C) Left panel: Performance metrics of the viral-prediction machine learning model based on the CHGV-HQ genomes (i.e., the CHGV-HQ model) in distinguishing the viral genomes in the IMG/VR database from the bacterial genomes in the)UHGG2 database (Experimental Section). *Y*–axis: Accuracy—correct predictions out of all predictions, precision—ratio of true positives over the sum of false positives and true negatives, recall—correctly predicted outcomes to all predictions, F1—combines the accuracy, precision, and recall metrics into one single metric that ranges from 0 to 1, TNR—true negative rate, AUC—Area under the ROC (receiver operating characteristic) curve. Right panel: Proportions of correctly recognized phage sequences in CHGV-HQ and public virome databases by the viral-cognition machine learning models. Top: proportions of the phage genomes in public virome databases that were correctly recognized as phages by the CHGV-HQ model. Bottom: proportions of the CHGV-HQ phage genomes that were correctly recognized by ML models based on the public virome databases (Experimental Section). D) DNA methylation sites and their prevalence in the 8848 phages, stratified by methylation types, including N6-methyl-adenine (m6A or 6 mA) and N4-methyl-cytosine (m4C or 4mC). E) Methylation motifs identified in this study and their overlaps with those in the REBASE (http://rebase.neb.com/rebase/rebase.html).

gut phage database (GPD),[6b] metagenomic gut virus catalog (MGV),[6c] cenote human virome database (CHVD),[6d] and Danish enteric virome catalog (DEVoC).[6e] The CHGV-HQ model correctly recognized ≈88% (≈73–95%) of the viral genomes at a false-discovery rate of 5.72%, suggesting that the CHGV-HQ catalog captured the key sequence characteristics of human gut vi-

rome/phageome. In addition, virus-detection ML models trained on the true positives from the public virome datasets correctly recognized ≈96% (≈95–97%) of our CHGV-HQ genomes as viral genomes (Figure 1C), further confirming their viral identity and demonstrating that the sequence signatures were sufficient for further analysis.

2.2. The DNA Methylation Landscape of the Human Gut DNA Phageome

We analyzed the long-read sequencing data in the 104 samples (Figure 1A; Experimental Section) and identified DNA methylation sites using the SMRTlink tool, which had an ≈85% detection rate and ≈95% accuracy according to its manual.[15] For each sample, we aligned the consensus reads (HIFI reads) of the subreads generated by the PacBio CCS mode to the 8848 CHGV-HQ phage genomes and then used the SMRTlink tool to detect the methylated bases which compared the mean Inter Pulse Duration ratio (IPDr) of all the subreads at a position of the reference genome with that of the unmethylated bases.[15] We obtained a total of 8 448 009 non-redundant DNA methylation sites, among which 50.73% and 49.27% were m4C and m6A modifications respectively. In total, 8630 (97.53%) out of the 8848 phages were methylated, and 8549 (96.6%) contained both types of modifications (Figure 1D).

We identified a total of 157 DNA methylation motifs using an established method MEME tool[25] based on the large number of phages and identified methylation sites. Among these, 115 and 100 were m4C and m6A motifs, respectively, 58 motifs were shared by both methylation types, that is, the same motif sequence with different methylated bases (Table S1, Supporting Information; Experimental Section). Of these, 155 motifs were identical to the REBASE motifs[26] (Figure 1E; Experimental Section).

We then characterized the methylation patterns of the phage genomes, and identified genomic and evolutionary features that favored them. To better capture signatures of the DNA methylation in different genomes, we combined the methylation sites of all the CHGV-HQ genomes from all samples (Figure S1, Supporting Information), assuming that all potential methylation sites could be methylated in a genome. We observed a lower density of the m4C methylations in the coding regions (as measured by the proportion of cytosines methylated per genome) than the non-coding regions, but an opposite pattern for the m6A methylations (**Figure 2**A, Table S1, Supporting Information; $p < 0.001$, Wilcoxon Rank Sum Test; Experimental Section). We further dissected these patterns by assigning the phages to four lifestyle groups, namely temperate, uncertain temperate, uncertain virulent, and virulent using a DeePhage tool,[27] and observed an increasing trend in the m4C densities with the increasing phage virulence, that is, increased from temperate to virulent in both coding and non-coding regions, but an opposite trend in the m6A methylation densities (Figure 2A). Overall, we found significantly higher methylation density of m4C than m6A in both coding and non-coding regions. However, further analysis revealed that the differences in the coding regions were caused by lower proportions of m6A-methylated coding genes (CDSs with at least one modified A base, Figure 2B; $p < 0.001$, Wilcoxon rank sum test) rather than lower m6A densities in each of the CDSs (Figure 2C; $p > 0.05$, Wilcoxon rank sum test). Additionally, we also observed a correlation between the phage lifestyle with the proportion of methylated CDSs. In particular, the proportion of m4C methylated CDSs increases with increasing phage virulence, while the opposite pattern was found for m6A (Figure 2B). These results indicate different methylation pattern preference in gut phages between lifestyle groups. Previous studies have

shown that the phage lifestyle is a determinant factor shaping the phage defense strategies against the bacterial hosts.[28] Our results thus suggested that the two methylation types (i.e., the m4C and m6A methylations) may play different roles in phages of different styles.

We also observed differential methylation patterns in genes with various functions (Figure 2D). For example, tRNA genes, and most genes encoding proteins with known viral functions such as roles in lysis, immune evasion, integration, assembly, packaging, and infection, were methylated by both m4C and m6A in most genomes, in contrast to significantly lower proportions observed for genes with less characterized functions such as those coding for hypothetical and unsorted proteins (Figure 2D and Figure S2, Supporting Information). These results suggest that functionally important genes are more likely to be methylated.

We also examined whether the methylation densities could correlate with the viability of the gut phages. The latter could be evaluated from two aspects, the ability of a phage to accumulate in one sample (i.e., abundance), and the ability to spread and survive across samples (i.e., prevalence). Because the methylation density is significantly affected by the sequencing depth (e.g., the accumulative coverage of the CHGV-HQ genomes in all samples) and the abundances (i.e., the higher the abundance, the higher the coverage), we thus calculated a partial correlation between the phage prevalence and the methylation densities, while controlling for the sequencing depth. We determined the prevalence of the 8848 CHGV-HQ phages in the 104 samples using an arbitrary relative abundance cutoff of 0.5 calculated from the short-read VLP sequencing data (Experimental Section), and observed a strong positive correlation between the prevalence and methylation densities (Figure S3A, Supporting Information; $p < 0.001$, Partial Pearson correlation $r = 0.35$ after the sequencing depth was controlled) as well as the methylation motifs ($p < 0.001$; Figure S3B, Supporting Information). Because virulent phages are generally more abundant and prevalent than the temperate phages,[6a] we also stratified our analysis according to the phage lifestyle groups and found similar trends (Figure 2E, Partial Pearson correlation after the sequencing depth was controlled). The same trends were found for both the m4C and m6A methylation types (Figure S3C,D, Supporting Information). In addition, changing the presence/absence threshold to other arbitrary abundance cut-off in our prevalence calculation did not affect our main results (Figure S3E,F, Supporting Information). To further remove discovery biases, we also limited our analysis to phages with more than 100× coverages, and obtained similar results (Figure S4A,B, Supporting Information). Thus, our results suggested that higher methylation densities and/or numbers of methylation motifs might correspond to higher phage viability.

Together, our results show that the DNA methylation is universally present in gut phages and may play important roles in their survival in the human gut.

