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

This manuscript (permalink) was automatically generated from fmaguire/mag_sim_paper@63af282 on January 17, 2020.

Authors

Please note the current author order is chronological and does not reflect the final order.

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

- **Baofeng Jia***
  XXXX-XXXX-XXXX-XXXX · imasianxd
  Department of Biochemistry and Molecular Biology, Simon Fraser University

- **Kristen Gray**
  XXXX-XXXX-XXXX-XXXX
  Department of Biochemistry and Molecular Biology, Simon Fraser University

- **Venus Lau**
  XXXX-XXXX-XXXX-XXXX
  Department of Biochemistry and Molecular Biology, Simon Fraser University

- **Robert G. Beiko**
  Faculty of Computer Science, Dalhousie University

- **Fiona S.L. Brinkman**
  XXXX-XXXX-XXXX-XXXX
  Department of Biochemistry and Molecular Biology, Simon Fraser University

Abstract

Metagenomic methods, in which all the DNA in sample is simultaneously sequenced, is an increasingly popular method in the life sciences. They have a major advantage over genomic or phenotypic methods, as they do not require time-intensive and bias-inducing culturing steps. This means a much greater diversity can be achieved with minimal *a priori* assumptions. Due to this strength, metagenomics is emerging as a key tool in public health microbiology for surveillance of virulence and antimicrobial resistance (AMR) genes. The most important sequences for surveillance purposes are those associated with mobile genetic elements such as plasmids and genomic islands (GIs). Unfortunately, metagenomic data, even when assembled, results in complex mixed set of DNA fragments rather than nicely resolved individual genomes. Recently, methods have been developed that attempt to group these DNA fragments into bins that likely to have been derived from the same underlying genome. These bins are referred to as metagenome-assembled genomes (MAGs). MAG based approaches have been used to great effect in revealing huge amounts of previously uncharacterised microbial diversity. These methods perform this grouping using aspects of the sequence composition and the relative abundance of that sequence in the dataset. Unfortunately, plasmids are often represented at different copy numbers than the corresponding chromosomes. Additionally, both plasmids and genomic islands often feature significantly different sequence composition than the rest of the source genome as a whole. Due to this we hypothesise, these types of sequences will be highly under-represented in MAG based approaches.

To evaluate this we generated a simulated metagenomic dataset comprised of 30 genomes with up to 16.65% of chrosomomal DNA consisting of GIs and 65 associated plasmids. MAGs were then recovered from this data using 12 different MAG pipelines. The recovery and correct binning of mobile genetic elements was then evaluated for each pipeline. 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 unsuited for analysis of mobile genes, especially vital groups such as AMR and VF genes. This underlines the utility of read-based and long-read approaches to thorougly evaluate the resistome in metagenomic data.

Introduction

Metagenomics, the untargeted sequencing of all DNA within a sample, has become the dominant approach for characterising viral and microbial communities over the last 17 years [1,2]. By sampling from the total genomic content these methods allow researchers to simultaneously profile the functional potential and the taxonomic identity of all organisms ina sample. This is in contrast to bar-coding based approaches such as 16S or 18S rRNA sequencing which only provide taxonomic information [3] (although you can attempt to predict functional potential from taxonomic data [4,5]). One of many areas where metagenomics has been very useful is in the analysis of antimicrobial resistance (AMR) and pathogen virulence. These approaches have been instrumental in developing our understanding of the distribution and evolutionary history of AMR genes [6,7,8]. It has also formed a key tool for pathogen tracking in public health outbreak analyses [9].

While 3rd generation long-read technology has begun to be adopted in metagenomics analyses [10,11] the majority of analyses still involve high-throughput 2nd generation sequencing. These 2nd generation platforms such as Illumina’s MiSeq provide high numbers (10s-100s of millions) of relatively short reads (150-250bp) randomly sampled from the underlying DNA in the sample. This sampling is, therefore, in proportion to the relative abundance of different organisms (i.e. more abundant organisms will be more represented in the reads). There are two main approaches for the analysis of 2nd generation metagenomic data: read homology and metagenome assembly. Read-based approaches involve using reference databases and BLAST-based sequence similarity search tools (e.g. DIAMOND [12]), read mapping (e.g. Bowtie 2 [13]), Hidden Markov Models (e.g. HMMER3 [14]) or k-mer hashing (e.g. CLARK [15]). 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. However, read-based methods are reliant on quality of the reference database (i.e. you don’t detect things you don’t already know about) and does not provide any information about the genomic organisation of the 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 [16], selective pressures [17], and how liable the sequence is to lateral gene transfer (LGT) [18].

In order to get more data about the relative genomic context and organisation of your genes of interest it is possible (although computationally demanding) to assemble the short reads into longer fragments of DNA (contigs). This approach has been used successfully since early metagenomic analysis papers [19]. There are a variety of specialised *de Bruijn* graph assemblers developed to handle the particular challenges of this type of assembly (such as metaSPAdes [20] , IDBA-UD [21], and megahit [22]) each with a range of different strengths and weaknesses [23]. While this approach does result in longer contigs it still leaves you with a large collection of fragmentary data derived from many different organisms.

An increasingly common way to deal with this is to attempt to group these assembled contigs into bins all derived from the same underlying genome in the sample. These resulting bins are 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 [24], MetaBAT 2 [25], and MaxBin 2 [26]. There is also the meta-binning tool DAS Tool [27] 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 [28,29,30].

Unfortunately, there is loss of information at both the metagenomic assembly step (e.g. repetitive DNA sequences that are difficult to correctly assemble with short reads) [31] and in binning. This compounded data loss means that only a relatively small proportion of reads are successfully assembled and binned in large complex metagenome datasets e.g. 24.2-36.4% of reads from permafrost [33] and soil metagenomes [34]. Additionally, a large number of detected genomes are not reconstructed at all with ~23% of all detected genomes recovered in some examples [34]. There have been attempts to benchmark and compare the assembly and binning tools such as the Critical Assessment of Metagenome Interpretation (CAMI) challenge’s (https://data.cami-challenge.org/) Assessment of Metagenome BinnERs (AMBER) [35] however these largely investigate the overall completeness and purity of recovered MAGs relative to the known genomes in the evaluation samples. 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. Two such genomic elements of great health and research importance are mobile genetic elements (MGEs) such as genomic islands (GIs) and plasmids.

