MICROBIAL GENOMICS

RESEARCH ARTICLE

Maguire *et al.*, *Microbial Genomics* 2020;6
DOI 10.1099/mgen.0.000436

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands

Finlay Maguire[1]†, Baofeng Jia[2]†, Kristen L. Gray[2], Wing Yin Venus Lau[2], Robert G. Beiko[1] and Fiona S. L. Brinkman[2,*]

Abstract

Metagenomic methods enable the simultaneous characterization of microbial communities without time-consuming and bias-inducing culturing. Metagenome-assembled genome (MAG) binning methods aim to reassemble individual genomes from this data. However, the recovery of mobile genetic elements (MGEs), such as plasmids and genomic islands (GIs), by binning has not been well characterized. Given the association of antimicrobial resistance (AMR) genes and virulence factor (VF) genes with MGEs, studying their transmission is a public-health priority. The variable copy number and sequence composition of MGEs makes them potentially problematic for MAG binning methods. To systematically investigate this issue, we simulated a low-complexity metagenome comprising 30 GI-rich and plasmid-containing bacterial genomes. MAGs were then recovered using 12 current prediction pipelines and evaluated. While 82–94% of chromosomes could be correctly recovered and binned, only 38–44% of GIs and 1–29% of plasmid sequences were found. Strikingly, no plasmid-borne VF nor AMR genes were recovered, and only 0–45% of AMR or VF genes within GIs. We conclude that short-read MAG approaches, without further optimization, are largely ineffective for the analysis of mobile genes, including those of public-health importance, such as AMR and VF genes. We propose that researchers should explore developing methods that optimize for this issue and consider also using unassembled short reads and/or long-read approaches to more fully characterize metagenomic data.

DATA SUMMARY

In keeping with FAIR principles (Findable, Accessible, Interoperable, Reusable data), all analyses presented in this paper can be reproduced and inspected with the associated GitHub repository (https://github.com/fmaguire/MAG_gi_plasmid_analysis) (10.5281/zenodo.4005062) and data repository (https://osf.io/nrejs/) (10.17605/OSF.IO/NREJS).

INTRODUCTION

Metagenomics, the sequencing of DNA from within an environmental sample, is widely used to characterize the functional potential and identity of microbial communities [1, 2]. These approaches have been instrumental in developing our understanding of the distribution and evolutionary history of antimicrobial resistance (AMR) genes [3–5], as well as tracking pathogen outbreaks [6]. Although long-read DNA technologies [e.g. Oxford Nanopore Technologies' nanopore sequencing [7] and Pacific Biosciences' (PacBio) single-molecule real-time sequencing [8] platforms] are now being used for metagenomic sequencing [9, 10], high-throughput sequencing of relatively short reads (150–250 bp) on platforms such as the Illumina MiSeq still dominates metagenomics. These reads can be directly analysed using reference databases and a variety of homology search tools [11–14]. Since these reads are shorter than most genes, however, read-based methods provide very little information about genomic organization. This lack of contextual information is particularly problematic in the study of AMR genes and virulence factors (VFs), as the genomic context plays a role in function [15], selective pressures [16] and likelihood of lateral gene transfer (LGT) [17].

Received 23 May 2020; Accepted 04 September 2020; Published 01 October 2020
Author affiliations: [1]Faculty of Computer Science, Dalhousie University, 6050 University Avenue, Halifax, Nova Scotia, B3H 4R2, Canada; [2]Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada.
*Correspondence: Fiona S. L. Brinkman, brinkman@sfu.ca
Keywords: antimicrobial resistance; genomic islands; metagenomics; metagenome-assembled genomes; mobile genetic elements.
Abbreviations: AMR, antimicrobial resistance; GI, genomic island; LGT, lateral gene transfer; MAG, metagenome-assembled genome; MGE, mobile genetic element; VF, virulence factor.
†These authors contributed equally to this work
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Three supplementary tables and seven supplementary figures are available with the online version of this article.

Sequence assembly using specialized metagenomic de Bruijn graph assemblers (e.g. metaSPAdes [18], IDBA-UD [19] and megahit [20]) is often used to try to recover information about genomic context [21]. To disentangle the resulting mix of assembled fragments, there has been a move to group these contigs based on the idea that those from the same source genome will have similar relative abundance and sequence composition [22]. These resulting groups or 'bins' are known as metagenome-assembled genomes (MAGs). A range of tools have been released to perform this binning, including CONCOCT [23], MetaBAT2 [24], MaxBin2 [25] and a tool that combines their predictions: DAS Tool [26]. These MAG binning methods have been used successfully in unveiling previously uncharacterized genomic diversity [27–29], but metagenomic assembly and binning has been shown to involve the loss of some information. This means as little as 24.2–36.4% of reads [30, 31] and ~23% of genomes [31] are successfully assembled and binned in some metagenomic analyses. The Critical Assessment of Metagenome Interpretation (CAMI) challenge's (https://data.cami-challenge.org/) Assessment of Metagenome BinnERs (AMBER) [32] benchmarks different MAG recovery methods in terms of global completeness and bin purity. Similarly, a recent study has also used the AMBER approach to evaluate 15 different binning methods applied to a common metaSPAdes assembly [33]. However, to the best of our knowledge, there has not been a specific assessment of MAG-based recovery of mobile genetic elements (MGEs) such as genomic islands (GIs) and plasmids, despite their health and research importance.

GIs are clusters of chromosomal genes that are known or predicted to have been acquired through LGT events. GIs can arise following the integration of MGEs, such as integrons, transposons, integrative and conjugative elements (ICEs), and prophages (integrated phages) [34, 35]. They are of high interest since VFs are disproportionately associated with these mobile sequences [36], as are certain AMR genes [37, 38]. GIs often have differing nucleotide composition compared to the rest of the genome [34], a trait exploited by GI prediction tools such as SIGI-HMM [39] and IslandPath-DIMOB [40], and integrative tools like IslandViewer [41]. GIs may also exist as multiple copies within a genome [42], leading to potential assembly difficulties and biases in the calculation of coverage statistics.

Plasmids are circular or linear extrachromosomal self-replicating pieces of DNA with variable copy numbers and repetitive sequences [43, 44]. Similar to GIs, the sequence composition (including G+C content, dinucleotide bias, etc.) of plasmids are often markedly different from the genome with which they are associated [45–47]. Plasmids are also of high interest as a major source of the lateral dissemination of AMR genes throughout microbial ecosystems [37, 48].

GIs and plasmids have proven particularly difficult to assemble from short-read sequencing data. Due to the history of their integration at specific insertion sites, GIs are commonly flanked by direct repeats [49, 50]. Repetitive sequences are known to complicate assembly from short reads, with repeats often found at contig break sites [51]. Given that assembly of closely related genomes in a metagenome is already challenging [52], the polymorphic nature of GIs and the known presence of flanking repeats would be expected to compound these separate assembly issues. Repeats also inhibit the assembly of plasmids from short-read sequencing data, particularly for longer plasmid sequences [53]. Additionally, the varying sequence composition and relative abundance features mean that GIs and plasmids pose significant challenges in MAG recovery.

Impact Statement

Metagenome-assembled genome (MAG) binning has become an increasingly common approach in environmental, microbiome and public-health studies that makes use of short-read metagenomic data. By examining 12 widely used MAG binning workflows, we demonstrate that these methods are not suitable for the analysis of mobile genetic elements. Given the potential human and animal health implications of antimicrobial resistance and virulence genes associated with these elements, inappropriate use of short-read MAGs has the potential to be misleading at best and harmful at worst. This result will hopefully stimulate a shift in MAG methods to focus on developing approaches optimized for these elements, as well as incorporating additional read-based and long-read analyses.

