Metagenome-Assembled Genome Binning Methods with Short Reads Disproportionately Fail for Plasmids and Genomic Islands

This manuscript (permalink) was automatically generated from fmaguire/mag_sim_paper@426cdfa on May 8, 2020.

Authors

- **Finlay Maguire***
  0000-0002-1203-9514 · fmaguire · finlaym
  Faculty of Computer Science, Dalhousie University

- **Baofeng Jia***
  0000-0002-4735-4709 · imasianxd · bfjia
  Department of Molecular Biology and Biochemistry, Simon Fraser University

- **Kristen Gray**
  0000-0002-1962-189X
  Department of Molecular Biology and Biochemistry, Simon Fraser University

- **Wing Yin Venus Lau**
  0000-0003-3884-4009
  Department of Molecular Biology and Biochemistry, Simon Fraser University

- **Robert G. Beiko**
  0000-0002-5065-4980
  Faculty of Computer Science, Dalhousie University

- **Fiona S.L. Brinkman**
  0000-0002-0584-4099
  Department of Molecular Biology and Biochemistry, Simon Fraser University

* authors contributed equally.

Abstract

Metagenome-assembled genome (MAG) binning methods have emerged as a promising approach to recover individual genomes of microbes simultaneously within a community from metagenomic data without time-consuming and bias-inducing culture. The transmission of mobile genetic elements (MGEs), such as plasmids and genomic islands (GIs), is a public health priority due to their clinical, agricultural, and environmental implications. Variable copy numbers and sequence compositions of MGEs creates a challenge for studying MGEs using MAG binning methods, and current tools has not been well assessed for its ability to recover MGEs. To systematically investigate this, we simulated a metagenome comprising 30 GI-rich and plasmid-containing bacterial genomes. MAGs were then recovered using 12 workflows and evaluated for recovery of MGE-associated antimicrobial resistance (AMR)/virulence factor (VF) genes. We showed that only 38-44% of GIs and 1-29% of plasmid were recovered and correctly binned. Most strikingly, no plasmid-borne VF or AMR genes were recovered and, only 0-45% within GIs were identified. We conclude that short-read MAGs are largely ineffective for the analysis of mobile genes, including those of public-health importance like AMR and VF genes.

We propose that microbiome researchers should instead primarily utilise unassembled short reads and/or long-read approaches to more accurately analyse metagenomic data.

Main

Metagenomics, the sequencing of DNA from within an environmental sample, is widely used to characterise 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 AMR genes [3,4,5], as well as tracking pathogen outbreaks [6]. Although long-read DNA technologies (e.g., Oxford Nanopore [7], PacBio [8]) are now being used for metagenomic sequencing [9,10], high-throughput sequencing of relatively short reads (150-250bp) in 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 (e.g., [11,12,13,14]). Since these reads are shorter than most genes, however, read-based methods provide very little information about their genomic organisation. This lack of contextual information is particularly problematic in the study of AMR genes and VFs as the genomic context plays a role in function [15], selective pressures [16], and likelihood of lateral gene transfer (LGT) [17].

Sequence assembly using specialised 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], MetaBAT 2 [24], MaxBin 2 [25], and a tool which combines their predictions: DAS Tool [26]. These MAG binning methods have been used in unveiling previously uncharacterised genomic diversity [27,28,29], but metagenomic assembly and binning results in the loss of some information. This compounded data loss 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. However, to the best of our knowledge, there has not been a specific assessment of MAG-based recovery of mobile genetic elements (MGEs) like genomic islands (GIs) and plasmids, despite their health and research importance.

Genomic islands (GIs) are clusters of 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) [33,34]. They are of high interest since VFs are disproportionately associated with mobile sequences [35] as well as certain AMR genes [36,37]. GIs often have differing nucleotide composition compared to the rest of the genome [33], a trait exploited by GI prediction tools such as SIGI-HMM [38], IslandPath-DIMOB [39], and integrative tools like IslandViewer [40]. GIs may also exist as multiple copies within a genome [41] 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 [42,43]. Similar to GIs, the sequence composition of plasmids are often markedly different from the genome with which they are associated [44,45]. Plasmids are also of high interest as a major source of the lateral dissemination of AMR genes throughout microbial ecosystems [36,46].

These varying composition and relative abundance features mean that GIs and plasmids pose significant challenges in MAG recovery. 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 GI and plasmid (and associated AMR/VF genes) recovery accuracy across a set of 12 state-of-the-art methods for short-read metagenome assemblies. We show that short-read MAG-based analyses are not suitable for the study of mobile sequences.

Material and Methods

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 github.com/fmaguire/MAG_gi_plasmid_analysis and data repository osf.io/nrejs/.

Metagenome Simulation

Thirty RefSeq genomes were selected using IslandPath-DIMOB [39] GI prediction data collated into the IslandViewer database www.pathogenomics.sfu.ca/islandviewer [40] (Supplemental Table 1). The selected genomes and associated plasmids (listed in Supplemental Table 2 and deposited at osf.io/nrejs/ under "data/sequences") were manually selected to satisfy the following criteria:

1. 10 genomes with 1-10 plasmids.

2. 10 genomes with >10% of chromosomal DNA predicted to reside in GIs.

3. 10 genomes with <1% of chromosomal DNA predicted to reside in GIs.

In accordance with the recommendation in the CAMI challenge [47] the genomes were randomly assigned a relative abundance following a log-normal distribution ($\mu = 1$, $\sigma = 2$). 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 ($\alpha = 4$, $\beta = 1$) 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 Supplemental Table 3. Sequences were then concatenated into a single FASTA file with the appropriate relative abundance. MiSeq v3 250bp paired-end reads with a mean fragment length of 1000bp (standard deviation of 50bp) were then simulated using art_illumina (v2016.06.05) [48] resulting in a simulated metagenome of 31,174,411 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) [49] resulting in 25,682,644 surviving read pairs. The trimmed reads were then assembled using 3 different metagenomic assemblers: metaSPAdes (v3.13.0) [18], IDBA-UD (v1.1.3) [19], and megahit (v1.1.3) [20]). The resulting assemblies were summarised using metaQUAST (v5.0.2) [50]. The assemblies were then indexed and reads mapped back using Bowtie 2 (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 MaxBin 2 (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 kilobase 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 10kb 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, MetaBAT 2, 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

Chromosomal Coverage

The MAG assessment for chromosomal coverage was performed by creating a BLASTN 2.9.0+ [51] 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 that contigs from a given MAG aligned to 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 close relatives in the metagenome on ability to bin chromosomes we generated a phylogenetic tree for all the input genomes. Specifically, single copy universal bacterial proteins were identified in the reference genomes using BUSCO v4.0.2 with the Bacteria Odb10 data [52]. The 86 of these proteins that were found in every reference genome were concatenated and aligned using MAFFT v7.427 [53] and masked with trimal v1.4.1-3 [54]. A maximum-likelihood phylogeny was then inferred with IQ-Tree v1.6.12 [55] with the in-built ModelFinder determined partitioning [56]. Pairwise branch distances were then extracted from the resulting tree using ETE3 v3.1.1 [57] and regressed using a linear model against coverage and contamination in seaborn v0.10.0 [58].

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 which 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.

Antimicrobial Resistance and Virulence Factors Assessment

Detection of AMR/VF Genes

For the reference genomes, as well as 12 sets of MAGs, prodigal [59] was used to predict open reading frames (ORFs) using the default 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) [60]. Virulence factors were predicted using the predicted ORFs and BLASTX 2.9.0+ [51] against the Virulence Factor Database (VFDB; obtained on Aug 26, 2019) with an e-value

cut-off of 0.001 and a minimum identity of 90% [61]. 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.

Protein subcellular localisation predictions

We then sought to assess what the impact of a protein's predicted subcellular localisation was on its recovery and binning in MAGs. The MAG bins from megahit-DAS Tool assembler-binner combination were selected (as generally best performing) and ORFs predicted using prodigal [59] as above. Subcellular localisation of these proteins were then predicted using PSORTb v3.0 with default parameters and the appropriate Gram setting for that bin's assigned taxa [62].

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 (see 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% to 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. S1).

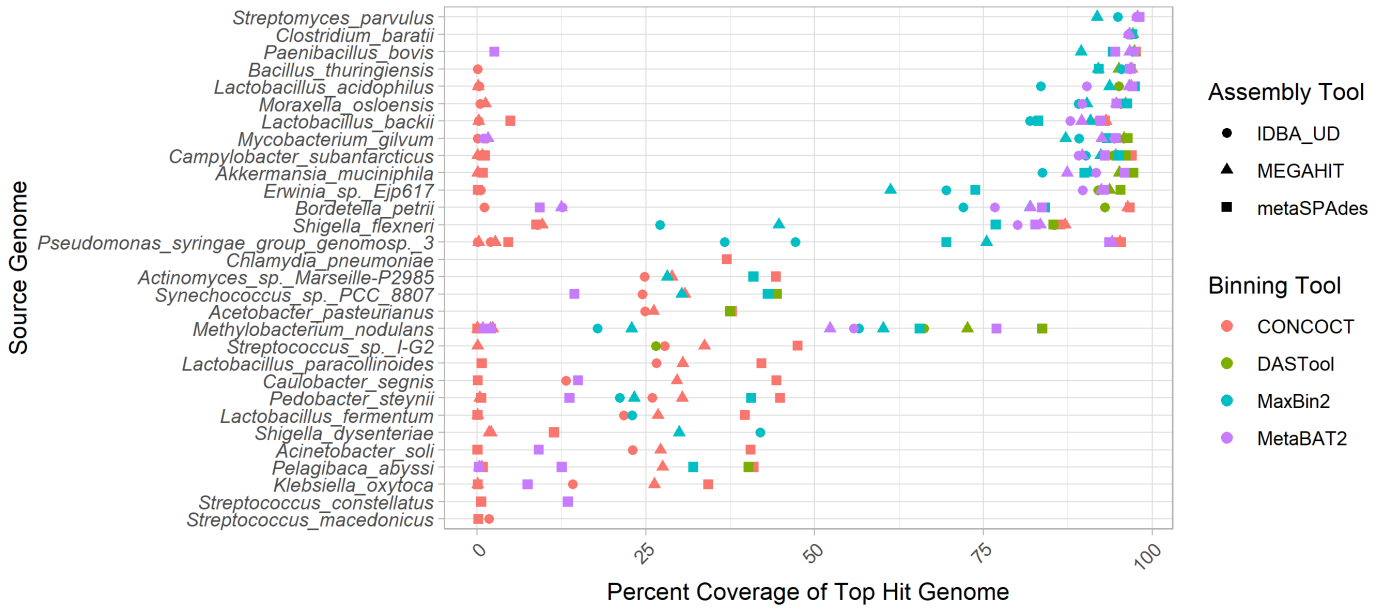

Figure 1: Top genome coverage for input genomes across MAG binners. Each dot represents the coverage of a specified genome when it comprised the plurality of the sequences in a bin. If a genome did not form the plurality of any bin for a specific binner-assembler pair no dot was plotted for that genome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