Genomic islands (GIs) are clusters of genes that are known or predicted to have been acquired through LGT events. These include integrons, transposons, integrative and conjugative elements (ICEs) and prophages (integrated phages) [36,37]. They have been shown to disproportionately encode virulence factors [38] and are a major mechanism of LGT of AMR genes [39,40]. However, these GIs often have different nucleotide composition compared to the rest of the genome [36]. This compositional difference is exploited by tools designed to detect GIs such as SIGI-HMM [41] and IslandPath-DIMOB [42]. GIs may exist as multiple copies within a genome [43] leading to potential difficulties in correctly assembling these regions in metagenome assemblies as well as likely biases in the calculation of coverage statistics. Similarly, plasmids, circular or linear extrachromosomal self-replicating pieces of DNA, are a major source of the dissemination of AMR genes throughout microbial ecosystems [39,44].

Due to their research importance, lots of work has identified the difficulty of assembling these sequences correctly from short-read data [32]. This is largely attributable to their repetitive sequences, variable copy number [46,47] and often markedly different sequence composition to the genome they are associated with [48,49]. As MAG binning is performed on the basis of sequence composition and relative abundance this suggests that these types of sequences are liable to being incorrectly binned or lost in the process of recovering MAGs. 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 broad-set of current state-of-the-art short-read metagenome assembly and binning approaches using a simulated medium complexity 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 [42] and collated into the IslandViewer database www.pathogenomics.sfu.ca/islandviewer [50]. Plasmid sequences and numbers were recovered for each genome using the linked GenBank Project IDs. Thirty genomes were manually selected to exemplify the following criteria:

```
1) 10 genomes with high numbers of plasmids.

2) 10 genomes with a very high proportion (>10%) of chromosomes
corresponding to GIs detected by compositional features.

3) 10 genomes with a very low proportion (<1%) of chromosomes corresponding
to GIs detected by compositional features.
```

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 to the recommendation in the CAMI challenge [51] 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 ($\alpha = 4$, $\beta = 1$) or the minimum edge of the regime.

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) [52] at a fold coverage of 2.9 resulting in a simulate 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) [53] resulting in 25,682,644 surviving read pairs. The trimmed reads were then assembled using 3 different metagenomic assemblers: metaSPAdes (v3.13.0)[20], IDBA-UD (v1.1.3) [21], and megahit (v1.1.3) [22]). The resulting assemblies were summarised using metaQUAST (v5.0.2) [54]. The assemblies were then indexed and reads mapped back using Bowtie 2 (v2.3.4.3) [13].

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) [25], and MaxBin 2 (v2.2.6) [26]. MAGs were also recovered using CONCOCT (v0.4.2) [24] 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 [27]. 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+ [55] 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 from a given MAG aligned to were tallied. Finally, the overall proportion of chromosomes that were not present in any MAG were tallied for each binner and assembler.

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. Any plasmid or GI with greater than 50% coverage in a MAG was retained. All plasmids or GIs which could be found in the unbinned contigs or the MAGs was recorded as having been successfully assembled. The subset of these which were found in the binned MAGs was then separately tallied. Finally, we evaluated the proportion of plasmids or GIs that were binned correctly in the bin that was maximally composed of chromosomes from the same source genome. This was determined using the bin “identities” from the chromosomal coverage analysis.

Antimicrobial Resistance and Virulence Factors Assessment

Detection of AMR/VF Genes

For each of the 12 MAGs sets, and the reference chromosome and plasmids, prodigal [56] 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) [57]. Virulence factors were predicted using the predicted ORFs and BLASTX 2.9.0+ [55] against the Virulence Factors Database (VFDB; obtained on Aug 26, 2019) with an e-value cut-off of 0.001 and a minimum identity of 90% [58]. 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 assembly and each MAG and compared this number 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 proteins predicted subcellular localization was on its recovery and binning in MAGs. The MAG bins from megahit-DAS Tool assembler-binner combination was selected (as generally best performing) and ORFs predicted using prodigal [56] 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 [59].

Results

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 by all methods for more than a 1/3rd of the chromosomes. The three Streptococcus genomes were particularly problematic with the best recovery for each ranging from 1.7% to 47.49%.