As these MGEs are key to the function and spread of pathogenic traits such as AMR and virulence, and with MAG approaches becoming increasingly popular within microbial and public-health research, it is both timely and vital that we assess the impact of metagenome assembly and binning on the recovery of these elements. Therefore, to address this issue, we performed an analysis of the recovery accuracy of GI and plasmid sequences, and associated AMR/VF genes, across a set of 12 state-of-the-art methods for short-read metagenome assemblies. We show that short-read MAG-based analyses alone are not suitable for the study of mobile sequences, including those of public-health importance.

METHODS

Metagenome simulation

Thirty RefSeq genomes were selected using IslandPath-DIMOB [40] GI prediction data collated into the IslandViewer database (www.pathogenomics.sfu.ca/islandviewer) [41] (Table S1, available with the online version of this article). The selected genomes and associated plasmids (listed in Table S2 and deposited at https://osf.io/nrejs/ under 'data/sequences') were manually selected to satisfy the following criteria: 10 genomes with 1–10 plasmids, 10 genomes with >10% of chromosomal DNA predicted to reside in GIs, and 10 genomes with <1% of chromosomal DNA predicted to reside in GIs.

In accordance with the recommendation in the CAMI challenge [52], the genomes were randomly assigned a relative abundance following a log-normal distribution (μ=1, σ=2). Plasmid copy number estimates could not be accurately found for all organisms. Therefore, plasmids were randomly assigned a copy number regime, low (1–20), medium (20–100) or high (500–1000) at a 2:1:1 rate. Within each regime, the exact copy number was selected using an appropriately scaled gamma distribution (α=4, β=1) truncated to the regime range.

Finally, the effective plasmid relative abundance was determined by multiplying the plasmid copy number with the genome relative abundance. The full set of randomly assigned relative abundances and copy numbers can be found in Table S3. Sequences were then concatenated into a single FASTA file with the appropriate relative abundance. MiSeq v3 250 bp paired-end reads with a mean fragment length of 1000 bp (standard deviation of 50 bp) were then simulated using art_illumina (v2016.06.05) [54], resulting in a simulated metagenome of 31174411 read pairs. The selection of relative abundance and metagenome simulation itself was performed using the 'data_simluation/simulate_metagenome.py' script.

MAG recovery

Reads were trimmed using sickle (v1.33) [55] resulting in 25682644 surviving read pairs. The trimmed reads were then assembled using three different metagenomic assemblers: metaSPAdes (v3.13.0) [18], IDBA-UD (v1.1.3) [19] and megahit (v1.1.3) [20]). The resulting assemblies were summarized using metaQUAST (v5.0.2) [56]. The assemblies were then indexed and reads mapped back using bowtie2 (v2.3.4.3) [12].

Samtools (v1.9) was used to sort the read mappings, and the read coverage was calculated using the MetaBAT2 accessory script (jgi_summarize_bam_contig_depths). The three metagenome assemblies were then separately binned using MetaBAT2 (v2.13) [24] and MaxBin2 (v2.2.6) [25]. MAGs were also recovered using CONCOCT (v0.4.2) [23], following the recommended protocol in the user manual. Briefly, the supplied CONCOCT accessory scripts were used to cut contigs into 10 kb fragments (cut_up_fasta.py) and read coverage calculated for the fragments (CONCOCT_coverage_table.py). These fragment coverages were then used to bin the 10 kb fragments before the clustered fragments were merged (merge_cutup_clustering.py) to create the final CONCOCT MAG bins (extra_fasta_bins.py). Finally, for each metagenome assembly, the predicted bins from these three binners (MaxBin2, MetaBAT2 and CONCOCT) were combined using the DAS Tool (v1.1.1) meta-binner [26]. This resulted in 12 separate sets of MAGs (one set for each assembler and binner pair).

MAG assessment

Synthetic read coverage and depth

The trimmed synthetic reads were mapped back to each reference replicon using bowtie2 (v2.4.1), and sorted and indexed using Samtools (v1.10). The coverage of each reference replicon is calculated using 'samtools coverage' and the per base sequencing depth calculated using 'samtools depth'. The mean and per base depth are then plotted using R (v.3.4.2).

Chromosomal coverage

The MAG assessment for chromosomal coverage was performed by creating a BLASTN 2.9.0+ [57] database consisting of all the chromosomes of the input reference genomes. Each MAG contig was then used as a query against this database and the coverage of the underlying chromosomes tallied by merging the overlapping aligning regions and summing the total length of aligned MAG contigs. The most represented genome in each MAG was assigned as the 'identity' of that MAG for further analyses. Coverage values of less than 5% were filtered out and the number of different genomes to which contigs from a given MAG aligned were tallied. Finally, the overall proportion of chromosomes that were not present in any MAG was tallied for each binner and assembler.

In order to investigate the impact of the presence of closely related genomes in the metagenome on the ability to bin chromosomes, we generated a phylogenetic tree for all the input genomes. Single copy universal bacterial proteins were identified in the reference genomes using BUSCO v4.0.2 with the Bacteria Odb10 data [58]. The 86 of these proteins that were found in every reference genome were concatenated and aligned using MAFFT v7.427 [59] and masked with trimal v1.4.1–3 [60]. A maximum-likelihood phylogeny was then inferred with IQ-Tree v1.6.12 [61] using 1000 ultrafast-bootstraps and the in-built ModelFinder determined partitioning [62]. The phylogeny was then visualized using the Interactive Tree of Life (iTOL) v4 [63]. Pairwise branch distances were extracted from the resulting tree using ETE3 v3.1.1 [64] and regressed using a linear model against coverage (via seaborn v0.10.0 [65]) and using a Poisson logistic regression model (via statsmodel v0.12.0 [66]) against contamination. R^2 and McFadden's pseudo-R^2 were calculated for each model using the statsmodel library.

Plasmid and GI coverage

Plasmid and GI coverage were assessed in the same way. Firstly, a BLASTN database was generated for each set of MAG contigs. Then, each MAG database was searched for plasmid and GI sequences with greater than 50% coverage. All plasmids or GIs that could be found in the unbinned contigs or MAGs were recorded as having been successfully assembled. The subset of these that were found in the binned MAGs was then separately tallied. Finally, we evaluated the proportion of plasmids or GIs that were correctly assigned to the bin that was maximally composed of chromosomes from the same source genome.

AMR and VF assessment

Detection of AMR/VF genes

For the reference genomes, as well as 12 sets of MAGs, prodigal [67] was used to predict ORFs using the default

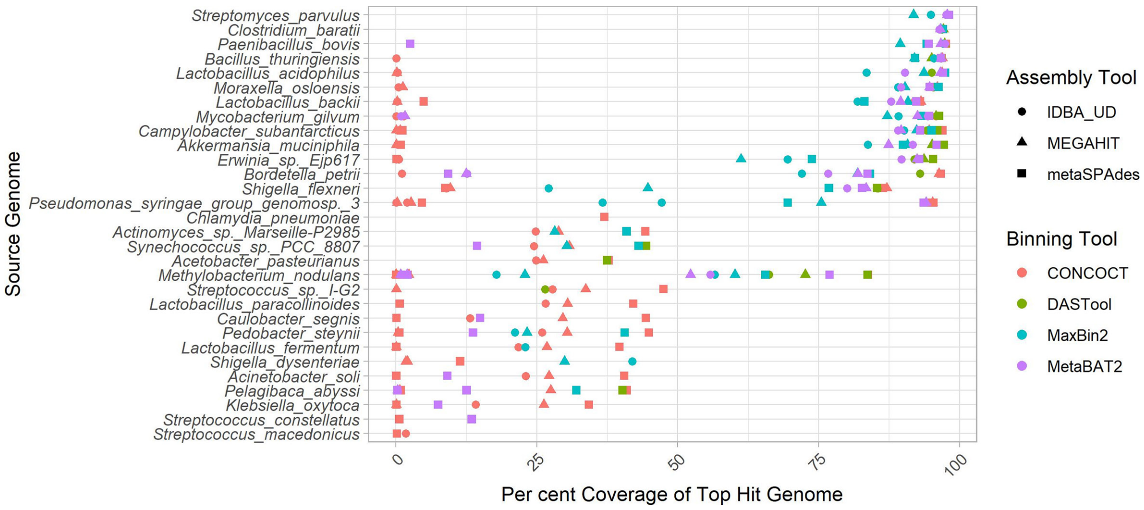

Fig. 1. Top genome coverage for input genomes across MAG binners. Each dot represents the coverage of a specified genome when it comprised the plurality of the sequences in a bin. If a genome did not form the plurality of any bin for a specific binner-assembler pair, no dot was plotted for that genome and binner-assembler. The binning tool is indicated by the colour of the dot as per the legend. Genomes such as *Clostridium baratii* were accurately recovered across all binner-assembler combinations, whereas genomes such as *Streptococcus macedonicus* were systematically poorly recovered.