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. 2 A). In terms of 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 DASTool that combines predictions from the other 3 binners (94.3% median top hit chromosome coverage per bin; (Fig. 2 A)).

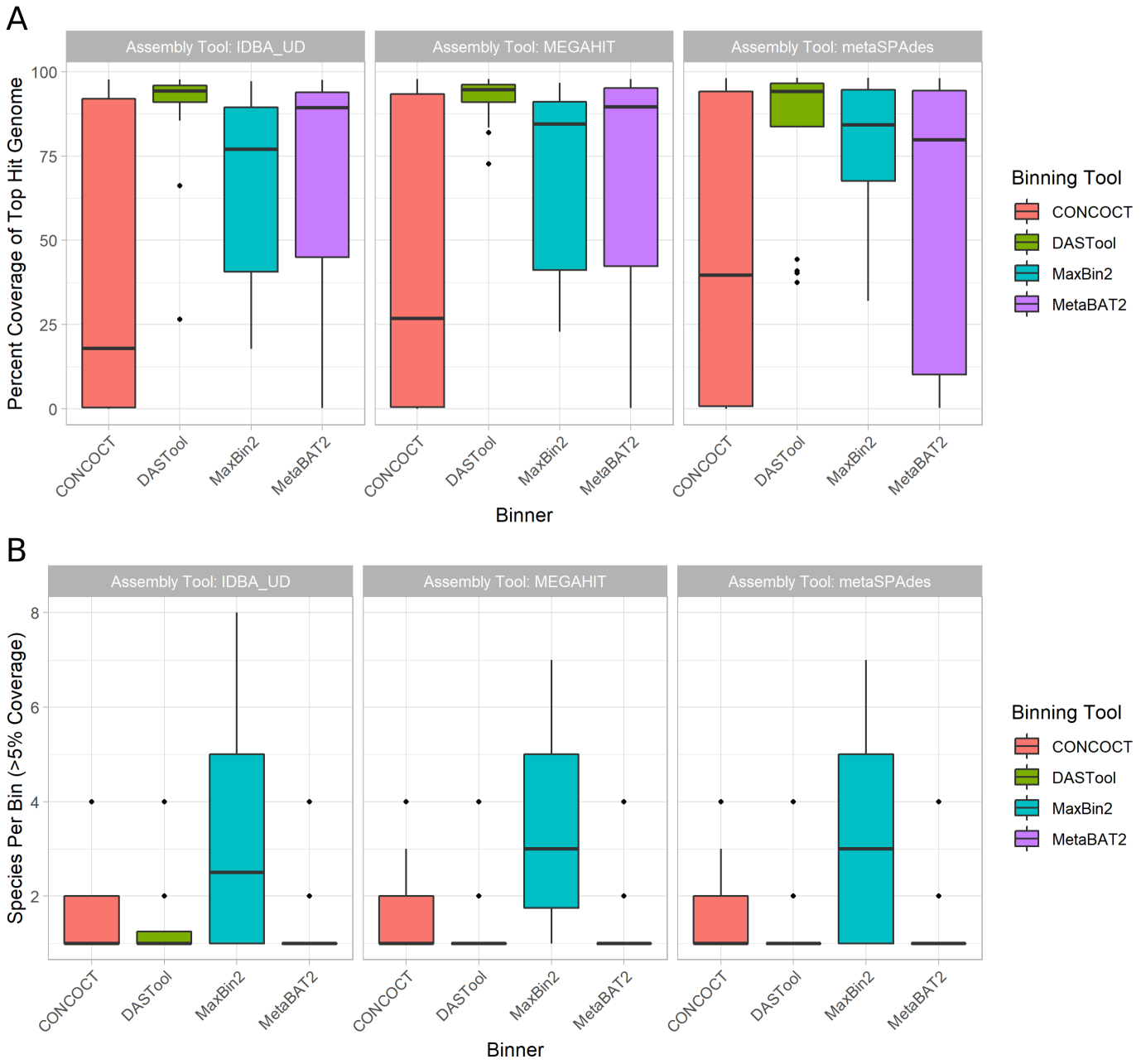

Figure 2: Overall binning performance for every combination of metagenome assembler (as indicated by pane titles) and MAG binning tool (x-axis and legend colours). Diamonds in the plots represent outliers (greater or lower than the interquartile range marked by the error bars) and the boxes represent the lower quartile, median, and upper quartile respectively. **(A)** Chromosomal coverage of the most prevalent genome in each bin across binners and metagenome assemblies. Of the 3 assemblers, megahit resulted in the highest median chromosomal coverage (y-axis) across all binners (colored bars) at 81.9% with metaSPAdes performing the worst (76.8%). Of the 4 binners, CONCOCT (red) performed poorly with a median coverage, followed by maxbin2 (blue), MetaBAT2 (purple) and DASTool (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 DASTool (green), which all produced chimeric bins at roughly the same rate.

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. Across binning tools maxbin2 proved an exception with nearly twice as many bins containing multiple species as the next binner (Fig. 2 B). The remaining binning tools were largely equivalent, producing chimeric bins at approximately the same rates. Unlike coverage, purity was strongly affected by the number of close relatives in the metagenome to a given input genome. Specifically, the closer the nearest relative the less pure the bin (Fig. S2).

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 DASTool 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 fraction were correctly binned (typically less than 1/3rd). Interestingly, the MetaBAT2 and DASTool binners were more conservative in assigning plasmid contigs to bins; of those assigned to bins, nearly all were correctly binned (Fig. 3).

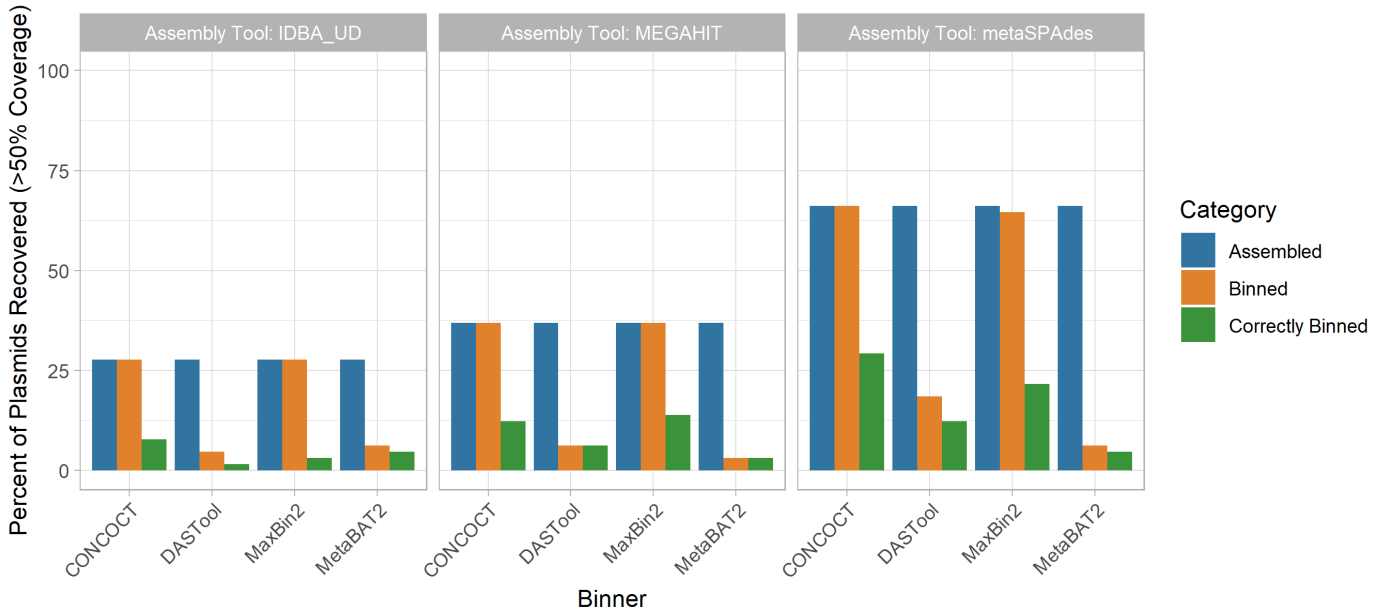

Figure 3: The performance of metagenomic assembly and binning to recover plasmid sequences. Each plot represents a different metagenome assembler, with the groups of bars along the x-axes showing the plasmid recovery performance of each binning tool when applied to the assemblies produced by that tool. 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.

