Published online 19 October 2020 *Nucleic Acids Research, 2021, Vol. 49, Database issue* ***D783–D791***
doi: 10.1093/nar/gkaa869

mMGE: a database for human metagenomic extrachromosomal mobile genetic elements

Senying Lai[1,†], Longhao Jia[1,†], Balakrishnan Subramanian[2,†], Shaojun Pan[1], Jinglong Zhang[1], Yanqi Dong[1], Wei-Hua Chen[2,3,*] and Xing-Ming Zhao[1,4,5,*]

[1]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China, [2]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, China, [3]College of Life Science, Henan Normal University, Xinxiang, Henan 453007, China, [4]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Ministry of Education, Shanghai 200433, China and [5]Research Institute of Intelligent Complex System, Fudan University, Shanghai 200433, China

Received August 14, 2020; Revised September 18, 2020; Editorial Decision September 22, 2020; Accepted September 24, 2020

ABSTRACT

Extrachromosomal mobile genetic elements (eMGEs), including phages and plasmids, that can move across different microbes, play important roles in genome evolution and shaping the structure of microbial communities. However, we still know very little about eMGEs, especially their abundances, distributions and putative functions in microbiomes. Thus, a comprehensive description of eMGEs is of great utility. Here we present mMGE, a comprehensive catalog of 517 251 non-redundant eMGEs, including 92 492 plasmids and 424 759 phages, derived from diverse body sites of 66 425 human metagenomic samples. About half the eMGEs could be further grouped into 70 074 clusters using relaxed criteria (referred as to eMGE clusters below). We provide extensive annotations of the identified eMGEs including sequence characteristics, taxonomy affiliation, gene contents and their prokaryotic hosts. We also calculate the prevalence, both within and across samples for each eMGE and eMGE cluster, enabling users to see putative associations of eMGEs with human phenotypes or their distribution preferences. All eMGE records can be browsed or queried in multiple ways, such as eMGE clusters, metagenomic samples and associated hosts. The mMGE is equipped with a user-friendly interface and a BLAST server, facilitating easy access/queries to all its contents easily. mMGE is freely available for academic use at: https://mgedb.comp-sysbio.org.

INTRODUCTION

Extrachromosomal mobile genetic elements (eMGEs), such as plasmids and bacteriophages, play critical roles in horizontal gene transfer (HGT) and microbial evolution within the microbial community by mediating intra- or intercellular DNA trafficking (1–4). Due to their high mobility and accessory genes related to antibiotic resistance (5–7), virulence factors (8,9) and auxiliary metabolic pathways (10–12), the eMGEs are essential for host fitness and the dissemination of drug resistance, which in turn shape microbial community structures. Furthermore, given that eMGEs frequently carry genes that encode toxins or other virulence factors, the prokaryotic hosts acquiring these genes have the potential to become deadly pathogens (13,14). Recently, disease-specific alterations of eMGEs have also been observed in several diseases (15–18), but the roles the eMGEs play in pathophysiology is still unclear, especially in a metagenomic setting.

With advances in sequencing technology, the accumulation of metagenomic data provides an unprecedented opportunity for detecting novel eMGEs (19). Recently, great progress has been made in identifying phages from metagenomic samples. For example, the human Gut Virome Database (GVD) (19) and the Integrated Microbial Genome/Virus (IMG/VR) database (20) detected phage genomes (and fragments) from assembled metagenomes. The Microbe Versus Phage (MVP) database established interactions between phages and prokaryotes based on a literature collection and a re-analysis of genomic and metagenomic sequences (21). Although those valuable resources significantly extend our knowledge of eMGEs, they focus only on phages from certain body sites, e.g. the gut, whereas plasmids were generally ignored (especially those derived

*To whom correspondence should be addressed. Tel: +86 21 55665546; Email: xmzhao@fudan.edu.cn
Correspondence may also be addressed to Wei-Hua Chen. Tel: +86 15827354263; Fax: +86 27 8779 2072; Email: weihuachen@hust.edu.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