parameters. AMR genes were predicted using Resistance Gene Identifier (RGI v5.0.0; default parameters) and the Comprehensive Antibiotic Resistance Database (CARD v3.0.3) [68]. VFs were predicted using the predicted ORFs and BLASTX 2.9.0+ [57] against the Virulence Factor Database (VFDB; obtained on August 26 2019) with an *E* value cut-off of 0.001 and a minimum identity of 90% [69]. Each MAG was then assigned to a reference chromosome using the above-mentioned mapping criteria for downstream analysis.

AMR/VF gene recovery

For each MAG set, we counted the total number of AMR/VF genes recovered in each metagenomic assembly and each MAG, and compared this to the number predicted in their assigned reference chromosome and plasmids. We then assessed the ability for MAGs to correctly bin AMR/VF genes of chromosomal, plasmid and GI origin by mapping the location of the reference replicon's predicted genes to the location of the same genes in the MAGs.

RESULTS

Recovery of genomic elements

Chromosomes

The overall ability of MAG methods to recover the original chromosomal source genomes varied widely. We considered the 'identity' of a given MAG bin to be that of the genome that comprises the largest proportion of sequence within that bin. In other words, if a bin is identifiably 70% species A and 30% species B, we consider that to be a bin of species A. Ideally, we wish to generate a single bin for each source genome consisting of the entire genome and no contigs from other genomes. Some genomes are cleanly and accurately binned regardless of the assembler and binning method used (Fig. 1). Specifically, greater than 90% of *Streptomyces parvulus* (minimum 91.8%) and *Clostridium baratii* (minimum 96.4%) chromosomes are represented in individual bins across all methods. However, no other genomes were consistently recovered at >30% chromosomal coverage across methods. The three *Streptococcus* genomes were particularly problematic with the best recovery for each ranging from 1.7–47.49%. Contrary to what might be expected, the number of close relatives to a given genome in the metagenome did not clearly affect the MAG coverage (Fig. S2).

In terms of the impact of different metagenome assemblers, megahit resulted in the highest median chromosomal coverage across all binners (81.9%), with metaSPAdes performing worst (76.8%) (Fig. 2a). Looking at binning tools, CONCOCT performed very poorly with a median 26% coverage for top hit per bin, followed by MaxBin2 (83.1%) and MetaBAT2 (88.5%). It is perhaps unsurprising that the best-performing binner in terms of bin top hit coverage was the metabinner DAS Tool that combines predictions from the other three binners (94.3% median top hit chromosome coverage per bin; Fig. 2a).

Bin purity, i.e. the number of genomes present in a bin at >5% coverage, was largely equivalent across assemblers, with a very marginally higher purity for IDBA-UD. Across binning tools, MaxBin2 proved an exception with nearly twice as many bins containing multiple species as the next binner (Fig. 2b). The remaining binning tools were largely equivalent, producing chimeric bins at approximately the same rates. Similar to coverage, there was a weak to non-existent relationship between bin purity and the number of

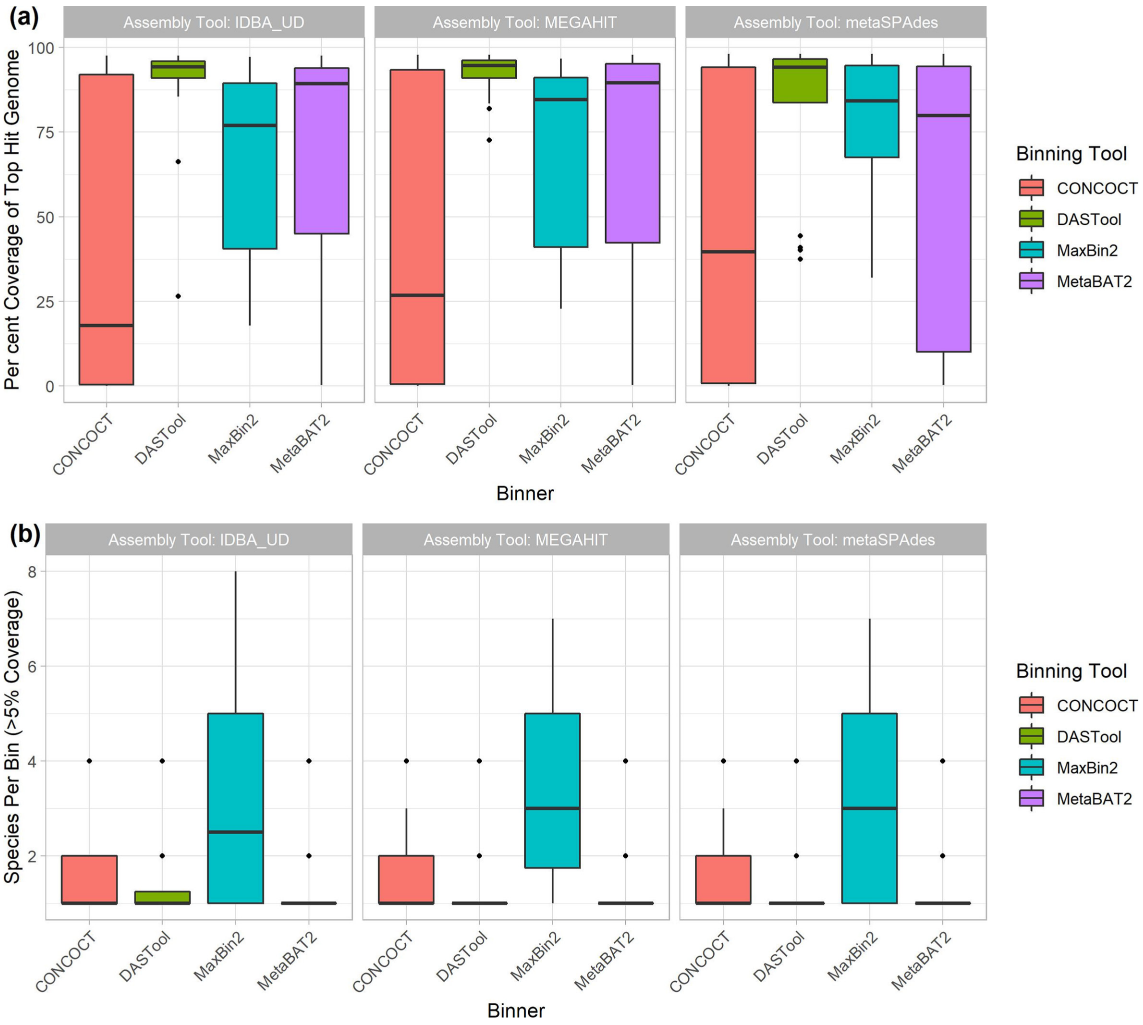

Fig. 2. Overall binning performance for every combination of metagenome assembler (as indicated by pane titles) and MAG binning tool (*x*-axis and legend colours). Diamonds in the plots represent outliers (greater or lower than the interquartile range marked by the error bars), and the boxes represent the lower quartile, median and upper quartile. (a) Chromosomal coverage of the most prevalent genome in each bin across binners and metagenome assemblies. Of the three assemblers, megahit resulted in the highest median chromosomal coverage (*y*-axis) across all binners (coloured bars) at 81.9%, with metaSPAdes performing the worst (76.8%). Of the four binners, CONCOCT (red) performed poorly with a median coverage, followed by MaxBin2 (blue), MetaBAT2 (purple) and DAS Tool (green) performing the best. (b) Distribution of bin purity across assemblers and binners. The total number of genomes present in a bin at >5% coverage (*y*-axis) was largely equivalent across assemblers (*x*-axis). For the binning tools, MaxBin2 (blue) produced nearly twice as many bins containing multiple species compared to CONCOCT (red), MetaBAT2 (purple) and DAS Tool (green), which all produced chimeric bins at roughly the same rate.

closely related genomes in the metagenome (Fig. S3). There was also not a clear relationship between coverage of a bin and purity, with low purity but high coverage bins observed, as well as high purity but low coverage bins.