Genomic Islands

GIs displayed a similar pattern of assembly and correct binning performance as plasmids (Fig. 4). 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 DASTool and MetaBAT2. DASTool, 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.

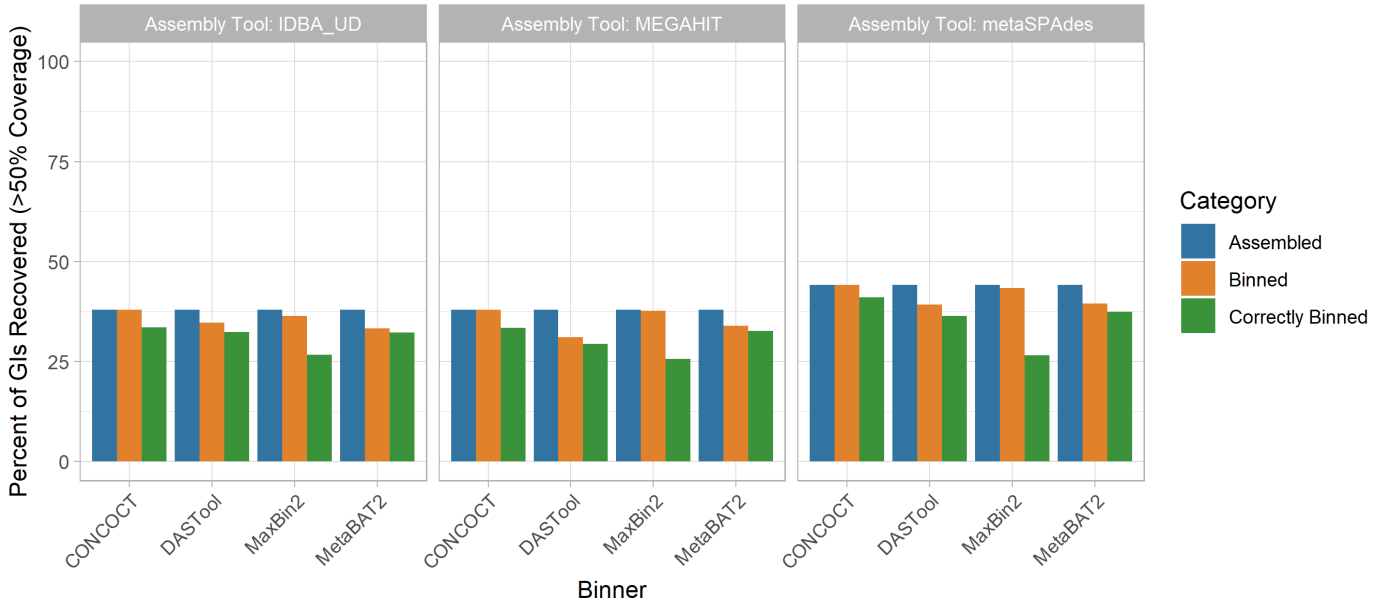

Figure 4: Impact of metagenomic assembly and MAG binning on recovery of GIs. GIs were recovered in a similarly poor fashion to plasmids. Generally, <40% were correctly assigned to the same bin majorly comprised of chromosomal contigs from the same source genome regardless of binning (x-axis) and assembly (panel) methods at >50% coverage. metaSPAdes performed the best at assembling GIs (blue). Maxbin2 and CONCOCT placed GIs in a bin majority of the time (orange) however a very small fraction was correctly binned (green). Generally, GIs were correctly binned better than plasmids with DASTool, MetaBAT2 and CONCOCT.

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. 5 A). 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 DASTool-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-UDA pair (24%) and highest in the CONCOCT-metaSPAdes pipe (40%).

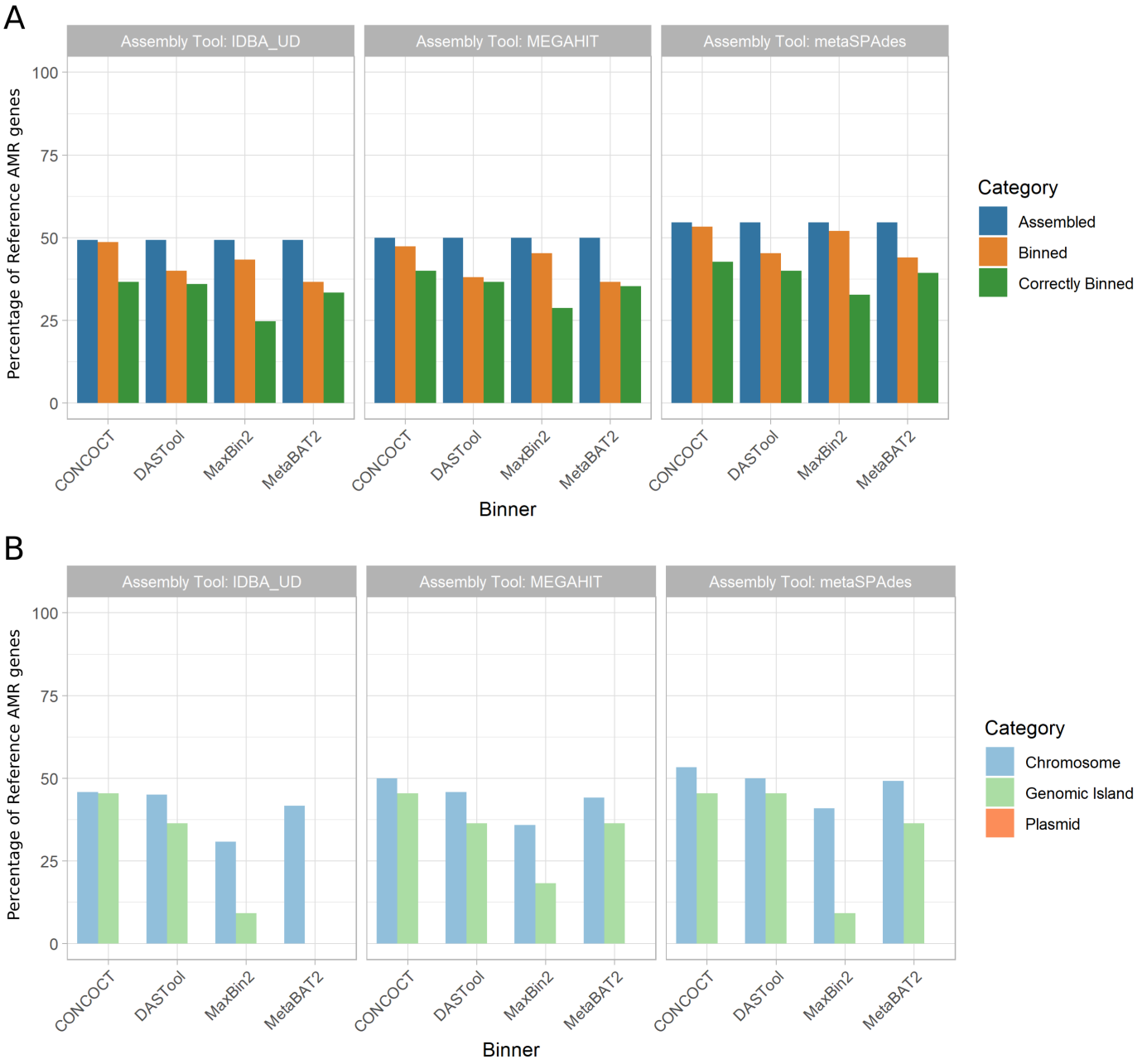

Figure 5: Recovery of AMR genes across assemblers, binners, and genomic context. **(A)** The proportion of reference AMR genes recovered (y-axis) was largely similar across assembly tools (panels as indicated by title) at roughly 50% with metaSPAdes performing marginally better overall. Binning tools (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)** Percent 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.

Moreover, focusing on only the AMR genes that were correctly binned (Fig. 5 B) 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-localised AMR genes (n=20) were correctly binned.

Virulence Factor Genes

We also examined the impact of MAG approaches on recovery of virulence factor (VF) genes as identified using the Virulence Factor Database (VFDB). We saw a similar trend as AMR genes (Fig. 6 A). 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% with DASTool-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.