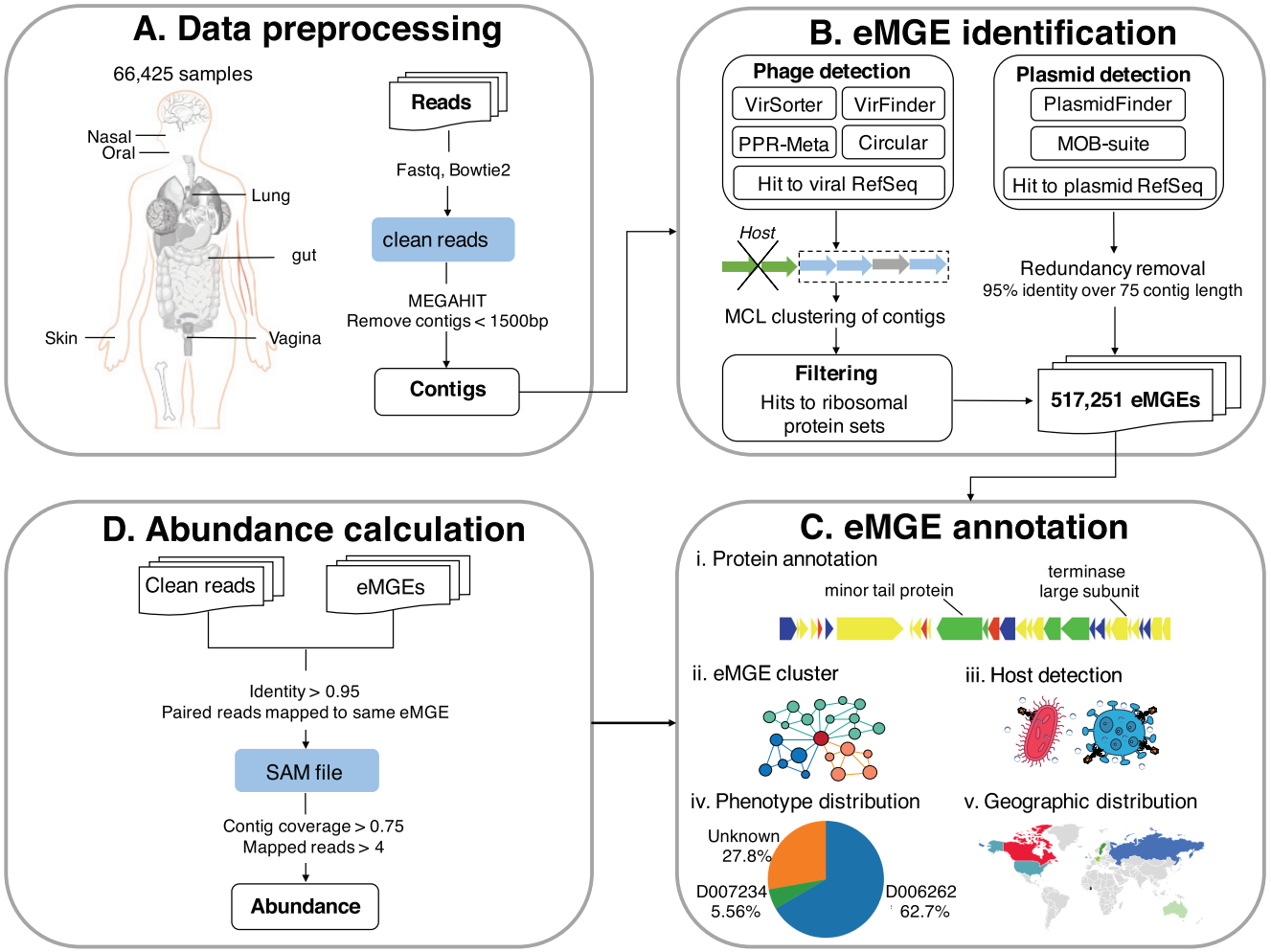

Figure 1. The overall workflow of mMGE. (**A**) Data pre-processing. A total of 66 425 human metagenomic samples and associated meta-data were collected, followed by pre-processing and assembly of raw sequencing reads. (**B**) eMGE identification. State-of-art toolsets were used to identify eMGEs. (**C**) eMGE annotation. Comprehensive annotations were provided for the eMGEs, including putative protein function and host information, etc; (**D**) Abundance calculation. Abundances and prevalence of the eMGEs across samples were also determined. See 'Materials and Methods' section for more details.

from metagenomes). The existing plasmid databases, including PlasmID, for plasmid clone information and distribution (22), Plasmid ATLAS, for plasmid visual analytics and identification (23), and PLSDB, for complete bacterial plasmids (24), collected information of existing plasmids without exploring the emerging large number of metagenomes. While the ACLAME database (25) provides mobile genetic elements, including phages, plasmids and prophages, it has not been updated in the ten years prior to this publication. Thus, a comprehensive eMGE database, including both phages and plasmids, as well as detailed sample metadata and their host information, will be of great use in understanding the diversity and putative functions of eMGEs in humans.

We have thus constructed mMGE, a database of human metagenomic extrachromosomal mobile genetic elements. Currently mMGE contains a total of 517 251 non-redundant eMGEs, including 92 492 plasmids and 424 759 phages, that we identified from 66 425 human metagenomic samples. In addition to basic information (including sequence characteristics, interactions with prokaryotic hosts, gene contents and taxonomic annotations), the extensive metadata of the samples, the abundances and distributions of the eMGEs across samples, phenotypes and populations are also available, allowing users to explore their biological functions, biogeographical patterns and habitat preferences. In addition, users can browse or query eMGE records in multiple ways, including eMGE clusters, metagenomic samples and putative hosts. mMGE is equipped with a user-friendly interface and a BLAST server, facilitating users to access and query all its contents easily.

DATABASE CONSTRUCTION

Figure 1 illustrates the overall workflow of mMGE. In brief, 66 425 metagenomic human samples were collected, followed by data preprocessing, eMGE identification, abundance calculation and eMGE annotation. Below we provide more details of materials and methods used in this study.

Table 1. The distribution of metagenomic samples included in mMGE across diverse human body sites

Body site	#samples	#projects	#associated phenotypes	#associated countries
Gut	41 841	233	63	42
Oral cavity	11 313	41	9	9
Skin	5384	30	7	7
Blood	2976	20	26	8
Nasopharyngeal	1930	16	9	6
Vagina	1028	6	1	3
Sputum	379	4	1	6
Eye	229	10	1	2
Urethra	123	5	3	3
Tooth	106	3	4	3
Reproductive system	76	1	0	1
Milk	60	2	0	2
Trachea	33	1	2	1
Lung	25	2	2	1
Liver	20	2	2	2
Circulatory system	12	1	1	0
Lymphatic system	11	3	3	2
Excretory system	1	1	1	1

Data collection and processing of metagenomic sequencing reads

Raw sequencing reads of 80 889 human metagenomic samples, from 370 datasets, were downloaded from the NCBI SRA (Sequencing Read Archive, https://www.ncbi.nlm.nih.gov/sra; Supplementary Table S1) database. Meta-data, including experimental conditions, dates of sampling and human host information, were also retrieved from corresponding publications and/or the NCBI SRA database. Phenotypes associated with samples were organized according to MeSH (Medical Subject Headings) (26), a hierarchically organized controlled vocabulary for biomedical information, while the metagenomic samples were organized according to the Genome Online Database classification system (27).

The FastQC (v0.11.8, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to check the overall quality of the downloaded sequences, followed by the use of Bowtie2 (28) to remove host-derived reads through mapping to the human reference genome (hg38). To trim sequence and remove low-quality bases, fastp (29) was utilized with the following parameters:'-l 50 -x -q 20 -u 5 -M 20 -W 4′. The samples containing less than 10 000 reads were removed from the subsequent analysis to ensure quality of the data. Then MEGAHIT (v1.2.8) (30) was used to assemble the high-quality clean reads per sample. After assembling, contigs that were less than 1.5 kb were discarded and the redundancy was removed with a threshold of 95% identity over 75% of their length.

In total, after quality control, we collected 66 425 human metagenomic samples associated with 110 phenotypes across 18 body sites from 49 countries. Table 1 summarizes the statistics of the samples we have collected.

eMGE identification, dereplication and clustering

After assembling and removing redundancy, all contigs were then piped through VirSorter (31), VirFinder (32) and PPR-Meta (33) for phage identification while went through PlasmidFinder (34) and MOB-suite (35) for plasmid identification. The phage sequences were identified by following the procedures described in (36) but with more stringent criteria. Firstly, the assemblies that met at least two of the following criteria were kept as putative phage contigs: (i) VirSorter positive (categories 1–2); (ii) PPR-Meta phage score > 0.7; (iii) VirFinder score > 0.6 and P-value < 0.05; (iv) Be circular; (v) Hit a phage genome from RefSeq with $>50\%$ identity and $>90\%$ coverage of contig length according to BLASTn (37). Subsequently, the candidate phage contigs obtained above were decontaminated using CheckV (38); those met the following criteria were discarded as described previously (36,39): (1) having more than three hits against ribosomal protein sets in COG (40) database or (2) having at least one ribosomal protein, VirSorter negative and non-circular and having less than three Hidden Markov Model (HMM) hits to the prokaryotic viral orthologous groups (pVOGs, E-value $<$ 1e-5) (41) per 10 kb. For the detection of plasmid sequences, the contigs satisfying at least one of the following criteria were selected as putative plasmid contigs: (i) Predicted by PlasmidFinder as positive; (ii) Predicted by MOB-suite as positive; (iii) Hit a plasmid genome from RefSeq with $>50\%$ identity and $>90\%$ coverage of contig length according to BLASTn.

