Plasmid detection and assembly in genomic and metagenomic datasets

Dmitry Antipov[1*†], Mikhail Raiko[1*], Alla Lapidus[1] and Pavel A. Pevzner[1,2]

[1]Center for Algorithmic Biotechnology, Institute for Translational Biomedicine,

St. Petersburg State University, St. Petersburg, Russia 199004;

[2]Department of Computer Science and Engineering,

University of California, San Diego, California 92093-0404, USA

* These authors contributed equally to this work.

† To whom correspondence should be addressed. E-mail d.antipov@spbu.ru

ABSTRACT. Although plasmids are important for bacterial survival and adaptation, plasmid detection and assembly from genomic, let alone metagenomic, samples remain challenging. The recently developed plasmidSPAdes assembler addressed some of these challenges in the case of isolate genomes but stopped short of detecting plasmids in metagenomic assemblies, an untapped source of yet to be discovered plasmids. We present the metaplasmidSPAdes tool for plasmid assembly in metagenomic datasets that reduced the false positive rate of plasmid detection as compared to the state-of-the-art approaches. We assembled plasmids in diverse datasets and have demonstrated that thousands of plasmids remained below the radar in already completed genomic and metagenomic studies. Our analysis revealed the extreme variability of plasmids and has led to the discovery of many novel plasmids

(including many plasmids carrying antibiotic-resistance genes) without significant similarities to currently known ones.

INTRODUCTION

Plasmids are extrachromosomal independently replicating DNA molecules that provide their bacterial hosts with additional genetic material important for their survival and adaptation. Prior to the sequencing era, plasmids were detected based on the various phenotypic changes they provide to their host, such as antibiotic resistance or ability to degrade recalcitrant organic compounds. Sequencing efforts, however, have revealed many cryptic plasmids that do not contribute to the phenotype of the host cell in any obvious way. Although there are about 10,000 plasmids listed in the RefSeq database (Pruitt et al, 2006), many plasmids remain undetected since the task of assembling plasmids from genomic and metagenomic datasets is far from trivial (Antipov et al., 2016, Rozov et al., 2017). We thus conjecture that many classes of plasmids continue to remain unknown the same way the many previously unknown classes of viruses that were found in recent studies (Paez-Espino et al, 2016, Roux et al., 2016).

Since plasmids exchange genetic material with the host chromosomes and vary in structure (circular or linear), size (from a thousand to millions of nucleotides), and gene content, it is not clear how to computationally define the concept of a plasmid in such a way that it would be possible to distinguish them from the chromosomes. Also, plasmid assembly is complicated by various repeats that are difficult to resolve using short reads sequencing technologies:

1. **An intra-plasmidic repeat** refers to a repeat within a plasmid. 34% of plasmids in the RefSeq database contain intra-plasmidic repeats longer than 300 nucleotides, the typical insert size in metagenomic studies.

2. **An inter-plasmidic repeat** refers to a repeat shared by multiple plasmids.
3. **A shared repeat** refers to a repeat shared between a plasmid and a chromosome. For many isolate samples shared repeats can be resolved if the plasmid coverage by reads significantly differs from the chromosome coverage (Antipov, 2016). It is, however, difficult to resolve such repeats in the case of metagenomic samples with a wide spectrum of chromosome and plasmid coverages across the bacterial community (Rozov et al., 2017), or in isolate samples sequenced during the growth phase (Antipov et al., 2016).

Circular plasmids form *uniformly covered cycles* within genomic and metagenomic assembly graphs, i.e., cycles that have a relatively uniform coverage by reads (with the exception of regions corresponding to intra-plasmidic, inter-plasmidic, and shared repeats). These cycles are difficult to detect since they are “hidden” within a large assembly graph that contains both *chromosomal edges* (originating from chromosomes) and *plasmidic edges* (originating from plasmids). Moreover, plasmids with inter-plasmidic repeats form self-overlapping cycles (that traverse edges corresponding to these repeats more than once) thus complicating their detection even further.

plasmidSPAdes (Antipov et al., 2016) and Recycler (Rozov et al, 2017) are plasmid assembly tools that identify plasmids as short uniformly covered cycles in the assembly graph constructed by the SPAdes assembler (Bankevich et al., 2012). Both tools address the complications caused by shared repeats using the difference between the plasmid and chromosome coverages (plasmidSPAdes is limited to isolate genomes, while Recycler can work with metagenomes). Although plasmidSPAdes and Recycler revealed a number of novel plasmids, they report many false-positives, especially in situations when the chromosome coverage is

non-uniform. Arredondo-Alonso et al., 2017 benchmarked these tools on 42 datasets containing short reads sampled from isolate bacterial genomes with 148 plasmids and estimated that plasmidSPAdes and Recycler have a precision 0.78 and 0.30, respectively. The low precision and reliance on the uniform coverage makes plasmidSPAdes inapplicable to metagenomic datasets with highly varying coverage across multiple genomes. This is unfortunate since metagenomic datasets represent an untapped source of yet to be discovered plasmids (Jørgensen et al, 2014, Li et al., 2015).

We present the metaplasmidSPAdes algorithm that improves on plasmidSPAdes and Recycler by (i) iteratively extracting subgraphs with gradually increasing coverage from the metagenome assembly graph, (ii) finding putative plasmids as uniformly-covered cycles in these subgraphs, and (iii) verifying the found putative plasmids using a new plasmidVerify tool. We applied plasmidSPAdes^{+} (plasmidSPAdes complemented by plasmidVerify) and metaplasmidSPAdes to diverse genomic and metagenomic samples and revealed 1000s of plasmids that were missed in previous studies, including many plasmids that share no significant similarities with currently known plasmids, and plasmids carrying antibiotic-resistance genes.

RESULTS

metaplasmidSPAdes workflow. plasmidSPAdes constructs the plasmid graph by removing all edges with coverage similar to the median coverage in the assembly graph. This approach does not work for metagenomes since they have highly non-uniform coverage across various bacterial genomes within a metagenome. metaplasmidSPAdes improves on plasmidSPAdes by resolving *dominant plasmids* in metagenomes, i.e., plasmids with coverage exceeding that of chromosomes and other plasmids, with which they share repeats with.

metaplasmidSPAdes utilizes metaSPAdes (Nurk et al, 2016) for transforming the de Bruijn graph into an assembly graph. It further detects plasmids in the assembly graph by iteratively constructing smaller and smaller subgraphs of the assembly graph and detecting plasmids in these subgraphs. metaplasmidSPAdes removes low-coverage edges (with increasing coverage cutoff at each iteration), uses exSPAnder (Prjibelski et al., 2014) to generate contigs, and detects putative plasmids as cyclic contigs (*cyclocontigs*) or small connected components in the generated subgraphs.

metaplasmidSPAdes sets a coverage cutoff *cov*, removes all edges with coverage below *cov* from the assembly graph, and searches either for a cycle (cyclocontig) supported by the paired-end reads or a small connected component in the resulting graph. Some of the found cyclocontigs and connected components represent dominant plasmids that were “hidden” in the assembly graph before the removal of low coverage edges. To reveal more and more hidden plasmids with progressively increasing coverage, metaplasmidSPAdes iteratively increases the coverage cutoff as $cov+cov_{add}$ or as $cov*cov_{mult}$ (Figure 1). Finally, it uses the plasmidVerify tool to check whether contigs and connected components found by metaplasmidSPAdes indeed represent plasmids. The Methods section describes the metaplasmidSPAdes workflow in further detail.