2.3. Phage-Encoded MTases Are Prevalent and Associated with Higher DNA Methylation Density and Phage Viability

To clarify the mechanisms underlying these frequent methylations, we examined whether gut phages encode their own
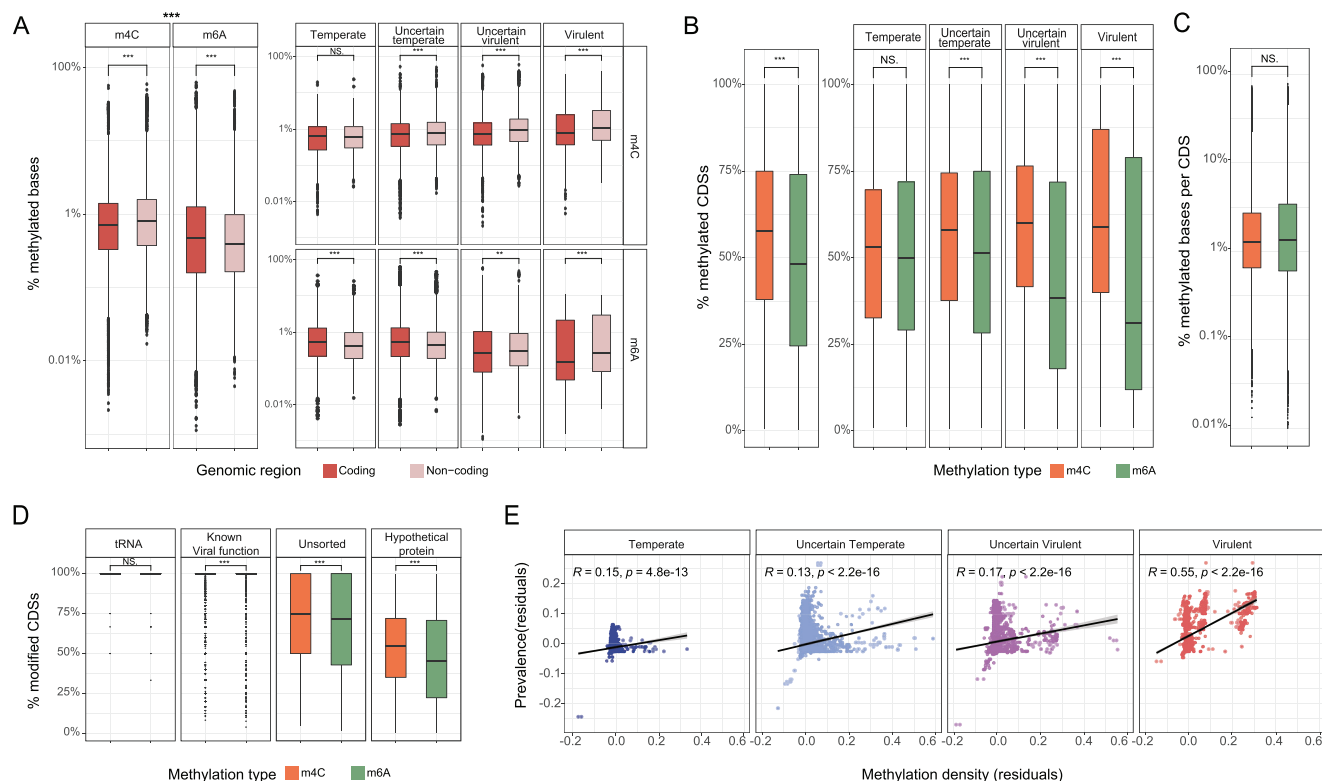
Figure 2. Distributions patterns of the DNA methylations in CHGV-HQ genomes as functions of phage lifestyles, gene functions, and phage prevalence. A) Methylation density with different distribution patterns of m6A and m4C in coding versus noncoding regions, and the impacts of phage lifestyles. The methylation density was calculated by the number of methylated bases divided by the total number of the related bases, for example, the number of mA bases/number of A bases. The overall trend of coding region being enriched in m4C methylation remained in each lifestyle group (right-top panel). Conversely, the m6A densities decreased with the increasing phage virulence (right-bottom panel); however, because the m6A densities decreased further in the coding than the noncoding regions in the two virulent phage groups (i.e., uncertain virulent and virulent), we found significantly higher m6A densities in the coding than the noncoding regions of the two virulent groups (right-bottom panel). We found significantly higher m4C densities in both coding and noncoding regions than that of m6A (left panel). B) Differential distribution patterns between m6A and m4C in coding sequences (CDS), and the impacts of phage lifestyles. The number methylated CDSs takes all the CDSs with at least one correlated methylation site into account. C) The ratio of modified bases/genome length per CDS between m6A and m4C. D) Differential distribution patterns of m6A and m4C modifications in coding genes with different functions; almost all tRNA genes and coding genes with known viral functions (Including Assembly, Immune evasion, Lysis, Integration, Replication, Regulation, Packaging, and Infection) are methylated, as compared with much lower methylation rates in phage genes coding for unsorted and hypothetical proteins. See Figure S2, Supporting Information for more detailed gene functional categories. E) The viability of the CHGV-HQ phages, that is, the prevalence across 104 fecal samples, was positively correlated (partial correlation using Pearson correlation, $p < 0.001$) with the overall DNA methylation density when the sequencing depth was controlled for (Experimental Section). Prevalence was calculated by using an abundance cutoff of 0.5 RPKM as the presence/absence threshold (Experimental Section); the results obtained when using other abundance cut-offs as the threshold were similar and can be found in Figure S3E,F, Supporting Information. Plotted here are the residuals of the prevalence (Y–axis) and methylation density (X–axis) after the sequence depth was controlled for (Experimental Section).

MTases. We searched all annotated proteins of the 8848 CHGV-HQ phage genomes against the conserved domain database (CDD)[29] using RPS-BLAST[30] (v2.12.0+) (Experimental Section), and identified a total of 4064 putative MTases with an E-value cut-off of 1E-5. Overall, phage-encoded MTases are prevalent, and could be found in 34.09% (3205, **Figure 3**A) of the CHGV-HQ phages (Table S1, Supporting Information), among which, 72% of the MTase-containing phages encoded only one MTase gene, whereas the others harbored multiple such genes (Figure S5, Supporting Information).

In addition to the RM systems, DNA MTases may exist without cognate REases, in which case they are referred to as the orphan MTases, and likely involved in processes other than anti-phage immunity such as gene expression regulation,[31] and DNA replication and repair.[32] We thus searched the phage-encoded MTases against those in the REBASE database that were already classified as orphan or non-orphan MTases. Out of the total 4064 MTases, 4030 had significant BLASTp hits with an E-value cut-off of 1E-5; among which, 93.25% (3758) had BLASTp hits with the non-orphan MTases, significantly higher than the overall 92.27% non-orphan MTases in the REBASE (207 303 out of 224 651; Figure 3C, $p < 0.001$, Hypergeometric test). Furthermore, the phage-encoded MTases with higher protein similarities (e.g., $\geq$ 90%) with those in the REBASE were further enriched with the non-orphan ones (97.71%, Figure 3C, $p < 0.001$, Hypergeometric test). Thus, most of the phages-encoded MTases are likely to be involved in the phage–bacterium defense processes.