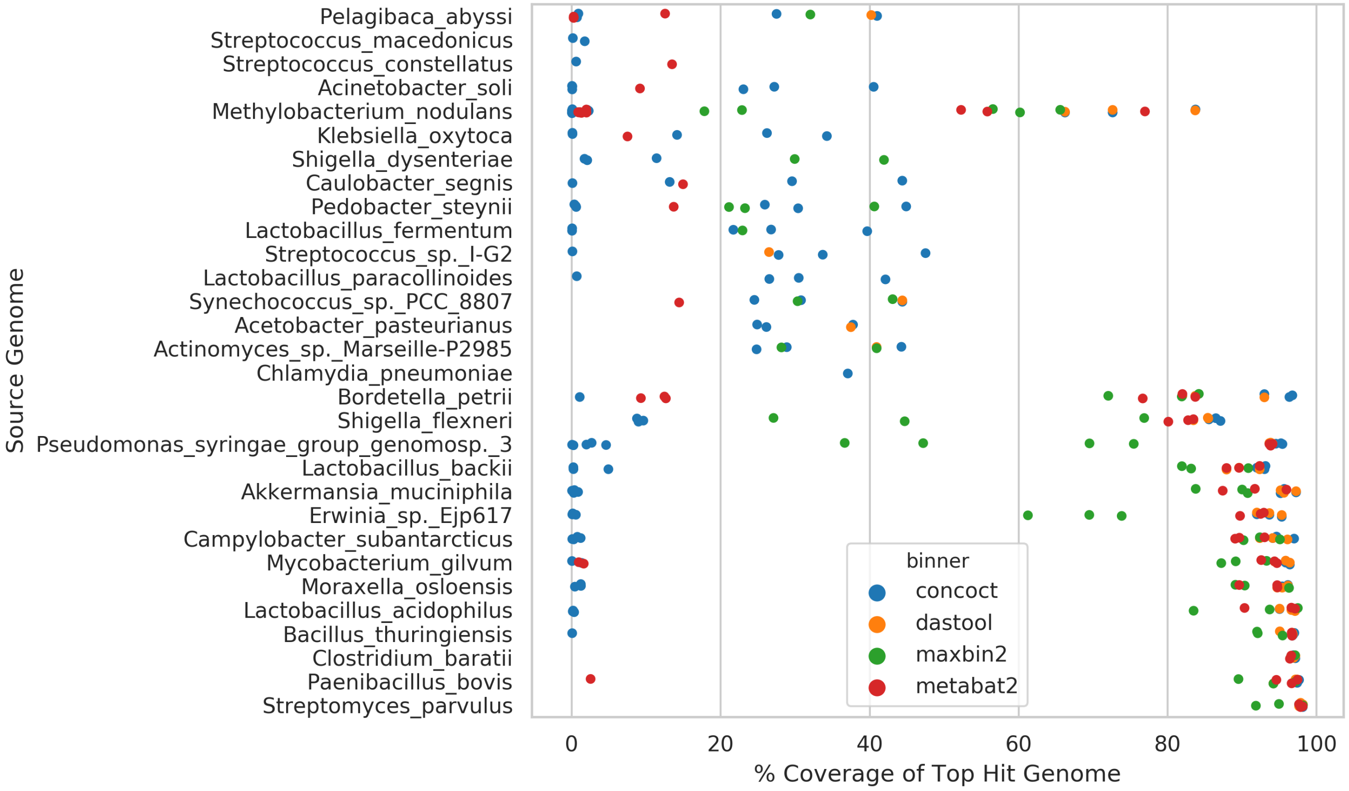

Figure 1: Top genome coverage for input genomes across MAG binners. Each dot represents the coverage of a specified genome when it comprised the majority of the sequences in a bin. The binning tool is indicated by the colour

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)).

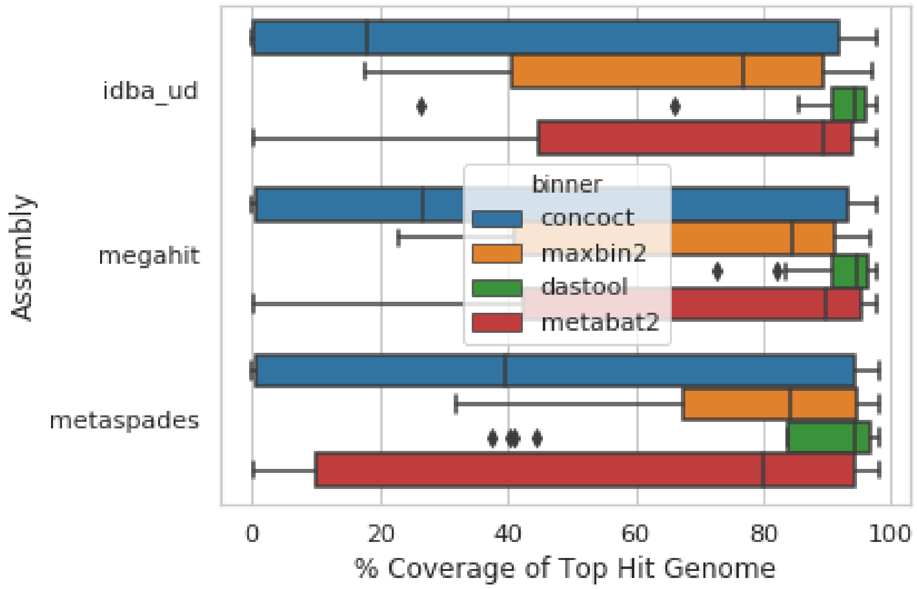

Figure 2: Chromosomal coverages of most prevalent genome in each bin across binners and metagenome assemblies. Of the 3 assemblers (y-axis), megahit resulted in the highest median chromosomal coverage (x-axis) across all binners (colored bars) at 81.9% with metaSPAdes performing the worst (76.8%). Of 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 represents 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 (Fig. 3), with a very marginally higher purity for IDBA. In terms of binning tool, however, maxbin2 proved an outlier with nearly twice as many bins containing multiple species as the next binner. The remaining binning tools were largely equivalent, producing chimeric bins at approximately the same rates.

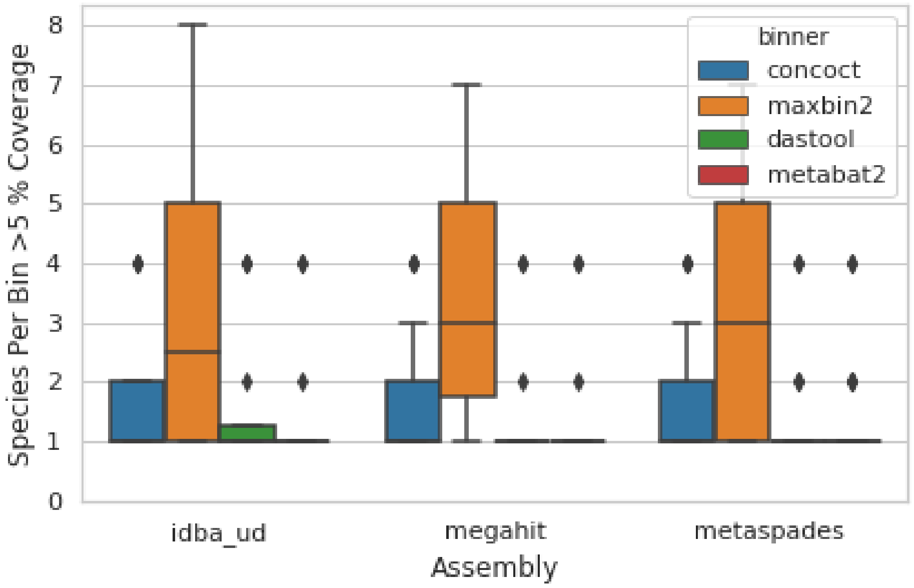

Figure 3: Distribution of bin purities across assemblers and binners. The total number of genomes present in a bin at >5% coverage (y-axis) was largely equivalent across assemblers (x-axis). In term of binning tools, maxbin2 (orange) produced nearly twice as many bins containing multiple species compared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

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 fairly consistently placed in a MAG bin by maxbin2 and CONCOCT, although a much smaller fraction were correctly binned (typically less than 1/3rd). Interestingly, MetaBAT2 and DASTool binners were a lot more conservative in assigning plasmid contigs to bins; however, of those assigned to bins nearly all were correctly binned (Fig. 4)

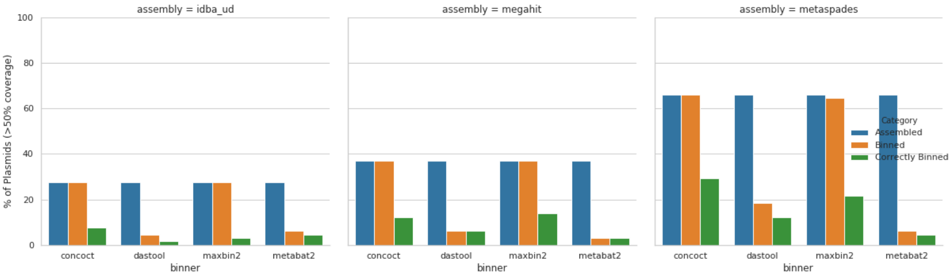

Figure 4: The performance of metagenomic assembly and binning to recover plasmid sequences. Each plot represents a different metagenome assembler, with the groups of bars along the x-axes showing the plasmid recovery performance of each binning tool when applied to the assemblies produced by that tool. For each of these 12 assembler-binner pair produced MAGs the grouped bars from left to right show the percentage of plasmids assembled, binned in 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). These sequences were assembled uniformly badly (37.8-44.1%) with metaSPAdes outperforming the other two assembly approaches. For 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.

