bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

This manuscript (permalink) was automatically generated from fmaguire/mag_sim_paper@6cc889b on March 31, 2020.

Authors

- **Finlay Maguire***
  0000-0002-1203-9514 · fmaguire · fmaguire
  Faculty of Computer Science, Dalhousie University · Funded by ['Genome Canada', 'Donald Hill Family Fellowship']

- **Baofeng Jia***
  · imasianxd
  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**

  Department of Molecular Biology and Biochemistry, Simon Fraser University

- **Robert G. Beiko**

  Faculty of Computer Science, Dalhousie University

- **Fiona S.L. Brinkman**

  Department of Molecular Biology and Biochemistry, Simon Fraser University

Abstract

Motivation

Metagenomic methods have emerged as a key tool in public-health microbiology for surveillance of virulence factor (VF) and antimicrobial resistance (AMR) genes. However, metagenomic data, even when assembled, typically results in complex, mixed sets DNA sequence fragments rather than fully resolved individual genomes. Recently, metagenome-assembled genomes (MAGs) have emerged as a promising approach that groups sequences into bins that are likely derived from the same underlying genome. However, MAGs have not been well assessed for their ability to identify some of the key sequences of interest for infectious disease surveillance purposes: AMR and VFs associated with mobile genetic elements (MGEs) such as plasmids and genomic islands (GIs). We hypothesized that due to the different copy number and sequence composition of plasmids and GIs compared to core genome sequence, such sequences will be under-represented in MAG-based approaches.

Results

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To evaluate the impact of MAG recovery methods on recovery of AMR genes and MGEs, we generated a simulated metagenomic dataset comprised of 30 genomes with up to 16.65% of the chromosomal DNA consisting of GIs and 65 associated plasmids. MAGs were then recovered from this data using 12 different MAG pipelines and evaluated for recovery accuracies. Across all pipelines, 81.9-94.3% of chromosomes were recovered and binned. However, only 37.8-44.1% of GIs and 1.5-29.2% of plasmids were recovered and correctly binned at >50% coverage. In terms of AMR and VF genes associated with MGEs, 0-45% of GI-associated AMR genes and 0-16% of GI-associated VF genes were correctly assigned. More strikingly, 0% of plasmid-borne VF or AMR genes were recovered. This work shows that regardless of the MAG recovery approach used, plasmid and GI dominated sequences will disproportionately be left unbinned or incorrectly binned. From a public-health perspective, this means MAG approaches are less suited for analysis of mobile genes, especially key groups such as AMR and VF genes. This underlines the utility of read-based and long-read approaches to thoroughly evaluate the resistome in metagenomic data.

Introduction

Metagenomics, the sequencing of fragments of DNA from within an environmental sample, is widely used for characterising microbial communities [1,2]. By randomly sampling from the total genomic content these methods allow researchers to simultaneously profile the functional potential and the taxonomic identity of a large proportion of the organisms in a sample. Metagenomic techniques are now being used to profile antimicrobial resistance (AMR) and pathogen virulence. 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].

While long-read DNA sequencing technology (e.g., Oxford Nanopore [7], PacBio [8]) is now being used for metagenomic sequencing [9,10], high-throughput sequencing of relatively short reads (150-250bp) in platforms such as the Illumina MiSeq currently dominate metagenomic analyses. Inference of taxonomic and functional diversity can be assessed directly from sequenced reads using reference databases and BLAST-based sequence similarity search tools (e.g. DIAMOND [11]), read mapping (e.g. Bowtie 2 [12]), Hidden Markov Models (e.g. HMMER3 [13]) or k-mer hashing (e.g. CLARK [14]). These read-based approaches allow analysis of all reads with detectable similarity to genes of interest even if the organism has relatively low abundance in the sample. Since these reads are shorter than most genes, however, read-based methods provide very little information about the genomic organisation of genes. This lack of contextual information is particularly problematic in the study of AMR genes and virulence factors as the genomic context plays a role in function [15], selective pressures [16], and then likelihood of lateral gene transfer (LGT) [17].

Sequence assembly is often used to generate information about genomic context [18]. de Bruijn graph-based assemblers have been developed to handle the particular challenges of this type of data including metaSPAdes [19] , IDBA-UD [20], and megahit [21]. A crucial challenge in metagenomic analysis is that reads from different organisms must be disentangled to avoid hybrid assemblies. A common way to deal with this challenge is to assign all contigs from a given source genomes to a cluster or “bin” based on similarities in the relative abundance and sequence composition. These resulting bins are often known as metagenome-assembled genomes (MAGs). This binning is typically performed by grouping all the contigs with similar abundance and similar sequence composition into the same bin. A range of tools have been released to perform this binning including CONCOCT [22], MetaBAT 2 [23], and MaxBin 2 [24]. There is also the meta-binning tool DAS Tool [25] which combines predictions from multiple binning tools together. These MAG approaches have been used to great effect in unveiling huge amounts of previously uncharacterised genomic diversity [26,27,28].

There is loss of information at both the metagenomic assembly and binning steps. This compounded data loss means that only a relatively small proportion of reads are successfully assembled and binned in large complex metagenome datasets, for example, 24.2-36.4% of reads from permafrost

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[29] and soil metagenomes [30]. Additionally, a large number of detected genomes are not reconstructed at all with ~23% of all detected genomes recovered in some examples [30]. The Critical Assessment of Metagenome Interpretation (CAMI) challenge's (https://data.cami-challenge.org/) Assessment of Metagenome BinnERs (AMBER) [31] assesses the global completeness and purity of recovered MAGs across methods. However, to our best knowledge, there hasn't been a specific assessment of the impact of metagenomic assembly and binning on the loss of specific genomic elements. In particular, the impact on mobile genetic elements (MGEs), such as genomic islands (GIs) and plasmids, which can be of great health and research importance, has not been evaluated.

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) [32,33]; they disproportionately encode virulence factors [34] and are a major mechanism of LGT of AMR genes [35,36]. GIs often have different nucleotide composition compared to the rest of the genome [32]. This compositional difference is exploited by tools designed to detect GIs such as SIGI-HMM [37] and IslandPath-DIMOB [38]. GIs may exist as multiple copies within a genome [39] leading to potential difficulties in correctly assembling these regions in metagenome assemblies as well as likely biases in the calculation of coverage statistics.

Plasmids are circular or linear extrachromosomal self-replicating pieces of DNA. Similar to GIs, plasmids's sequence composition are markley different compared to the genome they are associated with [40,41]. This is largely attributable to their repetitive sequences, variable copy number, and different selection pressures [42,43]. Plasmids are of great research importance, these elements are a major source of the lateral dissemination of AMR genes throughout microbial ecosystems [35,44]. Due to these reasons, the correct assembly of DNA sequence of plasmid origin has proven to be difficult from short read data [45].

GIs and plasmids pose significant challenges in MAG recovery due to their unusual sequence composition and relative abundance; as these MGEs are key to the function and spread of pathogenic traits such as AMR and virulence, it is vital that we assess the impact of metagenome assembly and binning on the representation of these specific elements. This is particularly important with the increasing popularity of MAG approaches within microbial and public-health research. Therefore, to address this issue we performed an analysis of GI and plasmid recovery accuracy across a set of state-of-the-art short-read metagenome assembly and binning approaches using a simulated metagenome comprised of GI- and plasmid-rich taxa.

Materials and Methods

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

All genomes were selected from the set of completed RefSeq genomes as of April 2019. Genomic islands for these genomes were previously predicted using IslandPath-DIMOB [38] and collated into the IslandViewer database www.pathogenomics.sfu.ca/islandviewer [47]. Plasmid sequences were recovered for each genome using the linked GenBank Project IDs. Thirty genomes 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

The data used to select the taxa is listed in Supplemental Table 1 and the details of the selected subset taxa are listed in Supplemental Table 2 with their NCBI accessions. The sequences themselves are available in the data repository osf.io/nrejs/ under "data/sequences".

In accordance with the recommendation in the CAMI challenge [48] 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) [49] 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.

Metagenome Assembled Genome Recovery

Reads were trimmed using sickle (v1.33) [50] resulting in 25,682,644 surviving read pairs. The trimmed reads were then assembled using 3 different metagenomic assemblers: metaSPAdes (v3.13.0) [19], IDBA-UD (v1.1.3) [20], and megahit (v1.1.3) [21]). The resulting assemblies were summarised using metaQUAST (v5.0.2) [51]. 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 calculated using the MetaBAT2 accessory script (jgi_summarize_bam_contig_depths). The three metagenome assemblies were then separately binned using MetaBAT2 (v2.13) [23], and MaxBin 2 (v2.2.6) [24]. MAGs were also recovered using CONCOCT (v0.4.2) [22] 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 [25]. 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+ [52] 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. Coverages less than 5% were filtered out and the number of different genomes that contigs

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 [53]. The 86 of these proteins that were found in every reference genome were concatenated and aligned using MAFFT v7.427 [54] and masked with trimal v1.4.1-3 [55]. A maximum-likelihood phylogeny was then inferred with IQ-Tree v1.6.12 [56] with the in-built ModelFinder determined partitioning [57]. Pairwise branch distances were then extracted from the resulting tree using ETE3 v3.1.1 [58] and regressed using a linear model against coverage and contamination in seaborn v0.10.0 [59].

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 [60] 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) [61]. Virulence factors were predicted using the predicted ORFs and BLASTX 2.9.0+ [52] 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% [62]. 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 localization predictions

We then sought to assess what the impact of a protein's predicted subcellular localization 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 [60] 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 [63].

Results

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Recovery of Genomic Elements

Chromosomes

The overall ability of MAG methods to recapitulate the original chromosomal source genome results varied widely. We considered the "identity" of a given MAG bin to be that of the genome that composes 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 comprised 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. 11).