Plasmids

Regardless of method, a very small proportion of plasmids were correctly grouped in the bin that was principally composed of chromosomal contigs from the same source genome. Specifically, between 1.5% (IDBA-UD assembly with DAS Tool bins) and 29.2% (metaSPAdes with CONCOCT bins) were correctly binned at over 50% coverage. In terms of metagenome assembly, metaSPAdes was by far the most successful assembler at assembling plasmids, with 66.2% of plasmids identifiable at greater than 50% coverage. IDBA-UD performed worst with 17.1% of plasmids recovered, and megahit recovered 36.9%. If the plasmid was successfully assembled it was, with one exception, placed in a MAG bin by MaxBin2 and CONCOCT, although a much smaller number were correctly binned (typically less than one third). Interestingly, the MetaBAT2 and DAS Tool binners were more conservative in assigning plasmid contigs to bins; but of those assigned to bins, nearly all were correctly binned (Fig. 3).

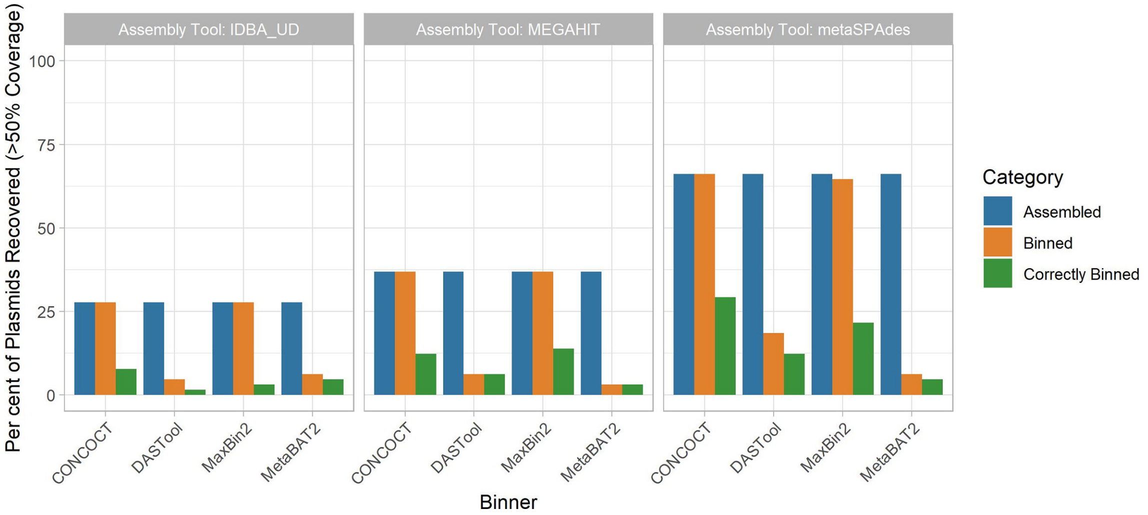

Fig. 3. The performance of metagenomic assembly and binning in recovery of plasmid sequences. Each plot represents a different metagenome assembler, with the groups of bars along the *x*-axes showing the plasmid recovery performance of each binning tool when applied to the assemblies produced by that tool. For each of these 12 assembler-binner-pair-produced MAGs, the grouped bars from left to right show the percentage of plasmids assembled, assigned to any bin and binned with the correct chromosomes. These stages of the evaluation are indicated by the bar colours as per the legend. Across all tools the assembly process resulted in the largest loss of plasmid sequences and only a small proportion of the assembled plasmids were correctly binned.

GIs

GIs were poorly assembled and binned across methods (Fig. 4). Unlike for plasmids, the performance of different methods was generally less variable, with no clear best-performing method. Assembly of GIs with >50% coverage was consistently poor (37.8–44.1%), with metaSPAdes outperforming the other two assembly approaches. For the CONCOCT and MaxBin2 binning tools, all GIs that were assembled were assigned to a bin, although the proportion of binned GIs that were correctly binned was lower than for DAS Tool and MetaBAT2. DAS Tool, MetaBAT2 and CONCOCT did not display the same precipitous drop between those assembled and those correctly binned as was observed for plasmids. In terms of overall correct binning with the chromosomes from the same genome, the metaSPAdes assembly with CONCOCT (44.1%) and MaxBin2 (43.3%) binners performed best.

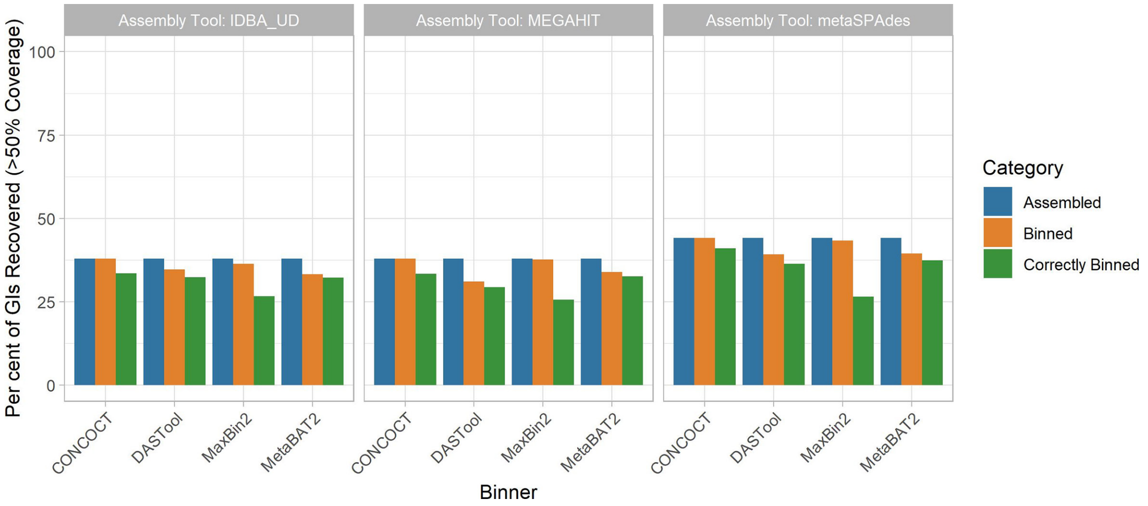

Fig. 4. Impact of metagenomic assembly and MAG binning on recovery of GIs. GIs were recovered in a similarly poor fashion to plasmids. Regardless of binning (x-axis) and assembly (panel) methods, <40% of GIs were correctly assigned to the correct source genome. MaxBin2 and CONCOCT placed GIs in a bin the majority of the time (orange); however, a very small fraction was correctly binned (green). Generally, GIs were correctly binned better than plasmids with DAS Tool, MetaBAT2 and CONCOCT.