We observed an increasing methylation density and number of motifs in the phages with more MTases genes (Figure 3D,E;
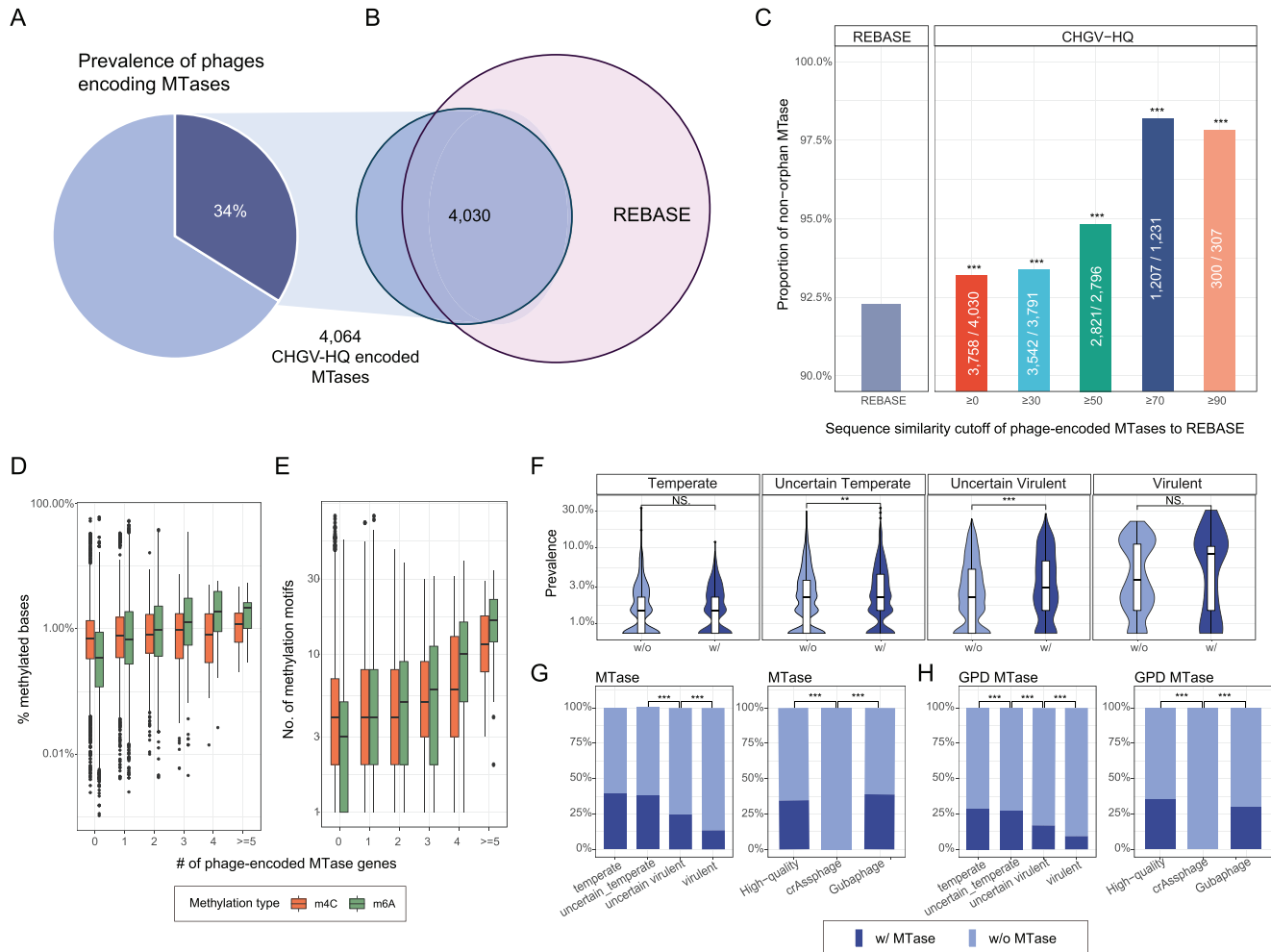
Figure 3. Phage-encoded MTases are prevalent in gut phages, and are associated with higher phage viability. A) Over one-third of the CHGV-HQ phages encode their own MTase genes. B) Most (4030 out of 4064, 98%) of the phage-encoded MTases share significant protein similarities with those in the REBASE (BLASTp e-value < 1e-5; Experimental Section). ≥0 stands for all phage-encoded MTases with BLASTp e-value < 1e-5. C) The phage-encoded MTases are mostly non-orphan MTases, and contain significantly higher proportion of non-orphan MTases than the REBASE. Chi-square tests were performed between REBASE and each of the CHGV-HQ groups with different protein similarity cutoffs; $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. D,E) Increasing numbers of phage-encoded MTase genes are associated with higher methylation densities and more methylation motifs in the phage genomes. F) Violin plots showing prevalence of the phages with/without MTases (X–axis), stratified by the phage lifestyles (i.e., from temperate, uncertain temperate, uncertain virulent, and virulent). Wilcoxon rank sum tests were performed between groups; $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. G) Bar plots on the left showing the prevalence of MTase genes in the CHGV-HQ phages as a function of phage virulence: a decreasing MTase prevalence is obvious with the increasing phage virulence (i.e., the highest in temperate phages, and lowest in virulent phages). Similar trends could be found in individual MTase types (Figure S6B,C, Supporting Information). Bar plots on the right showing the heterogeneous distributions of the MTases in phage subgroups such as the crAssphages and Gubaphages. H) The same trends could be found in GPD phages.

see also Figure S6A, Supporting Information for the trends for the individual modification types), suggesting that these phage-encoded MTases were indeed functional and contribute to the methylation of their encoding phages. In addition, we also observed that the phage-encoded MTases contributed to the significantly higher prevalence in two out of the four lifestyle groups (Figure 3F; the prevalence was measured by using an arbitrary abundance cut-off of an RPKM≥ 0.5 as the presence/absence threshold across the 104 samples; Experimental Section). These results suggest that the phage-encoded MTases may also contribute to the spreading of the corresponding phages across humans.

Phage lifestyle was also associated with the distribution of the MTase genes. For example, we found a significantly higher prevalence of MTase genes in temperate phages than in the virulent ones (Figure 3G,H). The trends in the individual MTase types, that is, MTases responsible for m4C and m6A modifications were mostly the same (Figure S6B,C, Supporting Information, line charts; Table S2, Supporting Information). The higher occurrence of MTases in temperate phages is likely due to their increased time spent within host cells and hence higher chances in exchanging genetic material horizontally.

However, heterogeneity in the lifestyle distributions was observed among different taxonomic groups. For example,
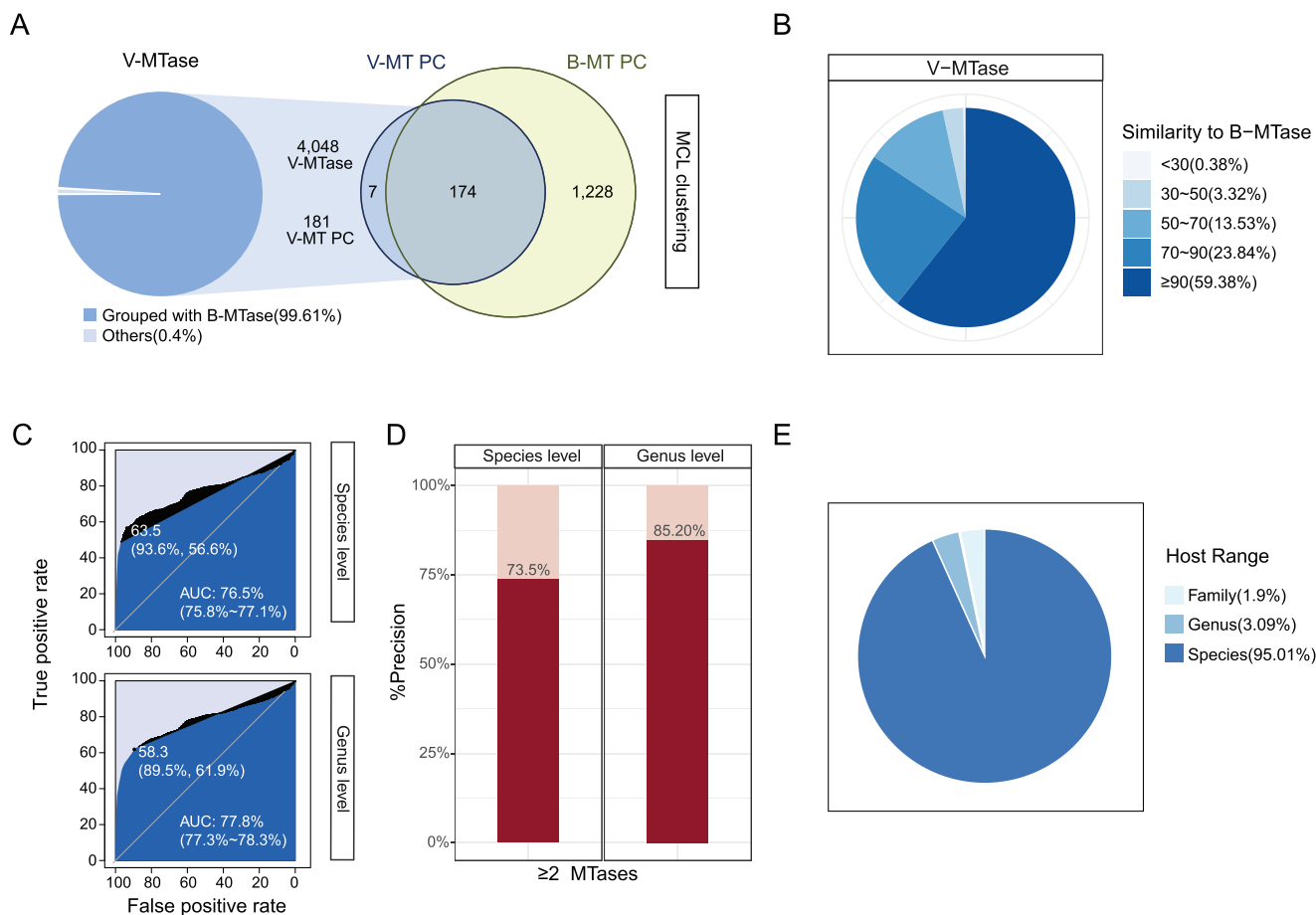
Figure 4. Phage-encoded MTases are homologous to bacteria-encoded ones and can be used to predict phage–host relationships with high precision. A) Most phage MTases (v-MTases) can be clustered into protein clusters (PCs) together with their bacterial homologs (b-MTase) using a Markov cluster algorithm (MCL). B) Distributions of protein sequence similarities of phage MTases. C) Phage–host prediction AUROC in the golden-standard dataset using at least one MTase at species- (upper) and genus- (lower) levels. D) Phage–host prediction precision in the golden-standard dataset using the two or more MTase at species- (left) and genus- (right) levels. E) Host ranges for phages with predicted hosts using MTase.

