Bioinformatics, 2016, 1–8
doi: 10.1093/bioinformatics/btw493
Advance Access Publication Date: 27 July 2016
Original Paper

OXFORD

Genome analysis

plasmidSPAdes: assembling plasmids from whole genome sequencing data

Dmitry Antipov[1,*], Nolan Hartwick[2], Max Shen[3], Mikhail Raiko[2], 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, [2]Department of Computer Science and Engineering, University of California, San Diego, CA, USA and [3]Bioinformatics and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on May 29, 2016; revised on July 11, 2016; accepted on July 16, 2016

Abstract

Motivation: Plasmids are stably maintained extra-chromosomal genetic elements that replicate independently from the host cell's chromosomes. Although plasmids harbor biomedically important genes, (such as genes involved in virulence and antibiotics resistance), there is a shortage of specialized software tools for extracting and assembling plasmid data from whole genome sequencing projects.

Results: We present the plasmidSPAdes algorithm and software tool for assembling plasmids from whole genome sequencing data and benchmark its performance on a diverse set of bacterial genomes.

Availability and Implementation: plasmidSPAdes is publicly available at http://spades.bioinf.spbau.ru/plasmidSPAdes/

Contact: d.antipov@spbu.ru

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Plasmids are common in Bacteria and Archaea, but have been detected in Eukaryotes as well (Gunge *et al.*, 1982). The cells often have multiple plasmids of varying sizes existing together in different numbers of copies per cell. Plasmids are important genetic engineering tools and the vectors of horizontal gene transfer that may harbor genes involved in virulence and antibiotic resistance. Thus, studies of plasmids are important for understanding the evolution of these traits and for tracing the proliferation of drug-resistant bacteria.

Since plasmids are difficult to study using Whole Genome Sequencing (WGS) data, biologists often use special biochemical methods for extracting and isolating plasmid molecules for further *plasmidome sequencing* (Kav *et al.*, 2012; Williams *et al.*, 2006). In the case of WGS, when a genome of a bacterial species is assembled, its plasmids often remain unidentified. Obtaining information about plasmids from thousands of genome sequencing projects (without preliminary plasmid isolation) is difficult since it is not clear which contigs in the genome assembly have arisen from plasmids.

Since the proliferation of plasmids carrying antimicrobial resistance and virulence genes leads to the proliferation of drug resistant-bacterial strains, it is important to understand the epidemiology of plasmids and to develop plasmid typing systems. Carattoli *et al.* (2014) developed PlasmidFinder software for detecting and classifying variants of known plasmids based on their similarity with plasmids present in plasmid databases. However, PlasmidFinder is unable to identify novel plasmids that have no significant similarities to known plasmids.

Lanza *et al.* (2014) developed the *plasmid constellation network (PLACNET)* tool for assembling plasmids from WGS data and applied it for analyzing plasmid diversity and adaptation (de Toro *et al.*, 2014) PLACNET uses three types of information to identify plasmids: (i) information about scaffold links and coverage in the

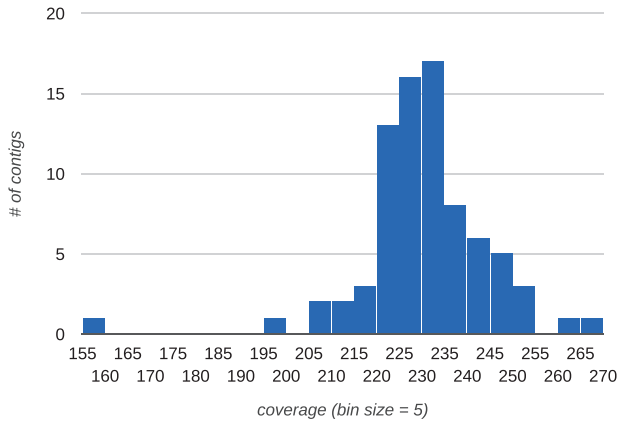

Fig. 1. *E. coli* coverage distribution for long contigs. The number of long contigs (greater than 10 kb in length) with a given *k*-mer coverage ($k = 55$) in the *E. coli* dataset (accession number ERA000206) assembled with SPAdes. The coverage of a *k*-mer in a genome is defined as the number of reads spanning this *k*-mer. The median coverage for long contigs is 231. Each bar in the histogram represents all contigs with the coverage in a bin of size 5. For example, the tallest bar in the histogram corresponds to the contigs with coverage between 230 and 235. The long contigs with minimum (160) and maximum (268) coverage have lengths 10438 and 13 078 bp, respectively