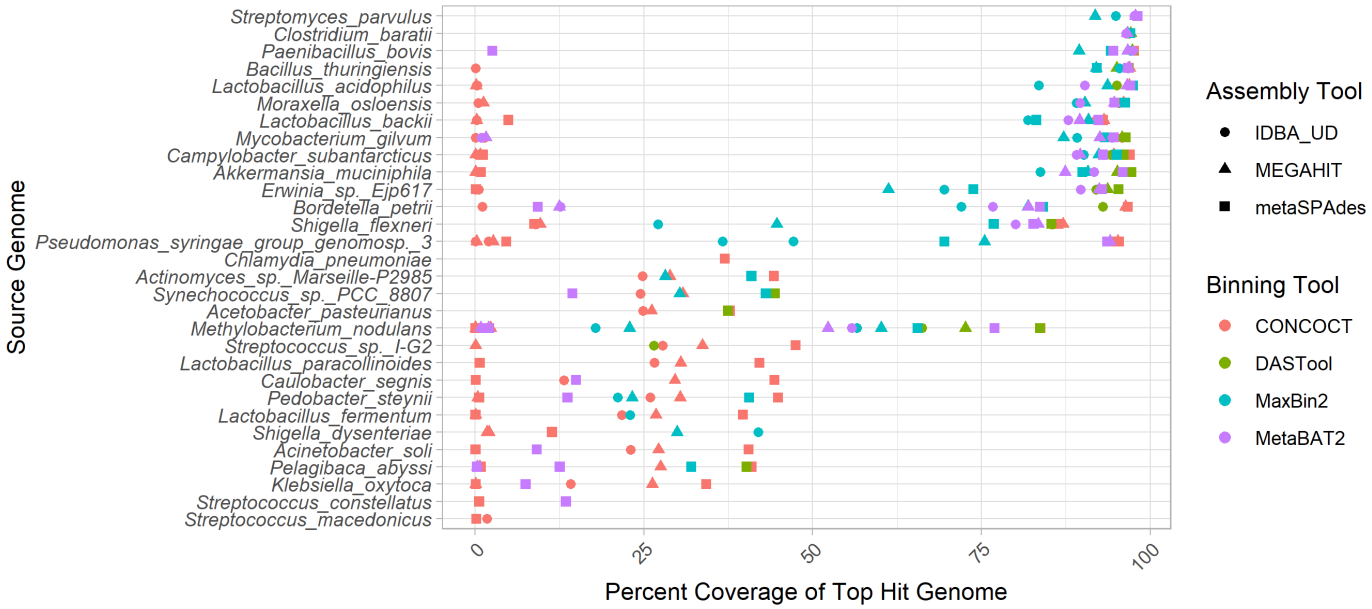


Figure 1: Top genome coverage for input genomes across MAG binners. Each dot represents the coverage of a specified genome when it comprised the plurality of the sequences in a bin. The binning tool is indicated by the colour of the dot as per the legend. Genomes such as *Clostridium baratti* were accurately 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). In terms of binning tool, 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)).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

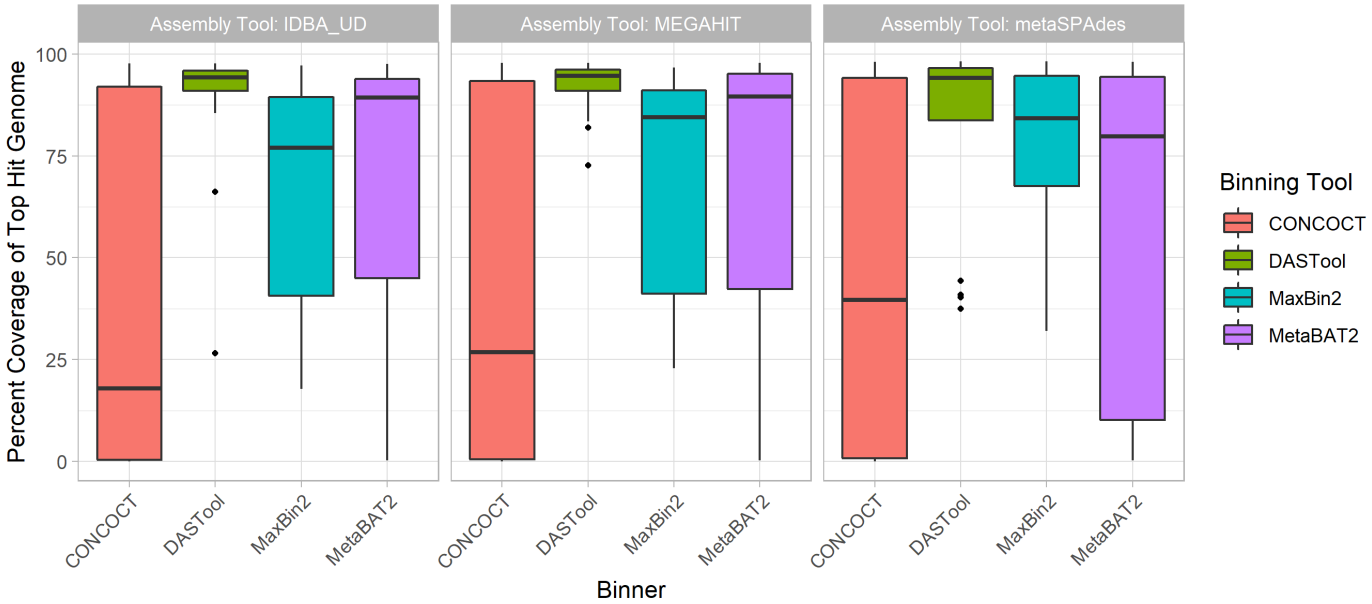


Figure 2: Chromosomal coverage of most prevalent genome in each bin across binners and metagenome assemblies. Of the 3 assemblers (y-axis), megahit resulted in the highest median chromosomal coverage (x-axis) across all binners (colored bars) at 81.9% with metaSPAdes performing the worst (76.8%). Of the 4 binners, CONCOCT (blue) performed poorly with a median coverage, followed by maxbin2 (yellow), MetaBAT2 (red) and DASTool (green) performing the best. Diamonds in the figure represent outliers (greater or lower than the interquartile range marked by the error bars) and box represents the lower quartile, median, and upper quartile respectively.

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. In terms of binning tools, however, maxbin2 proved an exception with nearly twice as many bins containing multiple species as the next binner (Fig. 3). 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. 12).

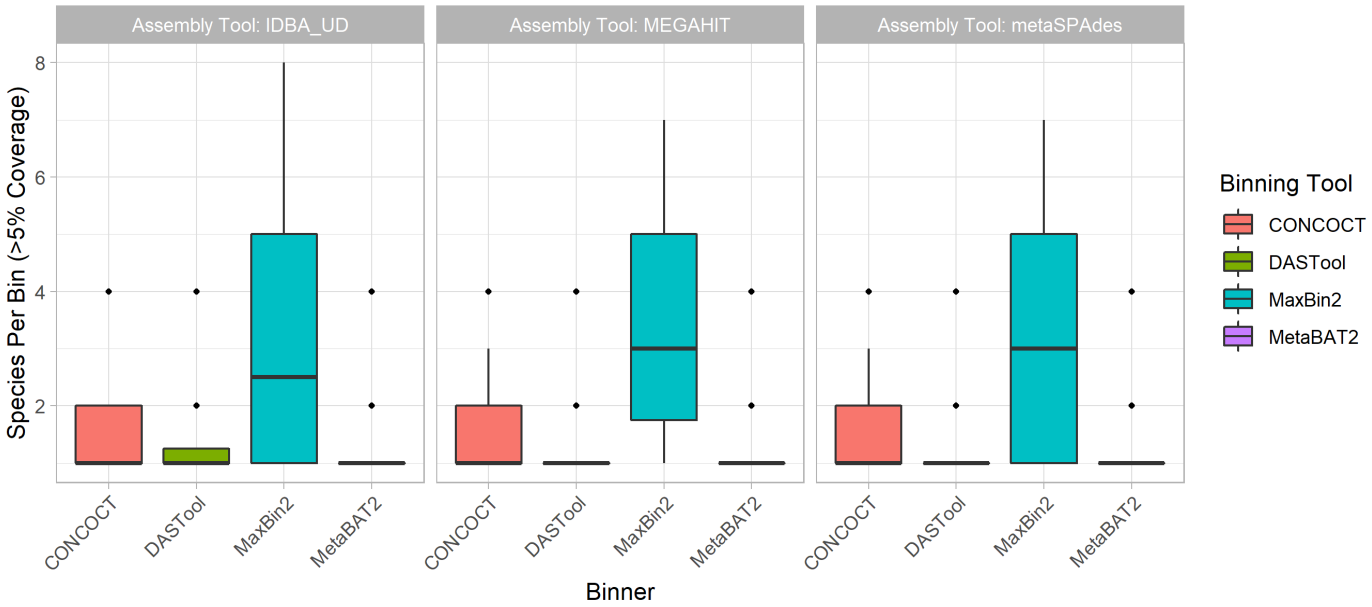


Figure 3: Distribution of bin purity across assemblers and binners. The total number of genomes present in a bin at >5% coverage (y-axis) was largely equivalent across assemblers (x-axis). In terms of binning tools, maxbin2 (orange) produced nearly twice as many bins containing multiple species compared to CONCOCT (blue), MetaBAT2 (red) and DASTool (green), which all produced chimeric bins at roughly the same rate. Similar to above, outliers outside the interquartile range marked by the error bars are shown as diamonds.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Plasmids

Regardless of method, a very small proportion of plasmids were correctly grouped in the bin that was principally comprised 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; however, of those assigned to bins nearly all were correctly binned (Fig. 4).