All the above identified eMGE contigs were dereplicated at the population level if they shared $>95\%$ nucleotide identity across $>70\%$ coverage according to Lincluster (42), resulting in 517 251 non-redundant eMGE populations (92 492 plasmids and 424 759 phages). Then, a sequence-based classification framework was adopted to group closely related eMGE genomes into clusters (43). As a result, a total of 70 074 eMGE clusters containing 316 926 eMGE fragments (ranging from 2 to 384 members per cluster) were obtained, with most clusters (46.96%) having only two members. In addition, the quality and completeness of each phage contig were evaluated with CheckV (38) and the 'Minimum Information about an Uncultivated Virus Genome'(MIUViG) framework (44) was utilized to classify phage contigs as 'Genome fragment' or 'High-quality draft genome'. Consequently, 23 738 high-quality genomes were obtained and 14 990 of them were complete.

Annotation of eMGE contigs

The open reading frames (ORFs) for each eMGE contig were predicted using prodigal (v2.6.3) (45). With the predicted proteins were subjected to all-vs-all Blastp with thresholds of E-value $<$ 1e-5 and bit score > 50, the proteins were clustered into families by using a Markov Clustering Algorithm (MCL) with log-transformed E-value as similarity score and two for MCL inflation (46). The HMM profile for each protein family (protein family is also the protein cluster) was built with MAFFT (47) and hmmbuild (48). The functional annotations of all proteins were achieved by querying against PFAM (49), VOGdb (http://vogdb.org/) and eggNOG (50) databases. Consequently, about 40.86% of all predicted proteins had hits to at least one of the public databases, leaving the majority of the eMGE proteins as having unknown functions. For each protein family, its functional annotation was the one where $>75\%$ of its members were annotated with the functions (51), while those that could not be annotated in the previous way were queried

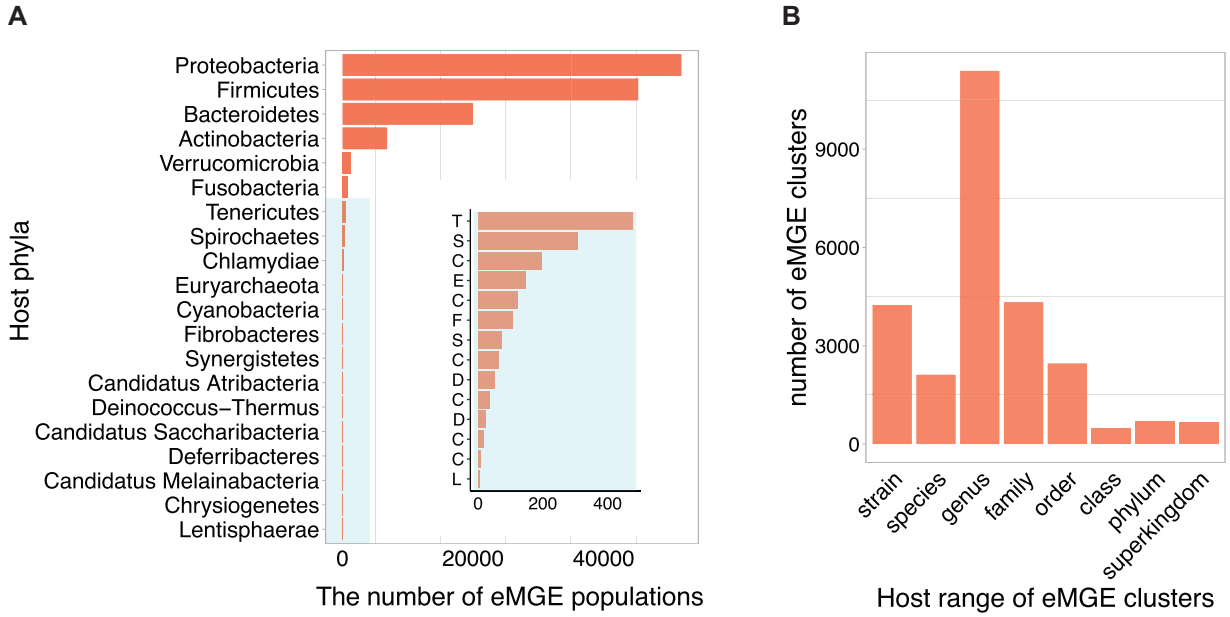

Figure 2. The distribution of identified eMGE-hosts. (**A**) The number of eMGE populations associated with their corresponding bacterial and archaeal host phyla; The inset with blue background provides resolution for the low frequency bacteria host phyla and each letter on the *y*-axis corresponds to the first letter of host phyla's name. (**B**) The number of eMGE clusters distributed across different host range levels.

against proteins from RefSeq with hmmsearch (52) (threshold of 1e-5 for *E*-value and 50 for bit score).

For each eMGE contig, taxonomy annotations were achieved through the following three steps. Firstly, 21 hallmark POGs recognized as taxon-specific signatures were used to assign a taxonomic lineage to eMGEs at different levels (order, family, subfamily and genus). Secondly, the eMGEs that clustered with genomes from RefSeq were able to be assigned to known taxonomic genera (53). Finally, for those contigs that could not be assigned to a specific taxonomy with above two steps, we tried to annotate them using a majority-rule approach. The proteins predicted in the contig were aligned against the proteins deposited in the UniProt database (54), and the taxonomy level was determined to a taxon if more than 75% of the proteins hit the same taxon (51).

eMGE-host identification

For each eMGE contig, its microbial hosts were determined by following the approaches described previously (19). Firstly, the eMGE fragments were aligned against NCBI RefSeq with BLASTn, and RefSeq genomes with >95% identity across more than 2500 bp were considered as putative hosts of eMGEs. Secondly, the bacterial genomes from NCBI RefSeq and metagenomic assemblies >1500 bp were used to build CRISPR-Cas spacer database. The CRISPR spacers in microbial genomes and assemblies were predicted using MinCED (55) with default parameters (4 110 100 spacers were obtained). The detected spacers were then aligned against phage fragments using BLASTn with the following options: -task blastn-short -word_size 5, *E*-value < 1e-5, bit score > 45, identity > 95% of full length, and a maximum of two mismatches was allowed. Finally, tRNAscan-SE (56) was separately used to identify tRNA genes from phage sequences and bacterial genomes, and bacterial genomes with tRNA genes matching phage tRNA genes at 95% identity across 100% of the length were considered as the corresponding host. The host of eMGE contigs can be determined in either of the three ways.