WGS assembly, (ii) comparison to reference plasmid sequences and (iii) plasmid-diagnostic sequence features such as replication initiator proteins. PLACNET combines these three types of data and outputs a network that needs to be further pruned by expert analysis to eliminate confounding data.

While combining all three types of data for plasmid sequencing is important, the focus of this paper is only on using WGS assembly for plasmid reconstruction in a fully automatic fashion. We argue that while the analysis of scaffolds in Lanza *et al.* (2014) is important, there is a wealth of additional information about plasmids encoded in the structure of the *de Bruijn graph* (constructed from *k*-mers in reads) that Lanza *et al.* (2014) do not consider. Recently, Rozov *et al.* (2015) demonstrated how to use the de Bruijn graphs constructed by the SPAdes assembler (Bankevich *et al.*, 2012) to significantly improve the plasmid assembly (focusing on data generated using plasmid isolation techniques) as well as reconstruction of plasmid sequences from metagenomics datasets. Below we describe a novel tool (plasmidSPAdes) aimed at assembling plasmids from the WGS data. Recently, this problem was addressed in the case of long SMRT reads (Conlan *et al.*, 2014) but it remains open for datasets containing short Illumina reads, which represent the lion's share of bacterial sequencing projects.

We show that plasmidSPAdes has the potential to massively increase the throughput of plasmid sequencing and to provide information about plasmids in thousands of sequenced bacterial genomes by re-assembling their genomes, identifying their plasmids and supplementing the corresponding GenBank entries with the plasmid annotations. Such plasmid sequencing efforts are important since many questions about plasmid function and evolution remain unanswered. For example, Anda *et al.* (2015) recently found a striking example of a bacterium (*Aureimonas* sp. AU20) that harbors the rRNA operon on a plasmid rather than on the chromosome. Thus, re-sequencing 1000s of bacterial genomes with the goal to reassemble their plasmids will help to answer important questions about plasmid evolution. We illustrate how plasmidSPAdes contributes to plasmid discovery by analyzing the *Citrobacter freundii* CFNIH1 genome with well-annotated plasmids and identifying a new previously overlooked plasmid in this genome. We also show how plasmidSPAdes was used to discover eight new plasmids in ten randomly chosen shotgun datasets derived from bacterial genomes and deposited in the Short Reads Archive. We further benchmark various tools for plasmid sequencing and provide the first analysis of accuracy of existing plasmid sequencing tools across diverse bacterial genomes.

2 Methods

2.1 Separating plasmids from chromosomes by read coverage

plasmidSPAdes uses the read coverage of contigs to assist in distinguishing between plasmids and chromosomes. Illumina DNA sequencing platform typically produces reads with highly uniform coverage of the bacterial chromosomes. Figure 1 and Table 1 illustrate that 92% (78 out of 85) of contigs greater than 10 kb in length in the assembly graph of the *E. coli* genome have coverage within 10% of the median value and 99% (84 out of 85) have coverage within 20% of the median value. The coverage of most genomes in this table is rather uniform with exception of *Bacillus anthracis* A1144, *Rhodococcus* J21s and *Thermus filiformis* ATT43280 (shown in bold).

Depending on the copy number of the plasmid, its coverage can be higher or lower than the chromosome coverage. For example, if a plasmid has a copy number 10, we expect it to have a much higher coverage than the chromosome coverage. Similarly, if a plasmid can only be found in 1/10 of the sampled cells, it will have a much lower coverage than the chromosome coverage. In order to distinguish plasmids and chromosomes by coverage, plasmidSPAdes first estimates the chromosome coverage. The naive strategy for estimating the chromosome coverage as the average coverage over all contigs often leads to an inflated estimate because some plasmids have very large copy numbers. Since such plasmids have high coverage, the average coverage may be skewed towards the plasmid coverage.

