Probing the mobilome: Discoveries in the dynamic microbiome

Victoria Carr[1,2,*], Andrey Shkoporov[3], Colin Hill[3], Peter Mullany[4], David Moyes[1,*]

[1]Centre for Host-Microbiome Interactions, Faculty of Dentistry, Oral and Craniofacial Sciences, King's College London, UK

[2]The Alan Turing Institute, British Library, London, UK

[3]APC Microbiome Ireland & School of Microbiology, University College Cork, Ireland

[4]Eastman Dental Institute, University College London, UK

Corresponding authors: victoria.carr@kcl.ac.uk and david.moyes@kcl.ac.uk

Abstract

There has been an explosion of metagenomic data representing human, animal and environmental microbiomes. This provides an unprecedented opportunity for comparative and longitudinal studies of many functional aspects of the microbiome that go beyond taxonomic classification, such as profiling genetic determinants of antimicrobial resistance, interactions with the host, potentially clinically relevant functions and the role of mobile genetic elements (MGEs). One of the most important but least studied of these aspects are the MGEs, collectively referred to as the "mobilome". Here we elaborate on the benefits and limitations of using different metagenomic protocols, discuss the relative merits of various sequencing technologies, and highlight relevant bioinformatics tools and pipelines to predict the presence of MGEs and their microbial hosts.

Introduction

The shift to high-throughput sequencing technologies in microbial genomics has radically changed our understanding of microbial communities in different habitats. The appreciation of the

complexity of these communities is now undergoing a further shift as more publicly available microbiome datasets based on shotgun metagenomic sequencing are becoming available. As well as establishing the taxonomy and relative abundance of microbial populations, these datasets are allowing individual genes and their variants to be characterised, including antimicrobial resistance genes (ARGs). Mobile genetic elements (MGEs) are critical to our understanding of how genes (and their associate functions) move within a community via horizontal gene transfer (HGT) within a community[1]. These elements can have a lasting impact on the composition of microbial communities, affecting their diversity and density, as well as their interaction with the environment[2]. The profile of these MGEs (mobilome) is thus likely to be a key player in influencing selection pressure-driven changes in the composition of microbial communities and their impact on the host organism or tissue. MGEs are also responsible for the movement of antimicrobial resistance determinants and virulence factors between microbes[3]. For example, the use of antimicrobials can increase the prevalence of MGEs carrying functioning ARGs that are integrated in microbial genomes[4]. Profiling the mobilome and its associated ARGs can provide insights into how ARGs move across multiple genomes within the microbiome. To characterise the mobilome in a microbial community, all MGEs sequences need to be identified from metagenomic data and ideally would be assigned to a microbial host. Although detecting MGEs from single isolates using whole genome sequencing is a common approach that is significantly more straightforward, metagenomic sequencing is increasingly being used to detect and classify multiple MGEs from microbial communities.

Mobilome composition

The microbial mobilome is defined as all MGEs within a given microbiome. MGEs themselves are segments of genetic material that are capable of moving within a genome or between genomes of different organisms. They include plasmids, transposable elements (both non-conjugative and conjugative transposons, the latter also called integrative conjugative elements [ICEs]) and

bacteriophages, which are covered primarily in this review (Box 1). Other MGEs include gene cassettes that are commonly part of integrons[5,6]. There are also mobilisable elements, both integrative and plasmid, that can utilise the conjugative functions of plasmids and/or ICEs but do not themselves encode a complete set of conjugative functions[7,8]. Finally, there are satellite viruses that can use phage machinery for induction and transfer[9,10].

Plasmids are extrachromosomal replicons present in bacteria and archaea. They range in size from less than a kilobase to the megabase size range[11], contain at least one replication origin, usually possess a gene expressing a replication initiation protein (Rep) and a series of direct, inverted and A-T rich repeats[12]. Some plasmids are cryptic, but many carry genes encoding important functions in the survival and fitness of their host. These include virulence traits and resistance to antimicrobials. In facilitating their transfer between microbes, conjugative plasmids include genes that encode proteins required for plasmid transfer. Furthermore, some plasmids can exploit the transfer of other conjugative elements without having to bear the large genetic load required to encode conjugation functions[13].

Insertion sequences (ISs) are short transposable elements containing genes that code for proteins involved in their own transposition. Most ISs contain a gene encoding a transposase, the most ubiquitous gene in prokaryotic and eukaryotic sequences[14], and are flanked by short inverted terminal repeat (ITR) sequences. Insertion of an IS leads to the duplication of the host-genome target site and formation of unique direct repeat (DR) sequences[15]. Two ISs can flank an accessory gene, such as an ARG, to form a composite transposon. More complex transposons, such as those of the Tn*3* family, transpose via the formation of a co-integrate. These still usually produce target site duplications. The most complex of transposons are the conjugative transposons, also known as integrative conjugative elements (ICEs)[16]. These genetic elements encode their own conjugation functions and can transfer between bacteria, usually using a similar mechanism as that employed by

conjugative plasmids. Unlike plasmids, ICEs are usually integrated into the host chromosome. ~~Another group of MGEs are the gene cassettes that are commonly part of integrons. The cassettes are typically between 0.5 and 1 kb and do not contain their own promotor.~~