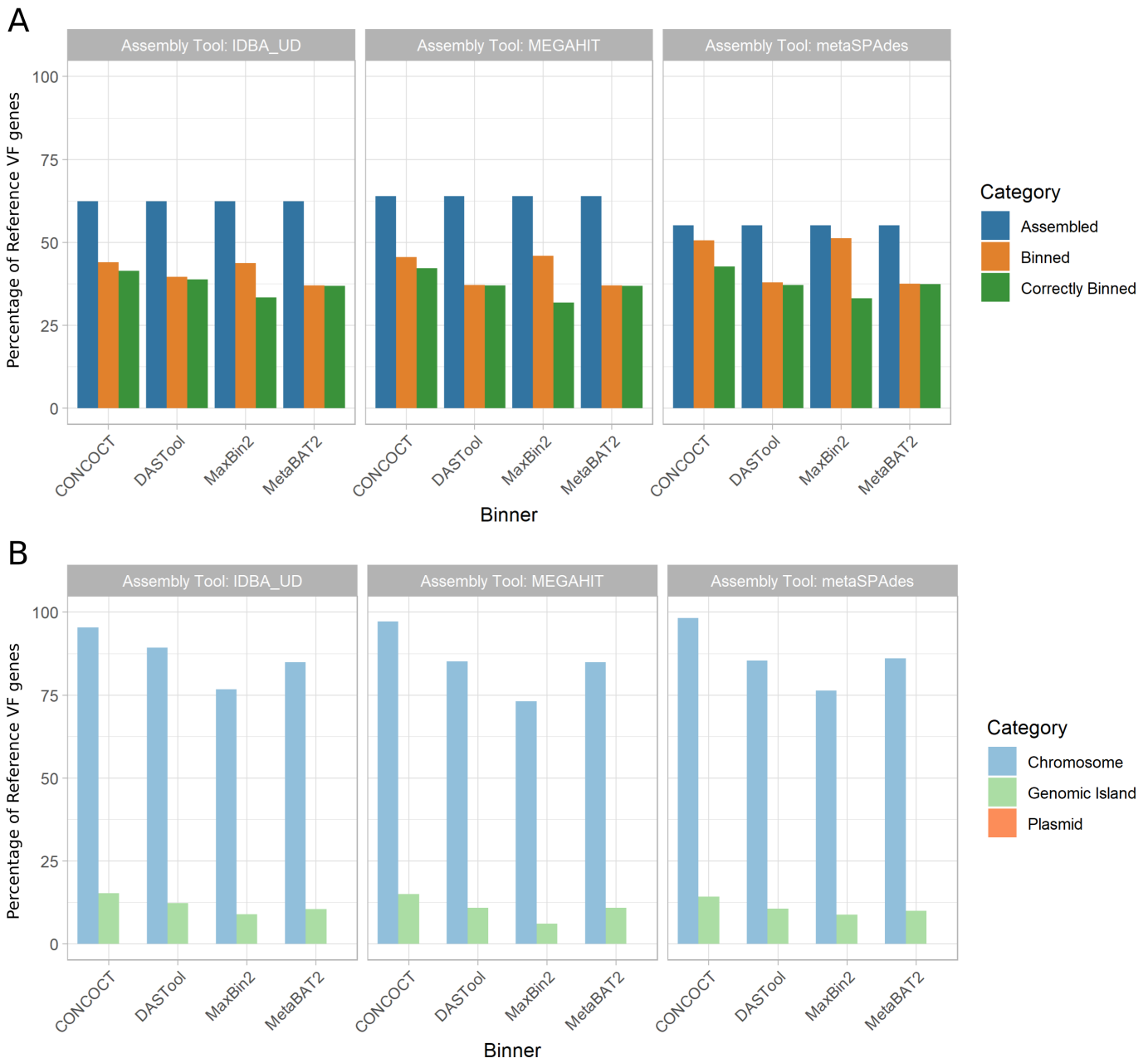

Figure 6: Recovery of VF genes across assemblers, binners, and genomic context. **(A)** Percent of reference virulence factor (VF) genes recovered across assemblers and binners. The proportion of reference VF genes recovered (y-axis) exhibited a similar trend as AMR genes. Recovery was greatest after 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 are binned are correctly binned (green). metaSPAdes-CONCOCT was again the best performing pair. **(B)** Percent of correctly binned VF genes recovered in each genomic region. Metagenome assembled genomes (MAGs) were again best at recovering chromosomally located VF genes (light blue), able to correctly bin majority of chromosomally located VFs. GIs recovered again performed very poorly (light green) and again none of the plasmid located AMR genes (orange) was correctly binned.

As with AMR genes, the genomic context (chromosome, plasmid, GI) of a given VF largely determined how well it was binned (Fig. 6 B). The majority (73%-98%) of all chromosomally located VF genes (n=757) were correctly binned. However, 0-16% of GI-localised 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, 4, 5, and 6 to compare the rates of loss of different components (see Fig. S4). 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.

Discussion

In this paper, we evaluated the ability of metagenome-assembled genome (MAG) binning methods to correctly recover mobile genetic elements (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-DASTool) however closely related genomes were consistently cross-contaminated with other sequences (e.g. *Streptococcus* species in Fig. S1, S2). Given the importance of MGEs in the function and spread of pathogenic traits, 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 lateral gene transfer 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. The particularly poor plasmid binning performance is likely attributable to the known difficulties in assembly of plasmids from short-read data [63]. Therefore, binning efficiency might improve with use of long-read sequencing or assembly methods optimised for recovering plasmids [63] (such as SCAPP [64]). Despite its lower effective sequencing depth and higher error rates, incorporating long-read sequencing has been shown to improve overall MAG binning [65] and facilitate metagenomic characterisation of plasmids [66]. Further research is needed to fully characterise the performance of different long-read protocols on the accuracy of recovering MGEs in metagenomic samples.

With the growing use of MAG methods in infectious disease research (e.g., [67,68,69,70,71]) and the public-health 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 the KPC and OXA carbapenemases. 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 suppl. discussion & Fig. S3). Previous studies have observed that ORF predictions in draft genomes are more fragmented, which can lead to downstream over- or under-annotation with functional labels depending on the approach used [72]. Although not yet developed, methods that combine the assembly/binning pipelines tested here with read-based inference would give 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 [73,74,75], other environments with relevance to public health such as the human gut, soil, and livestock can have 100-1000s of species [76,77,78,79]. 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 [32,80]. This simulation approach, combined with the development of an MGE/AMR-focused mock metagenome (similarly to the mockrobiota initiative [81]), could provide a key resource to develop and validate new binning approaches and different sequencing strategies.

This study has shown that MAG-based approaches provide a useful tool to study a bacterial species' core chromosomal elements, but 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 VFs and AMR genes. As many of these clinically relevant genes have a high propensity for lateral gene transfer between unrelated bacteria [35,36] it is critical to highlight that MAG approaches alone are insufficient to thoroughly profile them. Within public-health metagenomic research it is vital we utilise MAGs in conjunction with other methods (e.g. targeted AMR [82], long-read sequencing, plasmid specialised assembly approaches [64], and read-based sequence homology search [11]) before drawing biological or epidemiological conclusions.

Data availability

All datasets used or generated in this study are available at osf.io/nrejs All analysis and plotting code used is available at github.com/fmaguire/MAG_gi_plasmid_analysis

Supplementary Information

Impact of Related Genomes on MAG

By generating a phylogeny of universal single copy genes in our input genomes we analysed the relationship between the presence of closely related genomes and the ability of the different MAG-recovery methods to bin chromosomal sequences. Specifically, we regressed phylogenetic distance on this phylogeny with per-bin chromosomal coverage (Fig. S1) and bin purity (Fig. S2). This identified no clear relationship between chromosomal coverage and the phylogenetic distance to the nearest relative in the metagenome (Fig. S1), however, there did seem to be a negative correlation between phylogenetic distance to closest relative and the purity of a MAG bin (Fig. S2). In other words, across all methods, a MAG bin was more likely to have multiple genomes present if there were close relatives.

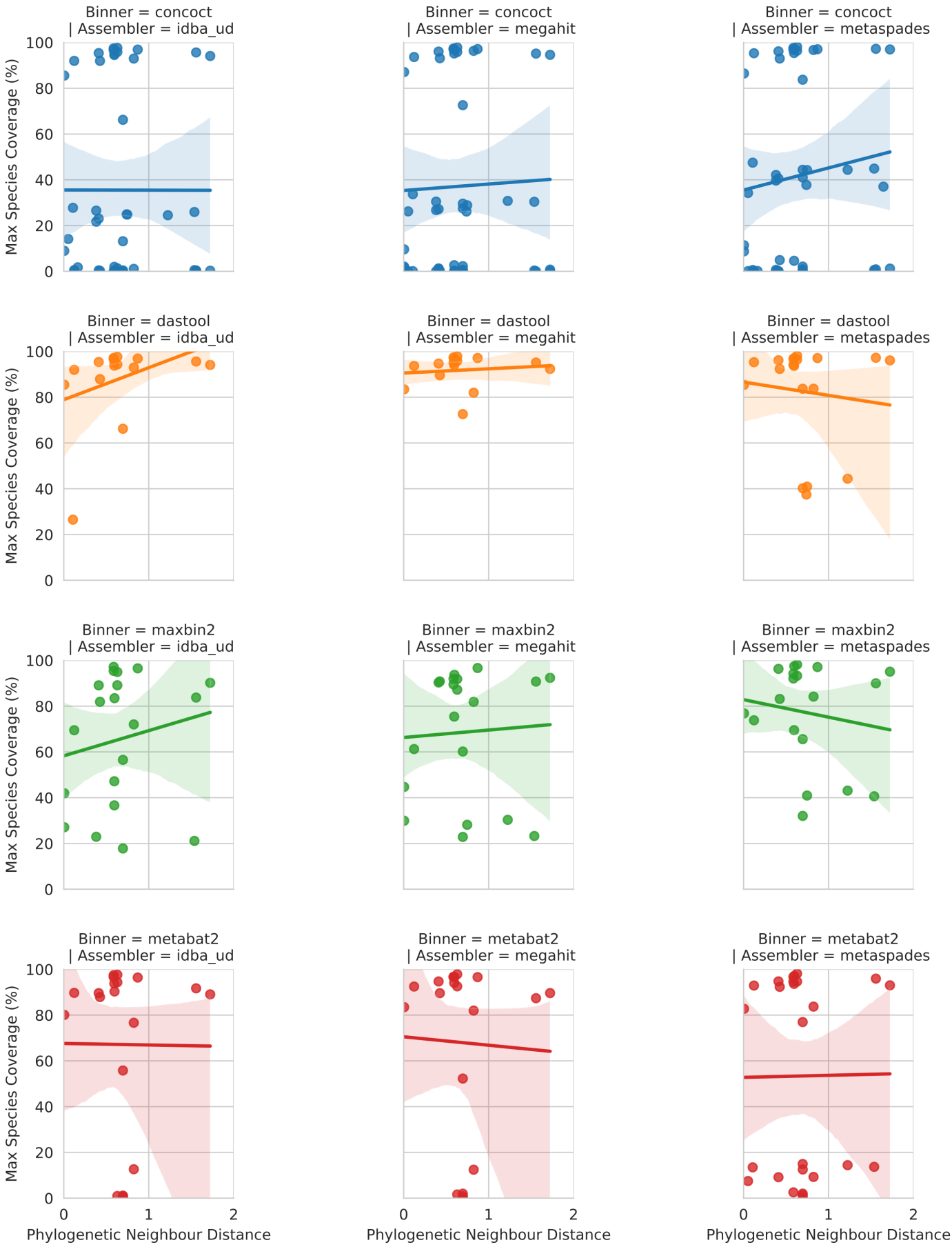

Figure S1: Relationship between phylogenetic distance to closest neighbour input genome on genomic coverage in MAG majority comprised of that taxa. Each dot represents the genomic coverage of a particular taxa and the branch distance on an 86-protein concatenated phylogeny between that taxa and its nearest neighbour. Rows indicate the binning software and columns the metagenomic assembler. Regression line is a simple linear model fitted in seaborn.