In total, we identified 2 032 843 eMGE-host associations, with the hosts spanning across 20 bacterial and archaeal phyla (Figure 2A). For each eMGE sequence and/or eMGE cluster in mMGE, its host range was determined by following the way described previously (21). Briefly, for eMGEs with only one host, the host range was assigned as the taxonomic rank of the host in the NCBI taxonomy database, while the host range was defined as the taxonomic rank of the Last Common Ancestor of all its hosts if an eMGE fragment infects multiple hosts. mMGE provides host information for 115 072 eMGE fragments (22.24% of all the eMGE fragments or genomes), and approximately 50% of eMGE clusters have host range at 'species' or 'genus' level (Figure 2B).

Abundance and prevalence of eMGEs across samples

To calculate the relative abundance of different eMGE fragments, the quality-filtered reads were mapped to eMGE contigs using BWA mem (57), where the reads mapped with <95% identity or paired-reads mapped to different locations were removed. The Bedtools (58) genomecov was then used to calculate the coverage over contigs and only the eMGE contigs with >75% of length covered by reads were considered to be present in that sample (59). Then the number of mapped reads to a contig were normalized by the total number of clean reads per metagenomic sample, which was used as the approximate relative abundance for that

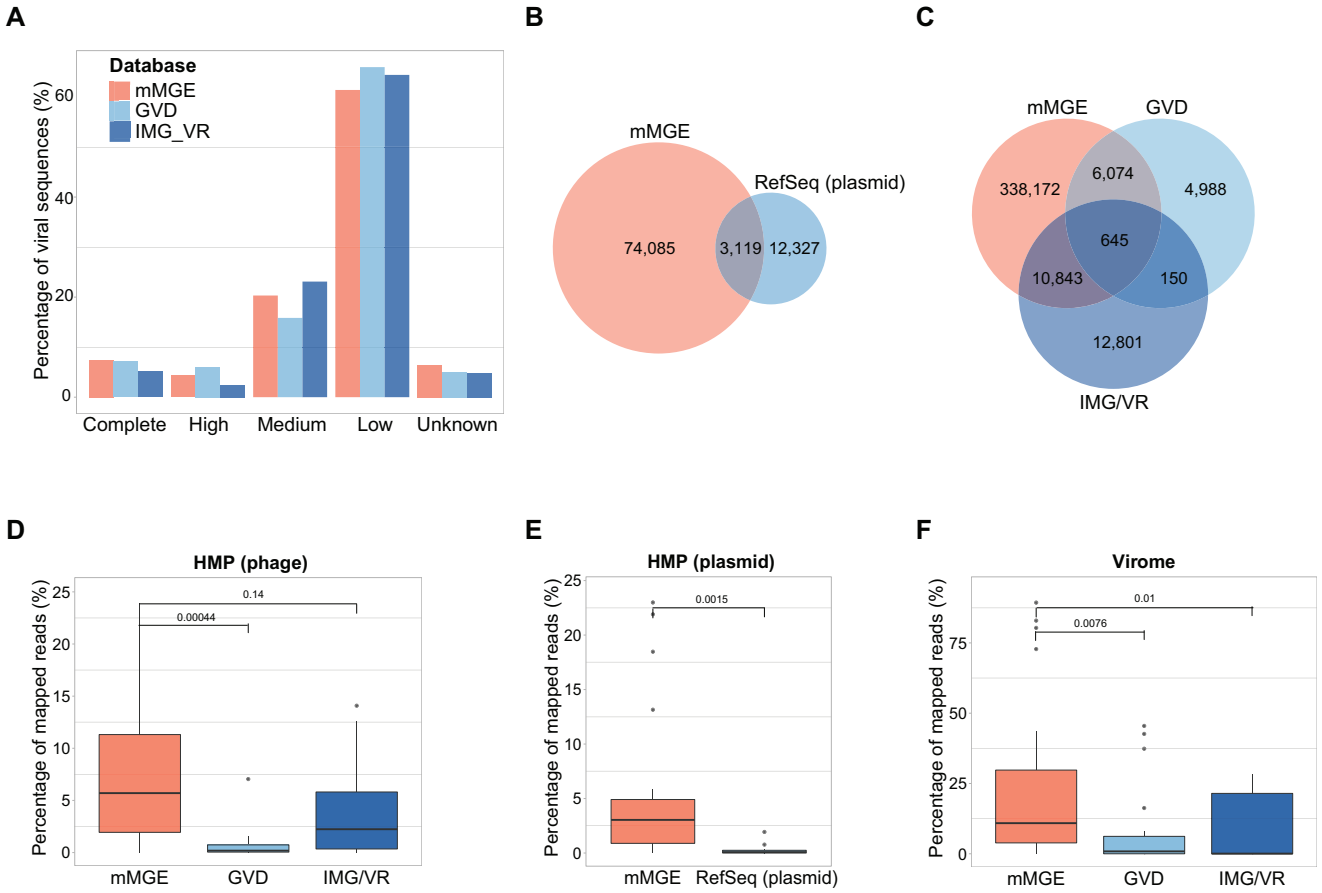

Figure 3. Contents of mMGE and comparisons with public databases. (**A**) The completeness and quality of phage contigs estimated by CheckV for mMGE, IMG/VR and GVD, where high denotes high quality and the same for medium and low; (**B**) and (**C**) The venn diagram of plasmids and phages from different sources, where all contigs were dereplicated at population level and decontaminated with CheckV and only phage populations from human samples were considered; (**D**–**F**) The percentage of mapped reads for phages or plasmids from HMP dataset (D and E) and Virome dataset (F), where the HMP dataset includes 20 samples from PRJNA48479 and the Virome dataset contains viral enriched samples that came from PRJNA588313.

eMGE fragment. The relative abundance of each eMGE cluster was calculated by summing the percentage of reads mapped to members belonging to that cluster.