Plasmid verification. Each cyclocontig/component reconstructed by metaplasmidSPAdes may contain some chromosomal edges (or even consist entirely of chromosomal edges) arising from phage sequences, transposons, repeats within bacterial chromosomes, etc. We thus developed a plasmidVerify tool that examines the gene content of a cyclocontig and classifies it as *plasmidic* (*chromosomal*) using a Naive Bayesian classifier. Since plasmids harbor a large

variety of genes, plasmidVerify uses a plasmid-specific profile-HMM database to detect remote similarities between cyclocontigs/components detected by metaplasmidSPAdes and known plasmid-specific genes (see Methods section). To construct a set of plasmid-specific HMMs, we formed the *PlasmidDatabase* dataset containing all 9,937 plasmids from the RefSeq database (total length 1,007 Mb) and the *nonPlasmidDatabase* dataset containing randomly selected 10% of complete bacterial chromosomes from RefSeq (837 bacterial genomes with total length 3,229 Mb).

Analysis of putative novel plasmids found by metaplasmidSPAdes. We annotated some putative novel plasmids found by metaplasmidSPAdes using Prodigal (Hyatt et al., 2010) in metagenomic mode for gene prediction (version 2.6.3), the *hmmsearch* tool (Finn et al., 2011) with PfamA 30.0 database for gene annotation (version 3.1b2), and the CARD database (Jia et al. 2017) for predicting antibiotic-resistance genes (only “Perfect” and “Strict” hits).

Benchmarking plasmid verification tools. We benchmarked plasmidVerify against three plasmid verification tools (Table 1):

1. cBar tool based on 5-mer frequencies (Zhou and Xu, 2010),
2. PlasFlow tool based on deep neural networks (Krawczyk et al., 2018),
3. repl_HMM approach based on manually curated plasmid replicase HMMs (Jørgensen et al., 2014).

We did not include PlasmidFinder (Carattoli et al., 2014) in the benchmarking because Arredondo-Alonso et al, 2017 recently showed that it has a very low recall rate (0.36).

To construct a true negative dataset for benchmarking, we randomly selected 10% of bacterial genomes from the RefSeq database using Python random.sample() function. Since most putative plasmids output by metaplasmidSPAdes are shorter than typical bacterial

chromosomes, we split all bacterial chromosomes into fragments of length 10 kb and used them as the true negative dataset. This procedure resulted in 323,362 sequences (partitioning of *PlasmidDatabase* into 10 kb long fragments) that we refer to as *nonPlasmidContigs*. We selected *PlasmidDatabase* as the true positive dataset for benchmarking.

Table 1 illustrates that plasmidVerify improved on both the true positive and the false positive rates compared to the cBar and PlasFlow tools. Although the repl_HMM approach (that uses a small manually curated set of plasmid replicase HMMs) has a lower false positive rate than plasmidVerify, it is not well suited for our goals since it has a low true positive rate and is limited in its ability to detect diverse plasmids, i.e., it fails to detect novel plasmid with replicases that significantly differ from the replicases in the curated dataset.

To evaluate plasmidVerify's performance on the unseen branches of the microbial tree of life, we performed the following procedure. For each of the four phyla (Firmicutes, Proteobacteria, Cyanobacteria, and Bacteroidetes) we removed all plasmids from the phylum from the training dataset, retrained plasmidVerify on the reduced training dataset, and tested it on the members of the removed phylum (Supplemental Table S1). The false negative (positive) rates varied from 14.6% to 19.6% (1.3% to 3.6%) across the four analyzed phyla.

We also tested various plasmid verification tools on the set of viral contigs that represent a major source of non-plasmidic circular DNA elements (Supplemental Table S2).

Datasets. We benchmarked metaplasmidSPAdes using one dataset with multiple isolate genomes, three mock metagenomic datasets with known bacterial genomes, four metagenomic datasets (with unknown genomes), and one plasmidome dataset (all datasets contain paired-

end Illumina reads). To infer the set of plasmids in each mock metagenomic dataset, we compiled the list of known plasmids from the genomes (including all strains with data present in RefSeq) present in this dataset. To check which plasmids from this list are indeed present in the mock sample, we mapped all metagenomic reads to each of this plasmids. We assume that a plasmid is present in the mock dataset (reference plasmid) if more than 95% of its length is covered by metaSPAdes assembly. We used metaSPAdes for this verification since all existing plasmid analysis tools use its assembly graph for plasmid assembly. See Supplemental Table S3 for information about plasmids in the mock datasets. It is worth noting that even though mock metagenomes are usually formed from well-studied genomes, metaplasmidSPAdes was able to reveal some still unknown plasmids even in the mock metagenomes.

Below we provide a brief description of each of the datasets (see Supplemental Table S4 "Information about benchmarking datasets" for detailed information).

ISOLATES. The ISOLATES dataset consists of 21,933 bacterial datasets from the JGI GOLD database (gold.jgi.doe.gov) representing isolate bacterial samples.

HMP. The HMP dataset is a mock community of 19 bacterial species, one archaea and one yeast species studied by the Human Microbiome Project Consortium (HMP Consortium, 2012). 20 plasmids were originally reported in this dataset but our more stringent approach reduced the number of reference plasmids to 14 (total length ≈854 kb).

MBARC. The MBARC (Mock Bacteria ARchaea Community) dataset is a mock microbial community of 23 bacterial and 3 archaeal species described in Singer et al, 2016. We identified 10 plasmids of total length ≈756 kb in the MBARC dataset.

SYNTH. The SYNTH dataset is a mock microbial community of 64 diverse bacterial and archaeal species described in Shakya et al., 2013. Shakya et al., 2013 identified 32 plasmids in

this dataset but our more stringent approach reduced the number of reference plasmids to 19 (total length ≈1450 kb).

INFANT. The INFANT is a human microbiome dataset from an infant's gut described in Bäckhed et al., 2017.

CROHN. The CROHN is a human gut microbiome dataset from a patient suffering from Crohn's disease (analyzed in Nurk et al., 2017).

PLASMIDOME. The PLASMIDOME is a plasmid-enriched dataset from a microbial community in a biological wastewater treatment reactor described in Shi et al., 2018.