Bacteriophages (phages) are viruses ranging in size from a few to hundreds of kilobases that replicate within bacteria and archaea[17]. They replicate rapidly, have huge genetic diversity and have genomes that can be comprised of single- or double-stranded DNA or RNA. Phages replicate through either the lytic or a lysogenic cycle. Virulent phages lyse their host at the completion of their replication cycle, whereas temperate phages integrate their genetic material into the host genome to become prophages as part of their replication cycle (lysogeny). Although temperate phages can sometimes carry virulence factors[18], it is still unclear whether ~~there is little evidence yet that~~ phages significantly contribute to the transfer of ARGs. Although there has been mounting evidence that phages very rarely contain ARGs[14], ~~ARGs are very rarely found in genomes~~ those phages that do contain them may be able to transfer these ARGs as frequently as plasmids[20]. ~~but generalised or lateral transduction may act as a mode of ARG transfer.~~

MGEs represent a highly heterogeneous group of elements, furthermore the difference between certain elements can be blurred. For example, there are phages that can transpose, plasmids that integrate like ICEs and ICEs that can replicate like plasmids. There are also MGEs that can mobilise a whole bacterial chromosome[21,22]. It is best to think of MGEs as a continuum rather than trying to place them in neat boxes. This continuum of MGEs within an individual bacterial species, never mind a community as a whole, is highly varied. Although these elements can be inherited vertically, their central role in HGT means that even within an individual species there is great heterogeneity.

Targeted metagenomic approaches and challenges in extracting MGEs

Despite having to overcome significant hurdles, metagenomic sequencing of microbial samples is increasingly being used to identify novel MGEs. Both targeted and whole metagenomic methods are now being used to identify and discover novel as well as known MGEs (Fig. 1). In contrast to whole metagenomic methods where all DNA extracts are sequenced, targeted metagenomics include a step that specifically selects a type of MGE prior to sequencing.

Targeted metagenomic methods currently include purifying MGEs prior to shotgun sequencing. For example, free phage particles, along with other virus-like particles (VLPs), are purified in several stages of physical and/or enzymatic treatments[23–25]. Nucleic acids extracted from VLPs are then sequenced and assembled into contiguous sequences for further annotation[25–27]. Circular plasmids are isolated using high-throughput transposon-aided capture (TRACA) from metagenomic DNA, which are then typically transformed into *Escherichia coli* for cloning[28], followed by shotgun sequencing and PCR-based approaches to close gaps in sequences[29]. However, these targeted approaches may misjudge the potential MGE load. Inefficiencies in the elution of VLPs from faecal samples have been shown to result in an underestimation of the viral load, and inconsistencies between protocols have led to discrepancies in results between studies[24]. Size-fractionation is an alternative technique involving enrichment of extracted DNA for novel viral particles by filtering the samples through a size exclusion membranes that has been applied to the cow rumen virome[30]. Of 148 viral genera enriched from the cow rumen, 75% had no counterpart in existing viral databases, highlighting the power of this technique to recover phages.

For plasmids, TRACA enriches metagenomic DNA for circular plasmids by using a DNAse that selectively removes linear DNA. Plasmids are subsequently “captured” by inserting a transposon (in

an *in vitro* transposition reaction) with an origin of replication and selection marker before transforming them into typically *Escherichia coli* for cloning[28]. This is followed by shotgun sequencing, with additional PCR to close gaps in sequences[29]. However, TRACA has a bias towards capturing smaller plasmids between 3-10 kb, excludes linearised plasmids, and potentially inactivates plasmid genes as a result of transposon insertion[31]. Alternatively, inverse-PCR together with multiple displacement amplification (another DNA amplification technique) has also been applied to identify small circular plasmids in metagenomic samples[32].

Finally, a targeted metagenomic approach using PCR amplification can be used to identify transposable elements by targeting the repeat regions[33]. Metagenomic DNA is amplified by PCR primers targeting transposable elements, purified and ligated into plasmid vectors, then transformed into host strains. After clonal expansion, the plasmids are isolated, sequenced and annotated for transposable elements.

Targeted metagenomic approaches are highly specific and therefore useful for extracting MGEs with distinct features, such as sequence composition. Given non-targeted MGEs would be excluded, these approaches would not be suitable for determining a more complete representation of MGEs within the whole metagenome. ~~All these approaches have a bias to preferentially detecting particular MGEs that may be more suited for that particular purification extraction protocol or a particular PCR primer set, thereby underestimating or missing other MGEs present in the whole metagenome.~~ However, recent advances in sequencing technology and data storage mean that whole metagenomic DNA sequencing is now a viable option for investigating the wider pool of MGEs, giving us a better representative picture of the mobilome[34–37].

Whole metagenomics

Whole metagenomic DNA sequencing has great potential for both identifying known and unknown MGEs and also for predicting the MGE hosts. However, there are several limiting factors, specifically with current next-generation sequencing technologies and bioinformatic software tools, that need to be considered.

Challenges in sequencing technologies

The current gold-standard for metagenome sequencing is using short-read sequencing methodologies, specifically Illumina and Ion Torrent technologies. Since short-read metagenomic sequencing produces reads that are too short to allow the identification of plasmids, phages and transposable elements, many bioinformatic pipelines involve assembling the metagenomic reads into longer contiguous sequences called contigs. However, assembling metagenomes is computationally intensive, and the choice of assembly tool has a significant impact on the accuracy of identifying MGEs[38–40]. Dealing with the microbial complexity of a metagenome with limited read depth and repeated regions is a challenge for current assembly algorithms. These tools are prone to generate erroneous inter-species chimeric contigs when processing complex metagenomic sequence datasets. Thus, plasmid and transposon contigs are often inaccurate or incomplete. Different plasmids often contain similar replication and conjugative elements[41], whilst transposable elements contain repeated regions[42]. For phages, assembly of short reads has further challenges including a high incidence of repeat regions and/or hypervariable regions[43], genetic diversity[44], frequent modular structures[45], and heterogeneity at strain level[43,46]. To circumvent these issues, many metagenomic assemblers attempt to produce shorter, less complete but more accurate contigs rather than longer, inaccurate ones. A direct consequence of this is that metagenomic contigs are often too short to accurately predict large MGEs.