although both crAssphages and Gubaphages, the two most prevalent viral clades in human gut, are known to be virulent,[6b] none of the crAssphage genomes in our CHGV-HQ collection encode MTase genes as compared to ≈38.9% of the Gubaphages ($p < 0.001$, Chi-squared test). Similar trends were found among the high-quality gut phage genomes in the GPD;[6b] Figure 3H; Experimental Section), implying that crAssphages developed other means of achieving high prevalence besides avoiding the RM-systems.

In summary, our results show that over one-third of the CHGV-HQ phages encode their own MTase genes, which are associated with higher DNA methylation density and increased phage viability.

2.4. Phage MTases Are Closely Homologous to Bacterium-Encoded Ones and Can Be Used for Accurate Phage–Host Prediction

To further characterize the phage-encoded MTases, and to quantify previously observed similarity with bacterial genes,[5,33] we annotated MTases from the Unified Human Gastrointestinal Genome v2.0 (UHGG2),[23] a metagenome-assembled human gut microbial genome collection by applying the same pipelines used for CHGV-HQ (Experimental Section). We then combined the 61 719 annotated MTases with the 4064 phage-encoded ones identified in the CHGV-HQ genomes, performed all-against-all protein similarity searches using BLASTp,[30] and built protein clusters (PCs) using a Markov clustering algorithm (MCL).[34] We obtained a total of 1409 MTase PCs (including singletons), among which 7, 1228, and 174 clusters contained phage, bacterial, and both bacterial and phage MTase genes; we referred to the last group of PCs as gene-sharing PCs (GS-PCs). Importantly, 4048 (99.61%) of the total phage MTases (v-MTases) were included in the GS-PCs, (**Figure 4**A) among which, 59.38% of the v-MTases share over 90% protein sequence identities with their bacterium-encoded homologs (Figure 4B). We observed similar results between phage- and bacterial-encoded MTases using experimentally validated phage–host relationship data from the microbe-versus-phage (MVP)[3] database (Figure S7A, Supporting Information). The high similarity of the gut phage MTases with bacterium-encoded ones thus points to

not only a frequent gene exchange but also potential bacterial hosts.

To validate that the exchange of MTases indeed occurred between phages and their hosts, we examined whether such genes could be used to accurately distinguish phage–host relationships in the MVP database from randomly generated phage–bacterium pairs (Experimental Section). By using the BLAST results between phage and bacterial MTases as the only feature, we achieved overall accuracies of 76.5% and 77.8% AUC values at the species and genus levels, respectively (Figure 4C). Using at least one MTase, the phage–host relationships could be predicted at 67.20% and 82.08% precision at the species and genus levels, respectively. The prediction precision could be further increased to 73.5% and 85.2% (Figure 4D), by requiring that at least two MTase genes were shared between the phage–bacterium pairs with protein similarity higher than 90% (Experimental Section). At the latter criteria, we could predict a total of 4.77% phages to their hosts (Table S1, Supporting Information). We further validated the prediction results by calculating the host range for each of the phages as the last common ancestor (LCA) of all predicted bacterial hosts passing the above criteria, and found that 95.01% of the phages whose hosts had LCA at the species level (Figure 4E; Figure S7B, Supporting Information).

Together, these results support the assumption of the frequent exchange of MTases between phages and their hosts. Since the exchanges are bidirectional,[35] further work is needed to quantify directionality, as acquisition by phages increases their viability, the transfer of MTases to host bacteria could also confer changes through the increased DNA methylation potential.

3. Conclusion

In this study, we use the SMRT technology to characterize the first DNA methylation landscape of human gut phages at the community scale. Our finding that the gut phages may use DNA methylation to escape bacterial anti-viral immune mechanisms has several important implications. First, the similar analysis could be applied to the phage communities in other environments[13a] to better illustrate the role of methylation and phage-encoded MTases in those environments. Since the RM systems are universally present in bacterial genomes, survived phages appear to encode the respective counter-mechanisms, among them the DNA methylation. Second, the finding that differential methylation patterns exist in genes with various functions suggests that for some genes, by shielding themselves with DNA methylation, are able to keep functioning in the bacterial genome after the phage infection. Furthermore, the higher methylation densities found in more virulent phages indicate that DNA methylation is crucial for phages to escape the restriction endonuclease recognization from their hosts. It has been suggested that the phage-encoded MTases, which are mostly orphan ones, can be functional methyltransfereases and help the phages to overcome the bacterial R-M Systems and even contribute to the emergence of phages with broad host ranges.[36] The fact that these genes are different in their methylation pattern further indicates that they are of important functions, such as those with known viral functions including the infection, immune evasion, and assembly. Third, the phage-encoded MTase shared significant protein similarity (often >90%) to

bacterial-encoded ones, indicating that they were frequently exchanged between phages and bacteria and thus could be used to infer phage–host relationships. Sequence similarity-based approaches have been used to establish such relationships in many environments;[37] consistently, our analysis showed high precision in predicting phage–host relationships using MTases in an experimentally validated dataset (Figure 4D). Last, the viability benefits that we observed in phages with self-encoded MTases might represent only a small part of the functional consequences of the large numbers of genes frequently exchanged between phage and their bacterial host genomes. We found that there are ≈20% genes share high similarity with bacterial genes (Figure S8, Supporting Information), suggesting that there might be other genes that help the virus survive in the bacteria, and these genes could also be used to predict the host.

Our estimation on the extent of the phage DNA methylation and annotation of MTases may suffer from a few technical drawbacks. First, although the PacBio CCS mode allowed us to more reliably detect methylation signals from the subreads, that is, a genomic fragment from a single virion was sequenced multiple times and thus the consensus signals were more reliable than using reads from multiple virions of the same species/strains, heterogeneous DNA methylation patterns due to the high micro-diversity caused by high viral mutation rates and/or partial methylation within a viral population could lead to increased false negative calls. The distribution of methylation density further confirmed that very few methylation-positive genomes with low methylation density (Figure S9, Supporting Information), thus reduces the possibilities of the overestimate methylation prevalence due to the misalignment of the identified methylation fragment. In addition, phages infecting multiple bacterial species may have different methylation patterns from different hosts, although such cases could be rare because most phages have rather narrow host ranges at the species and even strain levels.[3] Sequence depth is critical factor in methylation identification, too. Our rarefaction analysis indicated the number of unique methylation density increased with the increasing sequence depth and the rarefaction curve is far from saturation (Figure S4A, Supporting Information). Thus, our estimation may present only part of the global picture of the full extent of the phage DNA methylation landscape in the human gut. There is still in need of a more reliable bioinformatics pipeline to identify methylation sites in genomes outside the eukaryotic and prokaryotic worlds. Second, our annotation of MTases was rather conservative. In this study, we used an RPS-BLAST based search that is known to less sensitive than hidden Markov (HMM)-based search methods.[38] However, the latter might suffer from high false positives.[39] Last, we identified too many sequence motifs in the gut phages. For example, although we obtained on average ≈4–5 motifs per phage, some phages could contain more than 30 motifs. This was in part because that some MTases are not site specific,[40] or the less stringent search criteria used by MEME.