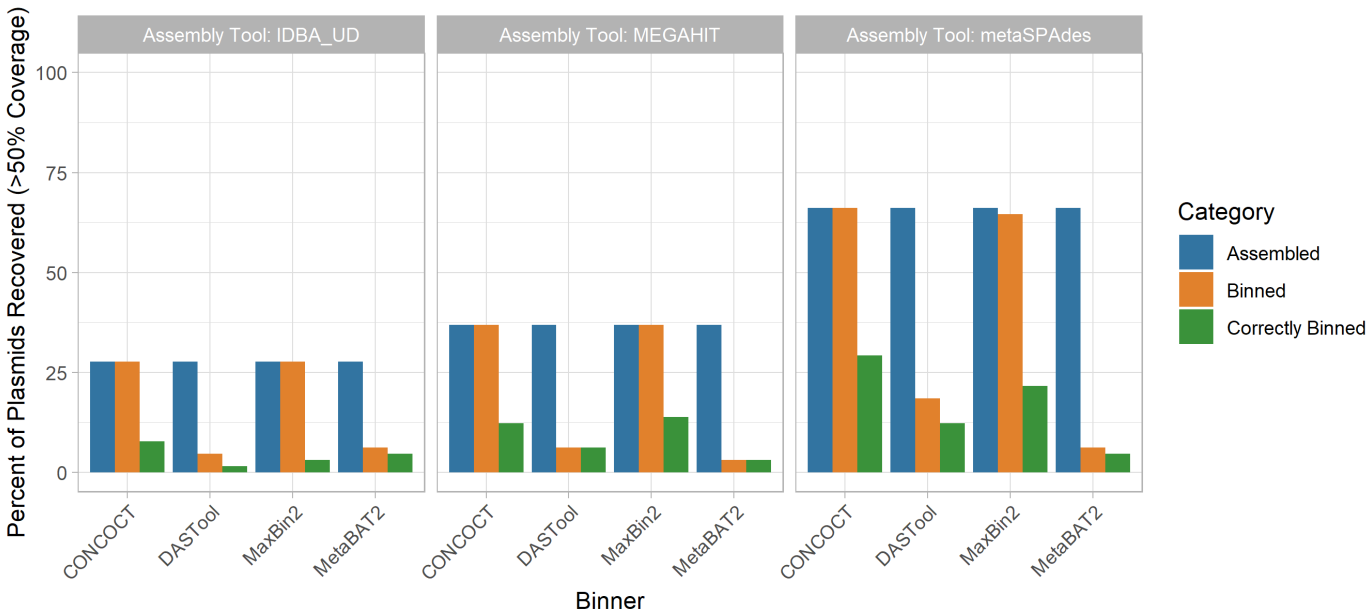


Figure 4: The performance of metagenomic assembly and binning to recover plasmid sequences. Each plot represents a different metagenome assembler, with the groups of bars along the x-axes showing the plasmid recovery performance of each binning tool when applied to the assemblies produced by that tool. 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. 5). 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

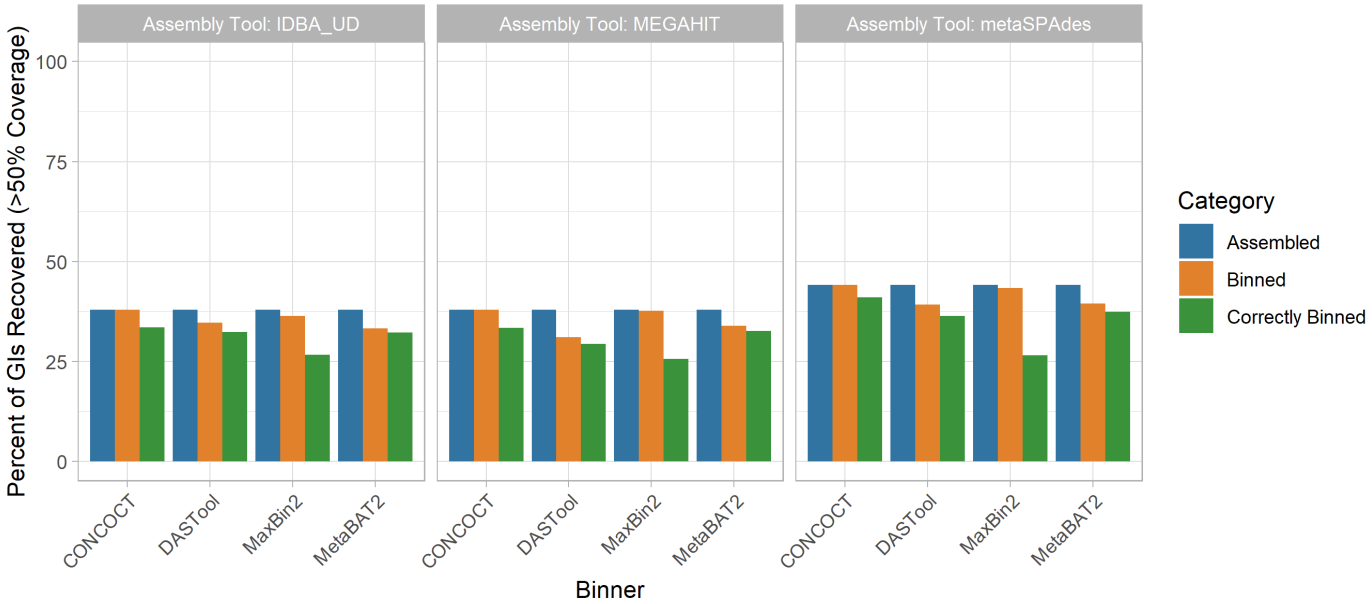


Figure 5: Impact of metagenomic assembly and MAG binning on recovery of genomic islands. GIs were recovered in a similarly poor fashion to plasmids. Generally, <40% were correctly assigned to the same bin majorly comprised of chromosomal contigs from the same source genome regardless of binning (x-axis) and assembly (facet) 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. 6). 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%).

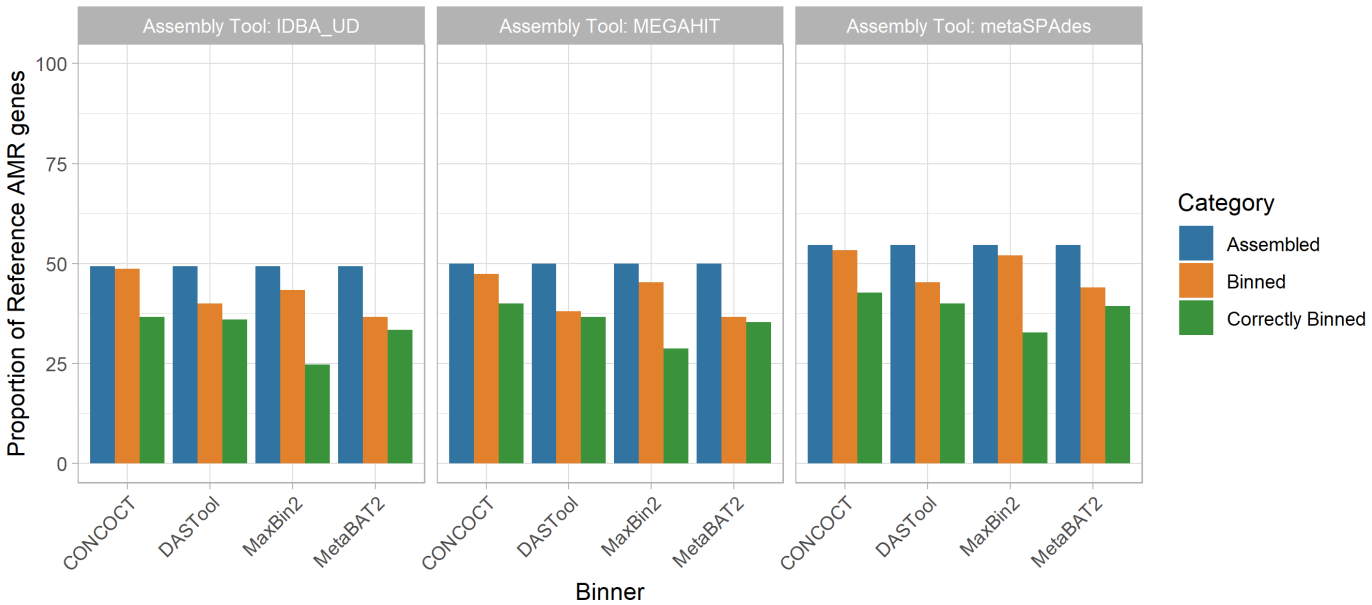


Figure 6: Recovery of AMR genes across assemblers and binners. The proportion of reference AMR genes recovered (y-axis) was largely similar across assembly tools (blue), at roughly 50% with metaSPAdes performing marginally better. Binning tools resulted in a small reduction in AMR genes recovered (orange), however only 24-40% of all AMR genes were correctly binned (green). metaSPAdes-CONCOCT was the best performing MAG binning pipeline.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Moreover, focusing on only the AMR genes that were correctly binned (Fig. 7) 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 genomic island located AMR genes (n=11) and none of the plasmid-localised AMR genes (n=20) were correctly binned.