MARINE. The MARINE is a marine sediment metagenome dataset collected near the field of active hydrothermal vents in the Atlantic Ocean (Spang et al., 2015).

LAKE. The LAKE is a lake metagenome dataset collected at an Indian lake subjected to industrial pollution with fluoroquinolone antibiotics.

Analyzing the ISOLATES dataset. We searched for plasmids in the ISOLATES dataset with the goal of identifying new plasmids that might have evaded detection in the already completed sequencing projects. We did not benchmark Recycler since Arredondo-Alonso et al, 2017 have already benchmarked plasmidSPAdes and Recycler on diverse isolate datasets.

plasmidSPAdes generated 44,172 plasmidic connected components, including 15,499 cyclocontigs that originated from 7,987 out of 21,933 genomes in the ISOLATES dataset. To simplify analysis, we limited benchmarking to cyclocontigs and ignored other connected components output by plasmidSPAdes^{+}.

To remove duplicated cyclocontigs from this set, we clustered them based on their *k*-mer content using Mash (Ondov et al. 2016) and classified plasmids as duplicates if their *k*-mer compositions differed by less than 1%. Once duplicates have been removed 6,694 out of the 15,499 identified cyclocontigs were classified as unique. 2,280 of these cyclocontigs (referred to as *plasmidic cyclocontigs*) were classified as plasmids by plasmidVerify (Figure 2). We compared these cyclocontigs against the CARD database of antibiotic-resistance genes (ARG) and detected 356 ARGs in 203 cyclocontig out of 2,280 cyclocontigs (see Figure 2B and Supplemental Table S5 for details).

To doublecheck whether a putative cyclocontig originated from a plasmid or a bacterial chromosome, we aligned it against the nt database using the BLAST tool (Altschul et al, 1990) with the e-value threshold 0.001. Cyclocontigs that aligned to the non-plasmidic sequences in the nt database (bacterial chromosomes, viruses, etc.) likely represent false positives, but cyclocontigs that aligned to plasmids (or do not align at all) may represent known or novel plasmids. Thus BLAST alignments can be used as an approximation for the ground truth for additional benchmarking of plasmidVerify, cBar, repl_HMM and PlasFlow (Supplemental Table S6).

If a cyclocontig aligned to multiple sequences in the nt database, we analyzed the one with the maximal BLAST score (alignments to sequences of unknown origin are ignored). BLAST generates either a single alignment that extends over the entire length of the cyclocontig or multiple local alignments. We defined the *span* of a cyclocontig as the ratio of the total alignment length over the cyclocontig length, and the *identity* of the cyclocontig as the average percent identity across all alignments.

1,134 and 603 out of 2,280 plasmidic cyclocontigs aligned to known plasmids with the span exceeding 10% and 90%, respectively. The remaining 2,280 - 1,134 = 1,146 cyclocontigs can be broken down into the following three categories (see Supplemental Table S5 for details):

1. 255 cyclocontigs ambiguously matched to plasmid/chromosome with span >10% (putative integrative plasmids)
2. 480 matched bacterial chromosomes (false positive bacterial segments)
3. 31 matched viral sequences (false positive phage segments)
4. 380 did not match any known plasmids/chromosomes with the span exceeding 10% and were classified as *novel plasmids*

We analyzed some of the newly identified plasmids in more details (see Supplemental Figure S1 for plasmid maps):

1. A 7,895 nucleotide long putative plasmid (from *Streptococcus pseudopneumoniae* clinical isolate) with span 28% and identity 96% carried an *Erm* 23S ribosomal RNA methyltransferase, providing resistance to macrolide antibiotics. It also carried a toxin-antitoxin system *PparE/relB* and *zeta* toxin that may inhibit the cell wall biosynthesis and act as a bacteriocin.
2. A 53,557 nucleotide long putative plasmid (from *Enterobacter sp*. CC120223-11) with span 12% and identity 90% carried an ATP-binding cassette (ABC) antibiotic efflux pump. It contained a toxin-antitoxin system *vapB-vapC*, genes related to pili and flagella development, and putative members of type IV conjugal transfer systems (Pfam families T4SS_TraI and TraI_2_C), indicating that it is likely self-transferable. It was similar to known plasmids only in the short region containing the *parA/parB* operon that ensures the accurate partitioning of plasmids after division.

3. The longest putative novel plasmid in the ISOLATES dataset (582 kb) belonged to the halophilic marine gammaproteobacteria *Ferrimonas marina*, strain DSM 16917. It encoded 685 predicted genes and contained the plasmid replication protein gene *repA,* as well as an outer membrane phospholipase A1 (*OMPLA*) essential for bacterial secretion, proteins for flagella formation, and *ydaS/ydaT* toxin-antitoxin system. It also had some phage signatures such as the phage integrase genes and bacteriophage T4-like capsid assembly protein (*Gp20*). However, the phage integrase genes do not represent a strong phage marker since they often occur in plasmids.
4. The shortest putative novel plasmid in the ISOLATES dataset (length 1284 bp) encoded a single protein (firmicute plasmid replication protein RepL) and belonged to the fish pathogen *Ca. Ichthyocystis* 2013Ark19i, a recently described novel intracellular β-proteobacteria (Seth-Smith et al., 2016).

Analyzing the HMP dataset. metaplasmidSPAdes reconstructed 21 cyclocontigs in the HMP dataset. plasmidVerify classified seven of them as plasmidic and all of them have corresponding reference plasmids. metaSPAdes and Recycler reconstructed 4 and 6 reference plasmids, respectively (Table 2 and Supplemental Table S3). metaplasmidSPAdes identified no small uniformly-covered connected components in the HMP dataset.

We analyzed why metaplasmidSPAdes missed 14-7=7 reference plasmids in the HMP dataset. Six of them were non-dominant plasmids that share repeats with their bacterial hosts or other plasmids (see Supplemental Table S3 for details). The remaining one (dominant plasmid NZ_CP015213.1) was not reconstructed as a single cyclocontig since it had a long intra-repeat This plasmid was not output as a uniformly covered connected component since it shares more than 50% of its length with another plasmid (NC_009007.1) and fails the test on the uniformity of

coverage as the total length of medial edges (see Methods section) exceeds 80% of the size of this component. For each plasmid that was not assembled in a single cyclocontig by metaplasmidSPAdes, we computed the size and the number of edge count of the largest connected component that contains this plasmid at each iteration of metaplasmidSPAdes (Supplemental Table S7).

Analyzing the MBARC dataset. metaplasmidSPAdes reconstructed 32 cyclocontigs and plasmidVerify classified 8 of them as plasmidic. metaplasmidSPAdes assembled eight out of ten reference plasmids in the MBARC dataset into a single cyclocontig (metaSPAdes and Recycler reconstructed six plasmids each). Two remaining plasmids were non-dominant plasmids that were missed by metaplasmidSPAdes because their coverage was close to the median coverages of their host chromosomes that share long repeats with these plasmids.

plasmidVerify erroneously classified 2 out of 8 assembled reference plasmids as non-plasmidic: (i) one plasmid from the archaea *Natronococcus occultus* was misclassified because plasmidVerify is not designed to verify archaeal plasmids and (ii) one short plasmid (of length 2,931 bp) did not yield any hits in the Pfam-A database.