DATABASE OVERVIEW AND FUNCTIONALITY

Overview of mMGE

The current version of the mMGE database contains 517 251 unique/non-redundant eMGEs (92 492 plasmids and 424 759 phages) identified from 66 425 metagenomic samples. Figure 3 shows the comparison of mMGE against multiple public databases, including IMG/VR (20), GVD (19) and plasmid RefSeq. Figure 3A shows the completeness of phage contigs from IMG/VR, GVD and mMGE estimated with CheckV, from which we can see that mMGE (Complete: 7.46%, high-quality: 4.36%, medium-quality: 20.32%, low-quality: 61.40%) has the highest percentage of complete genomes while has comparable quality compared with IMG/VR (Complete: 5.16%, high-quality: 2.45%, medium-quality: 23.11%, low-quality: 64.41%) and GVD (Complete: 7.10%, high-quality: 6.02%, medium-quality: 15.85%, low-quality: 65.97%). Compared with RefSeq, GVD and IMG/VR, mMGE significantly extends the number of phages and plasmids as shown in Figure 3B and C, and mMGE can successfully recover half of the phages detected by GVD or IMG/VR (56.67% in GVD and 47.01% in IMG/VR recovered by mMGE separately) which is much better compared with the overlap between GVD and IMG/VR (6.70 and 3.25% shared by GVD and IMG/VR separately).

Moreover, we evaluated the identification sensitivity of plasmids and phages from different databases by comparing their percentage of mapped reads. For this purpose, three datasets were used for a fair comparison, including 20 human gut metagenomic samples from PRJNA48419 (60) (HMP dataset), the viral enriched samples from PRJNA588313 (Virome dataset) and a plasmid dataset containing 131 plasmid contigs assembled from human metagenomes with metaplasmidSPAdes (61) (Metaplasmid data). Figure 3D–F separately shows the percentages of mapped reads for HMP and Virome datasets, from which we can clearly see that mMGE has the best identification sensitivity. For the Metaplasmid dataset, 63 of 131 plasmid contigs can be successfully recovered by mMGE

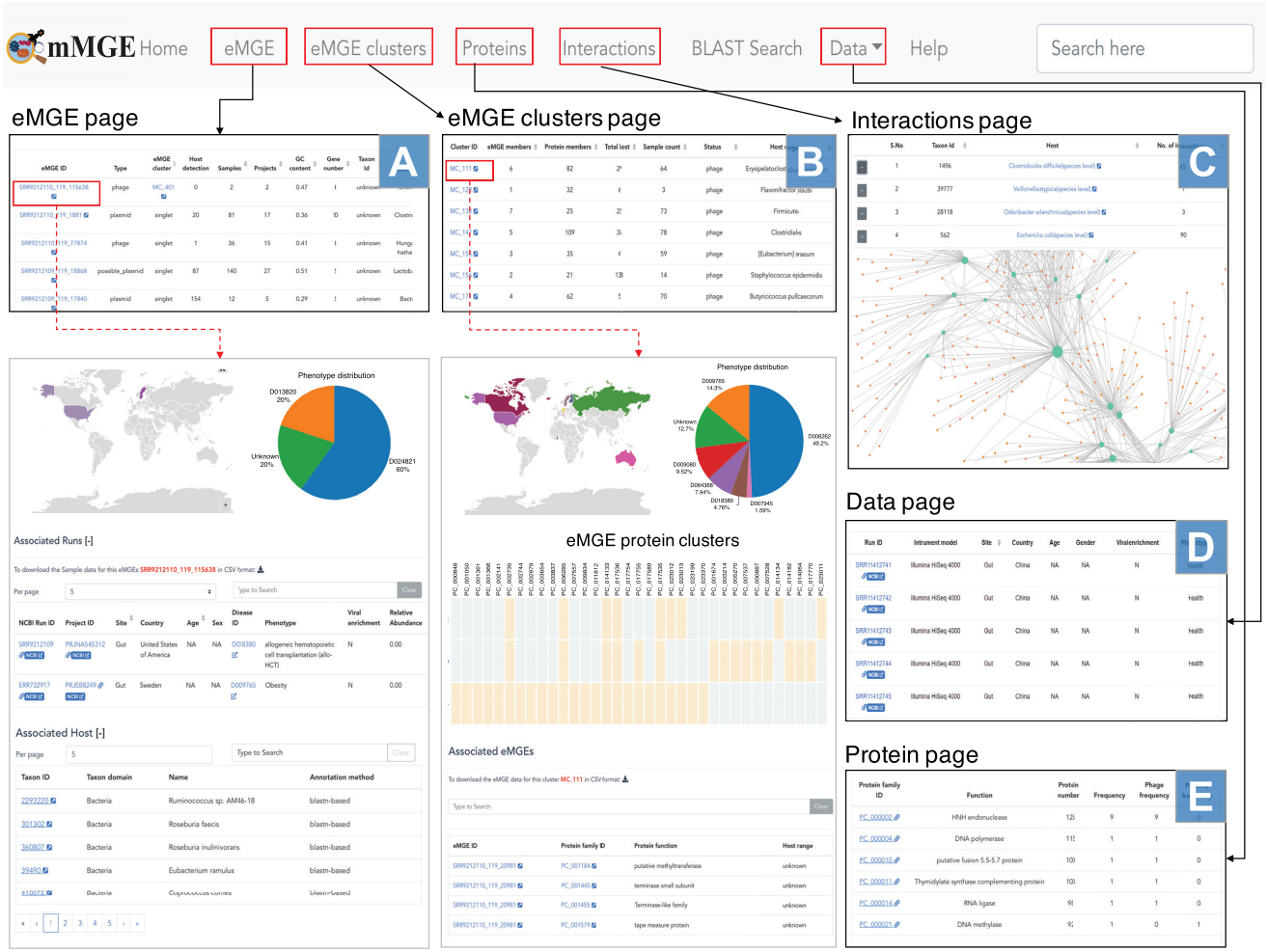

Figure 4. The user-friendly web interface of mMGE. (**A**) The 'eMGE' page shows the basic information of eMGEs; (**B**) The 'eMGE cluster' page shows the information about eMGE clusters; (**C**) The 'Interaction' page presents the interactions between eMGEs and their hosts; (**D**) The 'Data' page shows the information about each sample and project; (**E**) The 'Proteins' page presents the protein content of each eMGE. Those pages can be cross searched to provide more detailed information of eMGEs or eMGE clusters.

while RefSeq plasmid covers only 29 of them. From the results shown in Figure 3, we can see that mMGE significantly extends the number of eMGEs with comparable quality but much higher identification sensitivity compared with existing databases.