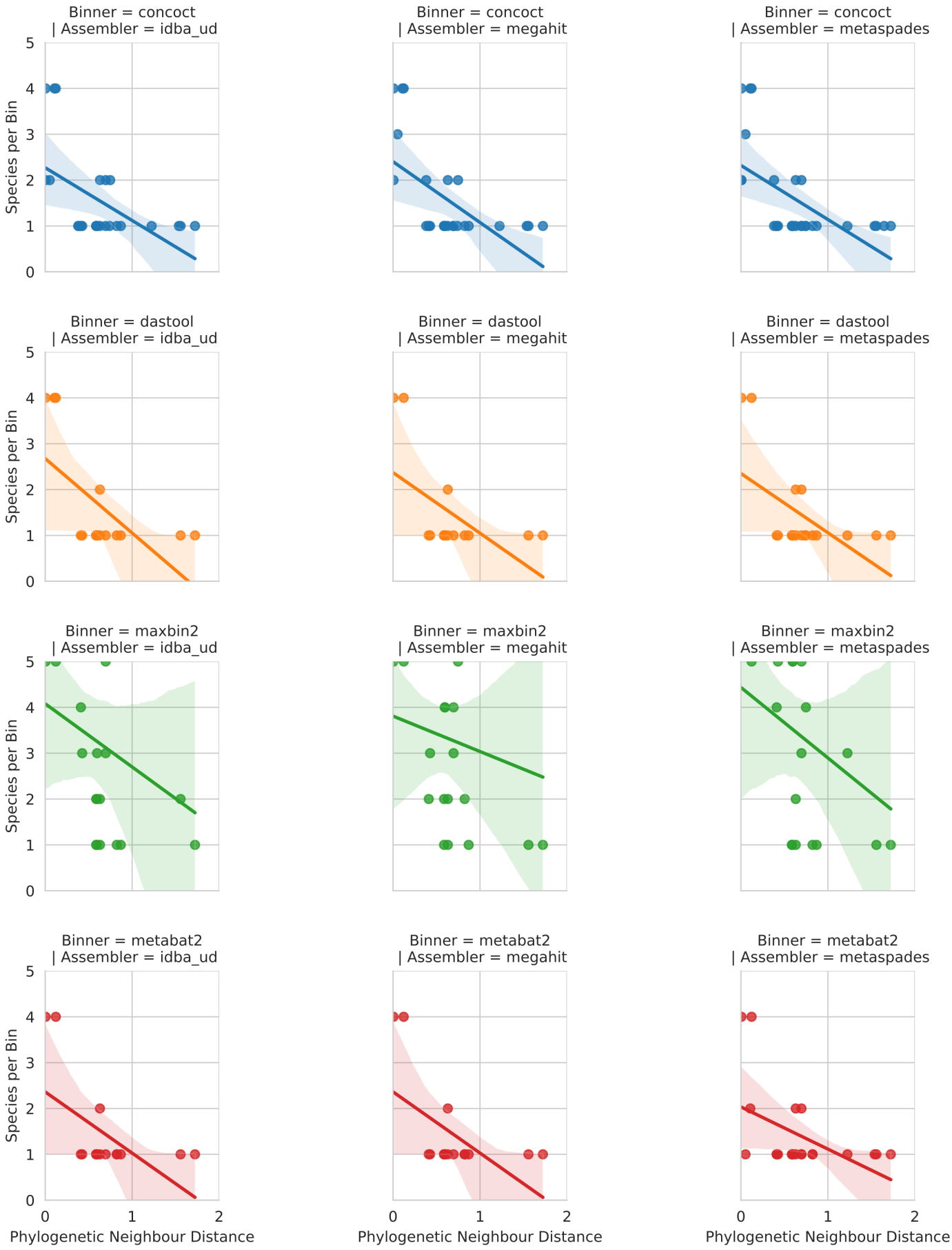

Figure S2: Relationship between phylogenetic distance to closest neighbour input genome on bin purity. Each dot shows the number of other input genomes detectable in a given MAG bin in relation to the branch distance on an 86-protein concatenated phylogeny between the majority taxa in that bin and its nearest neighbour.

Recovery of Specific Gene Content

We explored the ability of different approaches to find open reading frames (ORFs) within MAGs. Overall, the total number of predicted ORFs in MAGs followed a similar trend (Fig. S3) as the chromosomal coverage and purity (Fig. 2). Of the four binning tools, CONCOCT performed the worst, finding <30% of the number of ORFs in our reference genomes used to construct the synthetic data. MetaBAT2 performed second worst at ~80%. DASTool recovered a similar number to our reference and Maxbin2 detected 7-46% more genes. The Assembler method did not significantly impact the number of genes predicted with the exception of Maxbin2, in which IDBA_UD was the closest to reference and metaSPAdes predicted 46% more ORFs. Given that there is reason to suspect that there are some issues with the ORF calling in the MAGs. i.e. some tools produced more predicted ORFs than reference, it could be the case that some of these sequences are present in the assemblies (with errors/gaps), but are not being identified as ORFs, or are broken into multiple ORFs, leading to issues downstream labeling them correctly as AMR/VF genes. Regardless of different tools producing a different number of ORFs, the recovery of AMR/VF is pretty consistent regardless of how many ORFs are predicted.

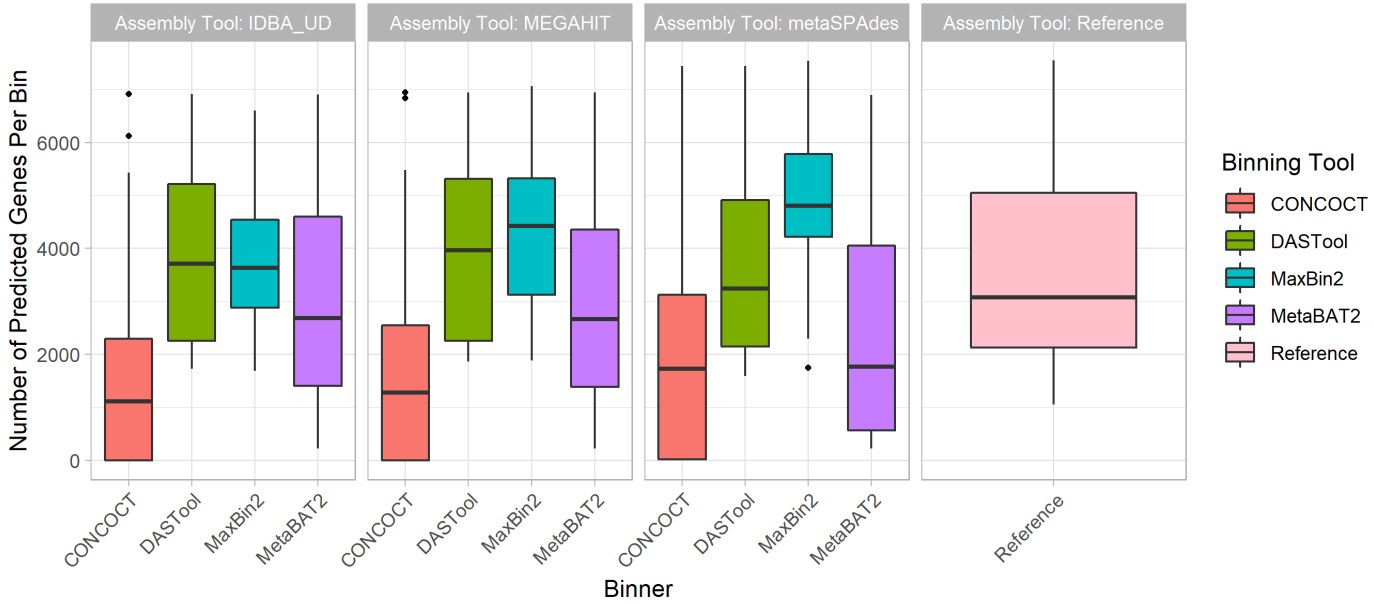

Figure S3: Predicted Gene Content. The total number of open reading frames (ORF) predicted followed the same trend as chromosomal coverage and purity. The assemblers (colored bars) did not contribute to variability in the number of ORFs detected. Of the 4 binners, CONCOCT recovered <30% of our reference genome ORFs. DASTool and MetaBAT2 predicted a similar number as our reference genomes.

Comparisons of Rates of Loss

Combining the performance metrics for Figs. 3, 4, 5, and 6 to compare the rates of loss of different components emphasises some of the observed patterns (see Fig. S4). This highlights that genomic components (GIs and plasmids) and plasmids in particular are lost at a higher rate than individual gene types during MAG recovery.