To avoid this pitfall, plasmidSPAdes computes the median coverage (denoted *medianCoverage*) using the assembly graph constructed by the SPAdes assembler (Bankevich *et al.*, 2012). SPAdes generates the assembly graph by first constructing the de Bruijn graph of all reads and further performs various *graph simplification* procedures (e.g. *bubble* and *tip* removals) to transform it into the assembly graph.

An edge in the assembly graph is classified as *long* if the length of the contig resulting from this edge exceeds the parameter *longEdgeLength* (the default value is 10 000 bp) and short otherwise. The median coverage is defined as the maximum coverage for which the collection of all long contigs of that coverage or greater covers at least half of the total length of the collection of all long contigs in the SPAdes assembly graph. We focus on long (rather than all) contigs for two reasons. First, analyzing long contigs allows us to exclude most repeats from consideration since the longest repeats in most bacterial genomes are shorter than 10 kb (Koren *et al.*, 2013). Second, long contigs have a lower variance in their coverage compared to short contigs. Figure 2 illustrates the larger variance in coverage for medium-sized (longer than 1 kb but shorter than 10 kb) contigs. These medium-sized contigs for the *E. coli* genome vary in coverage from 200 to 1702. For comparison, the long contigs (>10 kb) vary in coverage from 160 to only 268.

Given a parameter *maxDeviation* (the default value is 0.3), plasmidSPAdes classifies a long edge e in the assembly graph as a *chromosomal edge* if its coverage satisfies the following condition:

$$1 - maxDeviation < \frac{Coverage(e)}{medianCoverage} < 1 + maxDeviation$$

Table 1. Summary of contig coverage in various bacterial datasets. For each bacterial genome, each row shows the fraction of contigs with a coverage within 10, 20 and 30% of the median value

Genome	% contigs with cov. within x% of med.		
	10%	20%	30%
Bacillus cereus ATCC-10987	66(78)	84(100)	84(100)
Rhodobacter sphaeroides 2.4.1	53(73)	68(90)	73(97)
Providencia stuartii ATCC 33672	83(86)	98(100)	98(100)
Citrobacter freundii CFNIH1	47(55)	86(100)	86(100)
Corynebacterium callunae DSM 20147	88(91)	96(100)	100(100)
***Bacillus anthracis* A1144**	15(14)	23(23)	50(50)
Escherichia coli K12	92	99	99
Burkholderia cenocepacia DDS 22E-1	73	99	100
Acinetobacter sp. UNC434CL69	96	96	96
Butyrivibrio sp. INlla16	40	79	96
Lachnospiraceae sp. NK3A20	71	100	100
Luteibacter sp. UNC138MF	60	100	100
Prevotellaceae bacterium HUN156	100	100	100
Pseudoalteromonas sp. ND6B	67	98	100
***Rhodococcus* sp. J21s**	44	52	53
Ruminococcus flavefaciens YAD2003	64	100	100
Sphingomonas sp. UNC305MF	73	100	100
***Thermus filiformis* ATT43280**	45	73	86

For each dataset, we computed the median coverage and then counted the number of long (>10 kb) contigs within x% of the median coverage (for $x = 10$, 20 and 30%). For example, for *E. coli*, the median coverage was 216 and 85 out of 86 long contig had a coverage between within 20% of the median coverage. The table is divided into three parts corresponding to species with known plasmids (upper part), species that have no plasmids (middle part), and species for which it remains unknown whether they have plasmids. For the datasets with known plasmids we also computed the fraction of *chromosomal* contigs with a coverage within 10%, 20% and 30% of the median value among all chromosomal contigs (shown in parenthesis). The description of these datasets are provided in section 3.1 and Appendix.