Web interface

mMGE provides a user-friendly and interactive portal for browsing and querying all eMGEs and their associated information as shown in Figure 4. In the 'eMGE' page, the information of all eMGE fragments or genomes as well as their associated information can be easily browsed (Figure 4A). The detailed information page for a specific eMGE can be available by clicking on the eMGE ID of interest. With the relative abundance of eMGEs across samples, mMGE enables the users to investigate the distribution patterns of eMGEs across countries and human phenotypes. The detailed information of its hosts and the meta-data of metagenomic samples in which this eMGE can be detected is also available. All eMGE records can be browsed via eMGE clusters (Figure 4B), eMGE-host interactions (Figure 4C) and metagenomic samples (Figure 4D). In the 'eMGE cluster' page, mMGE enables users to browse the most relevant information of eMGE clusters, including the number of members in the cluster ('eMGE members'), the number of samples in which they are present ('Sample count'), the number of identified putative hosts ('Total hosts'), the number of protein families they contain ('Protein members') and the predicted host-range (Figure 4B). All identified eMGE-host interactions are listed in the 'Interactions' page, where the interactive visualization of the eMGE-host interaction network is also provided. The 'Data' page provides manually curated meta-data of the metagenomic samples used whenever possible (Figure 4D). The metagenomic samples can be viewed according to the collection sites, body sites or phenotypes. To facilitate researchers to download the raw sequencing data, additional links to samples and projects from NCBI were also provided. For each metagenomic sample or project, we also summarized the total number of associated eMGEs and the associated eMGE sequences of each metagenomic sample can be obtained. In addition, mMGE

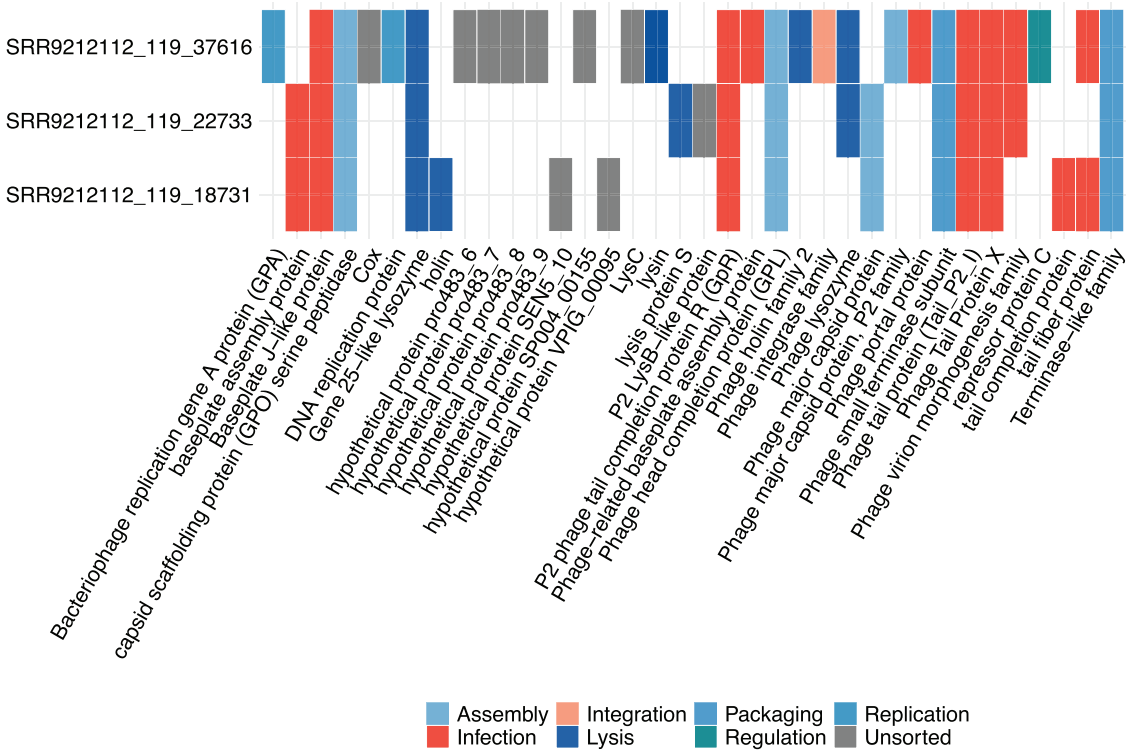

Figure 5. The example matrix view of protein clusters within the eMGE cluster 'MC_62'. The columns correspond to the protein clusters while the rows represent eMGE members within this cluster. The protein clusters were colored according to their functional annotations.

provides comprehensive annotations for protein families associated with eMGEs in the 'Proteins' page (Figure 4E).

Protein clusters evolutionarily conserved within eMGE clusters

Despite the diverse gene contents across different plasmids and phages, it has been found that conserved genetic modules maybe shared between related eMGE genomes (62). To facilitate users to explore the evolutionary relationships between eMGEs within eMGE clusters, the visualization of the matrix of protein clusters across eMGEs was also provided (Figure 4B). For example, Figure 5 shows an example of a matrix view of protein clusters within the eMGE cluster 'MC_62′, where the functions associated with each protein cluster are provided. We can see that the protein clusters shared by all members in 'MC_62′ are involved in assembly ('head', 'capsid' and 'neck'), DNA packaging ('terminase'), infection ('tail', 'baseplate' and 'virion', 'portal') and lysis ('lysozyme' and 'lysin'), indicating the essential functional roles of these proteins.

Querying eMGE with BLAST

To facilitate users to query mMGE with their own sequences, we provided a BLAST server for the users to align their sequences against the eMGEs deposited in the mMGE database (63). In this way, users may easily check what phages or plasmids their sequences are based on hits to the eMGEs from mMGE. The BLAST search could be accessed at: https://mgedb.comp-sysbio.org/#/submitBlast.

FUTURE DIRECTIONS

Since eMGEs are a major source of antimicrobial resistance genes, virulence and pathogenicity related genes, insertion sequences and other transposable elements, an annotation and discovery pipeline will be provided in a future version. In addition to collecting eMGEs detected in human metagenomes, the future version will complement the data with eMGEs derived from other source such as animals or ocean. We note that there is much room to improve mMGE in the following directions: (i) Including a binning method, such as MetaBat (64), which can be used to merge contigs derived from the same population and extend assembly completeness; (ii) Improving the discovery pipeline so it can detect more eMGEs with higher quality; (iii) Incorporating Long-read metagenomic datasets, which can then be used to improve the assembly and identify more eMGEs (4).