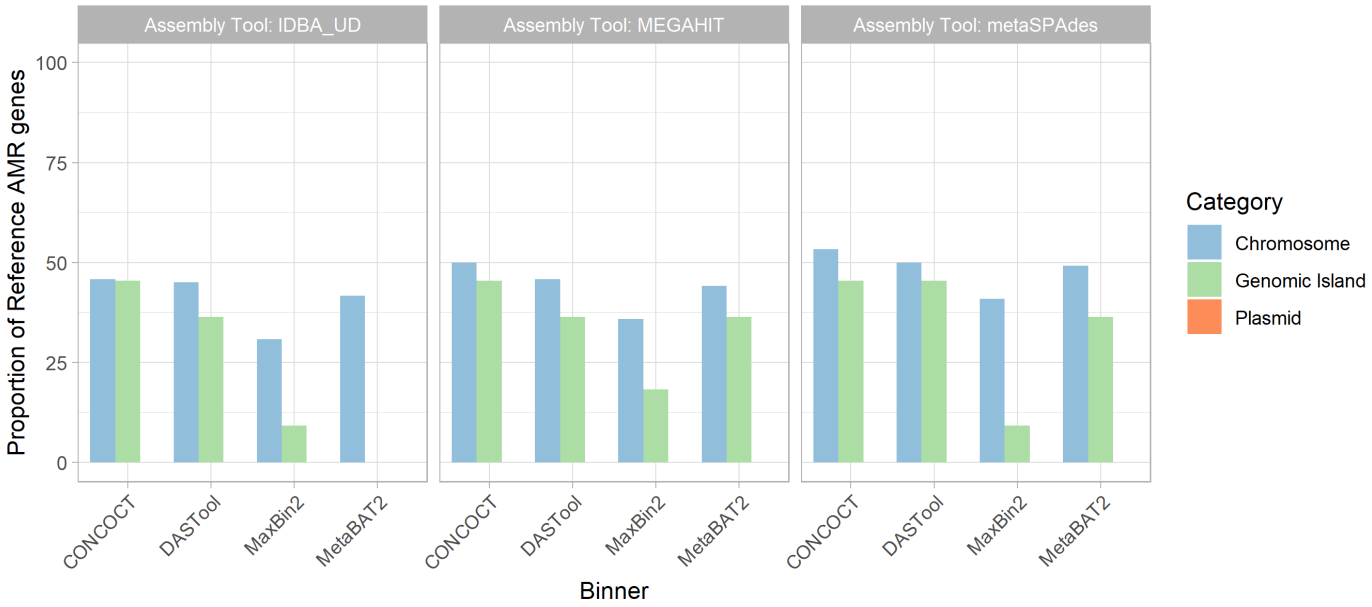
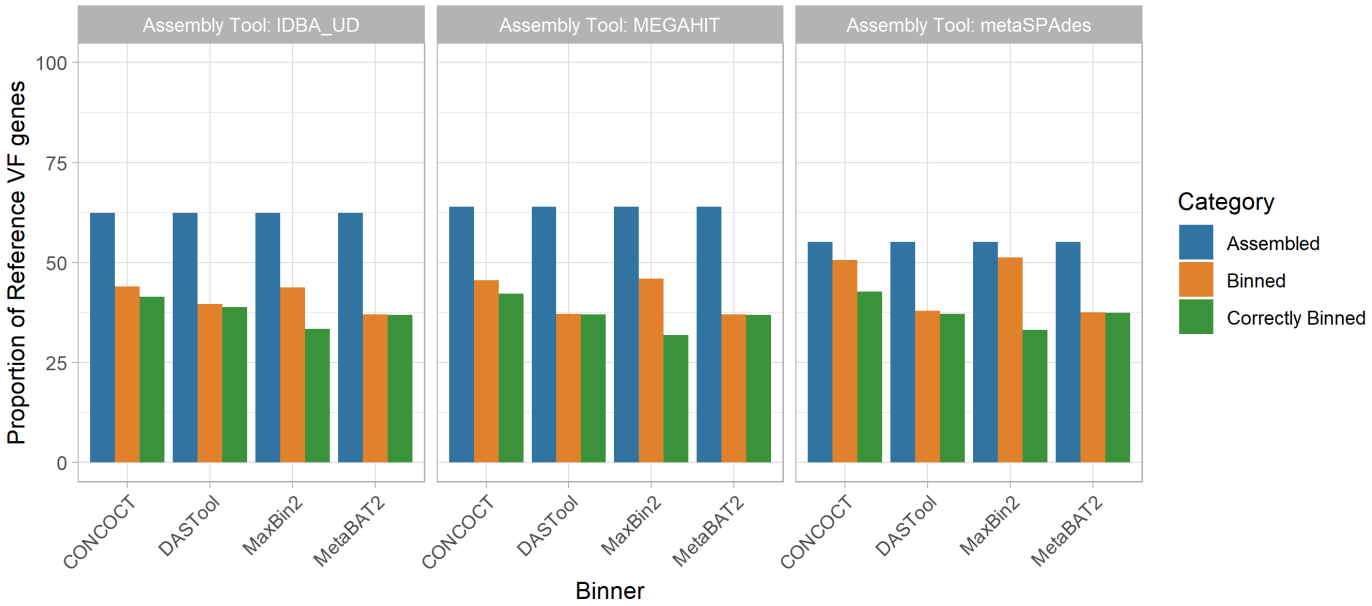


Figure 7: Percent of correctly binned AMR genes recovered by genomic context. MAG methods were best at recovering chromosomally located AMR genes (light blue) regardless of metagenomic assembler or binning tool used. Recovery of AMR genes in GIs showed a bigger variation between tools (light green). None of the 12 evaluated MAG recovery methods were able to recover plasmid located AMR genes (orange).

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. 9). 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%) and CONCOCT-metaSPAdes performing the best (4%). 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.



bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 8: 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 majority of cases, VF genes that are binned are correctly binned (green). metaSPAdes-CONCOCT was again the best performing pair.

As with AMR genes, the genomic context (chromosome, plasmid, GI) of a given VF largely determined how well it was binned (Fig. 9). 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.

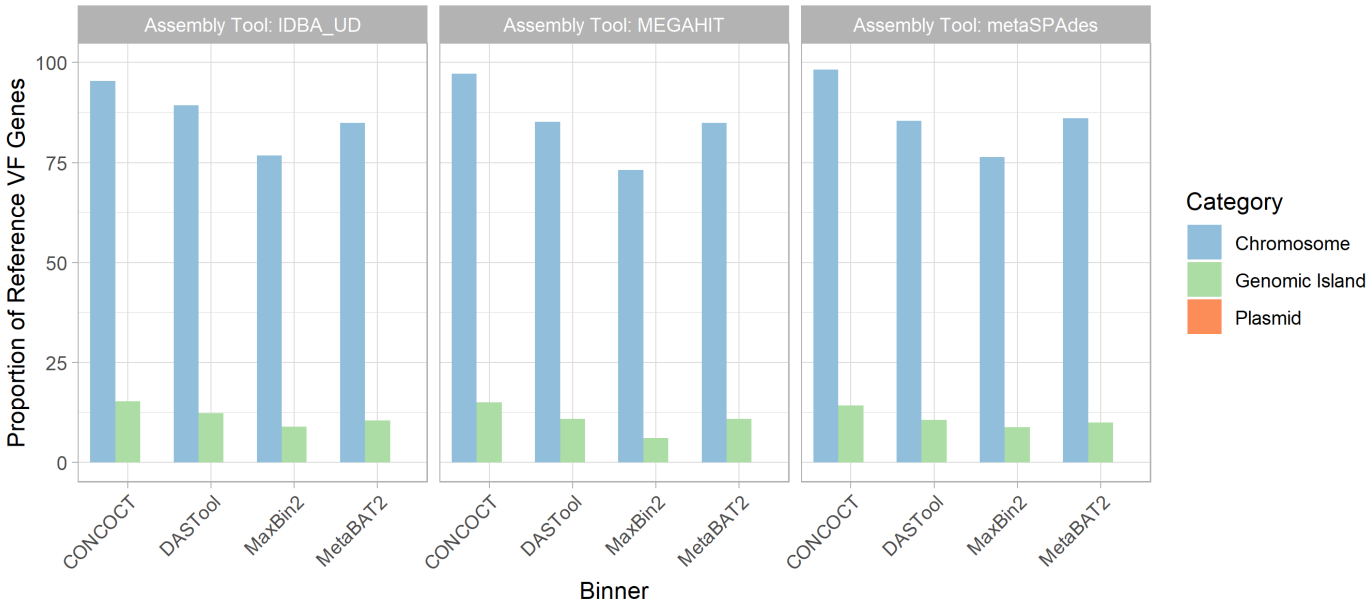


Figure 9: Percent of correctly binned VF genes recovered in each genomic region. Metagenome assembled genomes (MAGs) were again best at recovering chromosomally located VF genes (light blue), able to correctly bin majority of chromosomally located VFs. GIs recovered again performed very poorly (light green) and again none of the plasmid located AMR genes (orange) was correctly binned.

Comparisons of Rates of Loss

We combined the performance metrics for Figs. 4, 5, 6, and 9 to compare the rates of loss of different components (see Fig. 13). This highlighted that genomic components (GIs and plasmids) and plasmids in particular are lost at a higher rate than individual gene types during MAG recovery.

Discussion

In this paper, we evaluated the ability and accuracy of metagenome-assembled genome (MAGs) binning methods to correctly recover mobile genetic elements (i.e. genomic islands and plasmids) from metagenomic samples across different tools used to assemble and bin MAGs.

Overall, the best assembler-binner pair was megahit-DASTOOL in terms of both chromosomal coverage (94.3%) and bin purity (1). Looking at genomes with the lowest coverage, the three Streptococcus genomes that were recovered poorly are likely due to their similarity (Fig. 11, 12). This supports the intuition that MAG recovery approaches struggle to distinguish closely related species. While CONCOCT performed significantly worse than other binners in terms of chromosomal coverage and bin purity, we did notice that CONCOCT was prone to generating many small partial bins. Potentially, CONCOCT binning could be used to distinguish closely related species but at a cost of more fragmented genomes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

While the overall recovery and binning of chromosomes was likely sufficient for some use-cases, we were specifically interested in the ability of MAG methods to appropriately recover MGEs. This was due to the importance of MGEs in the function and spread of pathogen traits such as AMR and virulence, as well as our hypothesis that these sequences may prove difficult to bin. Regardless of the the metagenomic assembly approach or MAG binning method used, both plasmids and GIs were disproportionately lost compared to chromosomes in general. 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). The >50% coverage requirement set a high bar and more GIs and plasmids were likely recovered in more incomplete forms. Partial MGEs may be useful for some research, but for researchers interested in selective pressures and lateral gene transfer this may lead to inaccurate inferences.

This poor result is not unexpected as genomic islands and plasmids have known divergent compositional features and are often repetitive with variable copy numbers relative to the chromosome. Furthermore, the difference between the percentages suggests that binning plasmids is harder than binning GIs. This difference might be attributed to the known difficulties in assembly of plasmids from short-read data [64]. Therefore, binning efficiency might improve if we use DNA sequencing and assembly methods optimised for recovering plasmids [45] (such as SCAPP [65]).