Interestingly, we observed a significantly different distribution of the two modification types, namely m4C and m6A between the gut phageome and bacterial genomes. For example, the two methylation types accounted for 50.75% and 49.25% in our dataset, which was in sharp contrast to the bacteriome that the m6A is dominant and accounts for ≈75% of the modifications,[41] whereas m4C accounts for ≈20%.[42] The discrepancy could be

ADVANCED
SCIENCE NEWS

www.advancedsciencenews.com

ADVANCED
SCIENCE
Open Access

www.advancedscience.com

due to biological reasons such as that some bacteria are associated with more phages than others thus causing the overall methylation patterns in phages to be biased toward those bacteria, or technical reasons such as detection biases because of the lack of dedicated viral methylation identification tools. Of note, we did not analyze m5C modifications because of two reasons. First, currently the SMRT sequencing could not reliably detect them for prokaryotes and phages.[15] Second, m5C modifications are known to be rare in prokaryotic genomes.[42] Regardless, our knowledge on methylation patterns of the human gut bacteriome so far came from only a few studies with few fecal samples[41] and a comprehensive understanding is still lacking. In summary, we systematically characterized the DNA methylation landscape of human gut phages at the community scale using PacBio SMRT sequencing on 104 VLP-enriched samples. Our results suggest that the gut DNA phages universally use DNA methylation to escape from host defense systems with significant contribution from phage-encoded MTases. Together with recent studies exploring the identification of human phages at large scales [6a-e], our data and findings have only started to picture the survival tactics of the gut phages, the functional consequences of the frequent phage–bacterium gene exchanges to shape the gut microbial communities, and more daring, their implications in human health and diseases.[43]

4. Experimental Section

Generation of 8848 Near-Complete Double-Stranded DNA Phages from the Human Gut: The Chinese Human Gut Virome (CHGV) catalog was obtained from a previous study[16] that contains 21 646 non-redundant viral contigs generated by combined assembly of short- (Illumina) and long- (PacBio) reads. Briefly, a total of ≈500 g of feces from each of the 135 healthy volunteers of Chinese residence were collected and processed by a virome enrichment protocol to obtain a large number of VLPs. High-quality and high-molecular weight double-stranded DNAs were then extracted, and submitted to Illumina HiSeq2000 sequencer (Novogen, Beijing, China) for viral next generation sequencing (vNGS, short-reads). A subset of 104 samples with sufficient quantity of viral DNAs was also submitted to the PacBio RS II sequencer (Pacific Biosciences, Menlo Park, CA, USA) with CCS mode for viral third-generation sequencing (vTGS, long-reads). Human genome contaminations were identified and removed from both the vNGS and vTGS datasets, followed by a combined assembly pipeline to generate putative phage contigs (see ref. [16] for more details).

To identify viral genomes, the following tools were used: VirSorter v2.0[17] (–min-score 0.7), VirFinder v1.1[18] (default parameters), and PPR-Meta v1.1[19] (default parameters). A BLAST search against the Viral RefSeq genomes was also performed using BLASTn v.2.7.1[21] with the default parameters and an E-value cutoff of <1e-10; Release 201 (Jul 06, 2020) of the Viral RefSeq database contained 13 148 viral genomes. In addition, the annotated protein sequences were used for BLAST searches against the NCBI POG (Phage Orthologous Groups) database.[20]

A contig was annotated as a virus if it was circular/met at least two of the following criteria 1–5; the same criteria have been adopted by the GVD.[6a]

1. VirSorter score ≥ 0.7,
2. VirFinder score > 0.6,
3. PPR-Meta phage score > 0.7,
4. Hits to Viral RefSeq with > 50% identity & > 90% coverage,
5. Minimum of three ORFs, producing BLAST hits to the NCBI POG database 2013 with an E-value of ≤ 1e-5, with at least two per 10 kb of contig length.

Alternatively, contigs met one of the above criteria and were annotated as high-quality (≥ 90% completeness) by CheckV[44] and were also annotated as viruses.

To avoid bacterial contamination, possible prophage regions were first identified using PhageFinder[45] (v2.1) and they were removed from UHGG genomes to prevent over-estimation of the contamination. The resulting UHGG dataset was referred to as UHGG-Minus in this study. A BLAST search was then carried out against the UHGG-Minus sequences using BLASTn v.2.7.1[21] with the default parameters and an E-value cutoff of <1e-10, and contigs with blastn hit of 90% identity over 50% of its length were removed from further analysis. To avoid fragment genomes, the contigs were filtered with length longer than 5 kb or circular contigs longer than 1.5 kb.

In the end, a non-redundant set of 21 646 viral contigs was obtained and referred to as the Chinese Human Gut Virome (CHGV) catalog.

The 8848 near-complete double-stranded DNA phages were then screened from the CHGV catalog[16] using CheckV[44] based on a selection criterium of >90% completeness. In total, 5956 complete (with 100% completeness) and 2892 high-quality (with >90% completeness) phage genomes were selected for subsequent analyses.

Identification and Characterization of DNA Methylation Profiles of the 8848 Phages across 104 Fecal Samples: The Base Modification Analysis Application module of the SMRT Link tool (v10.1) from https://www.pacb.com/support/software-downloads was utilized on January 15, 2022, to identify methylated sites in the genomes of 8848 gut phages across 104 fecal samples using SMRT sequencing data.[15] For each sample, the consensus reads (HIFI reads) generated by the PacBio CCS mode were aligned to the 8848 CHGV-HQ phage genomes and then the SMRTlink tool was used to detect the methylated bases which compared the mean Inter Pulse Duration ratio (IPDr) of all the subreads belonging to the aligned HIFI reads at a position of the reference genome with that of the unmethylated bases.[15]

MotifMaker (SMRTLink, https://www.pacb.com/support/software-downloads/) was used to identify the methylated motifs, but did not produce any results. HOMER[46] (v4.11), a popular motif prediction tool for analyzing ChIP-seq data, was also used to identify motifs; although it identified a lot of motifs from the data, HOMER did not report the exact locations on the viral genome. It was speculated that these tools were designed for analyzing prokaryotic/eukaryotic genomes that were significantly longer than viruses, and thus might not be suitable for viral genome analysis. Methylation motifs were thus identified using MEME-ChIP[25b] (v5.4.1; default parameters). Motifs were then dereplicated with a customized R pipeline (https://github.com/whchenlab/gutphagemethylome). To further remove false positives, it was also required that the motifs should be found in at least 100 phage genomes. In total, 157 non-redundant motifs were identified (Table S1, Supporting Information). A customized R pipeline was used to compare the motifs with those in the REBASE.[26] A rather relaxed criteria was used to search for the overlap; for example, motifs that were one of the possibilities of the motifs in public database were considered overlapped, for example, AKCTCG was considered to be overlapped by BCNC (the former was one of the possibilities of the latter). In total, 155 were the same as compared with those in the REBASE (Table S1, Supporting Information).

Annotation of MTases in Phage Genomes as well as Bacterial/Archaeal Reference Genomes (BRGs): Prodigal[47] v2.6.3 with default parameters was used to predict genes from the 8848 gut phage genomes. The predicted protein sequences were searched against the CDD[29] using RPS-BLAST (v2.12.0+; part of the BLAST package;[21] e-value 1E-5). MTase proteins were identified with the following domain: cd21179.smp, COG0350.smp, KOG3191.smp, pfam05869.smp, pfam12047.smp, PRK10904.smp, PRK11524.smp, TIGR00589.smp, TIGR00675.smp, TIGR01712.smp and TIGR02987.smp. The MTases were further annotated by blast against the REBASE[26] with BLASTp (e-value < 1e-5).

The same methods were used to identify MTase proteins in bacterial/archaeal reference genomes in the Unified Human Gastrointestinal Genome collection v2.0 (UHGG2)[23] and public phage genomes in the GPD.[6b]

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
SCIENCE**
Open Access

www.advancedscience.com

The search for MTases might be limited by the CDD search method, for that it might underestimate/misclassify some of the MTase hits.