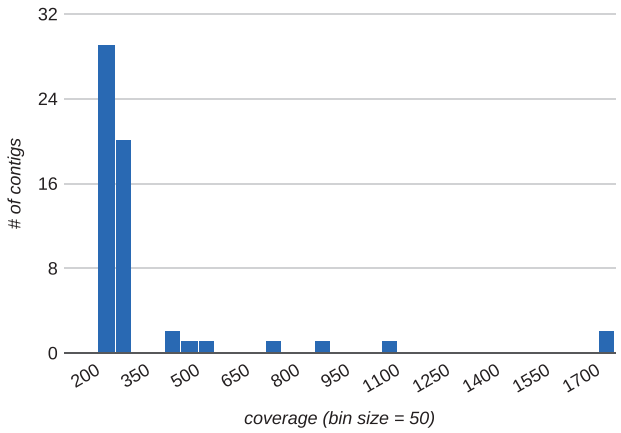

Fig. 2. *E. coli* coverage distribution for medium-sized contigs. Number of medium-sized contigs (longer than 1 kb but shorter than 10 kb in length) with a given k-mer coverage ($k = 55$) in the *E. coli* genome. The median coverage for medium-sized contigs is 231. Each bar in the histogram represents all contigs with the coverage in a bin of size 50. For example, the tallest bar in the histogram corresponds to the contigs with coverage between 200 and 250. The contigs with minimum (200) and maximum (1702) coverage have lengths 4122 and 1702 bp, respectively. Note that bars corresponding to repeats of various multiplicities are located near the projected coverage 462 (multiplicity 2), 693 (multiplicity 3), 924 (multiplicity 4), etc

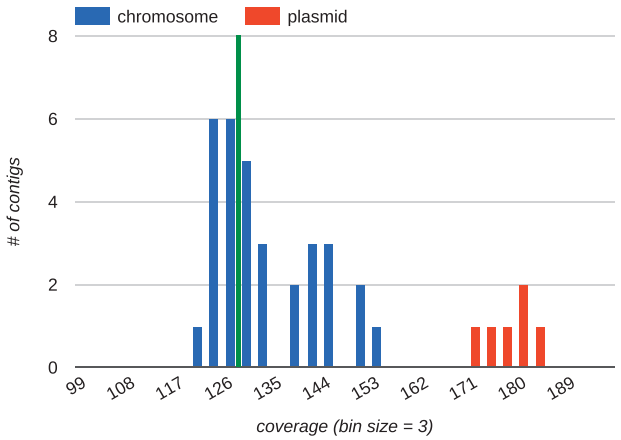

Fig. 3. *B. cereus* coverage distribution for long contigs. Number of contigs with a given k-mer coverage ($k = 55$) for all long contigs in *Bacillus cereus* (*medianCoverage* = 130 marked by green line). Blue bars represent contigs of chromosomal origin while red bars represent contigs of plasmid origin. Each bar in the histogram represents all edges with the coverage in a bin of size 2

Figure 3 shows the differences in coverage between *B. cereus* chromosome and its plasmid and illustrates the utility of using *medianCoverage* to identify chromosomal contigs. For this dataset, the long contigs can be separated with perfect sensitivity and specificity into chromosome contigs (coverage varying from 120 to 156) and plasmid contigs (coverage varying from 170 to 184) based on coverage. The *medianCoverage* (vertical green line) corresponds closely to the center of the bacterial contig distribution because bacterial genomes are typically much larger than plasmid genomes.

2.2 plasmidSPAdes algorithm

plasmidSPAdes utilizes SPAdes for transforming the de Bruijn graph into the assembly graph (Bankevich *et al.*, 2012) and finds a subgraph of the assembly graph that we refer to as the *plasmid graph*. It further uses exSPAnder (Prjibelski *et al.*, 2014) for repeat resolution in the plasmid graph using paired reads and generates *plasmidic contigs*.

We define the size of a connected component in the assembly graph as the sum of the lengths of the contigs resulting from its edges. An edge (v, w) in the assembly graph is called a *dead-end edge* if either the node v has indegree zero or the node w has outdegree zero (but not both). plasmidSPAdes classifies a connected component in an assembly graph as *plasmidic* if it is composed of a single loop edge of length at least *minCirc* (default value is 1 kb) or if its size exceeds *minCompSize* (default value is 10 kb).