Due to the importance of MGEs in the dissemination of clinically relevant AMR genes and VFs, we explored whether or not MAG approaches can be used to provide useful insight into the LGT of these genes. With respect to AMR genes, MAG methods were able to recover roughly 40% of all AMR genes present in our reference genomes. We noted a sharp drop in the number of AMR genes detected between assemblies and MAGs, suggesting that many of these genes were left in the unbinned portion. Overall, the CONCOCT-metaSPAdes combination, while it did not recover the highest amount of AMR genes at the assembly stage, performed the best in correctly binning an AMR gene to the right species. Regardless of tools, chromosomally located AMR genes were most frequently correctly binned (as expected from the relative performance of MAGs at recovering chromosomes). While there was variability in performance, AMR genes located on GIs were correctly binned slightly less well than chromosomally located AMR genes. This variability might be explained by the fact that there were only 11 AMR genes located on GIs in our reference genomes. All 20 of the plasmid-borne AMR genes were assembled, but none were placed into MAG bins. We included high-threat MGEs-associated AMR genes such as the KPC and OXA carbapenemases. We intended on a systematic review of which AMR genes are more or less likely to end up correctly binned, however, MAGs was not able to correctly bin enough AMR genes on plasmids or GIs to allow this.

Virulence factors showed a similar trend to the AMR genes, with a recovery of ~63% of virulence factors present in the reference genomes. There still is a sharp decline in the number of VF detected between assemblies and MAGs and CONCOCT-metaSPAdes again produced the highest binning accuracy. A majority (73%-98%) of chromosomally located VF genes were also able to be correctly binned to the right species for the MAGs. However, the MAG approach performed much worse in correctly recovering GI located and plasmid located VFs, with <16% of GI VFs (n=809) correctly recovered and none of the plasmid VFs (n=3). This drastic reduction in recovery accuracy of mobile elements, especially GIs, is expected. Previous studies have found that VFs are disproportionately present on GIs[34], which might be the reason why the recovery accuracy was worse compared to AMR genes. Together, this and the AMR gene results suggest that MAG-based methods might be of limited utility in public health research focused on the transmission and dissemination of AMR genes and VFs.

One potential caveat is that some AMR genes and VFs successfully assembled in the MAGs may no longer be annotated as such due to issues with ORF prediction (see suppl. discussion & Fig. 10). 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

used [66]. Similarly, if the ORFs predicted in the MAGs differ in sequence or degree of fragmentation from the corresponding ORFs predicted in the original reference genomes (or are no longer predicted at all), this could impact recovery of AMR/VF predictions, even though the sequences themselves may be partially or fully present in the assembly.

It should also be noted that while CONCOCT performed the best in terms of recovery of both chromosomes and MGEs, it created many relatively clean but fragmentary partial MAGs. While this might be ideal for some users, caution should be taken in using CONCOCT when assuming a bin represents a whole genome.

With the recovery of plasmids, GIs, VFs, and AMR genes the same pattern was observed, a progressive loss of data in each analytical step. The process of metagenomic assembly itself generally resulted in the loss of most of these elements/genes regardless of the assembly method used. With repetitive DNA sequence particularly difficult to correctly assemble from short reads [67]. Across binning tools, the binning process resulted in further loss with a large proportion of MGEs and genes left unbinned. Finally, only a very small proportion of these elements/genes were generally correctly binned with the appropriate host chromosomes. This follows the well known, but rarely explicitly stated, idea that the more analysis you perform the more of the original data gets lost. Indeed, this is one of the reasons why the huge amount of redundancy in metagenomic sequencing is necessary (i.e. many more base-pairs of DNA must be sequenced than are in the underlying sample).

Conclusions

Using a simulated medium-complexity metagenome, this study has shown that MAG-based approaches provide a useful tool to study a bacterial species' chromosomal elements, but have severe limitations in the recovery of MGEs. The majority of these MGEs will either both fail to assemble or be incorrectly binned. The consequence of this is the disproportionate loss of key public health priority genes like VF and AMR genes. This is particularly acute as the VF and AMR genes found on these poorly recovered MGEs are generally considered the most important due to their propensity for lateral gene transfer between unrelated bacteria. Therefore, it is vital that we utilize a combination of MAGs and other methods (e.g. read-based methods) in public health metagenomic research when short-read sequencing is used. For example, targeted AMR [68], plasmid specialised assembly approaches [65], and read-based sequence homology search [11]. Without this, MAG-based methods are insufficient to thoroughly profile the resistome and provide vital epidemiological data for metagenomic data.

Supplementals

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. 10) as the chromosomal coverage (Fig. 2) and purity (Fig. 3). 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 mulpitle ORFs, leading to issues downstream labeling them correctly as AMR/VF genes. Regardless of different tools producing

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

a different number of ORFs, the recovery of AMR/VF is pretty consistent regardless of how many ORFs are predicted.

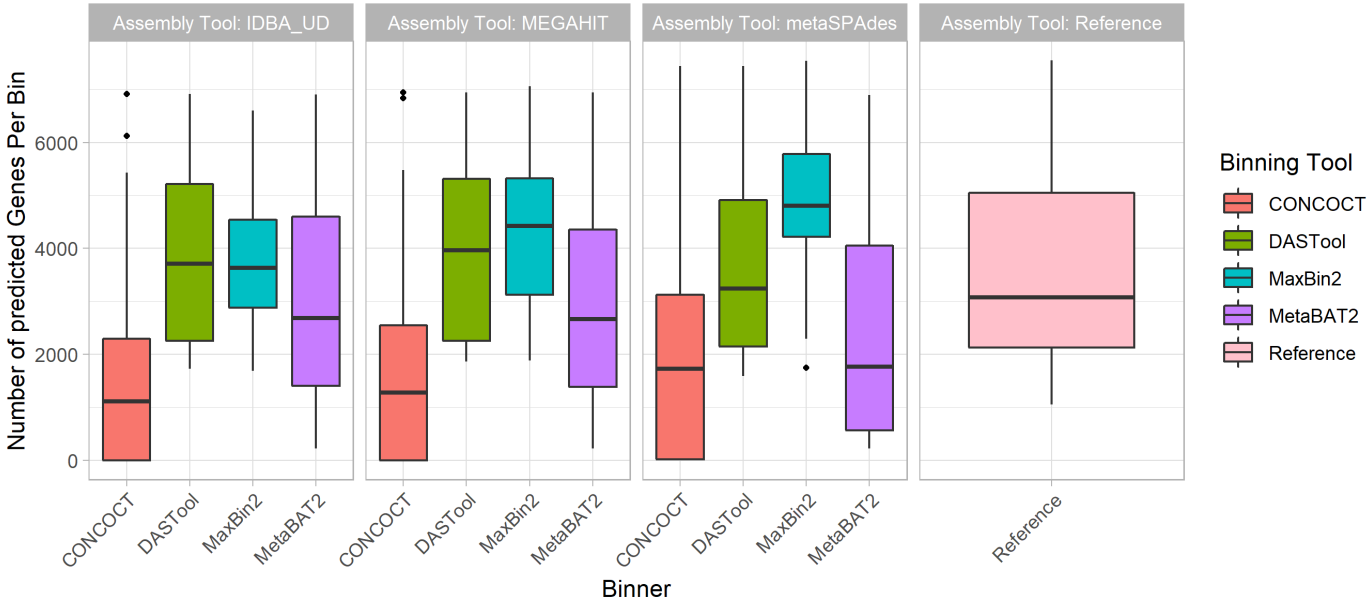


Figure 10: Predicted Gene Content. The total number of open reading frames (ORF) predicted followed the same trend as chromosomal coverage and purity. The assemblers (colored bars) did not contribute to variability in the number of ORFs detected. Of the 4 binners, CONCOCT recovered <30% of our reference genome ORFs. DASTool and MetaBAT2 predicted a similar number as our reference genomes.

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. 11) and bin purity (Fig. 12). This identified no clear relationship between chromosomal coverage and the phylogenetic distance to the nearest relative in the metagenome (Fig. ??), however, there did seem to be a negative correlation between phylogenetic distance to closest relative and the purity of a MAG bin (Fig. 12). In other words, across all methods, a MAG bin was more likely to have multiple genomes present if there were close relatives.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