Selection of High-Quality Phage Genomes from GPD, the GPD: For comparison purposes, 40 140 high-quality phage genomes (>90% completeness) were also selected from the GPD[6b] using CheckV.[44] Among these, 17 743 (44.2%) were of 100% completeness; this subset of GPD genomes was referred as to GPD-HQ in this study.

Machine Learning Models for Prediction of Human Gut Virome and Performance Evaluation: To check the quality and representativeness of CHGV-HQ and compare it with public human (gut) virome databases including GVD,[6a] GPD,[6b] MGV,[6c] CHVD,[6d] and DEVoC,[6e] a series of neural network models were trained with the same architecture as DeepVirFinder.[48]

For CHGV-HQ, GVD, and DEVoC, all sequences were kept as the true positive datasets, while for GPD and MGV, the longest sequence of each viral cluster (considered as genus level of viruses) was selected as representative genomes considering memory consumption and training time, and referred to as GPD-rep and MGV-rep, respectively. As for CHVD, the intestine-origin genomes were extracted. All the above-mentioned datasets were kept as positive datasets separately.

The bacterial genomes over 1.5 kbp from the human intestine were collected as negative samples from UHGG-Minus (Unified Human Gastrointestinal Genomes v2[23] without prophage sequence; prophage identified with Phage_Finder[45] with default parameter).

80% of each collection was randomly selected as the training set, and the other 20% as the test set. The DNA sequences in the training set were consecutively segmented into non-overlapping fragments (1 kbs), then encoded into numerical matrices with a one-hot encoding method. The testing dataset was also segmented into non-overlapping fragments, and for each sequence, the average score of its fragments was assigned. For that the bacterial fragments significantly out-numbered the viral ones; the negative bacterial training set was randomly down-sampled to match the positive viral one for every model.

The representativeness of CHGV-HQ was assessed with two experiments. The performance of the models built using the same method can be regarded as an indicator of the quality and representativeness of each dataset.

In the first experiment, a test on an independent dataset, IMG/VR database[24] (https://img.jgi.doe.gov/cgi-bin/vr/main.cgi) and bacterial sequences from the test set of UHGG-Minus, were carried out. To make sure that the CHGV-HQ training set was independent with this IMG/VR test set, the pairwise comparison was conducted, by blasting (BLASTn,[30] v2.2.26+; *E*-value < 1e-5) IMG/VR sequences against CHGV-HQ. The IMG/VR genome was kept only if it was with 90% nucleotide identity on less than 70% (calculated with bedtools v2.29.1[49]) of its genome. The same procedure was carried out for the bacterial test dataset.

Metrics such as accuracy, precision, true negative rate, recall, f1-score, and AUC score were then calculated to evaluate the CHGV-HQ model.

In the second experiment, the CHGV-HQ model was applied to the above human gut virome datasets to see if most of the latter sequences could be correctly recognized.

In addition, viral-recognition models trained on the public datasets were also applied to the CHGV-HQ genomes, to further validate the latter's viral identity.

Estimation of the Prevalence of the CHGV Genomes at the Viral Contig: To estimate the abundance of phage genomes, the VLP-Next generation sequencing clean reads were mapped to the CHGV-HQ genomes using Bowtie2. Then, the reads per kilobase million (RPKM) value of each phage genome was calculated. An arbitrary abundance cutoff of 0.5 was used to define the presence of a phage in each of the 104 samples. Changing the presence/absence threshold up to 5 RPKM did not affect the main results.

Partial correlation calculation between the phage prevalence and methylation densities with the sequencing depth control for partial correlation was a measure of the strength and direction of a linear relationship between two continuous variables whilst controlling for the effect of one or more other continuous variables. The partial correlations between phage prevalence and methylation densities, whilst controlling for sequencing depth were calculated using R (v 4.0.5) with "ppcor" Rpackage.[50] To plot

the results after the sequence depth was controlled for, the residuals of the prevalence and methylation densities were calculated using the "resid" function implemented in R with commands like 'resid(lm(prevalence ≈ sequencing_depth))" and "resid(lm(meth_density ≈ sequencing_depth))."

Annotation of crAssphages and Gubaphages: crAss-like phages were identified using the same method reported in a previous study (ref. [51]) First, phage nucleotide sequences were compared against the protein sequence of the polymerase (UGP_018) and the terminase (UGP_092) of the prototypical crAssphage (p-crAssphage, NC_024711.1) using BLASTx[21] (v2.2.26+). Second, the nucleotide sequence similarities between the phages and the p-crAssphage genome were assessed using BLASTn.[21] A phage was then classified as a crAssphage when it was longer than 70 kb and met at least one of the following criteria:

1. had a BLASTx hit with an *E*-value <1e-10 against either p-crAssphage polymerase or terminase
2. showed ≥95% nucleotide identity over 80% of the contig length with the p-crAssphage genome

Gubaphages were identified by searching for the large terminase gene of the Gubaphage genomes obtained from the GPD[6b] using BLASTp (v2.2.26+, *E*-value < 1e-5). Gubaphage was then classified into four genera (G1.1, G1.2, G1.3, and G2) according to the similarity of the large terminase genes to those identified in the GPD.

In total, 245 crAssphage and 180 Gubaphage genomes were identified in the dataset.

For comparison, 1490 Gubaphages and 2057 crAssphages were extracted according to the genome annotations in GPD-HQ.

Annotation of Phage Lifestyles: Phage lifestyles were predicted using DeePhage[27] v1.0 with default parameters. DeePhage uses a scoring system to classify phage genomes into four categories, including temperate (with scores < = 0.3), uncertain temperate (≈0.3–0.5), uncertain virulent (≈0.5–0.7), and virulent (>0.7). Higher scores indicate higher virulence.[27] Among the 8848 phages, 1211 (13.68%), 5272 (59.58%), 1647 (18.61%), and 718 (8.11%) were classified as temperate, uncertain temperate, uncertain virulent, and virulent respectively (Table S1, Supporting Information).

Clustering Analysis of the Phage- and Bacteria-Encoded MTase Proteins: Protein sequences of MTases from phages, bacterial and archaeal genomes were merged and a BLASTp (v2.2.26+; *E*-value 1e-5) algorithm was used to search the merged dataset against itself for homologous sequences. The query-hit pairs were further filtered with a coverage >75% on the query proteins. The filtered BLASTp results were used as input for a Markov Clustering Algorithm[34] (MCL v14-137) with default parameters to generate protein clusters (PCs). The clustering analysis was performed separately for MTase proteins.

Phage Host Prediction Using MTases: To check if MTases can be used to correctly predict phage hosts, and determine the best parameters for the predictions, a published phage–host dataset was obtained from the MVP database[3] and used as gold standard. Of note, only experimentally validated phage–host relationships and those inferred from prophages were retained for further analysis, resulting in 778 relationships between 422 phages and 1517 prokaryotic hosts (Table S1, Supporting Information). MTase proteins were identified for the phages and prokaryotic hosts, as described in the previous section. BLASTp (v2.2.26+)[21] was used to detect homologous relationships between phage- and bacteria-encoded MTases. A sequence similarity score (BLASTp q-identity × query coverage, Sim-Score) was calculated and AUC scores were used to evaluate the accuracy of the SimScores to distinguish true interactions from randomly generated non-interacting phage–host pairs. Such an analysis can help in determining the best SimScore cutoffs by taking specificity, sensitivity, or both into consideration.

To predict hosts for the phages, phage-encoded MTase protein sequences were searched against those of the UHGG2 using BLASTp (v2.2.26+). A bacterial/archaeal genome whose BLASTp SimScore to a phage MTase protein was higher than the threshold was assigned as the phage host. To avoid wrong host assignments because of the lack of bacterial reference genomes, a rather stringent threshold was adopted to

assign a bacterial host to a phage if i) at least two phage-encoded MTase proteins had SimScore higher than 90% with the host-encoded counterparts, and ii) at least two such sequence-matches agreed on the same host. After filtration, the annotation for each protein with the highest BLASTp SimScore was kept. At these criteria, the precision of 73.5% and 85.2% were achieved at the species and genus levels respectively on the MVP dataset (Figure 4D). The predicted phage–host relationships can be found in Table S1, Supporting Information.