To transform the assembly graph into a plasmid graph, plasmidSPAdes iteratively removes long chromosomal edges and short dead-end edges from the assembly graph. Chromosomal edges are removed because they are presumed to belong to chromosomes rather than plasmids. Dead-end edges are removed because plasmids are not expected to generate dead-end edges.

The plasmidSPAdes algorithm outlined below works best when the plasmids are circular and have a copy number significantly different from 1.plasmidSPAdes (*Reads*, *k*, *maxComponentSize*)

1. construct the de Bruijn graph using k-mers from *Reads* and transform it into the assembly graph.
2. compute *medianCoverage*
3. repeat
   a. remove each long chromosomal edge in the assembly graph unless it belongs to a connected component with no dead-

end edges and a size less than *maxComponentSize* (default value is 150 kb)
   b. if 3.a removes at least one edge, remove all short dead-end edges and replace each non-branching path in the resulting graph with a single edge
4. remove all non-plasmidic connected components from the assembly graph to construct a plasmid graph.
5. launch exSPAnder to perform repeat resolution on the plasmid graph
6. output all resultant plasmidic contigs and assign them to a connected component in the plasmid graph they originated from.

The *minCirc* and *minCompSize* parameters (implicit in step 4) serve an important goal of removing relatively short chromosomal contigs that evaded the step 3 of plasmidSPAdes.

For example, error-prone reads sometimes aggregate into short paths in the assembly graph that are represented by short isolated edge with low coverage. These erroneous contigs are not removed in step 3.a because they are short and their coverage differs from the genome coverage. They are not removed in step 3.b because they are not connected to any long edges. However, they are removed in step 4. Step 6 aggregates plasmid contigs into connected components (that are expected to originate from the same plasmid) rather than outputting all plasmid contigs as a set without attempting to assign them to individual plasmids.

Ideally, plasmidSPAdes should capture all plasmids and no chromosomal fragments in the plasmid graph (with the exception of chromosomal segments that share highly similar segments with chromosomes). However, it is not entirely true since some short segments of plasmids are sometimes missing from the plasmid graph and some short chromosomal segments are often present in the plasmid graph. Also, it is difficult to distinguish tandem repeats from plasmids by analyzing the assembly graph. Indeed, tandem repeats often form *whirls* in the assembly graph (Pevzner *et al.*, 2004) resulting in cycles with high coverage by reads. To distinguish plasmid from tandem repeats, one should perform the plasmid-diagnostic tests for each putative plasmid identified by plasmidSPAdes, e.g. a test on the presence of the replication initiation (Rep) and relaxase proteins, which are used for classification of plasmids into incompatibility groups and mobility types (Shintani *et al.*, 2015)

Ideally, plasmidSPAdes should assemble each plasmid into a separate contig. In practice, it is unlikely when plasmids contain long repeats (or share long repeats with other plasmids) with lengths exceeding the insert size.

3 Results

3.1 Datasets

Accession numbers and links to the datasets and reference genomes are available in the Appendix.

3.1.1 Genomes with annotated plasmids

Table 2 describes six datasets that we used for benchmarking plasmidSPAdes (see the next section for a detailed description of all columns). These datasets are composed from paired-end Illumina reads (at least 100 bp in length) from *Bacillus cereus* ATCC-10987, *Rhodobacter sphaeroides* 2.4.1, *Providencia stuartii* ATCC 33672, *Citrobacter freundii* CFNIH1, *Burkholderia cenocepacia* DDS 22E-1 and *Corynebacterium callunae* DSM 20147. These genomes, abbreviated as *Bce*, *Rsp*, *Pst*, *Cfr*, *Bcen* and *Cca*, represent well studied bacterial species with completed reference genomes and annotated plasmids. The number of plasmids of different types in these datasets varied from 0 for *Bcen* to 5 for *Rsp*. The average copy numbers (estimated as coverage ratios) varied from 1.4 for *Bce* to 14.0 for *Cfr*. The lengths of plasmids varied from 4109 bp for *Cca* to 272 297 bp for *Cfr*. Analysis of *B. anthracis* A1144 genome is excluded from analysis in Table 2 since it has highly non-uniform coverage.