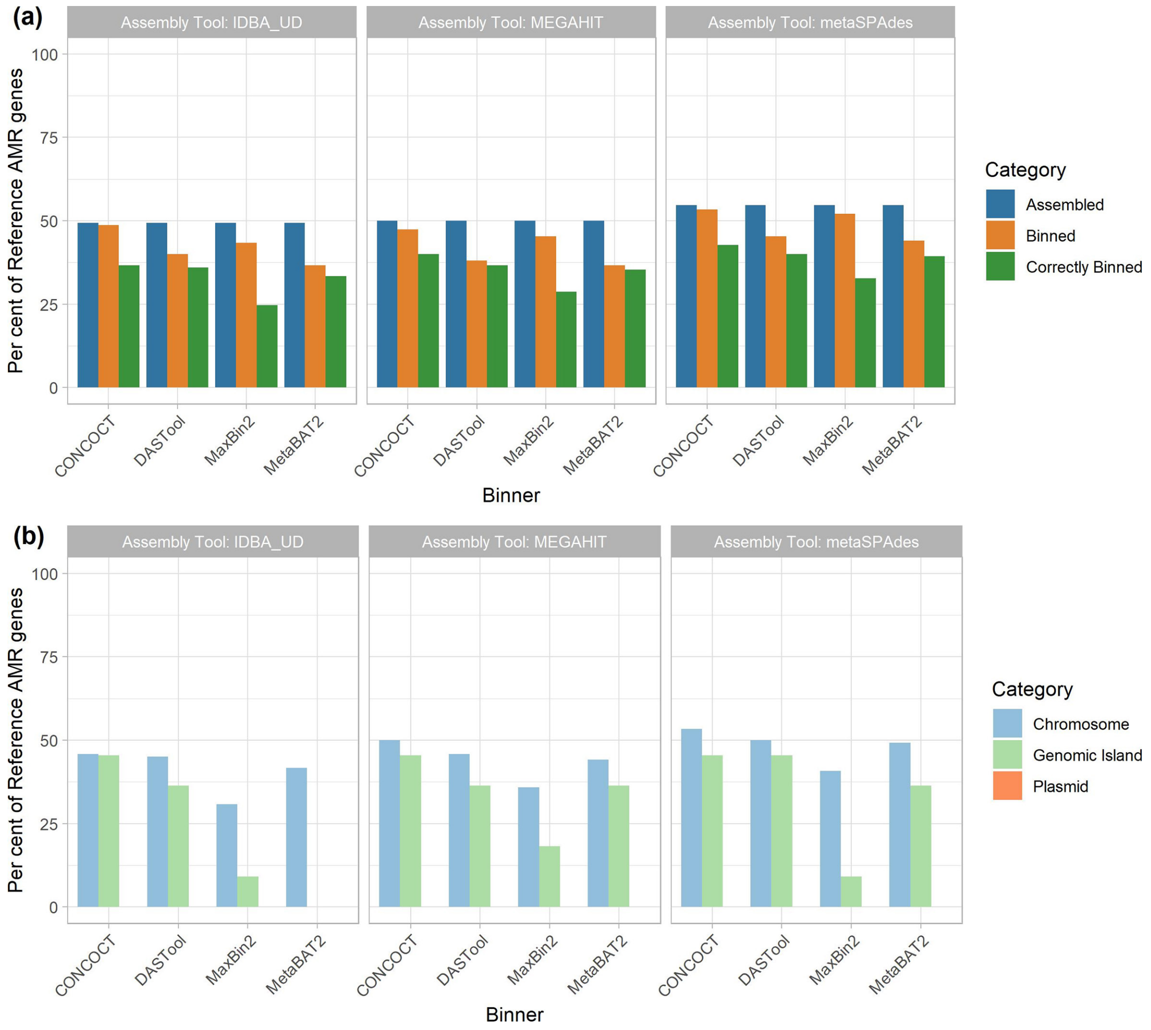

Fig. 5. Recovery of AMR genes across assemblers, binners and genomic context. (a) The proportion of reference AMR genes recovered (*y*-axis) was largely similar across assembly tools (panels as indicated by title) at roughly 50%, with `metaSPAdes` performing marginally better overall. Binning tools (*x*-axis) resulted in a small reduction in AMR genes recovered (orange); however, only 24–40% of all AMR genes were correctly binned (green). `metaSPAdes-CONCOCT` was the best performing MAG binning pipeline. (b) Per cent of correctly binned AMR genes recovered by genomic context. MAG methods were best at recovering chromosomally located AMR genes (light blue) regardless of metagenomic assembler or binning tool used. Recovery of AMR genes in GIs showed a bigger variation between tools (light green). None of the 12 evaluated MAG recovery methods were able to recover plasmid-located AMR genes.

AMR genes

The recovery of AMR genes in MAGs was poor with only ~49–55% of all AMR genes predicted in our reference genomes regardless of the assembly tool used, and `metaSPAdes` performing marginally better than other assemblers (Fig. 5a). Binning the contigs resulted in a ~1–15% loss in AMR gene recovery with the `concoct-metaSPAdes` pair performing best at only 1% loss and `DAS Tool-megahit` performing the worst at 15% reduction of AMR genes recovered. Overall, only 24–40% of all AMR genes were correctly binned. This was lowest with the `MaxBin2-IDBA-UD` pair (24%) and highest in the `CONCOCT-metaSPAdes` pipeline (40%).

Moreover, focusing on only the AMR genes that were correctly binned (Fig. 5b), we can evaluate the impact of different genomic contexts (i.e. chromosomal, plasmid, GI). Across all methods only 30–53% of all chromosomally located AMR genes (n=120), 0–45% of GI located AMR genes (n=11) and none of the plasmid-localized AMR genes (n=20) were correctly binned.

VF genes

We also examined the impact of MAG approaches on recovery of VF genes as identified using the Virulence Factor Database (VFDB). We saw a similar trend as AMR genes (Fig. 6a). Between 56 and 64% of VFs were identifiable in the metagenomic assemblies (with `megahit` recovering the greatest proportion). The binning process further reduced the number of recovered VFs by 4–26%,