Additionally, plasmidVerify classified two cyclocontigs as plasmidic - a 2,876 nucleotide-long cyclocontig with a plasmid replication protein that likely represents a novel plasmid (span 19% and identity 76%) and a 53 kb long cyclocontig that carries a plasmid-specific resolvase gene, and aligns to a bacterial chromosome and various plasmids.

Analyzing the SYNTH dataset. metaplasmidSPAdes reconstructed 87 cyclocontigs in the SYNTH dataset and plasmidVerify classified 13 of them as plasmidic. metaSPAdes, Recycler,

and metaplasmidSPAdes reconstructed 6, 7, and 8 out of the 19 reference plasmids, respectively. The remaining 11 reference plasmids in the SYNTH dataset evaded identification by metaplasmidSPAdes since:

1. 10 of them were non-dominant and share long repeats with chromosomes or plasmids with the same or higher coverage (see Supplemental Table S3 for details).
2. one dominant plasmid was not output as a cyclocontig because it has inter-plasmidic repeats larger than the library insert size. It was not output as a uniformly covered connected component either since its length (408 kb) exceeds the default threshold for the connected component length (200 kb).

Six out of 13 cyclocontigs that metaplasmidSPAdes classified as plasmidic likely represent still unknown plasmids in the SYNTH community:

1. three cyclocontigs have ~40% span and 80%-93% identity with known plasmids in various *Phaeobacter* genomes. Two of them (of lengths 22,035 and 5,444 nucleotides) were conjugative plasmids carrying mobilization proteins (MobA/MobC), and one of them (of length 11,215 bp) contained a plasmid replicase gene *repA*, a toxin-antitoxin system *parE/parD,* and a copper resistance operon *copAB*.
2. one cyclocontig (of length 38,668 nucleotides) did not match any known plasmid/bacterial genomes but carried a plasmid replicase gene.
3. two cyclocontigs (of length 22,963 and 4,103 nucleotides) both had short matches to known plasmids and chromosomes (with span 20% and identity 97-99%). Since they carry both a replicase gene and conjugal transfer proteins, they likely represent conjugative plasmids.

The remaining 2 out of 13 cyclocontigs that metaplasmidSPAdes classified as plasmidic aligned to bacterial chromosomes and likely represent false-positives (prophages or transposones). plasmidVerify misclassified three reference plasmids (of lengths 16,625, 8,368 and 8,362 nucleotides) as non-plasmidic since it did not detect any distinctively plasmidic genes within them.

Analyzing the INFANT dataset. metaplasmidSPAdes reconstructed 33 cyclocontigs in the INFANT dataset and plasmidVerify classified 5 of them as plasmidic (Table 3):

1. one of them (of length 4,234 nucleotides) matched the pRGFK1358 plasmid with 100% span and 95% identity.
2. one of them (of length 4,608 nucleotides) matched the pRGFK1348 plasmid with 56% span and 95% identity.
3. two of them (of length 3,687 and 3,338 nucleotides) did not match any known plasmids/chromosomes, but harbored the Mob plasmid recombination enzyme and the initiator of plasmid replication Rep3.
4. one of them (of length 1,553 nucleotides) matched bacterial chromosomes (likely a false positive).

Analyzing the CROHN dataset. metaplasmidSPAdes reconstructed 77 cyclocontigs in the CROHN dataset and plasmidVerify classified 28 of them as plasmidic (Table 3):

1. 4 of them matched known plasmids with 100% span and identity varying from 92% to 99%.
2. 14 of them matched known plasmids with spans varying from 21% to 79% and identity varying from 78% to 97%.

3. 9 of them had a span of under 10% and did not have significant matches with any sequences in the nr database.
4. 1 of them aligned to a bacterial chromosome with a span of 38% (likely a false positive).

A 1868-nucleotide long cyclocontig reconstructed by metaplasmidSPAdes and classified as non-plasmidic by plasmidVerify turned out to be a *Streptococcus* phage phiJH1301-2, carrying an aminoglycoside resistance gene (phages were recently shown to carry antibiotics-resistance genes (Balcazar, 2014)). Although plasmidSPAdes and metaplasmidSPAdes were not designed for viral assembly (there is still no specialized software for viral assembly from genomic and metagenomic datasets), our analysis demonstrates that they are able to detect viruses in genomic and metagenomic datasets.

Analyzing the PLASMIDOME dataset. Since the PLASMIDOME dataset did not contain information about the reference plasmids, we generated some references for this dataset by mapping the assembled PLASMIDOME contigs against the plasmid database (with Mash screen (Ondov et al., 2019), QUAST (Gurevich et al., 2013), and BLAST). This analysis revealed ten reference plasmids with a total length of ≈100 kb. The fact that the total length of the identified reference plasmids in the PLASMIDOME dataset was two orders of magnitude smaller than the total assembly length suggests that most plasmids in the PLASMIDOME dataset are not present in the plasmid database.

metaplasmidSPAdes reconstructed 103 cyclocontigs in the PLASMIDOME dataset and plasmidVerify classified 87 of them as plasmidic (Table 3). Seven of these 87 cyclocontigs

matched known plasmids with a span >90% (with identity varying from 82% to 99%) and 54 have a span exceeding 10% (with identity varying from 75% to 99%). 9 out of these 87 cyclocontigs matched a bacterial chromosome or a phage with a span exceeding 10% (likely false positives). The remaining 87-7-54-9=17 contigs have spans under 10% and were classified as putative novel plasmids.

Analyzing the MARINE dataset. metaplasmidSPAdes reconstructed 127 cyclocontigs in the MARINE dataset and plasmidVerify classified 21 of them as plasmidic (Table 3). Three of these cyclocontigs matched known plasmids (one with a 99% span and identity, two with spans of 20% and 60% and identity of 87% and 93%, respectively). Three others matched bacterial chromosomes with spans of 14%, 33%, and 48%, and identity of 75%, 100% and 74%, respectively. The remaining 15 cyclocontigs have spans under 10% and were classified as putative novel plasmids.

Analyzing the LAKE dataset. metaplasmidSPAdes reconstructed 1,860 cyclocontigs in the LAKE dataset and plasmidVerify classified 417 of them as plasmidic (Table 3). 7 of these cyclocontigs matched bacterial chromosomes, 13 matched viral sequences, 9 matched both chromosomes and plasmids and thus likely represent integrative plasmids. 59 cyclocontigs matched known plasmids with span exceeding 10%, and the remaining 329 cyclocontigs had no significant matches to the nt database. The large number of putative plasmids in the LAKE dataset (as compared to the other datasets we analyzed) may be explained by the fact that the lake was polluted with fluoroquinolones, making plasmids carrying antibiotics resistance and other genes particularly beneficial to the hosts.