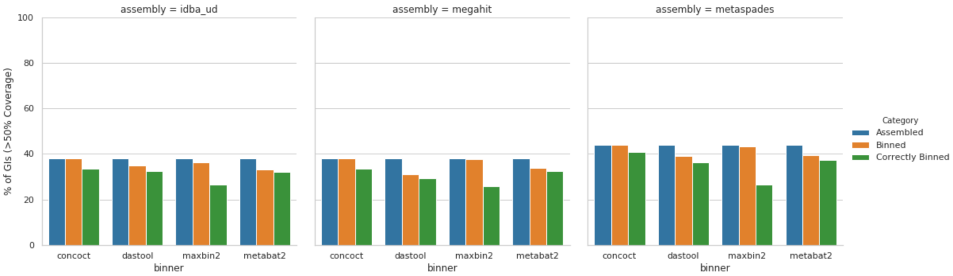

Figure 5: Impact of metagenomic assembly and MAG binning on recovery of genomic islands. GIs were recovered in a similarly poor fashion to plasmids. Generally, <40% were correctly binned to the same bin majority commprised of chromosomal contigs from the same source genome regardless of binning (x-axis)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

Recovery of Specific Gene Content

In term of gene content, we first explored the ability to find open reading frames (ORFs) within MAGs. Overall, the total number of predicted ORFs in MAGs followed a similar trend (Fig 6) 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. MetaBAT2 performed second worst at ~80%. DASTool recovered a similar number to our reference and Maxbin2 seemed to predicted 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.

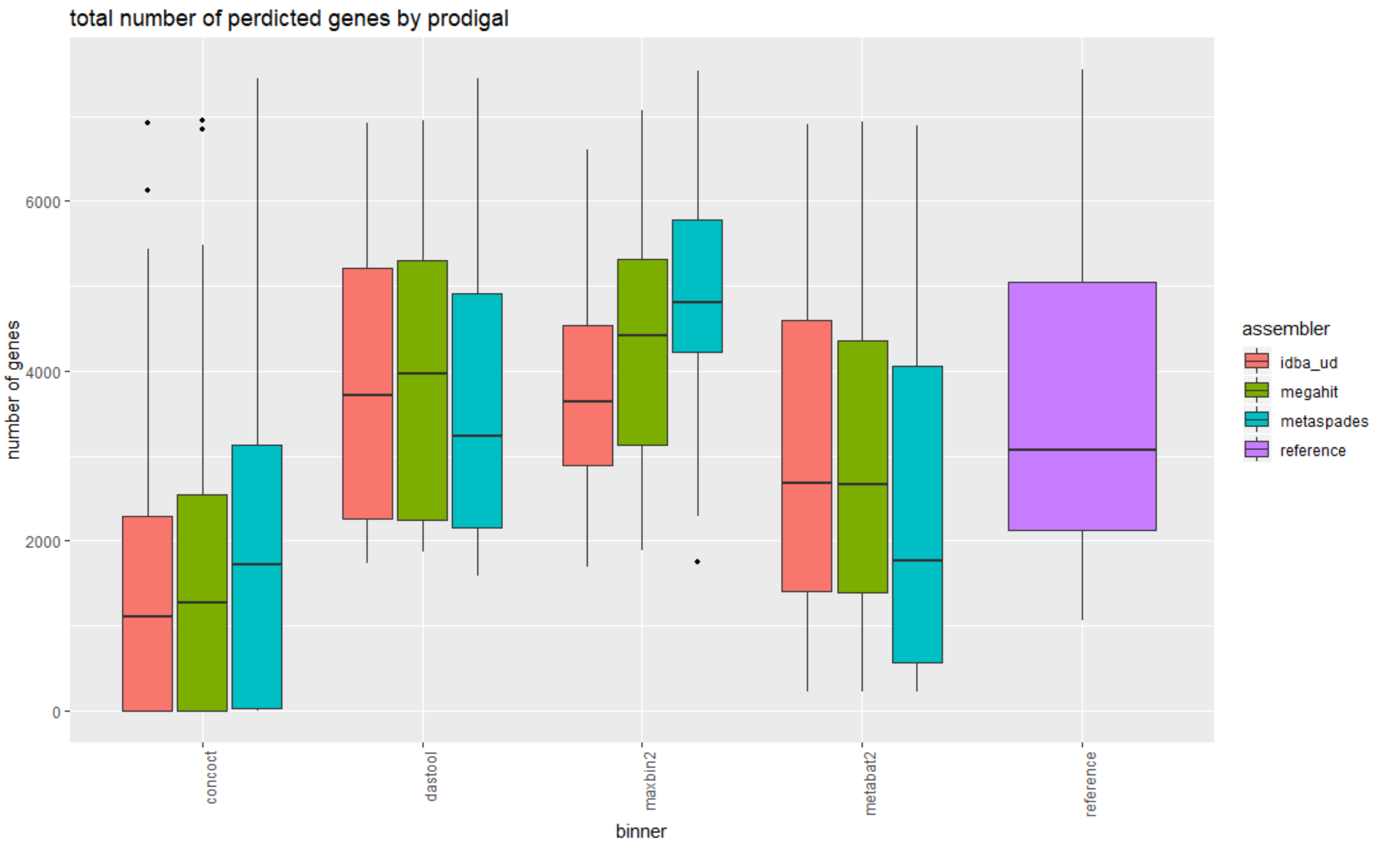

Figure 6: Predicted Gene Content. The total number of open reading frames (ORF) predicted followed the same trend as chromosomal coverage and purity. The assemblers (colored bars) did not contribute to a big variance in the number of ORFs. Of the 4 binners, CONCOCT recovered <30% of our reference genome ORFs. DASTool and MetaBAT2 predicted a similar number as our reference genomes.