Long-read sequencing technologies (such as Oxford Nanopore and PacBio's single-molecule real-time [SMRT] sequencing), produce longer sequence reads, meaning it is possible to more accurately

assemble much longer scaffolds and even complete genomes. Nanopore technology, for example, has been used to successfully recapitulate complete viral genomes from metagenomes[47,48]. However, the sequences generated contain more erroneous bases than short-read technology sequences due to technical defects in base calling[49,50]. PacBio has a higher accuracy rate in single-nucleotide and structural variants, but produces shorter reads than Nanopore and is more costly[51,52]. In addition, the limits in coverage depth from a run on a single Nanopore flowcell is a bottleneck for identifying lower abundant MGEs in metagenomes with high microbial diversity[49]. However, it is possible to improve and even complete the assembly of MGEs from complex whole metagenomes using an ensemble of short-read and long-read sequencing technologies[53].

Bioinformatic methods in MGE sequence annotation

When analysing microbiome composition, isolation and sequencing of DNA forms only part of the story – the subsequent computational analysis is every bit as important. This is also the case when mining sequencing data for MGEs and other genetic elements. Although advances in technology have markedly improved the accuracy of whole metagenomic sequencing, accurate and efficient bioinformatics software is required to resolve MGEs from a complex pool of fragmented microbial genomes.

Typically, genomic sequence features are identified broadly either by reference-based or *de novo* methods, or a combination of both. Reference-based methods generally use alignment algorithms, such as BLAST[54], to align query nucleotide or amino acid assemblies against a reference database or search tools against probability sequence models, such as HMMER for hidden Markov models (HMMs)[55]. Non-MGE-specific nucleotide sequence databases, such as RefSeq[56], and protein sequence databases, like Pfam[57] and UniProt[58], have been applied to detect HGT events in metagenomes[59,60]. Virus-specific sequence databases have more recently been established, such as the Prokaryotic Virus Orthologous Groups (pVOGs)[61], curated viral databases from RefSeq,

PATRIC[62] and IMG/VR[56]. ~~databases.~~ Databases suited for searching transposable elements in metagenomic assemblies include ISfinder for ISs[64], and ICEberg for ICEs and integrative and mobilisable elements (IMEs)[65]. PlasmidFinder is a popular database for identifying plasmids that contains plasmid replicon sequences from *Enterobacteriaceae* and gram positive bacteria[66]. In all cases, MGE containing databases contain a very narrow representation of the mobilome with incomplete coverage of element types, and do not reflect the actual MGE diversity. For instance, transposable elements are one of the most ubiquitous and genetically diverse elements in the microbiome[42,67], making cataloguing all of them an intractable task. Despite this obvious limitation, well-curated reference databases can be useful for discovering novel MGEs as they are often used in benchmarking new *de novo* bioinformatics tools[68].

Despite their utility, MGE reference databases ~~obviously~~ do not include all MGEs in existence. Further, it is difficult to find novel MGEs that are dissimilar in ~~their~~ sequence and structure to ~~the~~ known MGEs. ~~To f~~Finding these novel MGEs requires the use of *de novo* bioinformatics methods and tools to make predictions based on sequence data. There is a plethora of different algorithms used for discovering putative phages in assembled metagenomes, such as VirSorter[69], VirFinder[70], MARVEL[71], VirMiner[72] and ViraMiner[73] (Table 1). Apart from VirSorter that uses primarily HMMs, all these tools apply m~~M~~achine learning ~~is applied in all these tools apart from VirSorter (which uses Hidden Markov Models [HMMs])~~ to identify viral-like domains. A handful of tools have been developed for identifying plasmid sequences from metagenomes, including cBar[74], PlasFlow[40], Recycler[75] and metaplasmidSPAdes[76] (Table 1). ~~Similar to bioinformatic tools used for phage, m~~Machine learning approaches are also used in cBar and PlasFlow to predict linear and circular plasmids. ~~Despite the popularity of machine learning, caution must be taken in using such tools for whole metagenomes. Similar to reference-based tools, machine learning models struggle to classify genome signatures that have not been used to train the model, meaning it would be difficult to predict mobile elements with unique sequences. In addition, the accuracy of machine learning~~

~~predictions relies heavily on the quality of the sequenced and assembled metagenomes. Instead,~~ Other non-machine learning-based tools, Recycler and metaplasmidSPAdes, identify plasmids using~~se~~ D~~d~~e Bruijn graph assembly of ***k***-mers (small sequences of length *k*)~~to identify circular plasmids only~~. m~~M~~etaplasmidSPAdes ~~constructs assembly graphs from de Bruijn graphs and~~also includes a naïve~~naive~~ Bayesian classifier on custom plasmid-specific profile-HMMs to improve its accuracy. For discovery of ISs, only two *de novo* pipelines have been developed using existing algorithms to identify direct repeats and palindromic inverted terminal repeats (Table 1)[77].

When designing and building bioinformatic tools, it is valuable to benchmark them for specificity and sensitivity. For MGE identification tools applied to metagenomes, the ideal dataset for benchmarking predictions would include labels of known MGEs within real metagenomic sequences. Aside from VirMiner and metaplasmidSPAdes, these tools have not been ~~adequately~~ benchmarked using representative metagenomes. Since these ground truth datasets are difficult to obtain, many of these tools were benchmarked using simulated metagenomic sequences generated from a representative set of genomes from the most abundant species of a microbial community. ~~Instead, most of these tools were benchmarked using simulated read fragments generated from a representative set of the most abundant single species genomes of a microbial community.~~ Therefore, it is likely that when these tools are applied to complex whole metagenomic samples, they would not perform as well as their stated accuracy would suggest.