Interestingly, plasmidSPAdes assembled an additional previously unidentified short plasmid (5410 bp) in *Cfr* with high copy number (14) that is not listed in Table 2. This plasmid has a high-scoring BLAST hit to the plasmid pCAV1335-5410 in *Klebsiella oxytoca* strain CAV1335 (alignment length 4454 and percent identity 99.9).

3.1.2 Genomes with unannotated plasmids

Table 4 describes ten datasets that we used for benchmarking plasmidSPAdes for the cases when the plasmids have not been annotated yet (see the next section for a detailed description of all columns). These datasets are composed from paired-end Illumina reads (at least 100 bp in length) from *Acinetobacter* sp. UNC434CL69Tsu2S25, *Butyrivibrio* sp. INlla16, *Lachnospiraceae bacterium* NK3A20, *Lutei bacter* sp.UNC138MFCol5.1, *Prevotellaceae bacterium* HUN156, *Pseudoalteromonas* sp. ND6B, *Rhodococcus* sp. J21, *Ruminococcus flavefaciens* YAD2003, *Sphingomonas* sp. UNC305MFCol5.2 and *Thermus filiformis* ATT43280. These genomes are abbreviated as *Aci*, *But*, *Lac*, *Lut*, *Pre*, *Pse*, *Rho*, *Rum*, *Sph* and *Tfi* in Table 4. Datasets *Aci*, *But*, *Lac*, *Lut*, *Pre* and *Rum* were downloaded from JGI read archive while datasets *Pse*, *Rho*, *Sph* and *Tfi* were downloaded from NCBIs SRA.

3.2 Benchmarking plasmidSPAdes

3.2.1 Genomes with annotated plasmids

Table 2 lists the following statistics that were generated using QUAST software (Gurevich *et al.*, 2013). The first eight columns in this table refer to the annotated chromosomes and plasmids and the remaining columns refer to the predicted plasmids.

- Species name (*name*)
- Total chromosome length in kb (*chr length*)
- Number of plasmids (*plasm num*).
- Plasmid lengths in bp (*plasm len*). This field lists the length of the annotated plasmid.
- Number of distinct shared 50-mers between the chromosomes and the plasmids (*shared 50-mers*) computed using jellyfish tool Marçais and Kingsford (2011).
- Median coverage of the dataset (*med cov*)
- Median coverage of each annotated plasmid (*plasm cov*).
- Total length of contigs in each putative plasmid measured in bp (*plasm comp size*). We boldfaced the putative plasmid composed of a single circular contig (the edge representing the contig is a loop edge).
- Number of long contigs in each putative plasmid (shown in parenthesis) and the number of all contigs in each putative plasmid (*# contigs*)
- Longest contig in each putative plasmid (*max contig*)
- Coverage ratios for each putative plasmid, i.e. the coverage of each putative plasmid divided by the median coverage (*cov ratio*)
- Fraction of annotated plasmids (in percents) covered by contigs in the plasmid graph as found by QUAST (*plasm frac*). Ideally, plasmid fraction is 100%.
- Fraction of chromosome/chromosomes (in percents) covered by contigs in the plasmid graph as computed by QUAST (*chr frac*). Ideally, chromosome fraction is 0%.

Table 2. Benchmarking plasmidSPAdes on datasets with completed assemblies and annotated plasmids

name	chr length	plasm num	plasm length	# shared 50-mers	median cov	plasm cov	plasm comp size	# contigs	max contig	cov ratio	plasm frac(%)	chr frac(%)
Bce	5224283	1	208369	12953	130	186	208305	(1) 3	207886	1.4	100.0	0.0
Rsp	4131450	5	124310	1646	67	203	459442	(6) 23	123003	2.7	100.0	0.0
			114178			115					100.0	
			105281			174					100.0	
			100819			232					100.0	
			52135			271					92.4	
Pst	4285951	1	48866	0	231	699	**48874**	(1)1	48874	3.0	100.0	0.4
							16568	(0)11	7142	1.7		
Cfr	5099034	1	272297	18024	47	246	269720	(4)57	92135	5.6	99.6	0.1
							5410	(0)1	5410	14.0		
Bcen	8045250	0	0	0	136	0	N/A	0	0	N/A	N/A	0.0
Cca	2839551	2	4109	756	206	2421	**4109**	(0)1	4109	4.8	100.0	0.6
			85023			180	10352	(0)2	8160	11.8	0.0	

Correct and likely correct plasmids (plasmidic components) are shown in blue, while incorrect and likely incorrect plasmids (plasmidic components) are shown in red.