AMR Genes

First, we focused on the ability of MAGs to recover clinically relevant AMR genes (Fig. 7). After the assembly stage, we were only able to recover between ~49-55% of the AMR genes predicted in our reference genomes regardless of the assembly tool used, with metaSPAdes performing marginally better than other assemblers. Binning the contigs resulted in a ~1-15% loss in AMR gene recovery with the CONCOCT-metaSPAdes pair performing the 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%).

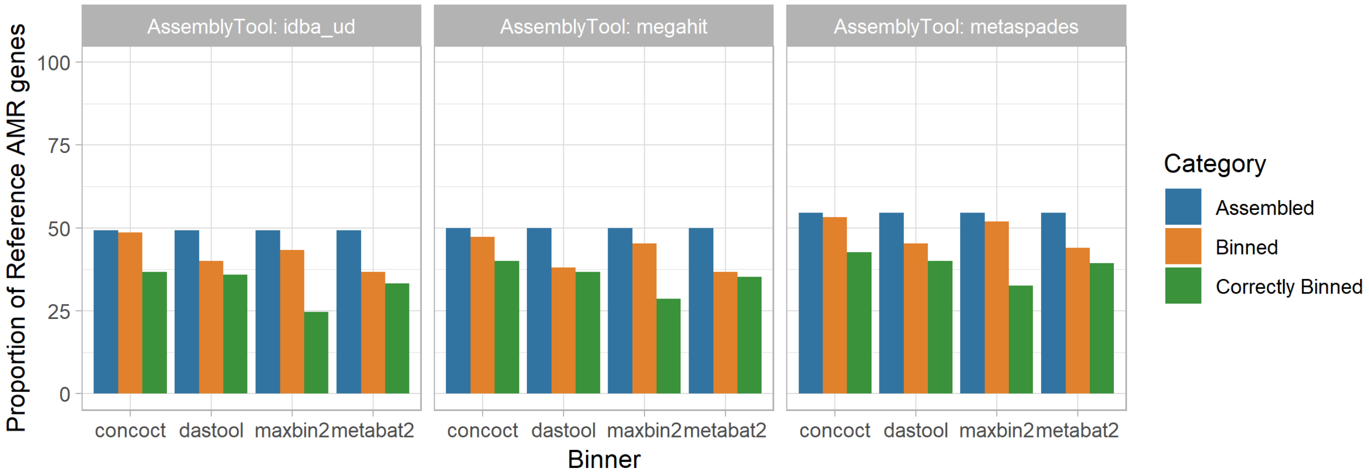

Figure 7: Percent of reference antimicrobial resistance genes (AMR) recovered across assemblers and binners. The proportion of reference AMR genes recovered (y-axis) was largely similar across assembly tools (blue), at roughly 50%

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.

Moreover, focusing on only the AMR genes that were correctly binned (Fig. 8) 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 located AMR genes (n=20) were correctly binned.

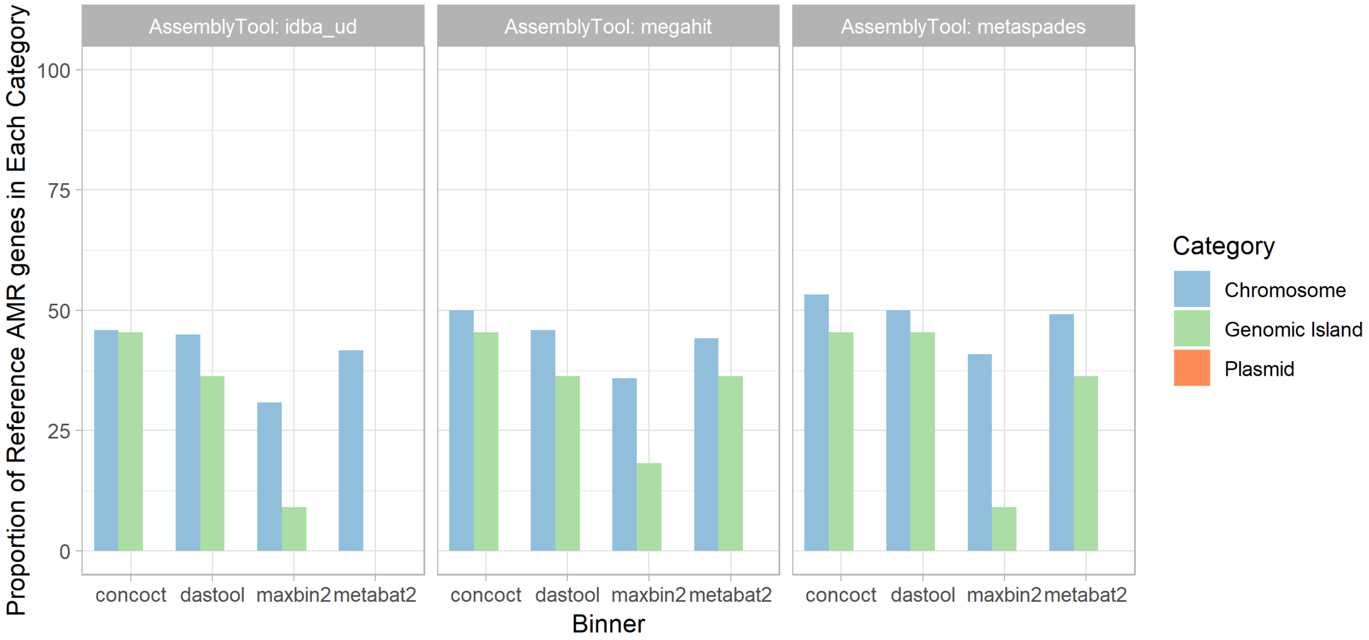

Figure 8: Percent of correctly binned AMR genes recovered by genomic context. MAG methods were best at recovering chromosomally located AMR genes (light blue) regardless of metagenomic assembler or binning tool used. Recovery of AMR genes in GIs showed a bigger variation between tools (light green). None of the 12 evaluated MAG recovery methods were able to recover plasmid located AMR genes (orange).

VF Genes

Aside from AMR 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. 10). 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.