Technological challenges in host prediction of MGEs

Identifying the microbial hosts of different MGEs will be central to developing our understanding of how MGEs shape microbial communities and *vice versa*. However, this is problematic for a variety of reasons, not least of which is our limited ability to find the specific microbial origin of MGEs in metagenomic samples. As technologies move forward, additional approaches such as wet-

lab protocols and bioinformatics tools are being applied with both short and long-read metagenomic sequencing to link MGEs with their host microbe.

Wet-lab technologies for microbial host prediction

Although associating genetic elements with individual organisms within a community initially seems insurmountable, there are promising laboratory-based techniques that can be exploited. Some of these can make use of features of different sequencing technologies, whilst other methods require pre-processing of samples prior to sequencing. Binning reads into groups prior to computational assembly is probably the simplest of these techniques. As SMRT sequencing can be applied to identify the methylation status of a nucleotide (Fig. 2a), metagenomic reads can be binned into species or subspecies based on methylation motifs[78]. SMRT sequencing can be applied to identify the methylation status of a nucleotide (Fig. 2a). Sequences are then clustered into groups based on the similarity of multiple methylation motifs. These motifs are usually shared by both chromosomes and plasmids within a microbe but are often unique to a microbial strain. However, as microbial communities become more complex, the methylation motifs become less unique as it becomes more likely that more than one strain or species contains the same motif.

An alternative approach is the use of proximity ligation methodologies, specifically Hi-C (Fig. 2b)[79]. DNA molecules in close proximity in the genome's three-dimensional structure are covalently bonded together. Thus MGEs that are in close proximity to their host genome are covalently bonded to the host genome. These connected sequences are then digested around the bond and ligated to form a continuous strand with ligation junctions. After this proximity ligation, the DNA is fragmented and sequenced as usual. Sequence information regarding these ligation junctions is used in downstream computational analysis pipelines to assign assembled metagenomic reads to their host microbe species. Hi-C has been used alongside short-read metagenomic sequencing to link plasmids to their hosts with strain-level resolution in synthetic metagenomes[80] and species-level

resolution in real metagenomic communities[81,82]. However, Hi-C has limited resolution capabilities for closely related organisms due to their high sequence similarity and uneven Hi-C link densities[83]. Proximity ligation has also been used to link phages to species from cattle rumen metagenomes[84]. ~~However, s~~Since proximity ligation relies on the three-dimensional structure of the host genome only, phages that do not integrate into the genome as prophages are largely undetected by this process. However, single-cell viral tagging with short-read metagenomic sequencing is an alternative approach specifically for predicting the hosts of both lytic and lysogenic phages[85].

Bioinformatic methods in microbial host prediction

Metagenomic reads and contigs containing MGEs and host genomes can be binned into groups using computational as well as wet-lab methods, allowing for two levels of identification and discrimination. There are many different algorithms for metagenomic binning, including analysing sequence composition features and coverage, sequence signature properties, k-mer frequencies and gene co-abundance across samples[37,86–92]. However, these binning algorithms, particularly gene co-abundance, can be computationally intensive.

An approach that can link MGEs with their hosts relies on distinct MGE sequences also found in microbial genomes[93–95]. When an MGE enters a bacterium, the bacterium uses a defence mechanism of Clustered Regularly Interspaced Palindromic Repeats (CRISPR). Fragments of the MGE sequence, known as spacers, are integrated between CRISPR loci in the bacterial genome. These spacers are transcribed into small RNA molecules and processed into a ribo-protein complex which targets and destroys invading genomes. The hosts of these MGEs can then be predicted by aligning the predicted MGE contigs against a reference database of candidate host genomes containing CRISPR spacers. This method has been previously used to identify phage and plasmid hosts in human gut metagenomes[96,97]. However, since many of these reference databases are incomplete, it may only be possible to assign a small proportion of MGE contigs to a host[93].

Conclusions and Further Perspectives

In general, there is currently no single sequencing, wet-lab or bioinformatics technique for whole metagenomes that can efficiently profile the entire mobilome and its microbial context. As we have shown here, employing a combination of approaches is the best solution to classifying novel MGEs and assigning these and known MGEs to their host microbes. In order to resolve longer MGEs such as plasmids and phages whilst maintaining accuracy, the ideal approach is to use a combination of short-read and long-read sequencing. Highly accurate short metagenomic reads can be assembled and scaffolded against more complete but less accurate contiguous sequences from long-read sequencing (see Outstanding Questions). Identifying the microbial hosts of the MGEs presents further problems. However, SMRT long-read sequencing used in combination with proximity ligation on short-read sequencing is a complementary approach that can be applied to all MGE types and will allow for association of these elements to host genomes with a reasonably high degree of certainty.

Having generated these sequences, many different bioinformatic methods can be highly effective at identifying and classifying MGEs in these sequences accurately, or binning MGEs with host sequences from the acquired metagenomic data. The bioinformatic tools listed are not evaluated computationally in this review, but cited reviews and papers have done so for tools identifying phages and plasmids[72,98]. Due to a rapid software developments, it is likely some tools outlined here will already be superseded by the time of publication, with one or a few tools that have been iterated and become standard. Popular approaches, such as machine learning, will still be important tools. However, a tool that has a high accuracy on simulated metagenomes may not perform well on real metagenomes and could be computationally expensive (see Outstanding Questions). Therefore, researchers will need to critically evaluate which tool is most suitable for their particular requirements.