DISCUSSION

We demonstrated that plasmidSPAdes^{+} and metaplasmidSPAdes improve on existing tools for plasmid reconstruction and identify many novel plasmids in diverse genomic and metagenomic datasets. However, even with the improved mechanism of identifying new plasmids, it is still likely that many more plasmids continue to evade detection (false negatives) and some non-plasmidic cyclocontigs end up being reported as plasmids (false positives).

Since some plasmids do not harbour any distinctively plasmidic genes (as defined based on the analysis of known plasmids), the corresponding cyclocontigs are not detected by metaplasmidSPAdes. Users have the option to switch off the plasmidVerify tool and manually analyze all cyclocontings that fall into this category.

Application of plasmidSPAdes^{+} and metaplasmidSPAdes to various datasets revealed that many plasmids remain undetected during genomic and metagenomic studies. Moreover, this analysis revealed the enormous variability of plasmids: a large fraction of the found plasmids did not match to any known ones. Even in the already completed sequencing projects (ISOLATES dataset) we found 1166 putative plasmidic cyclocontigs with <90% similarity to known ones and without significant hits to viruses or bacterial chromosomes. 91 of these putative plasmids contain antibiotic resistance genes, 246 contain carbohydrate-active enzymes (*CAZymes*), and 54 contain adhesion-related genes (possibly contributing to horizontal gene transfer). Expansion of the set of known plasmids can help classify them and reflects the evolutionary relationships between plasmids. One can compare plasmid phylogeny with host phylogeny and phenotypic traits, and analyze the relationships between resistance type, plasmid replication type, and host type. This information would also be relevant for epidemiological studies. For example, it remains unclear whether resistance dissemination involves a diverse set of plasmids or a single

dominant epidemic type. It may correlate with the host range and the type of the antibiotic resistance gene (Mathers et. al, 2015). metaplasmidSPAdes will help generate a comprehensive dataset of plasmids to help address these questions.

METHODS

metaplasmidSPAdes workflow.

metaplasmidSPAdes uses the default values cov_{add}=5x and cov_{mult}=1.3. The plasmidVerify module checks whether a cyclocontig or a connected component in the assembly graph originated from a plasmid using a Naive Bayesian classifier. To avoid time-consuming read alignments at each iteration, metaplasmidSPAdes aligns paired-end reads against the assembly graph only once and updates the information about the read alignments during the graph modifications. The concept of a plasmid-like connected component is described in the Methods section. metaplasmidSPAdes pseudocode can be presented as follows:

metaplasmidSPAdes(*Reads*, cov_{add}, cov_{mult})
Plasmids ← empty set
Graph ← assembly graph of *Reads* constructed by metaSPAdes
align paired-end reads to *Graph* and compute coverage of each edge by reads
cov_{max} ← maximum coverage of an edge in *Graph*
cov ← *0*
while *cov* < cov_{max}
  Contigs ← the set of all paths (contigs) in *Graph* generated by exSPAnder
  for each cyclocontig *Cycle* in *Contigs*

add *Cycle* to the set *Plasmids*

for each small *plasmid-like* connected components *Component* in *Graph*

if *Component* contains edges that do not belong to cyclocontigs in *Plasmids*

add *Component* to the set *Plasmids* and remove it from *Graph*

$cov \leftarrow max\{cov + c_{add}, cov * c_{mult}\}$

remove edges with coverage below *cov* from the assembly graph

Iteratively remove dead-end edges from *Graph* (Antipov et al, 2016)

replace each non-branching path in *Graph* with a single edge and recompute its coverage

for each cyclocontig or connected component *C* in *Plasmids*

If plasmidVerify(*C*)=0

remove *C* from *Plasmids*

return *Plasmids*

plasmidVerify workflow. We predicted genes with Prodigal v2.6.3 (Hyatt, 2010) and ran hmmsearch (part of HMMER 3.1b2, http://hmmerorg/) using Pfam-A database v. 30.0 (Finn et al, 2016) on the training datasets from both *PlasmidDatabase* and *nonPlasmidDatabase* (7550 plasmids and 242,681 "contigs", respectively). For each of the two runs and for each HMM, we counted the frequencies of matches (with the bit-score cutoff set to the "noise" level from the Pfam-A database) to *PlasmidDatabase* and *nonPlasmidDatabase*, respectively. These frequencies were used to train a Naive Bayesian classifier (Friedman et al., 2001). Supplemental Table S8 lists the HMM frequencies in the training dataset. Given a cyclocontig, plasmidVerify predicts genes in this contig using Prodigal in the metagenomic mode, runs

hmmsearch on the predicted proteins, and classifies the contig as plasmidic or chromosomal by applying the Naive Bayesian classifier.

plasmidVerify classified 1-2% of contigs in the analyzed metagenomic assemblies as plasmidic (Supplemental Table S9). However, since plasmidVerify incorrectly classified a number of chromosomal contigs as plasmidic, plasmidVerify (and other plasmid verification tools) by itself is unable to accurately classify plasmids and thus has to be combined with metaplasmidSPAdes for increased accuracy.

Plasmid-like connected components. We define the size of a connected component in the assembly graph as the total length of its edges. The connected component is called *small* if its size does not exceed $size_{max}$ (default value: 200 kb). For each connected component, we compute its median coverage by reads (cov_{med}) as described in Antipov et al., 2016. An edge in a connected component is called *medial* if its coverage exceeds cov_{med}/α and does not exceed $cov_{med}*\alpha$ (the default value α=1.3). A connected component is called *uniform* if the total length of its medial edges exceeds 80% of the size of this component. We classify a small uniform connected component as *plasmid-like* if its size exceeds 1 kb and if it contains at most two dead-end edges.

DATA ACCESS

metaplasmidSPAdes results on all mentioned datasets are available at http://data.cab.spbu.ru/index.php/s/tz7mCqDipgbcsbW as a Supplemental File S1.

Source code is available at https://github.com/ablab/spades/tree/metaplasmid_3.13.0 and as Supplemental File S2.

ACKNOWLEDGEMENTS

This study was supported by Russian Science Foundation (grant number 19-16-00049). We are grateful to Yevgeniya Mazur for many helpful comments.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410.

Anda, M., Ohtsubo, Y., Okubo, T., Sugawara, M., Nagata, Y., Tsuda, M., ... & Mitsui, H. (2015). Bacterial clade with the ribosomal RNA operon on a small plasmid rather than the chromosome. *Proceedings of the National Academy of Sciences,* 112(46), 14343-14347.

Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., & Pevzner, P. (2016). plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*, 32(22), 3380–3387.

Arredondo-Alonso, S., Willems, R. J., van Schaik, W., & Schürch, A. C. (2017). On the (im) possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial Genomics*, *3*(10).