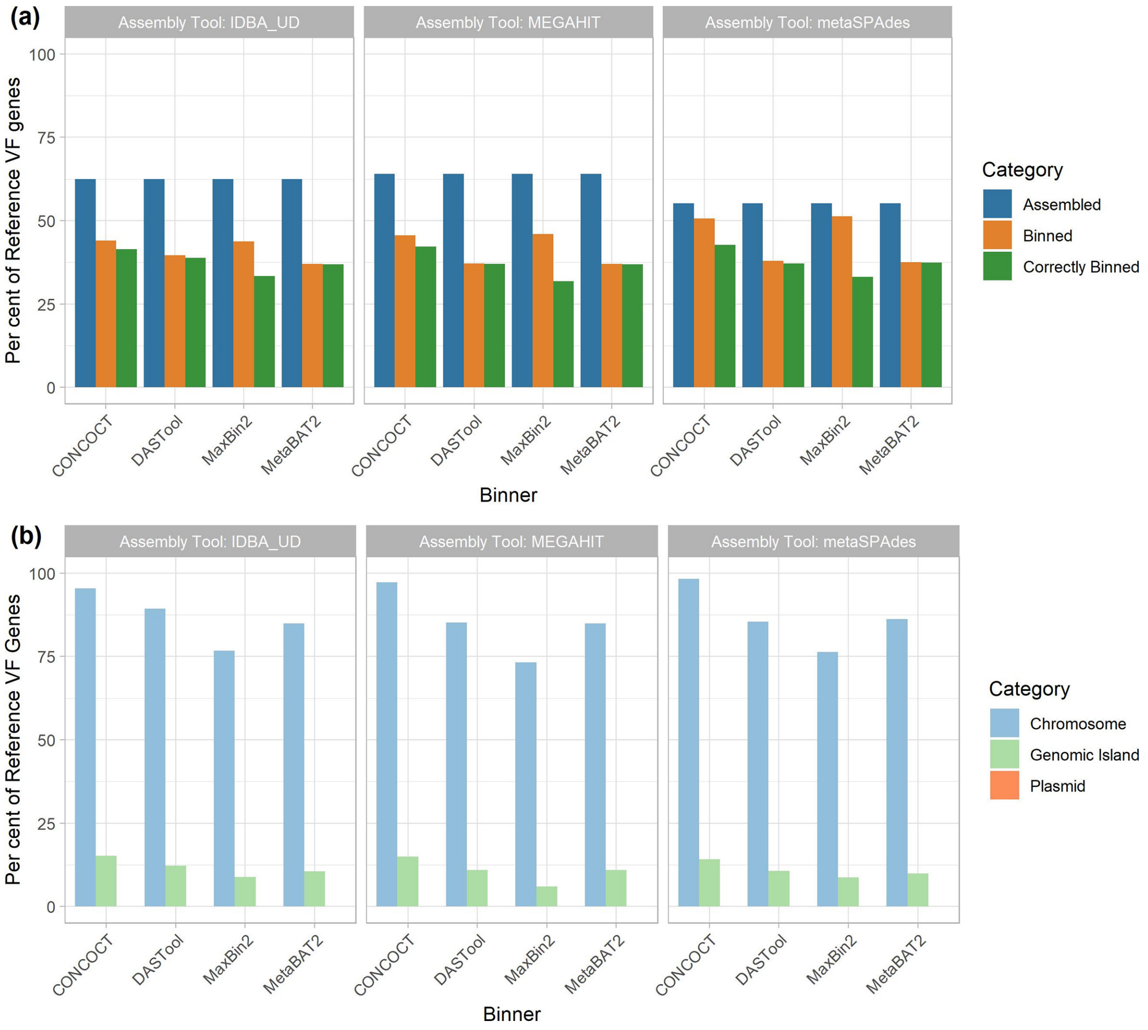

Fig. 6. Recovery of VF genes across assemblers, binners and genomic context. (a) Per cent of reference VF genes recovered across assemblers and binners. The proportion of reference VF genes recovered (*y*-axis) exhibited a similar trend as AMR genes. Recovery was greatest after the assembling stage (blue), with megahit performing best. Binning tools resulted in a larger reduction in VF genes recovered (orange) compared to AMR genes. However, in the majority of cases, VF genes that were binned were correctly binned (green). metaSPAdes-CONCOCT was again the best performing pair. (b) Per cent of correctly binned VF genes recovered in each genomic region. MAGs were again best at recovering chromosomally located VF genes (light blue), and able to correctly bin the majority of chromosomally located VFs. Again, there was very poor performance in terms of the recovery of GIs (light green), and none of the plasmid-located AMR genes (orange) were correctly binned.

with DAS Tool-megahit performing the worst (26% reduction) and CONCOCT-metaSPAdes performing the best (4% reduction). Unlike AMR genes, the majority of VF genes assigned to a bin were assigned to the correct bin (i.e. that bin largely made up of contigs from the same input genome). Overall, CONCOCT-metaSPAdes again performed best with 43% of all VFs correctly assigned.

As with AMR genes, the genomic context (chromosome, plasmid, GI) of a given VF largely determined how well it was binned (Fig. 6b). The majority (73–98%) of all chromosomally located VF genes (*n*=757) were correctly binned. However, 0–16% of GI-localized VF genes (*n*=809) and again none of the plasmid-associated VF genes (*n*=3) were recovered across all 12 MAG pipelines.

Comparisons of rates of loss

We combined the performance metrics for Figs 3–6 to compare the rates of loss of different components (Fig. S5). This highlighted that genomic components (GIs and plasmids) and plasmids in particular are lost at a disproportionately higher rate than individual gene types during MAG recovery. This also emphasizes that better metagenomic assembly does not necessarily result in better binning/recovery of GIs and plasmids.

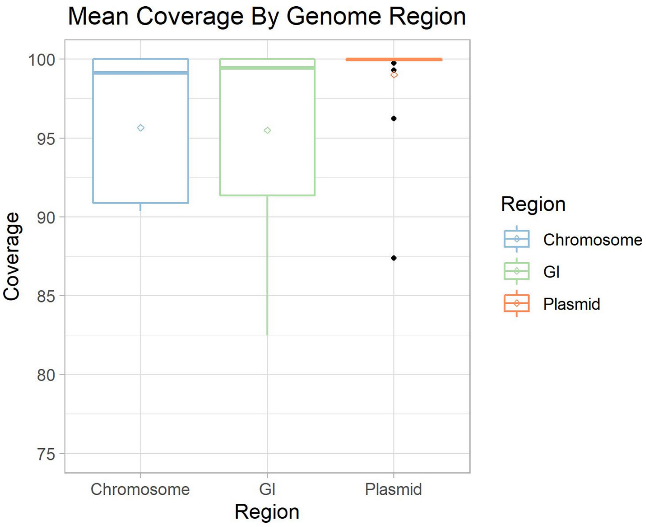

Fig. 7. Mean coverage by genomic region. The mean coverage of our synthetic reads to their source genome is plotted by their genomic region. Chromosome (blue) and GI (green) displayed a similar mean coverage of ~96.5%. Plasmids (orange) had a higher mean coverage at ~98%. The per genome coverage range of plasmids and GIs are higher than that of chromosomes. Diamond dots indicate the mean coverage of a region and black dots indicate outliers.

Simulated read analysis

To further explore the potential causes of poor assembly and binning of MGEs, we analysed the resultant coverage distribution from mapping our synthetically generated reads back to the original chromosomes, GIs and plasmids from which they were simulated. This analysis identified that while coverage of our synthetic metagenome was consistently between 91–100% across all reference chromosomes, the coverage of GIs and plasmids encompassed a larger range (Fig. 7). Inspecting individual genomes shows large spikes and drops in coverage and per base read depth in and around these elements (Figs S6 and S7). This variability in coverage might be attributed to repeated elements and the sequence composition differences that are commonly associated with MGEs. This issue is likely responsible for failures to accurately estimate the read-depth/coverage in these regions, upon which both assembly (in traversal of the assembly graph) and binning rely.

DISCUSSION

In this paper, we evaluated the ability of MAG binning methods to correctly recover MGEs (i.e. GIs and plasmids) from metagenomic samples. Overall, chromosomal sequences were binned well (up to 94.3% coverage, with perfect bin purity using `megahit-DAS Tool`). The presence of closely related genomes had unclear impacts on the coverage and cross-contamination of bins (e.g. *Streptococcus* species in Figs S2 and S3). Additionally, the trade-off between false positives and sensitivity in the binning of closely related sequences is an area in need of further exploration.

Given the importance of MGEs in the function and spread of virulence traits and AMR, it is particularly noteworthy that regardless of MAG binning method, plasmids and GIs were disproportionately lost compared to core chromosomal regions. At best (with `metaSPAdes` and `CONCOCT`), 29.2% of plasmids and 44.1% of GIs were identifiable at >50% coverage in the correct bin (i.e. grouped with a bin that was mostly made up of contigs from the same genome). While some MGEs were likely recovered in more partial forms (<50% coverage), use of these by researchers interested in selective pressures and LGT could lead to inaccurate inferences. This poor result is congruent with the intuition that the divergent compositional features and repetitive nature of these MGEs is problematic for MAG methods (a conclusion supported by the observed high coverage and read-depth variability of MGEs when mapping simulated reads back to the original genomes). The particularly poor plasmid binning performance is likely attributable to the known difficulties in assembly of plasmids from short-read data [53]. Therefore, binning efficiency might improve with use of long-read sequencing or assembly methods optimized for the assembly and binning of plasmid sequences [53] (such as SCAPP [70]). Incorporating long-read data has been shown to improve overall MAG binning [71] and facilitate metagenomic characterization of plasmids [72]. However, the low throughput and high error rate of current long-read technologies relative to widely used short-read approaches present a challenge when characterizing MGEs in metagenomes, especially those of greater complexity. Further research is needed to fully characterize the performance of different long-read protocols and analytical approaches (including hybrid approaches with short-reads) on the accuracy of recovering MGEs in metagenomic samples.