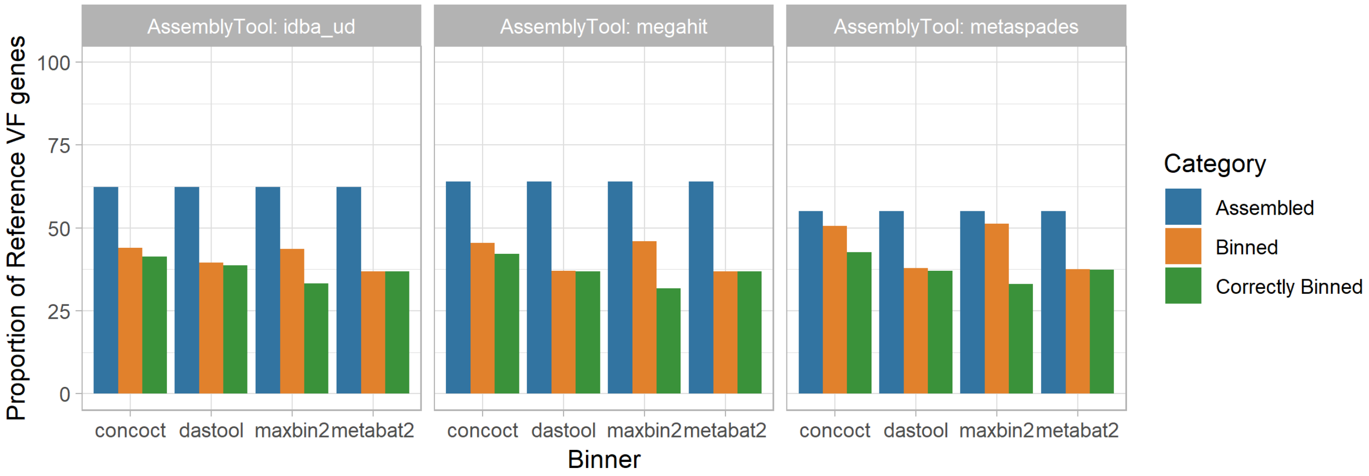

Figure 9: Percent of reference virulence factor (VF) genes recovered across assemblers and binners. The proportion of reference VF genes recovered (y-axis) exhibited a similar trend as AMR genes. Recovery was greatest after the assembling stage (blue), with megahit performing best. Binning tools resulted 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.

Again, the genomic context (chromosome, plasmid, GI) of a given VFs largely determined how well it was binned (Fig. 10). The majority (73%-98%) of all chromosomally located VF genes (n=757) were correctly binned. However, 0-16% of GI located VF genes (n=809) and again none of the plasmid located VF genes (n=3) were recovered across all 12 MAG pipelines.

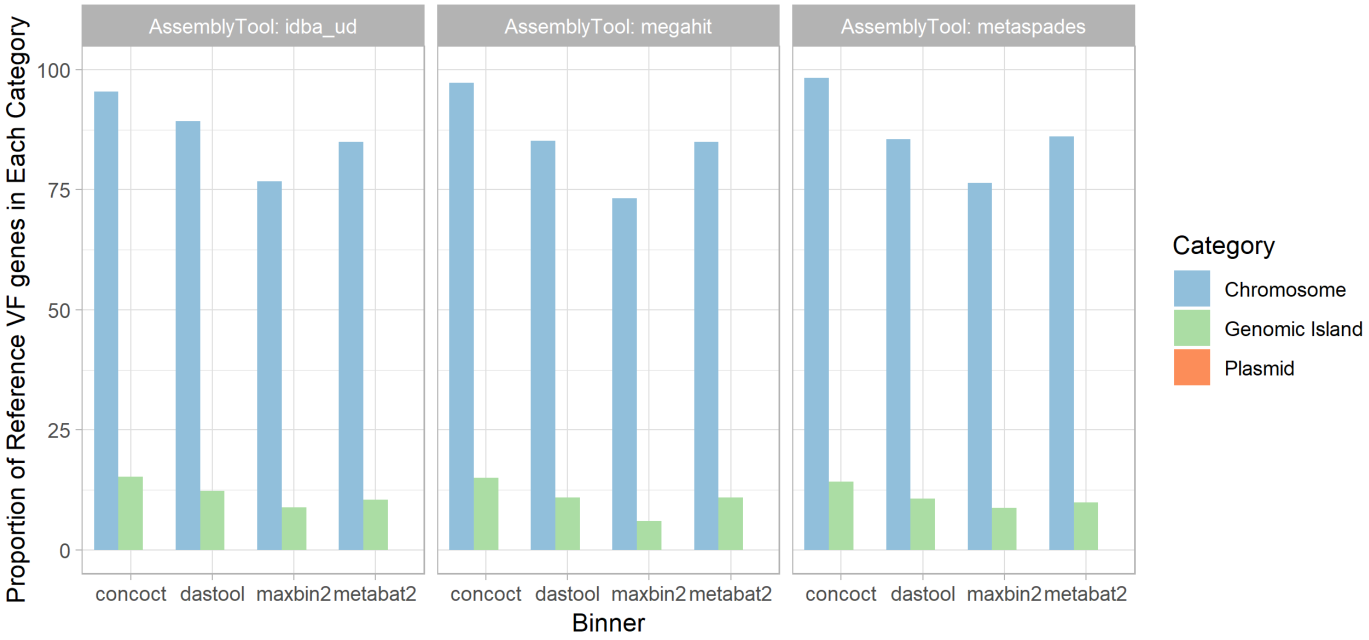

Figure 10: Percent of correctly binned VF genes Recovered in each genomic region. Metagenome assembled genomes (MAGs) were again best at recovering chromosomally located VF genes (light blue), able to correctly bin majority of chromosomally located VFs. GIs recovered again performed very poorly (light green) and again none of the plasmid located AMR genes (orange) was correctly binned.

Discussion

In this paper, we evaluated the ability and accuracy of metagenome-assembled genomes (MAGs) 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 term of both chromosomal coverage (94.3%) and bin purity (1). Looking at genomes with the lowest coverage, the three Streptococcus genomes that was recovered poorly are likely due to their similarity. 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 seems to display a trend of generating many small partial bins. Potentially, CONCOCT binning could be used to distinguish closely related species but at a cost of more fragmented genomes.

While the overall recovery and binning of chromosomes were acceptable, 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 our hypothesis that their sequence characteristics (composition and copy number) would prove difficult to bin. Unfortunately, our hypothesis was confirmed, despite the metagenomic assembly approach or MAG binning method used both plasmids and GIs were disproportionately lost compared to chromosomes in general. At best (via metaSPAdes and CONCOCT) 29.2% of plasmids and 44.1% 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 might have been a high-bar and there is a possibility that more GIs and plasmids were recovered but in very incomplete forms.