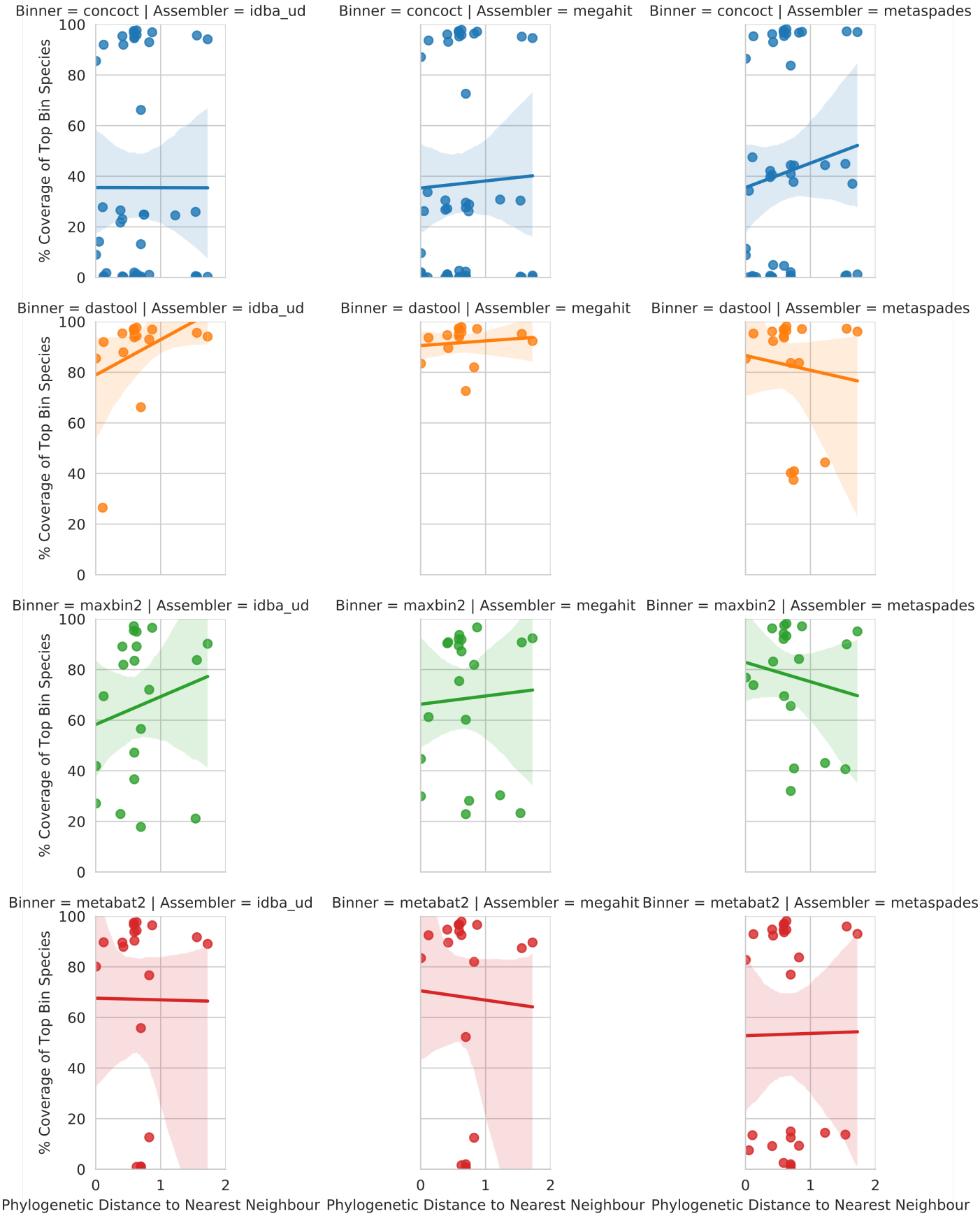


Figure 11: Relationship between phylogenetic distance to closest neighbour input genome on genomic coverage in MAG majority comprised of that taxa. Each dot represents the genomic coverage of a particular taxa and the branch distance on an 86-protein concatenated phylogeny between that taxa and its nearest neighbour. Rows indicate the binning software and columns the metagenomic assembler. Regression line is a simple linear model fitted in seaborn.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

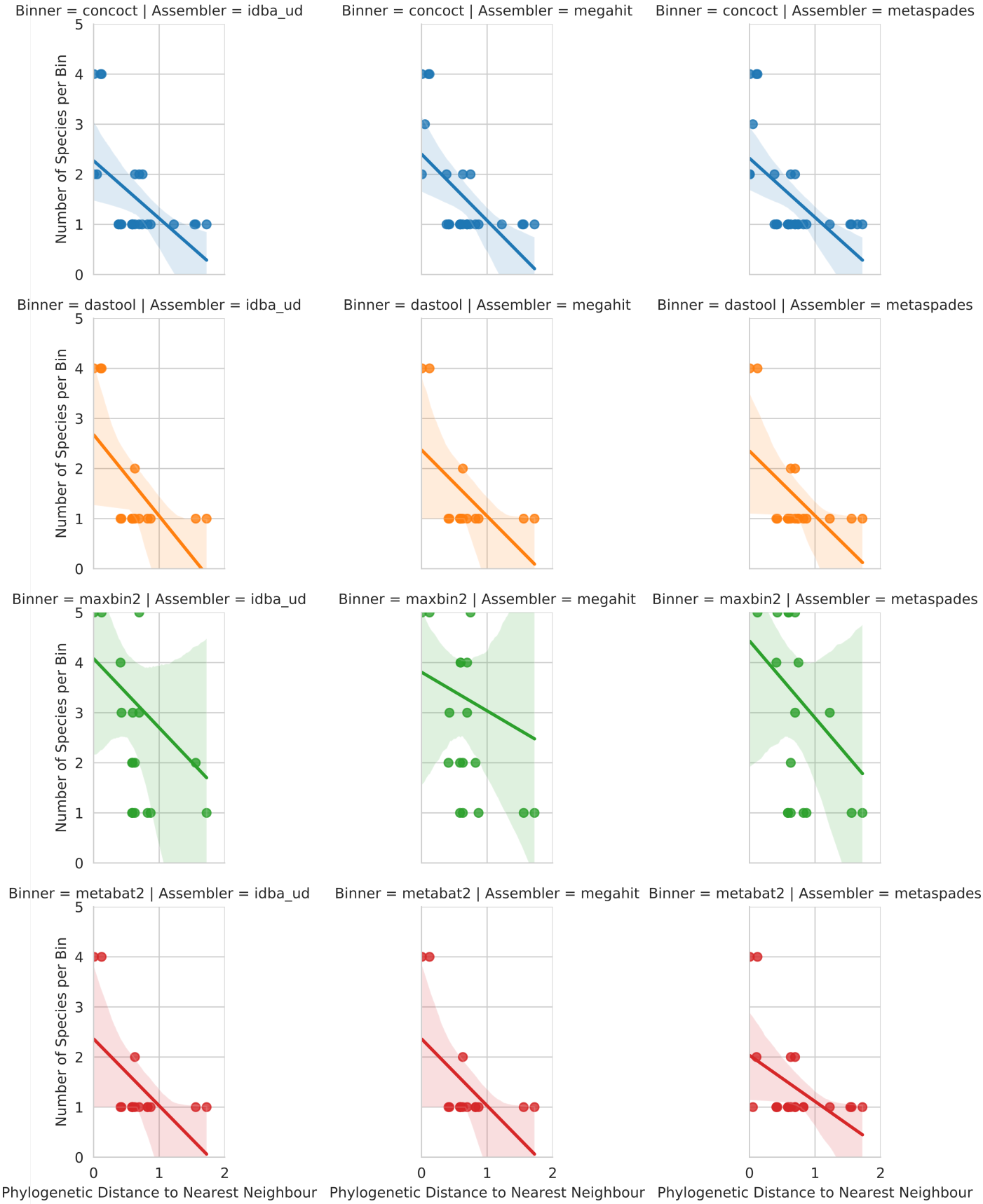


Figure 12: Relationship between phylogenetic distance to closest neighbour input genome on bin purity. Each dot shows the number of other input genomes detectable in a given MAG bin in relation to the branch distance on an 86-protein concatenated phylogeny between the majority taxa in that bin and its nearest neighbour.

Comparisons of Rates of Loss

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Combining the performance metrics for Figs. 4, 5, 6, and 9 to compare the rates of loss of different components emphasises some of the observed patterns (see Fig. 13). 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.

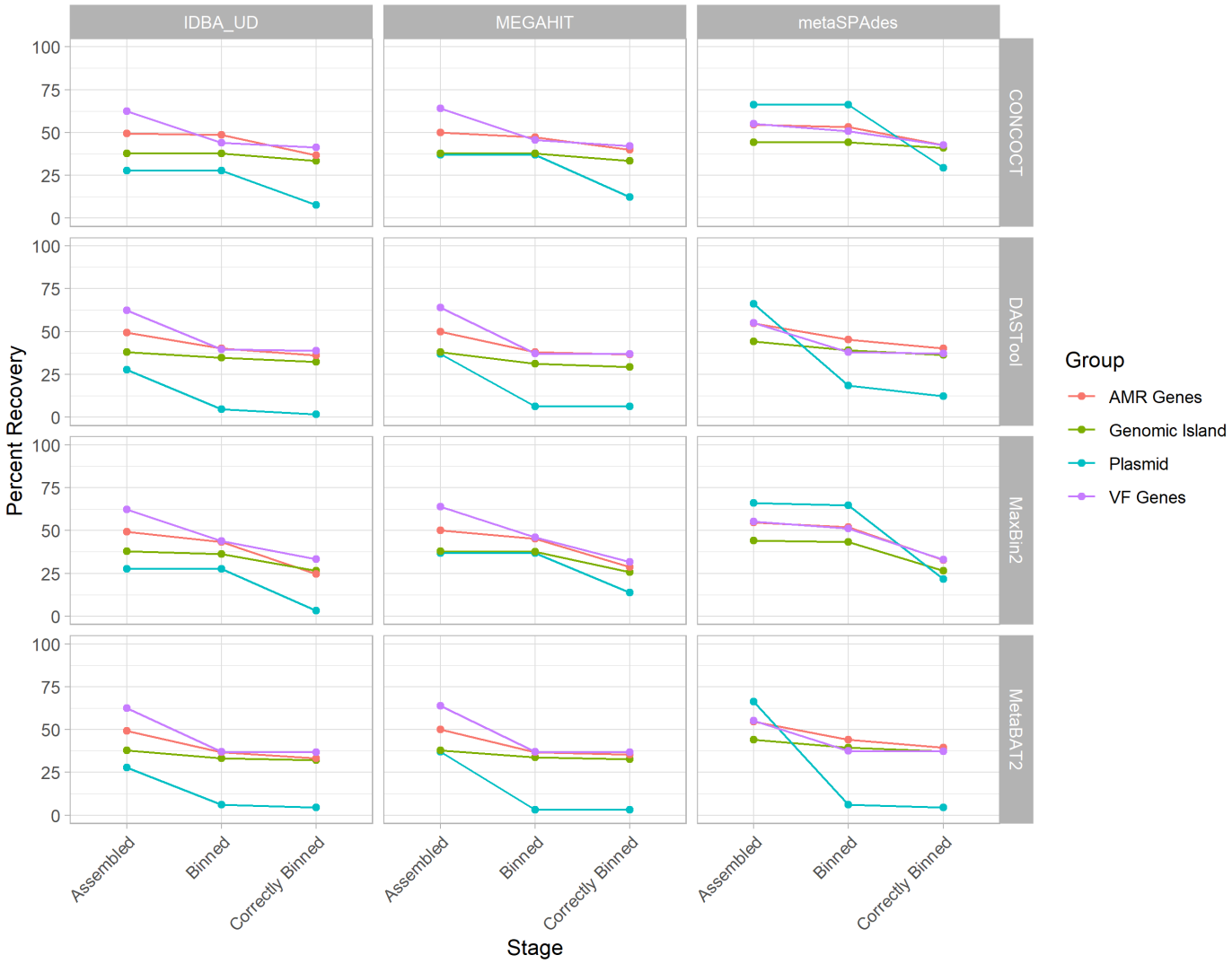


Figure 13: Comparison of rates of loss for different genomic components and gene types across assemblers and binning tools. Each line represents a different component as indicated by the legend with assemblers indicated by row and binning tool by column. This shows that regardless of approach genomic components (GIs and plasmids) are lost at a higher rate than individual VF or AMR genes.