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., ... & Khan, M. T. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. Cell host & microbe, 17(5), 690-703.

Balcazar, J. L. (2014) Bacteriophages as Vehicles for Antibiotic Resistance Genes in the Environment. *PLoS Pathogenes*, 10, e1004219

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. (2012).

SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology, 19*(5), 455-477.

Carattoli, A., Zankari, E., García-Fernández, A., Larsen, M.V., Lund,)., Villa, L., Aarestrup, F.M., Hasmanb, H. (2014) In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrob Agents Chemother*. 2014 *58*(7): 3895–3903.

Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26 (2010): 2460-2461.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., ... & Salazar, G. A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acid Research*, 44(D1), D279-D285.

Friedman, J., Hastie, T., Tibshirani R. *The elements of statistical learning*. Springer Series in Statistics, 2001.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072-1075.

Halary, S., Leigh, J. W., Cheaib, B., Lopez, P., & Bapteste, E. (2010). Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences*, 107(1), 127-132.

HMP Consortium. (2012). A framework for human microbiome research. Nature, 486(7402), 215-21.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1), 119.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., ... & Doshi, S. (2016). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 45(D1), D566–D573

Johnson, C. M., & Grossman, A. D. (2015). Integrative and conjugative elements (ICEs): what they do and how they work. *Annual Review of Genetics*, 49, 577-601.

Jørgensen, T. S., Xu, Z., Hansen, M. A., Sørensen, S. J., & Hansen, L. H. (2014). Hundreds of circular novel plasmids and DNA elements identified in a rat cecum metamobilome. *PLoS One*, *9*(2), e87924.

Krawczyk, P. S., Lipinski, L., & Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research, 46*(6), e35.

Li, A. D., Li, L. G., & Zhang, T. (2015). Exploring antibiotic resistance genes and metal resistance genes in plasmid metagenomes from wastewater treatment plants. *Frontiers in Microbiology, 6*, 1025.

Mathers, A. J., Peirano, G., & Pitout, J. D. (2015). The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant Enterobacteriaceae. *Clinical microbiology reviews*, *28*(3), 565-591.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research, 27*(5), 824-834.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome biology*, *17*(1), 132.

Ondov, B. D., Starrett, G. J., Sappington, A., Kostic, A., Koren, S., Buck, C. B., & Phillippy, A. M. (2019). Mash Screen: High-throughput sequence containment estimation for genome discovery. BioRxiv, 557314.

Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., ... & Stoesser, N. (2017). Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid*, *91*, 42-52.

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., ... & Kyrpides, N. C. (2016). Uncovering Earth's virome. *Nature*, *536*(7617), 425.

Petersen, J., Brinkmann, H., Berger, M., Brinkhoff, T., Päuker, O., & Pradella, S. (2010). Origin and evolution of a novel DnaA-like plasmid replication type in Rhodobacterales. *Molecular Biology and Evolution*, 28(3), 1229-1240.

Pevzner P, Tesler G. (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* 13(1), 37-45.

Prjibelski, A. D., I. Vasilinetc, A. Bankevich, A. Gurevich, T. Krivosheeva, S. Nurk, S. Pham, A. Korobeynikov, A. Lapidus, and P. A. Pevzner. (2014) ExSPAnder: a universal repeat resolver for DNA fragment assembly. *Bioinformatics* 30, 293-i301.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2006). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research, 35*(suppl 1), D61-D65.

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., ... & Pesant, S. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*,

537(7622), 689-693.

Rozov, R., Brown Kav, A., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I., & Shamir, R. (2017). Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics*, 33(4), 475-482.

Seth-Smith, H. M., Dourala, N., Fehr, A., Qi, W., Katharios, P., Ruetten, M., ... & Thomson, N. R. (2016). Emerging pathogens of gilthead seabream: characterisation and genomic analysis of novel intracellular β-proteobacteria. *The ISME Journal*, 10(7), 1791.

Shakya, M., Quince, C., Campbell, J. H., Yang, Z. K., Schadt, C. W., & Podar, M. (2013). Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental microbiology*, *15*(6), 1882-1899.

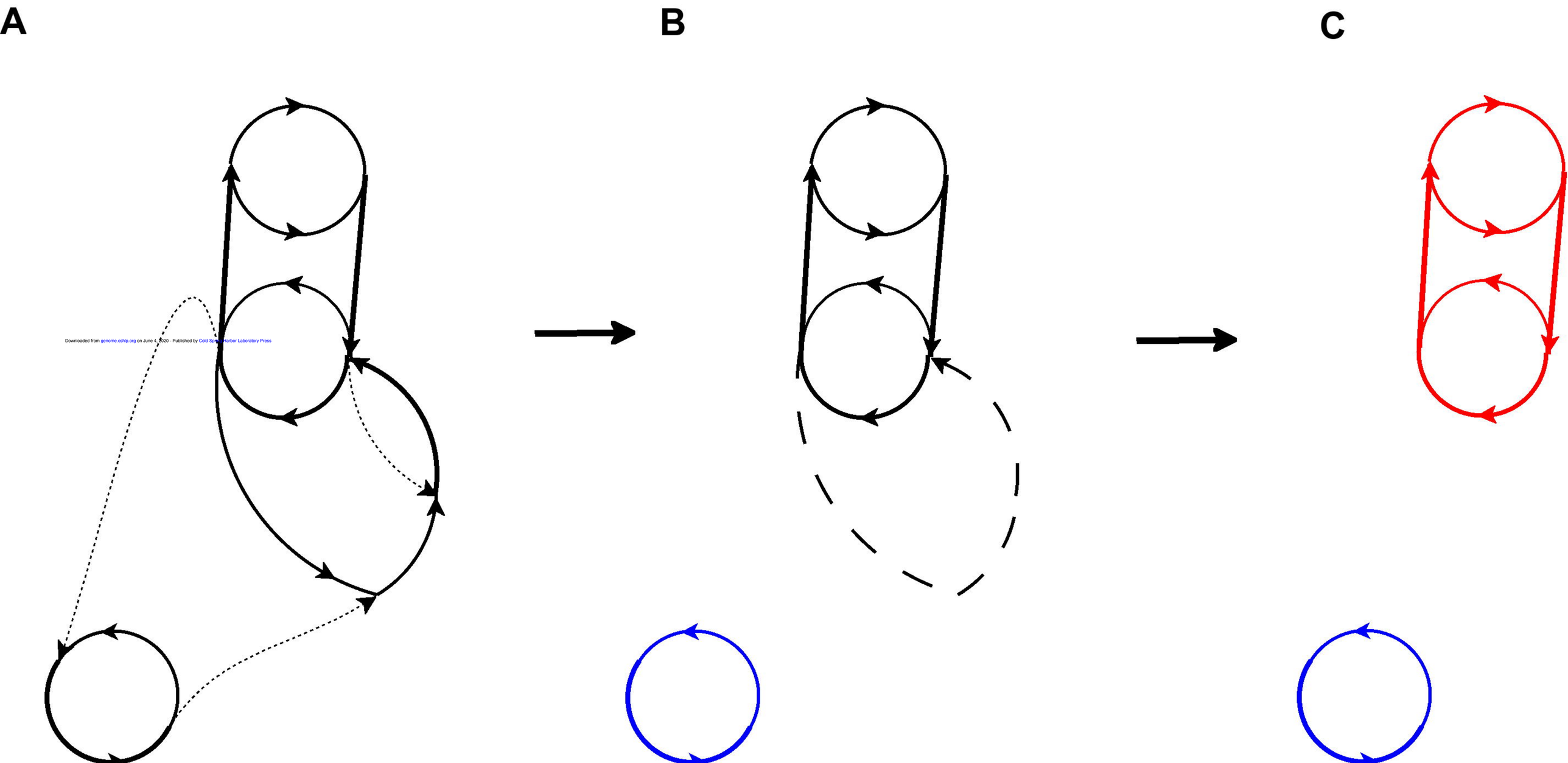
Shi, Y., Zhang, H., Tian, Z., Yang, M., & Zhang, Y. (2018). Characteristics of ARG-carrying plasmidome in the cultivable microbial community from wastewater treatment system under high oxytetracycline concentration. *Applied Microbiology and Biotechnology*, 1-12. 102(4), 1847–1858