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 chromosomes. Furthermore, the difference between the percentages suggest that binning plasmids are harder than binning GIs. This difference is partially attributal to the known difficulties in assembly of plasmids from short-read data [60]. Therefore, binning efficiency might improve if we use DNA sequencing and assembly methods optimised for recovering plasmids [32].

Due to the importance of mobile genetic elements to disseminate clinically relevant AMR genes and VFs, we explored whether or not MAG approaches can be used to provide useful insight 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, 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 a lot of 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 there were only 11 GI located AMR genes in our reference genomes. All 20 of the plasmid-borne AMR genes were assembled but none were placed into MAG bins. This included high-threat MGEs-based AMR genes such as the KPC and OXA carbapenemases.

Virulence factors had shown a similar trend as AMR genes, recovering ~63% of virulence factors present in the reference genome. 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. MAGs were also able to correctly bin majority (73%-98%) of chromosomally located VF genes to the right species. However, MAGs 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 GI, is expected. Previous studies has found that VFs are disproportionally present on GIs[38], which might be the reason to why the recovery accuracy was worse compared to AMR genes. Together, this and the AMR gene results suggests that MAG-based methods might be of limited utility in public health research focused on the transmission and dissemination of AMR genes and VFs.

It should also be noted that while CONCOCT performed the best in terms of recovery of both chromosomes and MGEs, it created lots of 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 act of metagenomic assembly itself generally resulted in the loss of the majority of these elements/genes regardless of assembly method used. 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. While the concept that analysis is “lossy” and that the more analysis you do the more of the input data you are likely to lose is fairly well known it is rarely explicitly stated. 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 than are in the underlying sample.

Conclusions

Using a simulated medium complexity metagenome, this study had shown that MAGs provides a great tool to study a bacterial species’ chromosomal elements but presented difficulties in the recovery of mobile genetic elements from metagenomic samples. These mobile genetic elements are liable to being incorrectly binned or lost in this process. Due to the importance of these mobile genomic components in the function and spread of pathogenic traits such as AMR and virulence, it is vital that we utilize a combination of MAGs and other methods (e.g. read-based methods) in public health metagenomic researches. This would allow both the detection of the sample microbial diversity and the thorough evaluation of resistome in metagenomic data to provide meaningful epidemiological information.

Supplementals

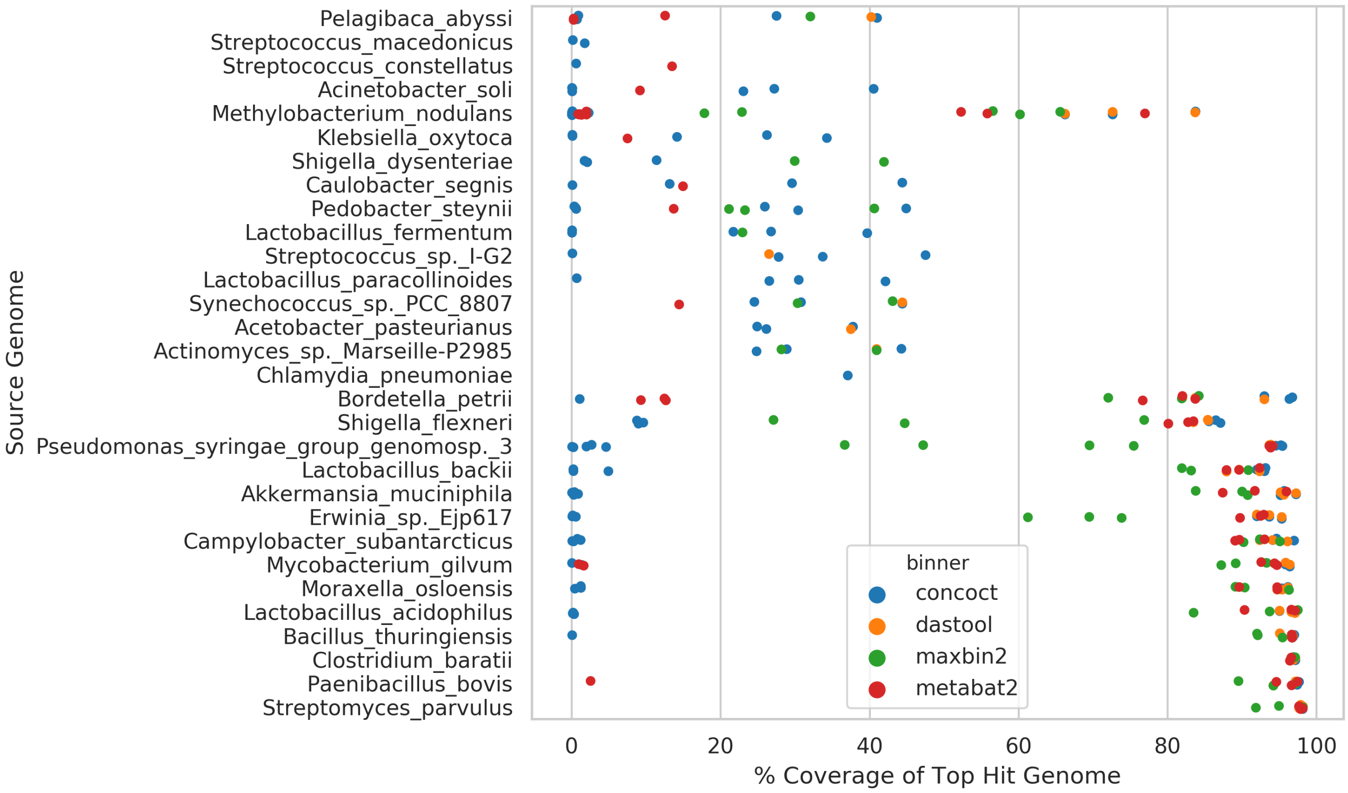

Figure 11: Top Species Coverage

We looked at the ability for MAGs to predict subcellular localization of proteins using PSORTb. Overall, the localization distribution of predicted proteins were very similar in MAGs compared to the reference genome (Fig. (12)). Previous works have shown that AMR genes that are on mobile genetic elements disproportionally encode secrete proteins. Given that the recovery of plasmid-borne genes were not great, we asked if MAGs would affect the ability to predict the subcellular localization of proteins. We found that the proportion of predicted localizations were very similar between MAGs and our reference genomes, suggesting that there is not a significant penalty to use MAGs as input for protein localization predictions.