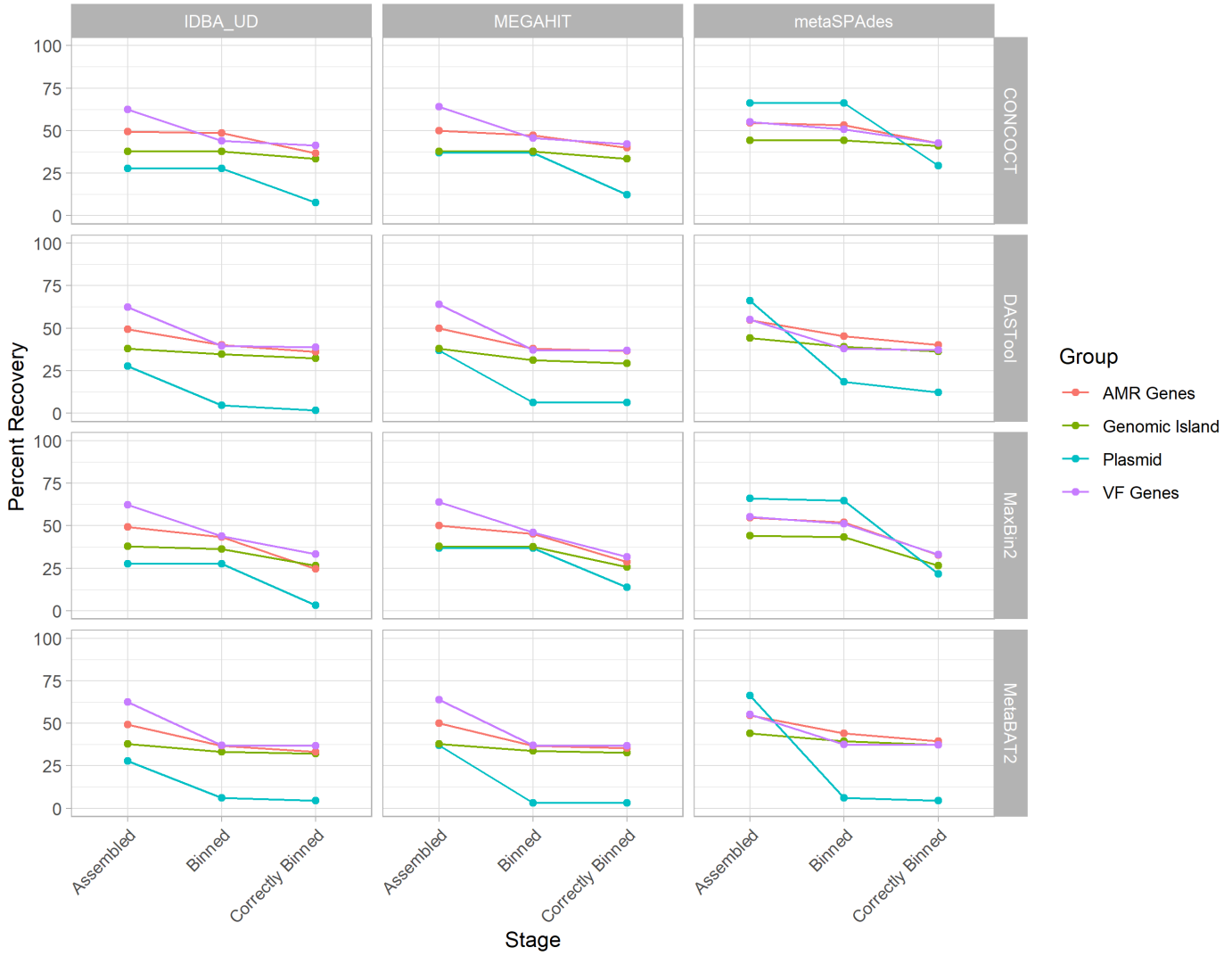

Figure S4: Comparison of rates of loss for different genomic components and gene types across assemblers and binning tools. Each line represents a different component as indicated by the legend with assemblers indicated by row and binning tool by column. This shows that regardless of approach genomic components (GIs and plasmids) are lost at a higher rate than individual VF or AMR genes.

Funding

This work was supported primarily by a Donald Hill Family Fellowship held by F.M. W.Y.V.L. and B.J. hold 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. holds an SFU Distinguished Professorship 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.[grant number XXXXX, yyyyy], and the authors thank the SFU Research Computing Group and Compute Canada for compute resource support.

Conflict of Interest

The authors declare no competing interests.

References

1. M. Breitbart, P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, F. Rohwer
(2002) Genomic analysis of uncultured marine viral communities *Proceedings of the National Academy of Sciences*
DOI: 10.1073/pnas.202488399 PMID: 12384570 PMCID: PMC137870

2. Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, Nicola Segata
(2017) Shotgun metagenomics, from sampling to analysis *Nature Biotechnology*
DOI: 10.1038/nbt.3935 PMID: 28898207

3. Mohamed S. Donia, Peter Cimermancic, Christopher J. Schulze, Laura C. Wieland Brown, John Martin, Makedonka Mitreva, Jon Clardy, Roger G. Linington, Michael A. Fischbach
(2014) A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics *Cell*
DOI: 10.1016/j.cell.2014.08.032 PMID: 25215495 PMCID: PMC4164201

4. Vanessa M D'Costa, Emma Griffiths, Gerard D Wright
(2007) Expanding the soil antibiotic resistome: exploring environmental diversity *Current Opinion in Microbiology*
DOI: 10.1016/j.mib.2007.08.009 PMID: 17951101

5. Vanessa M. D'Costa, Christine E. King, Lindsay Kalan, Mariya Morar, Wilson W. L. Sung, Carsten Schwarz, Duane Froese, Grant Zazula, Fabrice Calmels, Regis Debruyne, … Gerard D. Wright
(2011) Antibiotic resistance is ancient *Nature*
DOI: 10.1038/nature10388 PMID: 21881561

6. Nicholas J. Loman, Chrystala Constantinidou, Martin Christner, Holger Rohde, Jacqueline Z.-M. Chan, Joshua Quick, Jacqueline C. Weir, Christopher Quince, Geoffrey P. Smith, Jason R. Betley, … Mark J. Pallen
(2013) A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia coli O104:H4 *JAMA*
DOI: 10.1001/jama.2013.3231 PMID: 23571589

7. Alexander S. Mikheyev, Mandy M. Y. Tin
(2014) A first look at the Oxford Nanopore MinION sequencer *Molecular Ecology Resources*
DOI: 10.1111/1755-0998.12324 PMID: 25187008

8. J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, … S. Turner
(2009) Real-Time DNA Sequencing from Single Polymerase Molecules *Science*
DOI: 10.1126/science.1162986 PMID: 19023044

9. Samuel M Nicholls, Joshua C Quick, Shuiquan Tang, Nicholas J Loman
(2019) Ultra-deep, long-read nanopore sequencing of mock microbial community standards

GigaScience
DOI: 10.1093/gigascience/giz043 PMID: 31089679 PMCID: PMC6520541

10. Vincent Somerville, Stefanie Lutz, Michael Schmid, Daniel Frei, Aline Moser, Stefan Irmler, Jürg E. Frey, Christian H. Ahrens
(2019) 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 Microbiology*
DOI: 10.1186/s12866-019-1500-0 PMID: 31238873 PMCID: PMC6593500

11. Benjamin Buchfink, Chao Xie, Daniel H Huson
(2014) Fast and sensitive protein alignment using DIAMOND *Nature Methods*
DOI: 10.1038/nmeth.3176 PMID: 25402007

12. Ben Langmead, Steven L Salzberg
(2012) Fast gapped-read alignment with Bowtie 2 *Nature Methods*
DOI: 10.1038/nmeth.1923 PMID: 22388286 PMCID: PMC3322381

13. T. J. Wheeler, S. R. Eddy
(2013) nhmmer: DNA homology search with profile HMMs *Bioinformatics*
DOI: 10.1093/bioinformatics/btt403 PMID: 23842809 PMCID: PMC3777106

14. Rachid Ounit, Steve Wanamaker, Timothy J Close, Stefano Lonardi
(2015) CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers *BMC Genomics*
DOI: 10.1186/s12864-015-1419-2 PMID: 25879410 PMCID: PMC4428112

15. X. Xu, D. Lin, G. Yan, X. Ye, S. Wu, Y. Guo, D. Zhu, F. Hu, Y. Zhang, F. Wang, ... M. Wang
(2010) vanM, a New Glycopeptide Resistance Gene Cluster Found in Enterococcus faecium *Antimicrobial Agents and Chemotherapy*
DOI: 10.1128/aac.01710-09 PMID: 20733041 PMCID: PMC2976141

16. Craig Baker-Austin, Meredith S. Wright, Ramunas Stepanauskas, J. V. McArthur
(2006) Co-selection of antibiotic and metal resistance *Trends in Microbiology*
DOI: 10.1016/j.tim.2006.02.006 PMID: 16537105

17. Hatch W. Stokes, Michael R. Gillings
(2011) Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens *FEMS Microbiology Reviews*
DOI: 10.1111/j.1574-6976.2011.00273.x PMID: 21517914

18. Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, Pavel A. Pevzner
(2017) metaSPAdes: a new versatile metagenomic assembler *Genome Research*
DOI: 10.1101/gr.213959.116 PMID: 28298430 PMCID: PMC5411777

19.
Y. Peng, H. C. M. Leung, S. M. Yiu, F. Y. L. Chin
(2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth *Bioinformatics*
DOI: 10.1093/bioinformatics/bts174 PMID: 22495754

20.
Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, Tak-Wah Lam
(2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph *Bioinformatics*
DOI: 10.1093/bioinformatics/btv033 PMID: 25609793

21.
Gene W. Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar, Jillian F. Banfield
(2004) Community structure and metabolism through reconstruction of microbial genomes from the environment *Nature*
DOI: 10.1038/nature02340 PMID: 14961025