Singer E, Andreopoulos B, Bowers RM, Lee J, Deshpande S, Chiniquy J, Ciobanu D, Klenk HP, Zane M, Daum C, Clum A. (2016) Next generation sequencing data of a defined microbial mock community. *Scientific Data.* 27;3:160081.

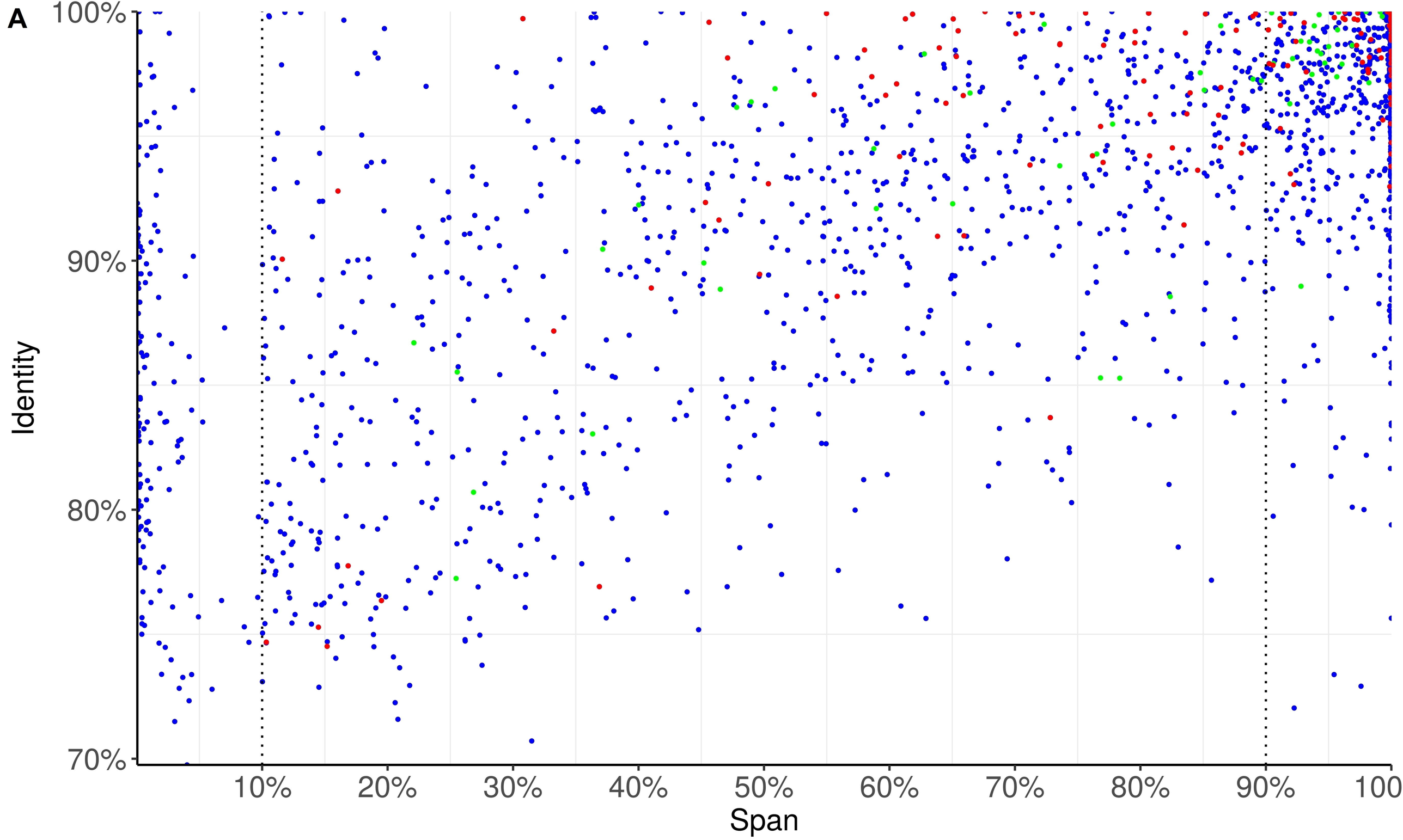
Spang, A., J.H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, and T.J.G. Ettema. "Complex archaea that bridge the gap between prokaryotes and eukaryotes." *Nature* 521, (2015): 173-179.

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational biology*, *7*(1-2), 203-214.

Zhou, F., & Xu, Y. (2010). cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics, 26*(16), 2051-2052.