References

1. **Genomic analysis of uncultured marine viral communities**
M. Breitbart, P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, F. Rohwer
Proceedings of the National Academy of Sciences (2002-10-16) https://doi.org/br7jq3
DOI: 10.1073/pnas.202488399 · PMID: 12384570 · PMCID: PMC137870

2. **Shotgun metagenomics, from sampling to analysis**
Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, Nicola Segata
Nature Biotechnology (2017-09-12) https://doi.org/gbv6nf
DOI: 10.1038/nbt.3935 · PMID: 28898207

3. **A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics**

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Mohamed S. Donia, Peter Cimermancic, Christopher J. Schulze, Laura C. Wieland Brown, John Martin, Makedonka Mitreva, Jon Clardy, Roger G. Linington, Michael A. Fischbach
Cell (2014-09) https://doi.org/f6k3fg
DOI: 10.1016/j.cell.2014.08.032 · PMID: 25215495 · PMCID: PMC4164201

4. **Expanding the soil antibiotic resistome: exploring environmental diversity**
Vanessa M D'Costa, Emma Griffiths, Gerard D Wright
Current Opinion in Microbiology (2007-10) https://doi.org/cfbpjj
DOI: 10.1016/j.mib.2007.08.009 · PMID: 17951101

5. **Antibiotic resistance is ancient**
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
Nature (2011-08-31) https://doi.org/b3wbvx
DOI: 10.1038/nature10388 · PMID: 21881561

6. **A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia coli O104:H4**
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
JAMA (2013-04-10) https://doi.org/f5rqft
DOI: 10.1001/jama.2013.3231 · PMID: 23571589

7. **A first look at the Oxford Nanopore MinION sequencer**
Alexander S. Mikheyev, Mandy M. Y. Tin
Molecular Ecology Resources (2014-11) https://doi.org/vmt
DOI: 10.1111/1755-0998.12324 · PMID: 25187008

8. **Real-Time DNA Sequencing from Single Polymerase Molecules**
J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, ... S. Turner
Science (2009-01-02) https://doi.org/cz7ndk
DOI: 10.1126/science.1162986 · PMID: 19023044

9. **Ultra-deep, long-read nanopore sequencing of mock microbial community standards**
Samuel M Nicholls, Joshua C Quick, Shuiquan Tang, Nicholas J Loman
GigaScience (2019-05) https://doi.org/gf39g3
DOI: 10.1093/gigascience/giz043 · PMID: 31089679 · PMCID: PMC6520541

10. **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**
Vincent Somerville, Stefanie Lutz, Michael Schmid, Daniel Frei, Aline Moser, Stefan Irmler, Jürg E. Frey, Christian H. Ahrens
BMC Microbiology (2019-06-25) https://doi.org/gf5ffc
DOI: 10.1186/s12866-019-1500-0 · PMID: 31238873 · PMCID: PMC6593500

11. **Fast and sensitive protein alignment using DIAMOND**
Benjamin Buchfink, Chao Xie, Daniel H Huson
Nature Methods (2014-11-17) https://doi.org/gftzcs
DOI: 10.1038/nmeth.3176 · PMID: 25402007

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

12. **Fast gapped-read alignment with Bowtie 2**
Ben Langmead, Steven L Salzberg
Nature Methods (2012-03-04) https://doi.org/gd2xzn
DOI: 10.1038/nmeth.1923 · PMID: 22388286 · PMCID: PMC3322381

13. **nhmmer: DNA homology search with profile HMMs**
T. J. Wheeler, S. R. Eddy
Bioinformatics (2013-07-09) https://doi.org/f5xm9x
DOI: 10.1093/bioinformatics/btt403 · PMID: 23842809 · PMCID: PMC3777106

14. **CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers**
Rachid Ounit, Steve Wanamaker, Timothy J Close, Stefano Lonardi
BMC Genomics (2015-03-25) https://doi.org/gb3h2t
DOI: 10.1186/s12864-015-1419-2 · PMID: 25879410 · PMCID: PMC4428112

15. **vanM, a New Glycopeptide Resistance Gene Cluster Found in Enterococcus faecium**
X. Xu, D. Lin, G. Yan, X. Ye, S. Wu, Y. Guo, D. Zhu, F. Hu, Y. Zhang, F. Wang, ... M. Wang
Antimicrobial Agents and Chemotherapy (2010-08-23) https://doi.org/cnpst5
DOI: 10.1128/aac.01710-09 · PMID: 20733041 · PMCID: PMC2976141

16. **Co-selection of antibiotic and metal resistance**
Craig Baker-Austin, Meredith S. Wright, Ramunas Stepanauskas, J. V. McArthur
Trends in Microbiology (2006-04) https://doi.org/fvkg6d
DOI: 10.1016/j.tim.2006.02.006 · PMID: 16537105

17. **Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens**
Hatch W. Stokes, Michael R. Gillings
FEMS Microbiology Reviews (2011-09) https://doi.org/fw543p
DOI: 10.1111/j.1574-6976.2011.00273.x · PMID: 21517914

18. **Community structure and metabolism through reconstruction of microbial genomes from the environment**
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
Nature (2004-02-01) https://doi.org/b85j5j
DOI: 10.1038/nature02340 · PMID: 14961025

19. **metaSPAdes: a new versatile metagenomic assembler**
Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, Pavel A. Pevzner
Genome Research (2017-05) https://doi.org/f97jkv
DOI: 10.1101/gr.213959.116 · PMID: 28298430 · PMCID: PMC5411777

20. **IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth**
Y. Peng, H. C. M. Leung, S. M. Yiu, F. Y. L. Chin
Bioinformatics (2012-04-11) https://doi.org/f3z7hv
DOI: 10.1093/bioinformatics/bts174 · PMID: 22495754

21. **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph**
Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, Tak-Wah Lam

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Bioinformatics (2015-05-15) https://doi.org/f7fb5z
DOI: 10.1093/bioinformatics/btv033 · PMID: 25609793

22. **COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge**
Yang Young Lu, Ting Chen, Jed A. Fuhrman, Fengzhu Sun
Bioinformatics (2016-06-02) https://doi.org/f9x7sc
DOI: 10.1093/bioinformatics/btw290 · PMID: 27256312

23. **MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies**
Dongwan Kang, Feng Li, Edward S Kirton, Ashleigh Thomas, Rob S Egan, Hong An, Zhong Wang (2019-02-06) https://doi.org/gf5fhv
DOI: 10.7287/peerj.preprints.27522v1

24. **MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets**
Yu-Wei Wu, Blake A. Simmons, Steven W. Singer
Bioinformatics (2016-02-15) https://doi.org/f8c9n2
DOI: 10.1093/bioinformatics/btv638 · PMID: 26515820

25. **Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy**
Christian M. K. Sieber, Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, Jillian F. Banfield
Nature Microbiology (2018-05-28) https://doi.org/gfwwfg
DOI: 10.1038/s41564-018-0171-1 · PMID: 29807988 · PMCID: PMC6786971

26. **Unusual biology across a group comprising more than 15% of domain Bacteria**
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
Nature (2015-06-15) https://doi.org/f7h5xj
DOI: 10.1038/nature14486 · PMID: 26083755

27. **Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life**
Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, Gene W. Tyson
Nature Microbiology (2017-09-11) https://doi.org/cczd
DOI: 10.1038/s41564-017-0012-7 · PMID: 28894102

28. **The genomic and proteomic landscape of the rumen microbiome revealed by comprehensive genome-resolved metagenomics**
Robert D. Stewart, Marc D. Auffret, Amanda Warr, Alan W. Walker, Rainer Roehe, Mick Watson
bioRxiv (2018-12-08) https://doi.org/gf5fhr
DOI: 10.1101/489443

29. **Genome-centric view of carbon processing in thawing permafrost**
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
Nature (2018-07-16) https://doi.org/gdth6p
DOI: 10.1038/s41586-018-0338-1 · PMID: 30013118

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

30. **Mediterranean grassland soil C–N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms**
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
Nature Microbiology (2019-05-20) https://doi.org/gf5fcx
DOI: 10.1038/s41564-019-0449-y · PMID: 31110364 · PMCID: PMC6784897

31. **AMBER: Assessment of Metagenome BinnERs**
Fernando Meyer, Peter Hofmann, Peter Belmann, Ruben Garrido-Oter, Adrian Fritz, Alexander Sczyrba, Alice C McHardy
GigaScience (2018-06) https://doi.org/gdptz9
DOI: 10.1093/gigascience/giy069 · PMID: 29893851 · PMCID: PMC6022608

32. **Detecting genomic islands using bioinformatics approaches**
Morgan G. I. Langille, William W. L. Hsiao, Fiona S. L. Brinkman
Nature Reviews Microbiology (2010-05) https://doi.org/d6ss55
DOI: 10.1038/nrmicro2350 · PMID: 20395967

33. **Horizontal gene transfer: building the web of life**
Shannon M. Soucy, Jinling Huang, Johann Peter Gogarten
Nature Reviews Genetics (2015-07-17) https://doi.org/f7j3d9
DOI: 10.1038/nrg3962 · PMID: 26184597