There is no single correct solution for characterising the mobilome. The performance of bioinformatics tools for *de novo* discovery is limited by the data quality which is dependent on the sequencing platform (see Outstanding Questions). Current sequencing technologies for whole metagenomes fall short of the levels required for a truly accurate and fully representative analysis of the mobilome. However, there is cause for optimism. The recent development of new methodologies, such as proximity ligation and SMRT sequencing technologies, means that we are rapidly evolving our ability to not only identify potential MGEs, but also to associate them with their host genomes. As these technologies improve, so too will bioinformatic tools be developed to make full use of these new datasets, and thus provide us with a more complete picture of the mobilome and how it spreads genetic elements through microbial communities.

References

1. Sitaraman, R. Prokaryotic horizontal gene transfer within the human holobiont: ecological-evolutionary inferences, implications and possibilities. *Microbiome* **6**, 163 (2018).
2. Hsu, B. B. *et al.* Dynamic Modulation of the Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model. *Cell Host & Microbe* **25**, 803-814.e5 (2019).
3. Penders, J., Stobberingh, E. E., Savelkoul, P. H. M. & Wolffs, P. The human microbiome as a reservoir of antimicrobial resistance. *Front. Microbiol.* **4**, (2013).
4. Bakkeren, E. *et al.* Salmonella persisters promote the spread of antibiotic resistance plasmids in the gut. *Nature* **573**, 276–280 (2019).
5. Gillings, M. R. Integrons: Past, Present, and Future. *Microbiol Mol Biol Rev* **78**, 257–277 (2014).
6. Cury, J., Jové, T., Touchon, M., Néron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res* **44**, 4539–4550 (2016).
7. Guédon, G., Libante, V., Coluzzi, C., Payot, S. & Leblond-Bourget, N. The Obscure World of Integrative and Mobilizable Elements, Highly Widespread Elements that Pirate Bacterial Conjugative Systems. *Genes (Basel)* **8**, (2017).

8. Osborn, A. M. & Böltner, D. When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. *Plasmid* **48**, 202–212 (2002).
9. Dokland, T. Molecular Piracy: Redirection of Bacteriophage Capsid Assembly by Mobile Genetic Elements. *Viruses* **11**, (2019).
10. Sun, J., Inouye, M. & Inouye, S. Association of a retroelement with a P4-like cryptic prophage (retronphage phi R73) integrated into the selenocystyl tRNA gene of Escherichia coli. *J. Bacteriol.* **173**, 4171–4181 (1991).
11. Hayes, F. The Function and Organization of Plasmids. in *E. coli Plasmid Vectors: Methods and Applications* (eds. Casali, N. & Preston, A.) 1–17 (Humana Press, 2003). doi:10.1385/1-59259-409-3:1.
12. Solar, G. del, Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M. & Díaz-Orejas, R. Replication and Control of Circular Bacterial Plasmids. *Microbiol. Mol. Biol. Rev.* **62**, 434–464 (1998).
13. Roberts, A. P., Allan, E. & Mullany, P. Chapter Two - The Impact of Horizontal Gene Transfer on the Biology of Clostridium difficile. in *Advances in Microbial Physiology* (ed. Poole, R. K.) vol. 65 63–82 (Academic Press, 2014).
14. Aziz, R. K., Breitbart, M. & Edwards, R. A. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res* **38**, 4207–4217 (2010).
15. Mahillon, J. & Chandler, M. Insertion Sequences. *Microbiol. Mol. Biol. Rev.* **62**, 725–774 (1998).
16. Salyers, A. A., Shoemaker, N. B., Stevens, A. M. & Li, L. Y. Conjugative transposons: an unusual and diverse set of integrated gene transfer elements. *Microbiol Rev* **59**, 579–590 (1995).
17. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
18. Fortier, L.-C. & Sekulovic, O. Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence* **4**, 354–365 (2013).
19. Enault, F. *et al.* Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J* **11**, 237–247 (2017).
20. Debroas, D. & Siguret, C. Viruses as key reservoirs of antibiotic resistance genes in the environment. *The ISME Journal* 1 (2019) doi:10.1038/s41396-019-0478-9.

21. Brouwer, M. S. M. *et al.* Horizontal gene transfer converts non-toxigenic *Clostridium difficile* strains into toxin producers. *Nature Communications* **4**, 2601 (2013).
22. Dordet-Frisoni, E. *et al.* Mycoplasma Chromosomal Transfer: A Distributive, Conjugative Process Creating an Infinite Variety of Mosaic Genomes. *Front Microbiol* **10**, 2441 (2019).
23. Kleiner, M., Hooper, L. V. & Duerkop, B. A. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* **16**, 7 (2015).
24. Conceição-Neto, N. *et al.* Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Scientific Reports* **5**, 16532 (2015).
25. Shkoporov, A. N. *et al.* Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* **6**, 68 (2018).
26. Milani, C. *et al.* Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX. *Microbiome* **6**, 145 (2018).
27. Aggarwala, V., Liang, G. & Bushman, F. D. Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mobile DNA* **8**, 12 (2017).
28. Jones, B. V. & Marchesi, J. R. Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nature Methods* **4**, 55–61 (2007).
29. Smalla, K., Jechalke, S. & Top, E. M. Plasmid Detection, Characterization, and Ecology. *Microbiology Spectrum* **3**, (2015).
30. Solden, L. M. *et al.* Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. *Nature Microbiology* **3**, 1274 (2018).
31. Dib, J. R., Wagenknecht, M., Farías, M. E. & Meinhardt, F. Strategies and approaches in plasmidome studies—uncovering plasmid diversity disregarding of linear elements? *Front. Microbiol.* **6**, (2015).
32. Jørgensen, T. S., Xu, Z., Hansen, M. A., Sørensen, S. J. & Hansen, L. H. Hundreds of Circular Novel Plasmids and DNA Elements Identified in a Rat Cecum Metamobilome. *PLOS ONE* **9**, e87924 (2014).
33. Tansirichaiya, S., Mullany, P. & Roberts, A. P. PCR-based detection of composite transposons and translocatable units from oral metagenomic DNA. *FEMS Microbiol Lett* **363**, (2016).