Figure 12: Distribution of Predicted Protein Subcellular Localization

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) https://doi.org/gbv6nf
DOI: 10.1038/nbt.3935 · PMID: 28898207

3. **Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing.**
TM Schmidt, EF DeLong, NR Pace
Journal of bacteriology (1991-07) https://www.ncbi.nlm.nih.gov/pubmed/2066334
DOI: 10.1128/jb.173.14.4371-4378.1991 · PMID: 2066334 · PMCID: PMC208098

4. **Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences**

Morgan GI Langille, Jesse Zaneveld, J Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A Reyes, Jose C Clemente, Deron E Burkepile, Rebecca L Vega Thurber, Rob Knight, … Curtis Huttenhower
Nature Biotechnology (2013-08-25) https://doi.org/f49xzd
DOI: 10.1038/nbt.2676 · PMID: 23975157 · PMCID: PMC3819121

5. **PICRUSt2: An improved and extensible approach for metagenome inference**
Gavin M. Douglas, Vincent J. Maffei, Jesse Zaneveld, Svetlana N. Yurgel, James R. Brown, Christopher M. Taylor, Curtis Huttenhower, Morgan G. I. Langille
Cold Spring Harbor Laboratory (2019-06-15) https://doi.org/gf5ffb
DOI: 10.1101/672295

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

7. **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

8. **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

9. **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

10. **Ultra-deep, long-read nanopore sequencing of mock microbial community standards**
Samuel M Nicholls, Joshua C Quick, Shuiquan Tang, Nicholas J Loman
GigaScience (2019-05-01) https://doi.org/gf39g3
DOI: 10.1093/gigascience/giz043 · PMID: 31089679 · PMCID: PMC6520541

11. **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

12. **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

13. **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

14. **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

15. **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

16. **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

17. **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

18. **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

19. **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

20. **metaSPAdes: a new versatile metagenomic assembler**
Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, Pavel A. Pevzner
Genome Research (2017-03-15) https://doi.org/f97jkv
DOI: 10.1101/gr.213959.116 · PMID: 28298430 · PMCID: PMC5411777

21. **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

22. **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
Bioinformatics (2015-01-20) https://doi.org/f7fb5z
DOI: 10.1093/bioinformatics/btv033 · PMID: 25609793

23. **Assembling metagenomes, one community at a time**
Andries Johannes van der Walt, Marc Warwick van Goethem, Jean-Baptiste Ramond, Thulani Peter Makhalanyane, Oleg Reva, Don Arthur Cowan
BMC Genomics (2017-07-10) https://doi.org/gf5fhs
DOI: 10.1186/s12864-017-3918-9 · PMID: 28693474 · PMCID: PMC5502489

24. **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

25. **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
PeerJ (2019-02-06) https://doi.org/gf5fhv
DOI: 10.7287/peerj.preprints.27522v1

26. **MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets**
Yu-Wei Wu, Blake A. Simmons, Steven W. Singer
Bioinformatics (2015-10-29) https://doi.org/f8c9n2
DOI: 10.1093/bioinformatics/btv638 · PMID: 26515820

27. **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

28. **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

29. **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

30. **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

Cold Spring Harbor Laboratory (2018-12-08) https://doi.org/gf5fhr
DOI: 10.1101/489443

31. **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 (2016-06-30) https://doi.org/f85vbc
DOI: 10.1093/jac/dkw184 · PMID: 27365186

32. **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

33. **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

34. **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

35. **AMBER: Assessment of Metagenome BinnERs**
Fernando Meyer, Peter Hofmann, Peter Belmann, Ruben Garrido-Oter, Adrian Fritz, Alexander Sczyrba, Alice C McHardy
GigaScience (2018-06-01) https://doi.org/gdptz9
DOI: 10.1093/gigascience/giy069 · PMID: 29893851 · PMCID: PMC6022608

36. **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

37. **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

38. **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

39. **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

40. **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

41.
Rainer Merkl
BMC Bioinformatics (2004) https://doi.org/bt5x8h
DOI: 10.1186/1471-2105-5-22 · PMID: 15113412 · PMCID: PMC394314

42. **Improved genomic island predictions with IslandPath-DIMOB**
Claire Bertelli, Fiona SL Brinkman
Bioinformatics (2018-02-23) https://doi.org/gdphgs
DOI: 10.1093/bioinformatics/bty095 · PMID: 29905770 · PMCID: PMC6022643

43. **Microbial genomic island discovery, visualization and analysis**
Claire Bertelli, Keith E Tilley, Fiona SL Brinkman
Briefings in Bioinformatics (2018-06-03) https://doi.org/gdnhfv
DOI: 10.1093/bib/bby042 · PMID: 29868902

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. **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

46. **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

47. **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 (2015-03-30) https://doi.org/f7k8bk
DOI: 10.1128/aac.00235-15 · PMID: 25824216 · PMCID: PMC4432117

48. **cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data**
Fengfeng Zhou, Ying Xu
Bioinformatics (2010-08-02) https://doi.org/cn7486
DOI: 10.1093/bioinformatics/btq299 · PMID: 20538725 · PMCID: PMC2916713

49. **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

50. **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 (2015-04-27) https://doi.org/f7n2xs
DOI: 10.1093/nar/gkv401 · PMID: 25916842 · PMCID: PMC4489224

51. **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

52. **ART: a next-generation sequencing read simulator**
Weichun Huang, Leping Li, Jason R. Myers, Gabor T. Marth
Bioinformatics (2011-12-23) https://doi.org/fzf84c
DOI: 10.1093/bioinformatics/btr708 · PMID: 22199392 · PMCID: PMC3278762

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

54. **MetaQUAST: evaluation of metagenome assemblies**
Alla Mikheenko, Vladislav Saveliev, Alexey Gurevich
Bioinformatics (2015-11-26) https://doi.org/f8jdjj
DOI: 10.1093/bioinformatics/btv697 · PMID: 26614127

55. **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

56. **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

57. **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 (2016-10-26) https://doi.org/f9wbjs
DOI: 10.1093/nar/gkw1004 · PMID: 27789705 · PMCID: PMC5210516

58. **VFDB 2019: a comparative pathogenomic platform with an interactive web interface**
Bo Liu, Dandan Zheng, Qi Jin, Lihong Chen, Jian Yang
Nucleic Acids Research (2018-11-05) https://doi.org/gf4zfr
DOI: 10.1093/nar/gky1080 · PMID: 30395255 · PMCID: PMC6324032

59. **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 (2010-05-13) https://doi.org/bz3q2w
DOI: 10.1093/bioinformatics/btq249 · PMID: 20472543 · PMCID: PMC2887053

60. **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