CONCLUSION

Due to the high mobility of eMGEs and their complex interactions with microbial hosts, the analysis of eMGEs is essential for characterization of microbial communities and exploring their potential roles in regulating microbial communities. Here we have introduced mMGE, an integrative resource for environmental uncultivated eMGEs derived from diverse human body sites coupled with extensive annotations. With 66 425 samples collected from 18 body sites, 110 human phenotypes and 49 countries, we manually curated meta-data of all samples and applied

stringent criteria to keep only high-confidence eMGE sequences. In total, 517 251 unique eMGEs were obtained, including 424 759 phages and 92 492 plasmids. Extensive comparisons with existing database indicated that mMGE contains more eMGEs with higher quality. To facilitate users to perform downstream analysis, mMGE provides precomputed relative abundance of eMGEs, their prevalence within and across samples as well as putative associations with phenotypes. Comprehensive annotations of each eMGE record including sequence characteristics, protein content, taxonomy affiliation and host-eMGE interactions are also available at the website. The web server allows users to browse the included eMGE records in multiple ways and uploaded nucleotide sequences can be searched in the database. mMGE provides a modern, interactive and user-friendly interface, enabling users to easily access and query all its contents. As metagenomic datasets continue to expand, we will continue developing mMGE in the near future by including eMGEs detected from more samples and more comprehensive annotations.

DATA AVAILABILITY

All data are freely available to all academic users. This work is licensed under a Creative Commons Attribution 3.0 Unported Licence (CC BY 3.0). In addition to downloading the data provided on certain web pages, the users can download all data from the 'Data download' section of the 'Help' page.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Natural Science Foundation of China [61932008, 61772368]; Shanghai Municipal Science and Technology Major Project [2018SHZDZX01]; Shanghai Science and Technology Innovation Fund [19511101404]; National Key Research and Development Program of China [2018YFC0910503 to X.M.Z., 2018YFC0910502, 2019YFA0905601 to W.H.C]. Funding for open access charge: National Key Research and Development Program of China; National Natural Science Foundation of China; Shanghai Municipal Science and Technology Major Project.

Conflict of interest statement. None declared.

REFERENCES

1. Frost,L.S., Leplae,R., Summers,A.O. and Toussaint,A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.*, **3**, 722–732.
2. Thomas,C.M. and Nielsen,K.M. (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.*, **3**, 711–721.
3. Sitaraman,R. (2018) Prokaryotic horizontal gene transfer within the human holobiont: ecological-evolutionary inferences, implications and possibilities. *Microbiome*, **6**, 163.
4. Suzuki,Y., Nishijima,S., Furuta,Y., Yoshimura,J., Suda,W., Oshima,K., Hattori,M. and Morishita,S. (2019) Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut. *Microbiome*, **7**, 119.
5. Calero-Caceres,W., Ye,M. and Balcazar,J.L. (2019) Bacteriophages as environmental reservoirs of antibiotic resistance. *Trends Microbiol.*, **27**, 570–577.
6. Wein,T., Hulter,N.F., Mizrahi,I. and Dagan,T. (2019) Emergence of plasmid stability under non-selective conditions maintains antibiotic resistance. *Nat. Commun.*, **10**, 2595.
7. Lopatkin,A.J., Meredith,H.R., Srimani,J.K., Pfeiffer,C., Durrett,R. and You,L. (2017) Persistence and reversal of plasmid-mediated antibiotic resistance. *Nat. Commun.*, **8**, 1689.
8. Kraushaar,B., Hammerl,J.A., Kienöl,M., Heinig,M.L., Sperling,N., Dinh Thanh,M., Reetz,J., Jäckel,C., Fetsch,A. and Hertwig,S. (2017) Acquisition of virulence factors in livestock-associated MRSA: Lysogenic conversion of CC398 strains by virulence gene-containing phages. *Sci Rep-Uk*, **7**, 2004.
9. Sarowska,J., Futoma-Koloch,B., Jama-Kmiecik,A., Frej-Madrzak,M., Ksiazczyk,M., Bugla-Ploskonska,G. and Choroszy-Krol,I. (2019) Virulence factors, prevalence and potential transmission of extraintestinal pathogenic Escherichia coli isolated from different sources: recent reports. *Gut Pathog.*, **11**, 10.
10. Hurwitz,B.L. and U'Ren,J.M. (2016) Viral metabolic reprogramming in marine ecosystems. *Curr. Opin. Microbiol.*, **31**, 161–168.
11. Rosenwasser,S., Ziv,C., Van Creveld,S.G. and Vardi,A. (2016) Virocell metabolism: metabolic innovations during host-virus interactions in the ocean. *Trends Microbiol.*, **24**, 821–832.
12. Kieft,K., Zhou,Z. and Anantharaman,K. (2020) VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, **8**, 90.
13. Davies,E.V., James,C.E., Williams,D., O'Brien,S., Fothergill,J.L., Haldenby,S., Paterson,S., Winstanley,C. and Brockhurst,M.A. (2016) Temperate phages both mediate and drive adaptive evolution in pathogen biofilms. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 8266–8271.
14. Meric,G., Mageiros,L., Pascoe,B., Woodcock,D.J., Mourkas,E., Lamble,S., Bowden,R., Jolley,K.A., Raymond,B. and Sheppard,S.K. (2018) Lineage-specific plasmid acquisition and the evolution of specialized pathogens in Bacillus thuringiensis and the Bacillus cereus group. *Mol. Ecol.*, **27**, 1524–1540.
15. Nakatsu,G., Zhou,H., Wu,W.K.K., Wong,S.H., Coker,O.O., Dai,Z., Li,X., Szeto,C.H., Sugimura,N., Lam,T.Y. *et al.* (2018) Alterations in Enteric Virome are associated with colorectal cancer and survival outcomes. *Gastroenterology*, **155**, 529–541.
16. Norman,J.M., Handley,S.A., Baldridge,M.T., Droit,L., Liu,C.Y., Keller,B.C., Kambal,A., Monaco,C.L., Zhao,G., Fleshner,P. *et al.* (2015) Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell*, **160**, 447–460.
17. Lawrence,D., Baldridge,M.T. and Handley,S.A. (2019) Phages and human health: more than idle hitchhikers. *Viruses*, **11**, 587.
18. Bedarf,J.R., Hildebrand,F., Coelho,L.P., Sunagawa,S., Bahram,M., Goeser,F., Bork,P. and Wullner,U. (2017) Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naive Parkinson's disease patients. *Genome Med.*, **9**, 39.
19. Ann C. Gregory, O.Z., Allison Howell. (2019) The human gut virome database. bioRxiv doi: https://doi.org/10.1101/655910, 02 July 2019, preprint: not peer reviewed.
20. Paez-Espino,D., Roux,S., Chen,I.A., Palaniappan,K., Ratner,A., Chu,K., Huntemann,M., Reddy,T.B.K., Pons,J.C., Llabres,M. *et al.* (2019) IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res.*, **47**, D678–D686.
21. Gao,N.L., Zhang,C., Zhang,Z., Hu,S., Lercher,M.J., Zhao,X.M., Bork,P., Liu,Z. and Chen,W.H. (2018) MVP: a microbe-phage interaction database. *Nucleic Acids Res.*, **46**, D700–D707.
22. Zuo,D., Mohr,S.E., Hu,Y., Taycher,E., Rolfs,A., Kramer,J., Williamson,J. and LaBaer,J. (d) PlasmID: a centralized repository for plasmid clone information and distribution. *Nucleic Acids Res.*, **35**, D680–D684.
23. Jesus,T.F., Ribeiro-Goncalves,B., Silva,D.N., Bortolaia,V., Ramirez,M. and Carrico,J.A. (2019) Plasmid ATLAS: plasmid visual analytics and identification in high-throughput sequencing data. *Nucleic Acids Res.*, **47**, D188–D194.
24. Galata,V., Fehlmann,T., Backes,C. and Keller,A. (2019) PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.*, **47**, D195–D202.