Host Range Calculation of Host-Prediction Results: Host prediction results were evaluated by calculating the host range for each phage. For phages with only one predicted host, the host range was at the species level. For those with multiple predicted hosts, the host range was calculated as the last common ancestor (LCA) of all the hosts on the NCBI taxonomic database using an in-house R script (https://github.com/whchenlab/gutphagemethylome.).

Statistics and Other Bioinformatics Analyses: All processed data, if not otherwise stated, were loaded into R (v4.0.5, https://www.r-project.org/), analyzed, or visualized.

Ethics Approval and Consent to Participate: This study was approved by the Ethics Committee of the Tongji Medical College of Huazhong University of Science and Technology (no., S1241) and the Human Ethics Committee of the School of Life Sciences of Fudan University (No, BE1940).

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Z.L., P.B., X.-M.Z., and W.-H.C. designed and directed the research. C.S. did most of the analysis. J.C. managed the sampling and did some of the experiments. M.J., X.Z., and Y.L. helped with the experiments. M.J. also helped with the sample collection and phage enrichment experiments. Y.D. did the machine learning analysis. C.S. wrote the paper with results from all authors. Z.L., P.B., X.-M.Z., and W.-H.C. polished the manuscript through multiple iterations of discussions with all authors. All authors have read and approved the final manuscript.

Data Availability Statement

The data that support the findings of this study are openly available in GSA at https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA008836, reference number PRJCA008836.

Keywords

[1] G. P. Salmond, P. C. Fineran, *Nat. Rev. Microbiol.* **2015**, *13*, 777.

[2] A. N. Shkoporov, A. G. Clooney, T. D. S. Sutton, F. J. Ryan, K. M. Daly, J. A. Nolan, S. A. McDonnell, E. V. Khokhlova, L. A. Draper, A. Forde, E. Guerin, V. Velayudhan, R. P. Ross, C. Hill, *Cell Host Microbe* **2019**, *26*, 527.

[3] N. L. Gao, C. Zhang, Z. Zhang, S. Hu, M. J. Lercher, X.-M. Zhao, P. Bork, Z. Liu, W.-H. Chen, *Nucleic Acids Res.* **2018**, *46*, D700.

[4] C. D. Johnston, S. L. Cotton, S. R. Rittling, J. R. Starr, G. G. Borisy, F. E. Dewhirst, K. P. Lemon, *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 11454.

[5] a) J. E. Samson, A. H. Magadan, M. Sabri, S. Moineau, *Nat. Rev. Microbiol.* **2013**, *11*, 675; b) J. Murphy, J. Mahony, S. Ainsworth, A. Nauta, D. v. Sinderen, *Appl. Environ. Microbiol.* **2013**, *79*, 7547.

[6] a) A. C. Gregory, O. Zablocki, A. A. Zayed, A. Howell, B. Bolduc, M. B. Sullivan, *Cell Host Microbe* **2020**, *28*, 724; b) L. F. Camarillo-Guerrero, A. Almeida, G. Rangel-Pineros, R. D. Finn, T. D. Lawley, *Cell* **2021**, *184*, 1098; c) S. Nayfach, D. Paez-Espino, L. Call, S. J. Low, H. Sberro, N. N. Ivanova, A. D. Proal, M. A. Fischbach, A. S. Bhatt, P. Hugenholtz, N. C. Kyrpides, *Nat. Microbiol.* **2021**, *6*, 960; d) M. J. Tisza, C. B. Buck, *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, e2023202118; e) L. V. Espen, E. G. Bak, L. Beller, L. Close, W. Deboutte, H. B. Juel, T. Nielsen, D. Sinar, L. D. Coninck, C. Frithioff-Bøjsøe, C. E. Fonvig, S. Jacobsen, M. Kjærgaard, M. Thiele, A. Fullam, M. Kuhn, J.-C. Holm, P. Bork, A. Krag, T. Hansen, M. Arumugam, J. Matthijnssens, *mSystems* **2021**, *6*, e00382; f) S. Lai, L. Jia, B. Subramanian, S. Pan, J. Zhang, Y. Dong, W. H. Chen, X. M. Zhao, *Nucleic Acids Res.* **2021**, *49*, D783.

[7] a) S. Roux, S. J. Hallam, T. Woyke, M. B. Sullivan, *Elife* **2015**, *4*, e08490; b) S. Martinez Arbas, S. Narayanasamy, M. Herold, L. A. Lebrun, M. R. Hoopmann, S. Li, T. J. Lam, B. J. Kunath, N. D. Hicks, C. M. Liu, L. B. Price, C. C. Laczny, J. D. Gillece, J. M. Schupp, P. S. Keim, R. L. Moritz, K. Faust, H. Tang, Y. Ye, A. Skupin, P. May, E. E. L. Muller, P. Wilmes, *Nat. Microbiol.* **2021**, *6*, 123.

[8] a) T. Borodovich, A. N. Shkoporov, R. P. Ross, C. Hill, *Gastroenterol. Rep.* **2022**, *10*, goac012; b) S. Khedkar, G. Smyshlyaev, I. Letunic, O. M. Maistrenko, L. P. Coelho, A. Orakov, S. K. Forslund, F. Hildebrand, M. Luetge, T. S. B. Schmidt, O. Barabas, P. Bork, *Nucleic Acids Res.* **2022**, *50*, 3155.

[9] C. Howard-Varona, M. M. Lindback, G. E. Bastien, N. Solonenko, A. A. Zayed, H. Jang, B. Andreopoulos, H. M. Brewer, T. Glavina Del Rio, J. N. Adkins, S. Paul, M. B. Sullivan, M. B. Duhaime, *ISME J.* **2020**, *14*, 881.

[10] a) J. S. Hao, J. F. Xing, X. Hu, Z. Y. Wang, M. Q. Tang, L. Liao, *Front. Plant Sci.* **2022**, *13*, 922152; b) X. Li, Z. Zhang, X. Luo, J. Schrier, A. D. Yang, T. P. Wu, *Protein Cell* **2021**, *12*, 756.

[11] a) C. Tellgren-Roth, M. Couturier, *Methods Mol. Biol.* **2022**, *2516*, 39; b) C. P. Davies, T. Jurkiw, J. Haendiges, E. Reed, N. Anderson, E. Grasso-Kelley, M. Hoffmann, J. Zheng, *Front. Microbiol.* **2022**, *13*, 970135; c) S. Han, J. Liu, M. Li, Y. Zhang, X. Duan, Y. Zhang, H. Chen, Z. Cai, L. Yang, Y. Liu, *mSystems* **2022**, *7*, e0043422.

[12] S. Jeudy, S. Rigou, J. M. Alempic, J. M. Claverie, C. Abergel, M. Legendre, *Nat. Commun.* **2020**, *11*, 2657.

[13] a) H. J. Seong, S. Roux, C. Y. Hwang, W. J. Sul, *Microbiome* **2022**, *10*, 157; b) S. Hiraoka, T. Sumida, M. Hirai, A. Toyoda, S. Kawagucci, T.

Yokokawa, T. Nunoura, *Nucleic Acids Res.* **2022**, *50*, 1531; c) J. Cao, Y. Zhang, M. Dai, J. Xu, L. Chen, F. Zhang, N. Zhao, J. Wang, *Med. Microecol.* **2020**, *4*, 100012.

[14] a) J. T. Simpson, R. E. Workman, P. C. Zuzarte, M. David, L. J. Dursi, W. Timp, *Nat. Methods* **2017**, *14*, 407; b) A. C. Rand, M. Jain, J. M. Eizenga, A. Musselman-Brown, H. E. Olsen, M. Akeson, B. Paten, *Nat. Methods* **2017**, *14*, 411.

[15] B. A. Flusberg, D. R. Webster, J. H. Lee, K. J. Travers, E. C. Olivares, T. A. Clark, J. Korlach, S. W. Turner, *Nat. Methods* **2010**, *7*, 461.

[16] W. Chen, J. Chen, C. Sun, Y. Dong, M. Jin, S. Lai, L. Jia, X. Zhao, N. L. Gao, Z. Liu, P. Bork, X.-M. Zhao, *bioRxiv 2022.07.03.498593* **2022**, https://doi.org/10.1101/2022.07.03.498593.

[17] S. Roux, F. Enault, B. L. Hurwitz, M. B. Sullivan, *PeerJ* **2015**, *3*, e985.