With the growing use of MAG methods in infectious disease research [73–77] and the public-health and agri-food importance of the LGT of AMR and VF genes, we also specifically evaluated the binning of these gene classes. The majority of these genes were correctly assembled across assemblers, but were either not assigned or incorrectly assigned to MAG bins during binning. At best across all binners, 40% of all AMR genes and ~63% of VF genes (`CONCOCT-metaSPAdes`) present in the reference genomes were assigned to the correct MAG. While a majority of chromosomally located VF genes (73–98%) and AMR genes (53%) were binned correctly, only 16% of GI VFs (*n*=809), 45% of GI AMR genes (*n*=11) and not a single plasmid-associated VF (*n*=3) or AMR gene (*n*=20) were correctly binned. This included critical high-threat MGE-associated AMR genes such as oxacillinases (OXA) and *Klebsiella pneumoniae* carbapenemases (KPC). One potential caveat of this is that some AMR genes and VFs may no longer be detectable in MAGs due to issues with ORF prediction (see Supplementary Information and Fig. S4). We also observed a higher variability in per base read depth and range of coverage in MGEs (Figs 7, S6 and S7). This, combined with previous studies observing fragmented ORF predictions in draft genomes, can lead to downstream over- or under-annotation with functional labels depending on the approach used [78]. Although not yet developed, methods that combine the assembly/binning pipelines tested here with read-based inference would provide a better sense of which functions are potentially being missed by the MAG reconstructions.

Our simulated metagenomic community comprised 30 distinct bacterial genomes with varying degrees of relatedness. While this diversity can be representative of certain clinical samples [79–81], other environments with relevance to public health, such as the human gut, soil and livestock, can have 100–1000s of species [82–85]. In addition, MGEs such as GIs and plasmids are known to recombine, producing closely related variants [86–88] that could further complicate assembly from a metagenomic sample. Polymorphic MGEs were not explicitly introduced in our simulated metagenome. Consequently, our analysis likely over-represents the effectiveness of the methods tested in a public-health setting. Metagenomic simulation is also unlikely to perfectly represent the noise and biases in real metagenomic sequencing, but it does provide the ground-truth necessary for evaluation [32, 89]. This simulation approach, combined with the development of an MGE/AMR-focused mock metagenome (similarly to the mockrobiota initiative [90]), could provide a key resource to develop and validate new binning approaches and different sequencing strategies. Additionally, it would provide a way to further optimize parameter settings of existing metagenomic assembly and binning tools beyond the default settings used in these analyses (considered representative of most 'real-world' usage [91]) without overfitting to a particular metagenome.

This study has shown that while short-read MAG-binning approaches provide a useful tool to study a bacterial species' core chromosomal elements, they have severe limitations in the recovery of MGEs. The majority of these MGEs will either fail to be assembled or be incorrectly binned. The consequence of this is the disproportionate loss of key public-health MGE-associated VF and AMR genes that may be crucial markers for monitoring the spread of virulence and resistance among clinically important pathogens. As many of these clinically relevant genes have a high propensity for LGT between unrelated bacteria [36, 37], it is critical to highlight that short-read MAG approaches are currently insufficient to thoroughly profile them. Within public-health metagenomic research, as well as other research areas that study MGEs, it is vital we utilize MAGs in conjunction with other methods (e.g. targeted AMR assembly [92], long-read sequencing, plasmid optimized assembly [70] and read-based sequence homology search [11]) before drawing biological or epidemiological conclusions.

Funding information
This work was supported primarily by a Donald Hill Family Fellowship held by F.M. W.Y.V.L. and B.J. received Canadian Institutes of Health Research (CIHR) doctoral scholarships. K.G. was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Collaborative Research and Training Experience (CREATE) Bioinformatics scholarship. B.J, W.Y.V.L. and K.G. also held Simon Fraser University (SFU) Omics and Data Sciences fellowships. F.S.L.B. is an SFU Distinguished Professor, and R.G.B. is a Professor and Associate Dean Research at Dalhousie University. Additionally, this work was partially supported by Genome Canada and NSERC grants to R.G.B. and F.S.L.B.

Acknowledgements
The authors would like to thank their funders, and the Simon Fraser University (SFU) Research Computing Group and Compute Canada for compute resource support.

Author contributions
F.M. and B.J.: conceptualization, investigation, validation, formal analysis, data curation, writing (original draft preparation – review and editing), visualization. W.Y.V.L. and K.G.: investigation, data curation writing (review and editing). F.S.L.B and R.G.B.: supervision, project administration, funding, writing (review and editing). All authors contributed to and approved the manuscript.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM *et al*. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 2002;99:14250–14255.
2. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 2017;35:833–844.
3. Donia MS, Cimermancic P, Schulze CJ, Wieland Brown LC, Martin J *et al*. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* 2014;158:1402–1414.
4. D'Costa VM, Griffiths E, Wright GD. Expanding the soil antibiotic resistome: exploring environmental diversity. *Curr Opin Microbiol* 2007;10:481–489.
5. D'Costa VM, King CE, Kalan L, Morar M, Sung WWL *et al*. Antibiotic resistance is ancient. *Nature* 2011;477:457–461.
6. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ-M *et al*. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA* 2013;309:1502.
7. Mikheyev AS, Tin MMY. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour* 2014;14:1097–1102.
8. Eid J, Fehr A, Gray J, Luong K, Lyle J *et al*. Real-time DNA sequencing from single polymerase molecules. *Science* 2009;323:133–138.
9. Nicholls SM, Quick JC, Tang S, Loman NJ. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* 2019;8:giz043.
10. Somerville V, Lutz S, Schmid M, Frei D, Moser A *et al*. Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiol* 2019;19:143.
11. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12:59–60.
12. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
13. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 2013;29:2487–2489.
14. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 2015;16:236.
15. Xu X, Lin D, Yan G, Ye X, Wu S *et al*. vanM, a new glycopeptide resistance gene cluster found in *Enterococcus faecium*. *Antimicrob Agents Chemother* 2010;54:4643–4647.
16. Baker-Austin C, Wright MS, Stepanauskas R, McArthur JV. Co-selection of antibiotic and metal resistance. *Trends Microbiol* 2006;14:176–182.
17. Stokes HW, Gillings MR. Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into gram-negative pathogens. *FEMS Microbiol Rev* 2011;35:790–819.

18. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017;27:824–834.
19. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 2012;28:1420–1428.
20. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;31:1674–1676.
21. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ *et al*. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 2004;428:37–43.
22. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* 2019;20:1125–1136.
23. YY L, Chen T, Fuhrman JA, Sun F. COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics* 2016:btw290.
24. Kang DD, Li F, Kirton ES, Thomas A, Egan RS *et al*. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;7:e7359.
25. Y-W W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2016;32:605–607.
26. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M *et al*. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 2018;3:836–843.
27. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ *et al*. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 2015;523:208–211.
28. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ *et al*. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2017;2:1533–1542.
29. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R *et al*. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat Biotechnol* 2019;37:953–961.
30. Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB *et al*. Genome-centric view of carbon processing in thawing permafrost. *Nature* 2018;560:49–54.
31. Diamond S, Andeer PF, Li Z, Crits-Christoph A, Burstein D *et al*. Mediterranean grassland soil C–N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms. *Nat Microbiol* 2019;4:1356–1367.
32. Meyer F, Hofmann P, Belmann P, Garrido-Oter R, Fritz A *et al*. AMBER: assessment of metagenome BinnERs. *Gigascience* 2018;7:giy069.
33. Yue Y, Huang H, Qi Z, Dou H-M, Liu X-Y *et al*. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics* 2020;21:334.
34. Langille MGI, Hsiao WWL, Brinkman FSL. Detecting genomic islands using bioinformatics approaches. *Nat Rev Microbiol* 2010;8:373–382.
35. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. *Nat Rev Genet* 2015;16:472–482.
36. Ho Sui SJ, Fedynak A, Hsiao WWL, Langille MGI, Brinkman FSL. The association of virulence factors with genomic islands. *PLoS One* 2009;4:e8094.
37. von Wintersdorff CJH, Penders J, van Niekerk JM, Mills ND, Majumder S *et al*. Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front Microbiol* 2016;7:173.
38. Brown-Jaque M, Calero-Cáceres W, Muniesa M. Transfer of antibiotic-resistance genes via phage-related mobile elements. *Plasmid* 2015;79:1–7.
39. Merkl R. SIGI: score-based identification of genomic islands. *BMC Bioinformatics* 2004;5:22.
40. Bertelli C, Brinkman FSL. Improved genomic island predictions with IslandPath-DIMOB. *Bioinformatics* 2018;34:2161–2167.
41. Dhillon BK, Laird MR, Shay JA, Winsor GL, Lo R *et al*. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Res* 2015;43:W104–W108.
42. Bertelli C, Tilley KE, Brinkman FSL. Microbial genomic island discovery, visualization and analysis. *Brief Bioinform* 2019;20:1685–1698.
43. San Millan A, Escudero JA, Gifford DR, Mazel D, MacLean RC. Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat Ecol Evol* 2016;1:10.
44. San Millan A, Santos-Lopez A, Ortega-Huedo R, Bernabe-Balas C, Kennedy SP *et al*. Small-plasmid-mediated antibiotic resistance is enhanced by increases in plasmid copy number and bacterial fitness. *Antimicrob Agents Chemother* 2015;59:3335–3341.
45. Zhou F, Xu Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* 2010;26:2051–2052.
46. Davis JJ, Olsen GJ. Modal codon usage: assessing the typical codon usage of a genome. *Mol Biol Evol* 2010;27:800–810.
47. Daubin V, Lerat E, Perrière G. The source of laterally transferred genes in bacterial genomes. *Genome Biol* 2003;4:R57.
48. Holmes AH, Moore LSP, Sundsfjord A, Steinbakk M, Regmi S *et al*. Understanding the mechanisms and drivers of antimicrobial resistance. *Lancet* 2016;387:176–187.
49. Williams KP. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res* 2002;30:866–875.
50. Schmidt H, Hensel M. Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev* 2004;17:14–56.
51. Acuña-Amador L, Primot A, Cadieu E, Roulet A, Barloy-Hubler F. Genomic repeats, misassembly and reannotation: a case study with long-read resequencing of *Porphyromonas gingivalis* reference strains. *BMC Genomics* 2018;19:54.
52. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S *et al*. Critical assessment of metagenome interpretation — a benchmark of metagenomics software. *Nat Methods* 2017;14:1063–1071.
53. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom* 2017;3:e000128.
54. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics* 2012;28:593–594.
55. Joshi N, Fass J. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files. *GitHub;* 2011.
56. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 2016;32:1088–1090.
57. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.
58. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31:3210–3212.
59. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 2018;34:2490–2492.
60. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972–1973.
61. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;32:268–274.
62. Lanfear R, Calcott B, Ho SYW, Guindon S. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol* 2012;29:1695–1701.

63. Letunic I, Bork P. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.

64. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* 2016;33:1635–1638.

65. Waskom M, Botvinnik O, Ostblom J, Lukauskas S, Hobson P. mwaskom/seaborn: v0.10.0 (January 2020). Zenodo; 2020.

66. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference* 2010;–92–96.

67. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW *et al*. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119.

68. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M *et al*. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2020;48:D517-D525.

69. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res* 2019;47:D687–D692.

70. Pellow D, Zorea A, Probst M, Furman O, Segal A *et al*. SCAPP: an algorithm for improved plasmid assembly in metagenomes. *bioRxiv* 2020.

71. Giguere DJ, Bahcheli AT, Joris BR, Paulssen JM, Gieg LM *et al*. Complete and validated genomes from a metagenome. *bioRxiv* 2020:10.1101/2020.04.08.032540.

72. Suzuki Y, Nishijima S, Furuta Y, Yoshimura J, Suda W *et al*. Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut. *Microbiome* 2019;7:119.

73. Ravi A, Halstead FD, Bamford A, Casey A, Thomson NM *et al*. Loss of microbial diversity and pathogen domination of the gut microbiota in critically ill patients. *Microb Genom* 2019;5:e000293.

74. Liu Z, Klümper U, Liu Y, Yang Y, Wei Q *et al*. Metagenomic and metatranscriptomic analyses reveal activity and hosts of antibiotic resistance genes in activated sludge. *Environ Int* 2019;129:208–220.

75. Newberry E, Bhandari R, Kemble J, Sikora E, Potnis N. Genome-resolved metagenomics to study co-occurrence patterns and intraspecific heterogeneity among plant pathogen metapopulations. *Environ Microbiol* 2020;22:2693–2708.

76. Zhang Y, Kitajima M, Whittle AJ, Liu W-T. Benefits of genomic insights and CRISPR-Cas signatures to monitor potential pathogens across drinking water production and distribution systems. *Front Microbiol* 2017;8:2036.

77. Huang AD, Luo C, Pena-Gonzalez A, Weigand MR, Tarr CL *et al*. Metagenomics of two severe foodborne outbreaks provides diagnostic signatures and signs of coinfection not attainable by traditional methods. *Appl Environ Microbiol* 2017;83:e02577-16.

78. Klassen JL, Currie CR. Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics* 2012;13:14.

79. Abayasekara LM, Perera J, Chandrasekharan V, Gnanam VS, Udunuwara NA *et al*. Detection of bacterial pathogens from clinical specimens using conventional microbial culture and 16S metagenomics: a comparative study. *BMC Infect Dis* 2017;17:631.

80. Rogers GB, Carroll MP, Serisier DJ, Hockey PM, Jones G *et al*. Characterization of bacterial community diversity in cystic fibrosis lung infections by use of 16S ribosomal DNA terminal restriction fragment length polymorphism profiling. *J Clin Microbiol* 2004;42:5176–5183.

81. Freitas AC, Chaban B, Bocking A, Rocco M, Yang S *et al*. The vaginal microbiome of pregnant women is less rich and diverse, with lower prevalence of Mollicutes, compared to non-pregnant women. *Sci Rep* 2017;7:9212.

82. Gołębiewski M, Deja-Sikora E, Cichosz M, Tretyn A, Wróbel B. 16S rDNA pyrosequencing analysis of bacterial community in heavy metals polluted soils. *Microb Ecol* 2014;67:635–647.

83. Youssef N, Sheik CS, Krumholz LR, Najar FZ, Roe BA *et al*. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl Environ Microbiol* 2009;75:5227–5236.

84. Claesson MJ, O'Sullivan O, Wang Q, Nikkilä J, Marchesi JR *et al*. Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS One* 2009;4:e6669.

85. Thomas M, Webb M, Ghimire S, Blair A, Olson K *et al*. Metagenomic characterization of the effect of feed additives on the gut microbiome and antibiotic resistome of feedlot cattle. *Sci Rep* 2017;7:12257.

86. Mulvey MR, Boyd DA, Olson AB, Doublet B, Cloeckaert A. The genetics of Salmonella genomic island 1. *Microbes Infect* 2006;8:1915–1922.

87. Arora SK, Wolfgang MC, Lory S, Ramphal R. Sequence polymorphism in the glycosylation island and flagellins of *Pseudomonas aeruginosa*. *J Bacteriol* 2004;186:2115–2122.

88. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M *et al*. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat Commun* 2020;11:3602.

89. Fritz A, Hofmann P, Majda S, Dahms E, Dröge J *et al*. CAMISIM: simulating metagenomes and microbial communities. *Microbiome* 2019;7:17.

90. Bokulich NA, Rideout JR, Mercurio WG, Shiffer A, Wolfe B *et al*. mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems* 2016;1:e00062-16.

91. Karimzadeh M, Hoffman MM. Top considerations for creating bioinformatics software documentation. *Brief Bioinform* 2018;19:693–699.

92. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J *et al*. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom* 2017;3:e000131.