25. Leplae,R., Lima-Mendez,G. and Toussaint,A. (2010) ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res.*, **38**, D57–D61.
26. Richter,R.R. and Austin,T.M. (2012) Using MeSH (medical subject headings) to enhance PubMed search strategies for evidence-based practice in physical therapy. *Phys. Ther.*, **92**, 124–132.
27. Reddy,T.B., Thomas,A.D., Stamatis,D., Bertsch,J., Isbandi,M., Jansson,J., Mallajosyula,J., Pagani,I., Lobos,E.A. and Kyrpides,N.C. (2015) The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.*, **43**, D1099–D1106.
28. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
29. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
30. Li,D., Luo,R., Liu,C.M., Leung,C.M., Ting,H.F., Sadakane,K., Yamashita,H. and Lam,T.W. (2016) MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, **102**, 3–11.
31. Roux,S., Enault,F., Hurwitz,B.L. and Sullivan,M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ*, **3**, e985.
32. 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.
33. Fang,Z., Tan,J., Wu,S., Li,M., Xu,C., Xie,Z. and Zhu,H. (2019) PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience*, **8**, giz066.
34. Carattoli,A. and Hasman,H. (2020) PlasmidFinder and in silico pMLST: identification and typing of plasmid replicons in whole-genome sequencing (WGS). *Methods Mol. Biol.*, **2075**, 285–294.
35. Robertson,J. and Nash,J.H.E. (2018) MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genome*, **4**, e000206.
36. Shkoporov,A.N., Clooney,A.G., Sutton,T.D.S., Ryan,F.J., Daly,K.M., Nolan,J.A., McDonnell,S.A., Khokhlova,E.V., Draper,L.A., Forde,A. *et al.* (2019) The human gut virome is highly diverse, stable, and individual specific. *Cell Host Microbe*, **26**, 527–541.
37. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
38. Nayfach,S., Camargo,A.P., Eloe-Fadrosh,E., Roux,S. and Kyrpides,N. (2020) CheckV: assessing the quality of metagenome-assembled viral genomes. bioRxiv doi: https://doi.org/10.1101/2020.05.06.081778, 08 May 2020, preprint: not peer reviewed.
39. Zolfo,M., Pinto,F., Asnicar,F., Manghi,P., Tett,A., Bushman,F.D. and Segata,N. (2019) Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.*, **37**, 1408–1412.
40. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
41. Grazziotin,A.L., Koonin,E.V. and Kristensen,D.M. (2017) Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.*, **45**, D491–D498.
42. Steinegger,M. and Soding,J. (2018) Clustering huge protein sequence sets in linear time. *Nat. Commun.*, **9**, 2542.
43. Paez-Espino,D., Pavlopoulos,G.A., Ivanova,N.N. and Kyrpides,N.C. (2017) Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nat. Protoc.*, **12**, 1673–1682.
44. Roux,S., Adriaenssens,E.M., Dutilh,B.E., Koonin,E.V., Kropinski,A.M., Krupovic,M., Kuhn,J.H., Lavigne,R., Brister,J.R., Varsani,A. *et al.* (2019) Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.*, **37**, 29–37.
45. 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.
46. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
47. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
48. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
49. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
50. Huerta-Cepas,J., Szklarczyk,D., Heller,D., Hernandez-Plaza,A., Forslund,S.K., Cook,H., Mende,D.R., Letunic,I., Rattei,T., Jensen,L.J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.
51. Roux,S., Hallam,S.J., Woyke,T. and Sullivan,M.B. (2015) Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife*, **4**, e08490.
52. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.
53. Roux,S., Brum,J.R., Dutilh,B.E., Sunagawa,S., Duhaime,M.B., Loy,A., Poulos,B.T., Solonenko,N., Lara,E., Poulain,J. *et al.* (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, **537**, 689–693.
54. UniProt,C. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
55. Bland,C., Ramsey,T.L., Sabree,F., Lowe,M., Brown,K., Kyrpides,N.C. and Hugenholtz,P. (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.
56. Chan,P.P. and Lowe,T.M. (2019) tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol. Biol.*, **1962**, 1–14.
57. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv doi: https://arxiv.org/abs/1303.3997, 26 May 2013, preprint: not peer reviewed.
58. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
59. Roux,S., Emerson,J.B., Eloe-Fadrosh,E.A. and Sullivan,M.B. (2017) Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ*, **5**, e3817.
60. Lloyd-Price,J., Mahurkar,A., Rahnavard,G., Crabtree,J., Orvis,J., Hall,A.B., Brady,A., Creasy,H.H., McCracken,C., Giglio,M.G. *et al.* (2017) Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*, **550**, 61–66.
61. Antipov,D., Raiko,M., Lapidus,A. and Pevzner,P.A. (2019) Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res.*, **29**, 961–968.
62. Bin Jang,H., Bolduc,B., Zablocki,O., Kuhn,J.H., Roux,S., Adriaenssens,E.M., Brister,J.R., Kropinski,A.M., Krupovic,M., Lavigne,R. *et al.* (2019) Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.*, **37**, 632–639.
63. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
64. Kang,D.D., Froula,J., Egan,R. and Wang,Z. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, **3**, e1165.