22.
Florian P Breitwieser, Jennifer Lu, Steven L Salzberg
(2019) A review of methods and databases for metagenomic classification and assembly *Briefings in Bioinformatics*
DOI: 10.1093/bib/bbx120 PMID: 29028872 PMCID: PMC6781581

23.
Yang Young Lu, Ting Chen, Jed A. Fuhrman, Fengzhu Sun
(2016) COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge *Bioinformatics*
DOI: 10.1093/bioinformatics/btw290 PMID: 27256312

24.
Dongwan Kang, Feng Li, Edward S Kirton, Ashleigh Thomas, Rob S Egan, Hong An, Zhong Wang
(2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies
DOI: 10.7287/peerj.preprints.27522v1

25.
Yu-Wei Wu, Blake A. Simmons, Steven W. Singer
(2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets *Bioinformatics*
DOI: 10.1093/bioinformatics/btv638 PMID: 26515820

26.
Christian M. K. Sieber, Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, Jillian F. Banfield
(2018) Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy *Nature Microbiology*
DOI: 10.1038/s41564-018-0171-1 PMID: 29807988 PMCID: PMC6786971

27.
Christopher T. Brown, Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, Michael J. Wilkins, Kelly C. Wrighton, Kenneth H. Williams, Jillian F. Banfield

(2015) Unusual biology across a group comprising more than 15% of domain Bacteria *Nature*
DOI: 10.1038/nature14486 PMID: 26083755

28.
Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, Gene W. Tyson
(2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life *Nature Microbiology*
DOI: 10.1038/s41564-017-0012-7 PMID: 28894102

29.
Robert D. Stewart, Marc D. Auffret, Amanda Warr, Alan W. Walker, Rainer Roehe, Mick Watson
(2018) The genomic and proteomic landscape of the rumen microbiome revealed by comprehensive genome-resolved metagenomics *bioRxiv*
DOI: 10.1101/489443

30.
Ben J. Woodcroft, Caitlin M. Singleton, Joel A. Boyd, Paul N. Evans, Joanne B. Emerson, Ahmed A. F. Zayed, Robert D. Hoelzle, Timothy O. Lamberton, Carmody K. McCalley, Suzanne B. Hodgkins, … Gene W. Tyson
(2018) Genome-centric view of carbon processing in thawing permafrost *Nature*
DOI: 10.1038/s41586-018-0338-1 PMID: 30013118

31.
Spencer Diamond, Peter F. Andeer, Zhou Li, Alexander Crits-Christoph, David Burstein, Karthik Anantharaman, Katherine R. Lane, Brian C. Thomas, Chongle Pan, Trent R. Northen, Jillian F. Banfield
(2019) Mediterranean grassland soil C–N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms *Nature Microbiology*
DOI: 10.1038/s41564-019-0449-y PMID: 31110364 PMCID: PMC6784897

32.
Fernando Meyer, Peter Hofmann, Peter Belmann, Ruben Garrido-Oter, Adrian Fritz, Alexander Sczyrba, Alice C McHardy
(2018) AMBER: Assessment of Metagenome BinnERs *GigaScience*
DOI: 10.1093/gigascience/giy069 PMID: 29893851 PMCID: PMC6022608

33.
Morgan G. I. Langille, William W. L. Hsiao, Fiona S. L. Brinkman
(2010) Detecting genomic islands using bioinformatics approaches *Nature Reviews Microbiology*
DOI: 10.1038/nrmicro2350 PMID: 20395967

34.
Shannon M. Soucy, Jinling Huang, Johann Peter Gogarten
(2015) Horizontal gene transfer: building the web of life *Nature Reviews Genetics*
DOI: 10.1038/nrg3962 PMID: 26184597

35.
Shannan J. Ho Sui, Amber Fedynak, William W. L. Hsiao, Morgan G. I. Langille, Fiona S. L. Brinkman
(2009) The Association of Virulence Factors with Genomic Islands *PLoS ONE*
DOI: 10.1371/journal.pone.0008094 PMID: 19956607 PMCID: PMC2779486

36.
Christian J. H. von Wintersdorff, John Penders, Julius M. van Niekerk, Nathan D. Mills, Snehali Majumder, Lieke B. van Alphen, Paul H. M. Savelkoul, Petra F. G. Wolffs
(2016) Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer *Frontiers in Microbiology*
DOI: 10.3389/fmicb.2016.00173 PMID: 26925045 PMCID: PMC4759269

37.
Maryury Brown-Jaque, William Calero-Cáceres, Maite Muniesa
(2015) Transfer of antibiotic-resistance genes via phage-related mobile elements *Plasmid*
DOI: 10.1016/j.plasmid.2015.01.001 PMID: 25597519

38.
Rainer Merkl
(2004) SIGI: score-based identification of genomic islands. *BMC bioinformatics*
DOI: 10.1186/1471-2105-5-22 PMID: 15113412 PMCID: PMC394314

39.
Claire Bertelli, Fiona SL Brinkman
(2018) Improved genomic island predictions with IslandPath-DIMOB *Bioinformatics*
DOI: 10.1093/bioinformatics/bty095 PMID: 29905770 PMCID: PMC6022643

40.
Bhavjinder K. Dhillon, Matthew R. Laird, Julie A. Shay, Geoffrey L. Winsor, Raymond Lo, Fazmin Nizam, Sheldon K. Pereira, Nicholas Waglechner, Andrew G. McArthur, Morgan G. I. Langille, Fiona S. L. Brinkman
(2015) IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis: Figure 1. *Nucleic Acids Research*
DOI: 10.1093/nar/gkv401 PMID: 25916842 PMCID: PMC4489224

41.
Claire Bertelli, Keith E Tilley, Fiona SL Brinkman
(2019) Microbial genomic island discovery, visualization and analysis *Briefings in Bioinformatics*
DOI: 10.1093/bib/bby042 PMID: 29868902 PMCID: PMC6917214

42.
Alvaro San Millan, Jose Antonio Escudero, Danna R. Gifford, Didier Mazel, R. Craig MacLean
(2016) Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria *Nature Ecology & Evolution*
DOI: 10.1038/s41559-016-0010 PMID: 28812563

43.
Alvaro San Millan, Alfonso Santos-Lopez, Rafael Ortega-Huedo, Cristina Bernabe-Balas, Sean P. Kennedy, Bruno Gonzalez-Zorn
(2015) Small-Plasmid-Mediated Antibiotic Resistance Is Enhanced by Increases in Plasmid Copy Number and Bacterial Fitness *Antimicrobial Agents and Chemotherapy*
DOI: 10.1128/aac.00235-15 PMID: 25824216 PMCID: PMC4432117

44.
Fengfeng Zhou, Ying Xu
(2010) cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data *Bioinformatics*
DOI: 10.1093/bioinformatics/btq299 PMID: 20538725 PMCID: PMC2916713

45.
J. J. Davis, G. J. Olsen
(2009) Modal Codon Usage: Assessing the Typical Codon Usage of a Genome *Molecular Biology and Evolution*
DOI: 10.1093/molbev/msp281 PMID: 20018979 PMCID: PMC2839124

46.
Alison H Holmes, Luke SP Moore, Arnfinn Sundsfjord, Martin Steinbakk, Sadie Regmi, Abhilasha Karkey, Philippe J Guerin, Laura JV Piddock
(2015) Understanding the mechanisms and drivers of antimicrobial resistance. *Lancet (London, England)*
DOI: 10.1016/s0140-6736(15)00473-0 PMID: 26603922

47.
Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, ... Alice C McHardy
(2017) Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software *Nature Methods*
DOI: 10.1038/nmeth.4458 PMID: 28967888 PMCID: PMC5903868

48.
Weichun Huang, Leping Li, Jason R. Myers, Gabor T. Marth
(2012) ART: a next-generation sequencing read simulator *Bioinformatics*
DOI: 10.1093/bioinformatics/btr708 PMID: 22199392 PMCID: PMC3278762

49.
NA Joshi, JN Fass
(2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files *GitHub*

50.
Alla Mikheenko, Vladislav Saveliev, Alexey Gurevich
(2016) MetaQUAST: evaluation of metagenome assemblies *Bioinformatics*
DOI: 10.1093/bioinformatics/btv697 PMID: 26614127

51.
Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, Thomas L Madden
(2009) BLAST+: architecture and applications *BMC Bioinformatics*
DOI: 10.1186/1471-2105-10-421 PMID: 20003500 PMCID: PMC2803857

52.
Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, Evgeny M. Zdobnov
(2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs *Bioinformatics*
DOI: 10.1093/bioinformatics/btv351 PMID: 26059717

53.
Tsukasa Nakamura, Kazunori D Yamada, Kentaro Tomii, Kazutaka Katoh
(2018) Parallelization of MAFFT for large-scale multiple sequence alignments *Bioinformatics*
DOI: 10.1093/bioinformatics/bty121 PMID: 29506019 PMCID: PMC6041967

54.
S. Capella-Gutierrez, J. M. Silla-Martinez, T. Gabaldon
(2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses *Bioinformatics*
DOI: 10.1093/bioinformatics/btp348 PMID: 19505945 PMCID: PMC2712344

55.
Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, Bui Quang Minh
(2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies *Molecular Biology and Evolution*
DOI: 10.1093/molbev/msu300 PMID: 25371430 PMCID: PMC4271533

56.
R. Lanfear, B. Calcott, S. Y. W. Ho, S. Guindon
(2012) PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses *Molecular Biology and Evolution*
DOI: 10.1093/molbev/mss020 PMID: 22319168

57.
Jaime Huerta-Cepas, François Serra, Peer Bork
(2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data *Molecular Biology and Evolution*
DOI: 10.1093/molbev/msw046 PMID: 26921390 PMCID: PMC4868116