[18] J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, F. Z. Sun, *Microbiome* **2017**, *5*, 69.

[19] Z. C. Fang, J. Tan, S. F. Wu, M. Li, C. M. Xu, Z. J. Xie, H. Q. Zhu, *GigaScience* **2019**, *8*, giz066.

[20] D. M. Kristensen, A. S. Waller, T. Yamada, P. Bork, A. R. Mushegian, E. V. Koonin, *J. Bacteriol.* **2013**, *195*, 941.

[21] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **1990**, *215*, 403.

[22] S. Roux, D. Paez-Espino, I. A. Chen, K. Palaniappan, A. Ratner, K. Chu, T. B. K. Reddy, S. Nayfach, F. Schulz, L. Call, R. Y. Neches, T. Woyke, N. N. Ivanova, E. A. Eloe-Fadrosh, N. C. Kyrpides, *Nucleic Acids Res.* **2021**, *49*, D764.

[23] A. Almeida, S. Nayfach, M. Boland, F. Strozzi, M. Beracochea, Z. J. Shi, K. S. Pollard, E. Sakharova, D. H. Parks, P. Hugenholtz, N. Segata, N. C. Kyrpides, R. D. Finn, *Nat. Biotechnol.* **2021**, *39*, 105.

[24] S. Roux, D. Páez-Espino, I. M. A. Chen, K. Palaniappan, A. Ratner, K. Chu, T. B. K. Reddy, S. Nayfach, F. Schulz, L. Call, R. Y. Neches, T. Woyke, N. N. Ivanova, E. A. Eloe-Fadrosh, N. C. Kyrpides, *Nucleic Acids Res.* **2021**, *49*, D764.

[25] a) A. Tourancheau, E. A. Mead, X. S. Zhang, G. Fang, *Nat. Methods* **2021**, *18*, 491; b) P. Machanick, T. L. Bailey, *Bioinformatics* **2011**, *27*, 1696.

[26] R. J. Roberts, T. Vincze, J. Posfai, D. Macelis, *Nucleic Acids Res.* **2015**, *43*, D298.

[27] S. Wu, Z. Fang, J. Tan, M. Li, C. Wang, Q. Guo, C. Xu, X. Jiang, H. Zhu, *Gigascience* **2021**, *10*, giab056.

[28] a) T. N. Mavrich, G. F. Hatfull, *Nat. Microbiol.* **2017**, *2*, 17112; b) J. A. Moura de Sousa, E. Pfeifer, M. Touchon, E. P. C. Rocha, *Mol. Biol. Evol.* **2021**, *38*, 2497.

[29] a) S. Lu, J. Wang, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, M. Yang, D. Zhang, C. Zheng, C. J. Lanczycki, A. Marchler-Bauer, *Nucleic Acids Res.* **2020**, *48*, D265; b) A. Marchler-Bauer, S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng, S. H. Bryant, *Nucleic Acids Res.* **2011**, *39*, D225.

[30] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, *BMC Bioinf.* **2009**, *10*, 421.

[31] S. Adhikari, P. D. Curtis, *FEMS Microbiol. Rev.* **2016**, *40*, 575.

[32] a) W. Messer, U. Bellekes, H. Lother, *EMBO J.* **1985**, *4*, 1327; b) C. D. Putnam, *DNA Repair* **2016**, *38*, 32.

[33] A. B. Isaev, O. S. Musharova, K. V. Severinov, *Biochemistry* **2021**, *86*, 319.

[34] A. J. Enright, S. Van Dongen, C. A. Ouzounis, *Nucleic Acids Res.* **2002**, *30*, 1575.

[35] a) C. L. Schneider, in *Bacteriophages: Biology, Technology, Therapy*, (Eds: D. Harper, S. Abedon, B. Burrowes, M. McConville), Springer, Cham, **2017**; b) C. Canchaya, G. Fournous, S. Chibani-Chennoufi, M.-L. Dillmann, H. Brüssow, *Curr. Opin. Microbiol.* **2003**, *6*, 417.

[36] J. Murphy, J. Mahony, S. Ainsworth, A. Nauta, D. van Sinderen, *Appl. Environ. Microbiol.* **2013**, *79*, 7547.

[37] a) S. Nayfach, S. Roux, R. Seshadri, D. Udwary, N. Varghese, F. Schulz, D. Wu, D. Paez-Espino, I. M. Chen, M. Huntemann, K. Palaniappan, J. Ladau, S. Mukherjee, T. B. K. Reddy, T. Nielsen, E. Kirton, J. P. Faria, J. N. Edirisinghe, C. S. Henry, S. P. Jungbluth, D. Chivian, P. Dehal, E. M. Wood-Charlson, A. P. Arkin, S. G. Tringe, A. Visel, H. Abreu, S. G. Acinas, E. Allen, M. A. Allen, et al., *Nat. Biotechnol.* **2021**, *39*, 499; b) A. P. Camargo, S. Nayfach, I. M. A. Chen, K. Palaniappan, A. Ratner, K. Chu, S. J. Ritter, T. B. K. Reddy, S. Mukherjee, F. Schulz, L. Call, R. Y. Neches, T. Woyke, N. N. Ivanova, E. A. Eloe-Fadrosh, N. C. Kyrpides, S. Roux, *Nucleic Acids Res.* **2023**, *51*, D733.

[38] J. Söding, *Bioinformatics* **2005**, *21*, 951.

[39] W. C. Wong, C. K. Yap, B. Eisenhaber, F. Eisenhaber, *Biol. Direct* **2015**, *10*, 39.

[40] E. Murray Noreen, *Microbiol. Mol. Biol. Rev.* **2000**, *64*, 412.

[41] a) J. Beaulaurier, S. Zhu, G. Deikus, I. Mogno, X.-S. Zhang, A. Davis-Richardson, R. Canepa, E. W. Triplett, J. J. Faith, R. Sebra, E. E. Schadt, G. Fang, *Nat. Biotechnol.* **2017**, *36*, 61; b) J. Beaulaurier, E. E. Schadt, G. Fang, *Nat. Rev. Genet.* **2019**, *20*, 157.

[42] M. J. Blow, T. A. Clark, C. G. Daum, A. M. Deutschbauer, A. Fomenkov, R. Fries, J. Froula, D. D. Kang, R. R. Malmstrom, R. D. Morgan, J. Posfai, K. Singh, A. Visel, K. Wetmore, Z. Zhao, E. M. Rubin, J. Korlach, L. A. Pennacchio, R. J. Roberts, *PLoS Genet.* **2016**, *12*, e1005854.

[43] J. Li, F. Yang, M. Xiao, A. Li, *Cell Host Microbe* **2022**, *30*, 908.

[44] S. Nayfach, A. P. Camargo, F. Schulz, E. Eloe-Fadrosh, S. Roux, N. C. Kyrpides, *Nat. Biotechnol.* **2021**, *39*, 578.

[45] D. E. Fouts, *Nucleic Acids Res.* **2006**, *34*, 5839.

[46] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, C. K. Glass, *Mol. Cell* **2010**, *38*, 576.

[47] D. Hyatt, G. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, L. J. Hauser, *BMC Bioinf.* **2010**, *11*, 119.

[48] J. Ren, K. Song, C. Deng, N. A. Ahlgren, J. A. Fuhrman, Y. Li, X. Xie, R. Poplin, F. Sun, *Quant. Biol.* **2020**, *8*, 64.

[49] A. R. Quinlan, I. M. Hall, *Bioinformatics* **2010**, *26*, 841.

[50] S. Kim, *Commun. Stat. Appl. Methods* **2015**, *22*, 665.

[51] K. Fujimoto, Y. Kimura, M. Shimohigoshi, T. Satoh, S. Sato, G. Tremmel, M. Uematsu, Y. Kawaguchi, Y. Usui, Y. Nakano, T. Hayashi, K. Kashima, Y. Yuki, K. Yamaguchi, Y. Furukawa, M. Kakuta, Y. Akiyama, R. Yamaguchi, S. E. Crowe, P. B. Ernst, S. Miyano, H. Kiyono, S. Imoto, S. Uematsu, *Cell Host Microbe* **2020**, *28*, 380.