34. Ghai, R., Mehrshad, M., Mizuno, C. M. & Rodriguez-Valera, F. Metagenomic recovery of phage genomes of uncultured freshwater actinobacteria. *The ISME Journal* **11**, 304–308 (2017).
35. Waller, A. S. *et al.* Classification and quantification of bacteriophage taxa in human gut metagenomes. *The ISME Journal* **8**, 1391–1402 (2014).
36. Ogilvie, L. A. *et al.* Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences. *Nature Communications* **4**, 2420 (2013).
37. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology* **32**, 822–828 (2014).
38. Sutton, T. D. S., Clooney, A. G., Ryan, F. J., Ross, R. P. & Hill, C. Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **7**, 12 (2019).
39. Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**, e3817 (2017).
40. Krawczyk, P. S., Lipinski, L. & Dziembowski, A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* **46**, e35 (2018).
41. Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & Cruz, F. de la. Mobility of Plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434–452 (2010).
42. Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol Rev* **38**, 865–891 (2014).
43. Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. Hypervariable loci in the human gut virome. *PNAS* **109**, 3962–3966 (2012).
44. Manrique, P. *et al.* Healthy human gut phageome. *PNAS* **113**, 10400–10405 (2016).
45. Lima-Mendez, G., Toussaint, A. & Leplae, R. A modular view of the bacteriophage genomic space: identification of host and lifestyle marker modules. *Research in Microbiology* **162**, 737–746 (2011).
46. Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nature Communications* **8**, 15892 (2017).
47. Warwick-Dugdale, J. *et al.* Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* **7**, e6800 (2019).

48. Beaulaurier, J. *et al.* Assembly-free single-molecule nanopore sequencing recovers complete virus genomes from natural microbial communities. *bioRxiv* 619684 (2019) doi:10.1101/619684.
49. Tyler, A. D. *et al.* Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Scientific Reports* **8**, 10931 (2018).
50. Somerville, V. *et al.* Long read-based de novo assembly of low complex metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *bioRxiv* 476747 (2018) doi:10.1101/476747.
51. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* **13**, 278–289 (2015).
52. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* 1–8 (2019) doi:10.1038/s41587-019-0217-9.
53. Bertrand, D. *et al.* Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nature Biotechnology* 1 (2019) doi:10.1038/s41587-019-0191-2.
54. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
55. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29–W37 (2011).
56. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).
57. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427–D432 (2019).
58. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506–D515 (2019).
59. Li, C., Jiang, Y. & Li, S. LEMON: a method to construct the local strains at horizontal gene transfer sites in gut metagenomics. *BMC Bioinformatics* **20**, 702 (2019).
60. Jiang, X., Hall, A. B., Xavier, R. J. & Alm, E. J. Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. *PLOS ONE* **14**, e0223680 (2019).

61. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res* **45**, D491–D498 (2017).
62. Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* **42**, D581–D591 (2014).
63. Paez-Espino, D. *et al.* IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res* **47**, D678–D686 (2019).
64. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32-36 (2006).
65. Liu, M. *et al.* ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res* **47**, D660–D665 (2019).
66. Carattoli, A. *et al.* In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrob Agents Chemother* **58**, 3895–3903 (2014).
67. Filée, J., Siguier, P. & Chandler, M. Insertion Sequence Diversity in Archaea. *Microbiol Mol Biol Rev* **71**, 121–157 (2007).
68. Mangul, S. *et al.* Systematic benchmarking of omics computational tools. *Nat Commun* **10**, 1–11 (2019).
69. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
70. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
71. Amgarten, D., Braga, L. P. P., da Silva, A. M. & Setubal, J. C. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Front. Genet.* **9**, (2018).
72. Zheng, T. *et al.* Mining, analyzing, and integrating viral signals from metagenomic data. *Microbiome* **7**, 42 (2019).
73. Tampuu, A., Bzhalava, Z., Dillner, J. & Vicente, R. ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLOS ONE* **14**, e0222271 (2019).
74. Zhou, F. & Xu, Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* **26**, 2051–2052 (2010).

75. Rozov, R. *et al.* Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* **33**, 475–482 (2017).
76. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res.* **29**, 961–968 (2019).
77. Kamoun, C., Payen, T., Hua-Van, A. & Filée, J. Improving prokaryotic transposable elements identification using a combination of de novo and profile HMM methods. *BMC Genomics* **14**, 700 (2013).
78. Beaulaurier, J. *et al.* Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nature Biotechnology* **36**, 61–69 (2018).
79. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289–293 (2009).
80. Beitel, C. W. *et al.* Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2**, e415 (2014).
81. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature Communications* **9**, 870 (2018).
82. Stalder, T., Press, M. O., Sullivan, S., Liachko, I. & Top, E. M. Linking the resistome and plasmidome to the microbiome. *The ISME Journal* 1 (2019) doi:10.1038/s41396-019-0446-4.
83. Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. *G3: Genes, Genomes, Genetics* **4**, 1339–1346 (2014).
84. Bickhart, D. *et al.* Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *bioRxiv* 491175 (2018) doi:10.1101/491175.
85. Džunková, M. *et al.* Defining the human gut host–phage network through single-cell viral tagging. *Nat Microbiol* 1–12 (2019) doi:10.1038/s41564-019-0526-2.
86. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology* **31**, 533–538 (2013).
87. Herath, D., Tang, S.-L., Tandon, K., Ackland, D. & Halgamuge, S. K. CoMet: a workflow using contig coverage and composition for binning a metagenomic sample with high precision. *BMC Bioinformatics* **18**, 571 (2017).