58.
Michael Waskom, Olga Botvinnik, Joel Ostblom, Saulius Lukauskas, Paul Hobson, MaozGelbart, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, ... Constantine Evans
(2020) mwaskom/seaborn: v0.10.0 (January 2020) *Zenodo*
DOI: 10.5281/zenodo.3629446

59.
Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, Loren J Hauser
(2010) Prodigal: prokaryotic gene recognition and translation initiation site identification *BMC Bioinformatics*
DOI: 10.1186/1471-2105-11-119 PMID: 20211023 PMCID: PMC2848648

60.
Brian P Alcock, Amogelang R Raphenya, Tammy TY Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V Nguyen, Annie A Cheng, Sihan Liu, ... Andrew G McArthur
(2019) CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database *Nucleic Acids Research*
DOI: 10.1093/nar/gkz935 PMID: 31665441 PMCID: PMC7145624

61.
Bo Liu, Dandan Zheng, Qi Jin, Lihong Chen, Jian Yang
(2019) VFDB 2019: a comparative pathogenomic platform with an interactive web interface *Nucleic Acids Research*
DOI: 10.1093/nar/gky1080 PMID: 30395255 PMCID: PMC6324032

62.
Nancy Y. Yu, James R. Wagner, Matthew R. Laird, Gabor Melli, Sébastien Rey, Raymond Lo, Phuong Dao, S. Cenk Sahinalp, Martin Ester, Leonard J. Foster, Fiona S. L. Brinkman
(2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization

subcategories and predictive capabilities for all prokaryotes *Bioinformatics*
DOI: 10.1093/bioinformatics/btq249 PMID: 20472543 PMCID: PMC2887053

63.
Sergio Arredondo-Alonso, Rob J. Willems, Willem van Schaik, Anita C. Schürch
(2017) On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data *Microbial Genomics*
DOI: 10.1099/mgen.0.000128 PMID: 29177087 PMCID: PMC5695206

64.
David Pellow, Maraike Probst, Ori Furman, Alvah Zorea, Arik Segal, Itzik Mizrahi, Ron Shamir
(2020) SCAPP: An algorithm for improved plasmid assembly in metagenomes *bioRxiv*
DOI: 10.1101/2020.01.12.903252

65.
Daniel J Giguere, Alexander T Bahcheli, Benjamin R Joris, Julie M Paulssen, Lisa M Gieg, Martin W Flatley, Gregory B Gloor
(2020) Complete and validated genomes from a metagenome *bioRxiv*
DOI: 10.1101/2020.04.08.032540

66.
Yoshihiko Suzuki, Suguru Nishijima, Yoshikazu Furuta, Jun Yoshimura, Wataru Suda, Kenshiro Oshima, Masahira Hattori, Shinichi Morishita
(2019) Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut *Microbiome*
DOI: 10.1186/s40168-019-0737-z PMID: 31455406 PMCID: PMC6712665

67.
Anuradha Ravi, Fenella D Halstead, Amy Bamford, Anna Casey, Nicholas M. Thomson, Willem van Schaik, Catherine Snelson, Robert Goulden, Ebenezer Foster-Nyarko, George M. Savva, ... Beryl A. Oppenheim
(2019) Loss of microbial diversity and pathogen domination of the gut microbiota in critically ill patients *Microbial Genomics*
DOI: 10.1099/mgen.0.000293 PMID: 31526447 PMCID: PMC6807385

68.
Zongbao Liu, Uli Klümper, Yang Liu, Yuchun Yang, Qiaoyan Wei, Jih-Gaw Lin, Ji-Dong Gu, Meng Li
(2019) Metagenomic and metatranscriptomic analyses reveal activity and hosts of antibiotic resistance genes in activated sludge *Environment International*
DOI: 10.1016/j.envint.2019.05.036 PMID: 31129497

69.
Eric Newberry, Rishi Bhandari, Joseph Kemble, Edward Sikora, Neha Potnis
(2020) Genome-resolved metagenomics to study co-occurrence patterns and intraspecific heterogeneity among plant pathogen metapopulations *Environmental Microbiology*
DOI: 10.1111/1462-2920.14989 PMID: 32207218

70.
Ya Zhang, Masaaki Kitajima, Andrew J. Whittle, Wen-Tso Liu
(2017) Benefits of Genomic Insights and CRISPR-Cas Signatures to Monitor Potential Pathogens across Drinking Water Production and Distribution Systems *Frontiers in Microbiology*
DOI: 10.3389/fmicb.2017.02036 PMID: 29097994 PMCID: PMC5654357

71.
Andrew D. Huang, Chengwei Luo, Angela Pena-Gonzalez, Michael R. Weigand, Cheryl L. Tarr, Konstantinos T. Konstantinidis
(2016) Metagenomics of Two Severe Foodborne Outbreaks Provides Diagnostic Signatures and Signs of Coinfection Not Attainable by Traditional Methods *Applied and Environmental Microbiology*
DOI: 10.1128/aem.02577-16 PMID: 27881416 PMCID: PMC5244306

72.
Jonathan L Klassen, Cameron R Currie
(2012) Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation *BMC Genomics*
DOI: 10.1186/1471-2164-13-14 PMID: 22233127 PMCID: PMC3322347

73.
Lalanika M. Abayasekara, Jennifer Perera, Vishvanath Chandrasekharan, Vaz S. Gnanam, Nisala A. Udunuwara, Dileepa S. Liyanage, Nuwani E. Bulathsinhala, Subhashanie Adikary, Janith V. S. Aluthmuhandiram, Chrishanthi S. Thanaseelan, ... Janahan Ilango
(2017) Detection of bacterial pathogens from clinical specimens using conventional microbial culture and 16S metagenomics: a comparative study *BMC Infectious Diseases*
DOI: 10.1186/s12879-017-2727-8 PMID: 28927397 PMCID: PMC5606128

74.
G. B. Rogers, M. P. Carroll, D. J. Serisier, P. M. Hockey, G. Jones, K. D. Bruce
(2004) Characterization of Bacterial Community Diversity in Cystic Fibrosis Lung Infections by Use of 16S Ribosomal DNA Terminal Restriction Fragment Length Polymorphism Profiling *Journal of Clinical Microbiology*
DOI: 10.1128/jcm.42.11.5176-5183.2004 PMID: 15528712 PMCID: PMC525137

75.
Aline C. Freitas, Bonnie Chaban, Alan Bocking, Maria Rocco, Siwen Yang, Janet E. Hill, Deborah M. Money, The VOGUE Research Group
(2017) The vaginal microbiome of pregnant women is less rich and diverse, with lower prevalence of Mollicutes, compared to non-pregnant women *Scientific Reports*
DOI: 10.1038/s41598-017-07790-9 PMID: 28835692 PMCID: PMC5569030

76.
Marcin Gołębiewski, Edyta Deja-Sikora, Marcin Cichosz, Andrzej Tretyn, Borys Wróbel
(2014) 16S rDNA Pyrosequencing Analysis of Bacterial Community in Heavy Metals Polluted Soils *Microbial Ecology*
DOI: 10.1007/s00248-013-0344-7 PMID: 24402360 PMCID: PMC3962847

77.
N. Youssef, C. S. Sheik, L. R. Krumholz, F. Z. Najar, B. A. Roe, M. S. Elshahed
(2009) Comparison of Species Richness Estimates Obtained Using Nearly Complete Fragments and Simulated Pyrosequencing-Generated Fragments in 16S rRNA Gene-Based Environmental Surveys *Applied and Environmental Microbiology*
DOI: 10.1128/aem.00592-09 PMID: 19561178 PMCID: PMC2725448

78.
Marcus J. Claesson, Orla O'Sullivan, Qiong Wang, Janne Nikkilä, Julian R. Marchesi, Hauke Smidt, Willem M. de Vos, R. Paul Ross, Paul W. O'Toole
(2009) Comparative Analysis of Pyrosequencing and a Phylogenetic Microarray for Exploring

Microbial Community Structures in the Human Distal Intestine *PLoS ONE*
DOI: 10.1371/journal.pone.0006669 PMID: 19693277 PMCID: PMC2725325

79.
Milton Thomas, Megan Webb, Sudeep Ghimire, Amanda Blair, Kenneth Olson, Gavin John Fenske, Alex Thomas Fonder, Jane Christopher-Hennings, Derek Brake, Joy Scaria
(2017) Metagenomic characterization of the effect of feed additives on the gut microbiome and antibiotic resistome of feedlot cattle *Scientific Reports*
DOI: 10.1038/s41598-017-12481-6 PMID: 28947833 PMCID: PMC5612972

80.
Adrian Fritz, Peter Hofmann, Stephan Majda, Eik Dahms, Johannes Dröge, Jessika Fiedler, Till R. Lesker, Peter Belmann, Matthew Z. DeMaere, Aaron E. Darling, … Alice C. McHardy
(2019) CAMISIM: simulating metagenomes and microbial communities *Microbiome*
DOI: 10.1186/s40168-019-0633-6 PMID: 30736849 PMCID: PMC6368784

81.
Nicholas A. Bokulich, Jai Ram Rideout, William G. Mercurio, Arron Shiffer, Benjamin Wolfe, Corinne F. Maurice, Rachel J. Dutton, Peter J. Turnbaugh, Rob Knight, J. Gregory Caporaso
(2016) mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking *mSystems*
DOI: 10.1128/msystems.00062-16 PMID: 27822553 PMCID: PMC5080401

82.
Martin Hunt, Alison E Mather, Leonor Sánchez-Busó, Andrew J Page, Julian Parkhill, Jacqueline A Keane, Simon R Harris
(2017) ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads *Microbial Genomics*
DOI: 10.1099/mgen.0.000131 PMID: 29177089 PMCID: PMC5695208