A



B

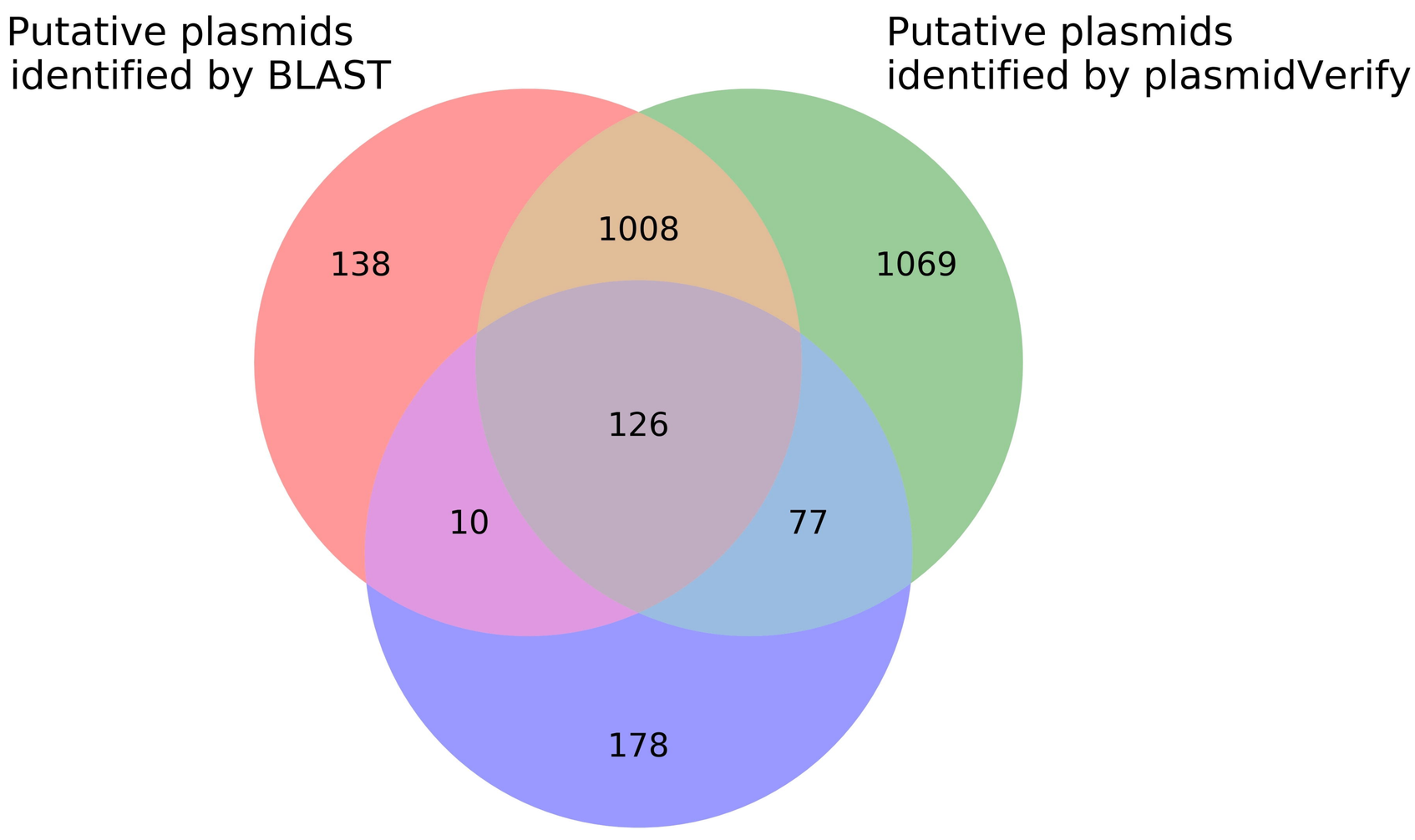


Figure 1: Iterative plasmid detection in the assembly graph. (A) The assembly graph *Graph* with three dotted edges representing edges with the lowest coverage. (B) Removal of three edges with the lowest coverage from *Graph* reveals a plasmid (cyclocontig) shown in blue. The three edges on the graph in panel A now represent a single dashed edge that has the lowest coverage in *Graph*. (C) The same graph after the second iteration of metaplasmidSPAdes that removes the dashed edge with the lowest coverage and reveals a plasmid (connected component) shown in red.

Figure 2. **The scatter plot of the span and identity for all 2,280 unique cyclocontigs in the ISOLATES dataset reconstructed by plasmidSPAdes^{+} (A) and the Venn diagram for cyclocontigs identified as plasmids by plasmidVerify, cyclocontigs identified as plasmids by BLAST (span over 10%) and cyclocontigs containing ARGs (B).** (A) Each dot represents a cyclocontig reported by plasmidSPAdes and verified by plasmidVerify. Red dots represent cyclocontigs containing antibiotic-resistance genes. Green dots represent cyclocontigs classified as viral sequences. (B) The Venn diagram illustrates that the HMM-based approach in metaplasmidSPAdes identifies many plasmids with important phenotypes that are missed by a straightforward BLAST-based approach.

	cBar	PlasFlow	repl_HMM	plasmidVerify
PlasmidDatabase test dataset true positive, 2,484 plasmids	2,117 (85.2%)	1,959 (78.9%)	1,298 (52,3%)	2,208 (88.9%)
nonPlasmidContigs test dataset true negative, 80,840 contigs	15,810 (19.5%)	16,526 (20.4%)	580 (0.7%)	2,463 (3.1%)

Table 1. Benchmarking various plasmid verification tools. *PlasmidDatabase* (9937 plasmids) and *nonPlasmidContigs* (323362 contigs of length 10 kb) were divided into a training (75%) and a test (25%) datasets. plasmidVerify was trained on the training dataset. All plasmid verification tools were benchmarked on the test dataset. Since our goal is to distinguish *complete* plasmids from *short* chromosomal fragments output by metaplasmidSPAdes, our benchmarking datasets differ from the ones described in Zhou et al., 2010 and Krawczyk et al., 2018 where various plasmid verification tools were benchmarked on full plasmids/chromosomes or plasmidic/chromosomal contigs of varying lengths.

dataset	#reference plasmids	#reconstructed reference plasmids		
		metaSPAdes	Recycler	metaplasmidSPAdes
HMP	14	4	6	**7**
MBARC	10	6	6	**8**
SYNTH	19	6	7	**8**

Table 2. Information about reference plasmids reconstructed as cyclocontigs by metaSPAdes, Recycler, and metaplasmidSPAdes (HMP, MBARC, and SYNTH datasets).

Dataset	Assembly length (metaSPAdes)	# of cyclocontigs (# of cyclocontigs verified by plasmidVerify)		
		metaSPAdes	Recycler	metaplasmidSPAdes
INFANTGUT	230 Mb	11(2)	**49 (5)**	33 (5)
CROHN	596 Mb	45 (15)	**-**	**77 (28)**
PLASMIDOME	18 Mb	56 (35)	71 (49)	**103 (87)**
MARINE	234 Mb	175 (24)	**210(28)**	127(21)
LAKE	119 Mb	1,882 (277)	1,609 (370)	**1,860 (417)**

Table 3. Number of cyclocontigs reconstructed by metaSPAdes, Recycler, and metaplasmidSPAdes in the INFANTGUT, CROHN, PLASMIDOME, MARINE, and LAKE datasets. We did not provide the Recycler results on the most complex CROHN dataset since it ran for over a month, but did not output any putative plasmids.

Plasmid detection and assembly in genomic and metagenomic datasets

Dmitry Antipov, Mikhail Raiko, Alla Lapidus, et al.

Genome Res. published online May 2, 2019
Access the most recent version at doi:10.1101/gr.241299.118