88. Girotto, S., Pizzi, C. & Comin, M. MetaProb: accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics* **32**, i567–i575 (2016).
89. Plaza Oñate, F. *et al.* MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics* doi:10.1093/bioinformatics/bty830.
90. Yu, G., Jiang, Y., Wang, J., Zhang, H. & Luo, H. BMC3C: binning metagenomic contigs using codon usage, sequence composition and read coverage. *Bioinformatics* **34**, 4172–4179 (2018).
91. Wang, Z., Wang, Z., Lu, Y. Y., Sun, F. & Zhu, S. SolidBin: improving metagenome binning with semi-supervised normalized cut. *Bioinformatics* doi:10.1093/bioinformatics/btz253.
92. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nature Methods* **11**, 1144–1146 (2014).
93. Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* **22**, 1985–1994 (2012).
94. Wang, J., Gao, Y. & Zhao, F. Phage–bacteria interaction network in human oral microbiome. *Environmental Microbiology* **18**, 2143–2158 (2016).
95. Zhang, Q., Rho, M., Tang, H., Doak, T. G. & Ye, Y. CRISPR-Cas systems target a diverse collection of invasive mobile genetic elements in human microbiomes. *Genome Biol.* **14**, R40 (2013).
96. Gogleva, A. A., Gelfand, M. S. & Artamonova, I. I. Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs. *BMC Genomics* **15**, 202 (2014).
97. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host & Microbe* **26**, 527-541.e5 (2019).
98. Arredondo-Alonso, S., Willems, R. J., van Schaik, W. & Schürch, A. C. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom* **3**, (2017).
99. Boucher, Y. *et al.* Recovery and evolutionary analysis of complete integron gene cassette arrays from Vibrio. *BMC Evol Biol* **6**, 3 (2006).
100. Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clinical Microbiology Reviews* **31**, (2018).

101. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-358 (2005).

102. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

103. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).

104. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* **7**, e1002195 (2011).

105. Roux, S., Tournayre, J., Mahul, A., Debroas, D. & Enault, F. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**, 76 (2014).

106. Noguchi, H., Taniguchi, T. & Itoh, T. MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. *DNA Res* **15**, 387–396 (2008).

107. Brown Kav, A., Benhar, I. & Mizrahi, I. A method for purifying high quality and high yield plasmid DNA for metagenomic and deep sequencing approaches. *Journal of Microbiological Methods* **95**, 272–279 (2013).

346

347

348 **Acknowledgements**

349 The research was supported by the Centre for Host-Microbiome Interactions, King's College
350 London, funded by the Biotechnology and Biological Sciences Research Council (BBSRC) grant
351 BB/M009513/1 awarded to D.L.M., The Alan Turing Institute under the Engineering and Physical
352 Sciences Research Council (EPSRC) grant EP/N510129/1, and APC Microbiome Ireland funded by
353 the Research Centre grant from Science Foundation Ireland (SFI) under Grant Number
354 SFI/12/RC/2273.

355

356 **Author contributions**

V.R.C. and D.L.M. conceived the presented idea and wrote the manuscript with support from A.S., C.H. and P.M.

Conflicts of Interest

The authors declare no conflicts of interest.

Figure Legends

Figure 1: Targeted and whole metagenomic technologies for extracting MGEs

Figure 2: Wet-lab protocols for microbial host identification of MGEs (applicable to plasmids and prophages) using a) SMRT sequencing and b) Hi-C

Definition of types of MGEs

Plasmids are replicons that are distinct from chromosomal DNA found in bacteria and archaea.
Length: less than a kilobase to megabases
Main function: They are highly heterogeneous elements and the simplest just encode their own replication functions. Some also encode conjugation functions. They commonly contain cargo DNA that encode functions for survival in different environments e.g. antibiotic resistance genes and virulence factors.
HGT mechanism: conjugation, transduction and transformation.

Insertion sequences are short transposable elements containing genes that code for proteins involved in transposition
Length: kilobases
Main function: The simplest code for proteins involved in transposition only. They often have cargo genes that encode functions for survival in different environments e.g. antibiotic resistance genes and virulence factors
HGT mechanism: They can be spread by transposing to conjugative elements and by transformation and transduction.

Integrative conjugative elements (ICE) also called conjugative transposons
Length: 18 kilobases and upwards
Main function: They are highly heterogeneous elements that have the capability of inserting into bacterial genomes and transferring by conjugation between bacteria. They commonly contain cargo DNA that encode functions for survival in different environments e.g. antibiotic resistance genes and virulence factors.
HGT mechanism: conjugation

Mobilisable genetic elements
Length: less than a kilobase to megabases
Main function: They are highly heterogeneous elements that do not contain enough genetic information for independent conjugative transfer but can utilise the transfer functions of conjugative plasmids or ICEs. They can exist as plasmids or as integrative elements; the latter are sometimes called integrative and mobilisable elements (IMEs). They commonly contain cargo DNA that encode functions for survival in different environments e.g. antibiotic resistance genes and virulence factors.
HGT mechanism: conjugation