34. **The Association of Virulence Factors with Genomic Islands**
Shannan J. Ho Sui, Amber Fedynak, William W. L. Hsiao, Morgan G. I. Langille, Fiona S. L. Brinkman
PLoS ONE (2009-12-01) https://doi.org/c7hsvv
DOI: 10.1371/journal.pone.0008094 · PMID: 19956607 · PMCID: PMC2779486

35. **Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer**
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
Frontiers in Microbiology (2016-02-19) https://doi.org/gf5fht
DOI: 10.3389/fmicb.2016.00173 · PMID: 26925045 · PMCID: PMC4759269

36. **Transfer of antibiotic-resistance genes via phage-related mobile elements**
Maryury Brown-Jaque, William Calero-Cáceres, Maite Muniesa
Plasmid (2015-05) https://doi.org/f7dvxv
DOI: 10.1016/j.plasmid.2015.01.001 · PMID: 25597519

37.**:{unav)**
Rainer Merkl
BMC Bioinformatics (2004) https://doi.org/bt5x8h
DOI: 10.1186/1471-2105-5-22 · PMID: 15113412 · PMCID: PMC394314

38. **Improved genomic island predictions with IslandPath-DIMOB**
Claire Bertelli, Fiona SL Brinkman
Bioinformatics (2018-07-01) https://doi.org/gdphgs
DOI: 10.1093/bioinformatics/bty095 · PMID: 29905770 · PMCID: PMC6022643

39. **Microbial genomic island discovery, visualization and analysis**
Claire Bertelli, Keith E Tilley, Fiona SL Brinkman

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Briefings in Bioinformatics (2019-09) https://doi.org/gdnhfv
DOI: 10.1093/bib/bby042 · PMID: 29868902 · PMCID: PMC6917214

40. **cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data**
Fengfeng Zhou, Ying Xu
Bioinformatics (2010-08-15) https://doi.org/cn7486
DOI: 10.1093/bioinformatics/btq299 · PMID: 20538725 · PMCID: PMC2916713

41. **Modal Codon Usage: Assessing the Typical Codon Usage of a Genome**
J. J. Davis, G. J. Olsen
Molecular Biology and Evolution (2009-12-17) https://doi.org/bhsmq5
DOI: 10.1093/molbev/msp281 · PMID: 20018979 · PMCID: PMC2839124

42. **Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria**
Alvaro San Millan, Jose Antonio Escudero, Danna R. Gifford, Didier Mazel, R. Craig MacLean
Nature Ecology & Evolution (2016-11-07) https://doi.org/bs76
DOI: 10.1038/s41559-016-0010 · PMID: 28812563

43. **Small-Plasmid-Mediated Antibiotic Resistance Is Enhanced by Increases in Plasmid Copy Number and Bacterial Fitness**
Alvaro San Millan, Alfonso Santos-Lopez, Rafael Ortega-Huedo, Cristina Bernabe-Balas, Sean P. Kennedy, Bruno Gonzalez-Zorn
Antimicrobial Agents and Chemotherapy (2015-06) https://doi.org/f7k8bk
DOI: 10.1128/aac.00235-15 · PMID: 25824216 · PMCID: PMC4432117

44. **Understanding the mechanisms and drivers of antimicrobial resistance.**
Alison H Holmes, Luke SP Moore, Arnfinn Sundsfjord, Martin Steinbakk, Sadie Regmi, Abhilasha Karkey, Philippe J Guerin, Laura JV Piddock
Lancet (London, England) (2015-11-18) https://www.ncbi.nlm.nih.gov/pubmed/26603922
DOI: 10.1016/s0140-6736(15)00473-0 · PMID: 26603922

45. **On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data**
Sergio Arredondo-Alonso, Rob J. Willems, Willem van Schaik, Anita C. Schürch
Microbial Genomics (2017-10-01) https://doi.org/gf6b63
DOI: 10.1099/mgen.0.000128 · PMID: 29177087 · PMCID: PMC5695206

46. **MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies**
James Robertson, John H. E. Nash
Microbial Genomics (2018-08-01) https://doi.org/ggcm6q
DOI: 10.1099/mgen.0.000206 · PMID: 30052170 · PMCID: PMC6159552

47. **IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis: Figure 1.**
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
Nucleic Acids Research (2015-07-01) https://doi.org/f7n2xs
DOI: 10.1093/nar/gkv401 · PMID: 25916842 · PMCID: PMC4489224

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

48. **Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software**
Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, ... Alice C McHardy
Nature Methods (2017-10-02) https://doi.org/gbzspt
DOI: 10.1038/nmeth.4458 · PMID: 28967888 · PMCID: PMC5903868

49. **ART: a next-generation sequencing read simulator**
Weichun Huang, Leping Li, Jason R. Myers, Gabor T. Marth
Bioinformatics (2012-02-15) https://doi.org/fzf84c
DOI: 10.1093/bioinformatics/btr708 · PMID: 22199392 · PMCID: PMC3278762

50. **Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files**
NA Joshi, JN Fass
GitHub (2011) https://github.com/najoshi/sickle

51. **MetaQUAST: evaluation of metagenome assemblies**
Alla Mikheenko, Vladislav Saveliev, Alexey Gurevich
Bioinformatics (2016-04-01) https://doi.org/f8jdjj
DOI: 10.1093/bioinformatics/btv697 · PMID: 26614127

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

53. **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**
Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, Evgeny M. Zdobnov
Bioinformatics (2015-10-01) https://doi.org/gfznpw
DOI: 10.1093/bioinformatics/btv351 · PMID: 26059717

54. **Parallelization of MAFFT for large-scale multiple sequence alignments**
Tsukasa Nakamura, Kazunori D Yamada, Kentaro Tomii, Kazutaka Katoh
Bioinformatics (2018-07-15) https://doi.org/gc4th3
DOI: 10.1093/bioinformatics/bty121 · PMID: 29506019 · PMCID: PMC6041967

55. **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses**
S. Capella-Gutierrez, J. M. Silla-Martinez, T. Gabaldon
Bioinformatics (2009-06-08) https://doi.org/bjhdh7
DOI: 10.1093/bioinformatics/btp348 · PMID: 19505945 · PMCID: PMC2712344

56. **IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies**
Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, Bui Quang Minh
Molecular Biology and Evolution (2015-01) https://doi.org/f3srtd
DOI: 10.1093/molbev/msu300 · PMID: 25371430 · PMCID: PMC4271533

57. **PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses**
R. Lanfear, B. Calcott, S. Y. W. Ho, S. Guindon

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Molecular Biology and Evolution (2012-01-20) https://doi.org/fzgsw3
DOI: 10.1093/molbev/mss020 · PMID: 22319168

58. **ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data**
Jaime Huerta-Cepas, François Serra, Peer Bork
Molecular Biology and Evolution (2016-06) https://doi.org/gfzpph
DOI: 10.1093/molbev/msw046 · PMID: 26921390 · PMCID: PMC4868116

59. **mwaskom/seaborn: v0.10.0 (January 2020)**
Michael Waskom, Olga Botvinnik, Joel Ostblom, Saulius Lukauskas, Paul Hobson, MaozGelbart, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, ... Constantine Evans
Zenodo (2020-01-24) https://doi.org/ggkff7
DOI: 10.5281/zenodo.3629446

60. **Prodigal: prokaryotic gene recognition and translation initiation site identification**
Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, Loren J Hauser
BMC Bioinformatics (2010-03-08) https://doi.org/cktxnm
DOI: 10.1186/1471-2105-11-119 · PMID: 20211023 · PMCID: PMC2848648

61. **CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database**
Baofeng Jia, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, Biren M. Dave, Sheldon Pereira, Arjun N. Sharma, ... Andrew G. McArthur
Nucleic Acids Research (2017-01-04) https://doi.org/f9wbjs
DOI: 10.1093/nar/gkw1004 · PMID: 27789705 · PMCID: PMC5210516

62. **VFDB 2019: a comparative pathogenomic platform with an interactive web interface**
Bo Liu, Dandan Zheng, Qi Jin, Lihong Chen, Jian Yang
Nucleic Acids Research (2019-01-08) https://doi.org/gf4zfr
DOI: 10.1093/nar/gky1080 · PMID: 30395255 · PMCID: PMC6324032

63. **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes**
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
Bioinformatics (2010-07-01) https://doi.org/bz3q2w
DOI: 10.1093/bioinformatics/btq249 · PMID: 20472543 · PMCID: PMC2887053

64. **On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data.**
Sergio Arredondo-Alonso, Rob J Willems, Willem van Schaik, Anita C Schürch
Microbial genomics (2017-08-18) https://www.ncbi.nlm.nih.gov/pubmed/29177087
DOI: 10.1099/mgen.0.000128 · PMID: 29177087 · PMCID: PMC5695206

65. **SCAPP: An algorithm for improved plasmid assembly in metagenomes**
David Pellow, Maraike Probst, Ori Furman, Alvah Zorea, Arik Segal, Itzik Mizrahi, Ron Shamir
bioRxiv (2020-01-14) https://doi.org/ggkt4f
DOI: 10.1101/2020.01.12.903252

66. **Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation**
Jonathan L Klassen, Cameron R Currie
BMC Genomics (2012) https://doi.org/fzg6gg
DOI: 10.1186/1471-2164-13-14 · PMID: 22233127 · PMCID: PMC3322347

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.31.997171; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

67. **Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data**
Philip T. L. C. Clausen, Ea Zankari, Frank M. Aarestrup, Ole Lund
Journal of Antimicrobial Chemotherapy (2016-09) https://doi.org/f85vbc
DOI: 10.1093/jac/dkw184 · PMID: 27365186

68. **ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads**
Martin Hunt, Alison E Mather, Leonor Sánchez-Busó, Andrew J Page, Julian Parkhill, Jacqueline A Keane, Simon R Harris
Microbial Genomics (2017-10-01) https://doi.org/gf5fd9
DOI: 10.1099/mgen.0.000131 · PMID: 29177089 · PMCID: PMC5695208