Integrons and gGene cassettes
Length: 0.5 to hundreds of ~~–1~~ kilobases[99]
Main function: Mobilise ~~integrons~~gene cassettes that are associated with a variety of functions (including antimicrobial resistance genes and virulence factors)[100]
HGT mechanism: Site specific recombination. Gene cassettes can be moved between integrons (through an intermediate form of circular DNA molecule) and assembled in large arrays. ~~Present as circular DNA molecules which can be captured and integrated into i~~Integrons themselves can in turn be mobilised via action of composite ~~and transferred by~~ transposons, ~~ing to~~ conjugative elements, plasmids ~~and~~ or by transformation

Bacteriophages (phages) are viruses that replicate within bacteria and archaea
Length: few to hundreds of kilobases
Main function: Replicate and destroy (lytic phages) or integrate DNA into host genome (lysogenic phages)
HGT mechanism: transduction

Box 1: Mobile Genetic Elements Definitions

MGE	Tool	Authors and Year	Data Type	Search algorithm	Advantages	Disadvantages
Insertion sequence	Pipelines: Two *de novo* and one profile HMM search	Kamoun et al., 2013[77]	Raw fragments	*De novo* "Repeat search": RepeatScout algorithm [101] *De novo* "IR search": palindrome software of the EMBOSS package[102] Profile HMM: MUSCLE[103] and HMMER2 package[104]	*De novo* methods do not rely on incomplete ISfinder database Profile HMM search performsed significantly better than BLAST on simulated and real metagenomic datasets	Repeat search had high false positive rate IR search has lower true positive rate Repeat search and IR search not tested on metagenomic datasets
Bacteriophage	MARVEL	Amgarten et al., 2018[71]	Raw fragments in metagenomic bins	Random forest machine learning	Better sensitivity and similar specificity to VirSorter and VirFinder	No option in software to retrain on alternative training data Only testsed algorithm on simulated metagenomic bins Does not consider prophages
	VirSorter	Roux et al., 2015[69]	Contigs	Prediction of circular sequences[105] Gene predicting using MetaGeneAnnotator[106] HMMER3 for pHMMs and BLASTP for unclustered proteins	Prediction of novel prophages from reference-independent prediction of viral domains	Not tested on metagenomics of whole microbial communities, only viral metagenomes Does not have complete prophage prediction, as optimised for assemblies of fragments
	VirFinder	Ren et al., 2017[70]	Raw fragments	*k*-mer-based Logistic regression model with lasso regularisation machine learning	Outperforms VirSorter Do not need to assemble metagenomes before using tool	Model limited to learning from training data before 1st January 2014 so may not be appropriate for recently discovered viral sequences, and no option in software to retrain on alternative training data, Only testsed algorithm on simulated metagenomes Need to filter out eukaryotic host sequences, as may mis-classify as viral
	VirMiner	Zheng et al., 2019[72]	Raw fragments	Random forest machine learning on phage contigs	Validatesd algorithm and comparesd with VirSorter and VirFinder using metagenomic data from human gut samples. Better sensitivity than and similar specificity to VirSorter and VirFinder Also extends the pipeline to include raw read processing and assembly, sequence and functional annotation of phage contigs, and phage-host prediction using CRISPR-spacer recognition, and two-group comparison (e.g. case and control) User-friendly website	Does not have a command-line or API tool, making it difficult to analyse multiple metagenomes No option in software to use alternative tools in pipeline or retrain random forest on alternative training data
	ViraMiner	Tampuu et al., 2019[73]	Contigs	Deep Learning using Convolutional Neural Networks	Model can be retrained on alternative data unlike MARVEL or VirFinder	Does not directly compare performance against other tools The accuracy of the model on human metagenomic contigs is likely to be an overestimate because reference-based alignment is used to benchmark these contigs that would likely contain many false negatives
Plasmid	Recycler	Rozov et al., 2016[75]	Raw fragments	Circular de Bruijn graphs with coverage filters	Even though lack of metagenome benchmark, tool comparesd plasmid prediction from cow rumen metagenomic data[107] with plasmids extracted usingfrom PCR validation from a previous study[32]	Ignores linear plasmids, and those integrated in chromosomes Performance metrics, i.e. precision and recall, only calculated from applying to a Recycler simulated plasmidome, not whole metagenomes Only 35% of plasmid predictions from metagenomes matched plasmids reported in PCR validation
	cBar	Zhou and Xu, 2010[74]	Contigs	Sequential minimal optimization-based model on pentamer frequencies	First tool that attempts to distinguish plasmids from chromosomal DNA from whole metagenomes	Achieves 88.29% accuracy with independent test set. but dDoes not describe how the independent test set was generated. Does not attempt to bin plasmids.
	PlasFlow	Krawczyk et al., 2018[40]	Contigs	Machine learning model trained using a deep neural network on genome signatures	Outperforms cBar on plasmidome data	Applied and Ccomparesd PlasFlow to cBar, Recycler and PlasmidFinder on whole metagenomes, but could not evaluate performance Assemblies required to be longer than 1 kb
	metaplasmidSPAdes	Antipov et al., 2019[76]	Raw fragments	Circular assembly graphs with coverage filters. Includes a verification tool, plasmidVerifty, which uses a naive Bayesian classifier on plasmid-specific profile-HMMs	plasmidVerify outperforms cBar and PlasFlow annotation of custom a plasmid and non-plasmid sequences from RefSeq Generally identifiesd more plasmids than Recycler using metagenomic data, mock data, multiple genomic isolates and plasmidome data	Ignores linear plasmids

Table 1: Published tools for de novo MGE discovery intended for whole metagenomes