Table 3 benchmarked plasmidSPAdes, SPAdes and Recycler (Rozov *et al.*, 2015) on genomes with annotated plasmids. Although SPAdes was not intended to be used for plasmid sequencing, we classified isolated cycles in the assembly graph (of length at least 1000 nt) as putative plasmids output by SPAdes. Thus, Table 3 fully automated tool since it requires some manual analysis inside the Cytoscape program.

We classify a plasmid (or a plasmidic component) as:

- correct if its contigs were previously annotated as plasmidic (at least 90% plasmid fraction).
- incorrect if its contigs were previously annotated as chromosomal.
- likely correct if it matches an annotated plasmid from the NCBI webBLAST nucleotide database (at least 80% percent sequence identity) or carries plasmid-specific genes

Table 3. Comparing plasmidSPAdes with SPAdes and Recycler on datasets with completed assemblies and annotated plasmids

Assembler	plasmidSPAdes				SPAdes				Recycler			
name	plasm comp size	# contigs	plasm frac(%)	chr frac(%)	plasm comp size	# contigs	Plasm frac(%)	chr frac(%)	plasm comp size	# contigs	plasm frac(%)	chr frac(%)
Bce	208305	(1)3	100.0	0.0	0	0	0	0	**128266**	(1)	0	30.4
									1365548	(1)		
									96421	(1)		
Rsp	459442	(6) 23	100.0	0.0	0	0	0.0	0.0	**6131**	(0)	22.6	0.0
			100.0				0.0		**28193**	(1)	0.0	
			100.0				0.0				0.0	
			100.0				6.0				6.0	
			92.4				0.0				0.0	
Pst	**48874**	(1)1	100.0	0.4	**48874**	(1)	100.0	0.0	**48874**	(1)	100.0	0.1
	16568	(0)11							**2843**	(0)		
Cfr	269720	(4)57	99.6	0.1	**5410**	(0)	99.6	0.0	**5410**	(0)	99.6	0.0
	5410	(0)1										
Bcen	N/A	0	N/A	0.0	N/A	0	N/A	0.0	N/A	0	N/A	0.0
Cca	**4109**	(0)1	100.0	0.6	**4109**	(0)	100.0	0.0	**4109**	(0)	100.0	0.0
	10352	(0)2	0.0				0.0				0.0	

Correct and likely correct plasmids (plasmidic components) are shown in blue, while incorrect and likely incorrect plasmids (plasmidic components) are shown in red/.

should be taken with caution since all bechmarked tools were developed with somewhat different goals. We have not included PLACNET (de Toro *et al.*, 2014) in the benchmarking of various plasmid sequencing tools since it is not a truly *de novo* tool for plasmid finding, e.g. it requires RefSeq database and uses some proprietary databases which are not publicly available. Also, it is not a

- likely incorrect if it does not match an annotated plasmid from the plasmid database and does not carry plasmid-specific genes

As Table 3 illustrates, SPAdes and Recycler showed somewhat similar results but Recycler generated more false positive plasmids. plasmidSPAdes detected more plasmids than SPAdes and Recycler but often assembled them in multiple contigs rather than intact plasmids. We believe that plasmidSPAdes and Recycler may complement each other since plasmidSPAdes focuses on sequencing data from genomes of cultivated bacteria while Recycler focuses on plasmidome and metagenome data. Appendix 'Evaluating plasmidSPAdes on annotated plasmids' provides additional information about this benchmarking.

3.2.2 Genomes with unannotated plasmids

Table 4 lists the following statistics that represent the plasmidSPAdes output.

- Species name (*name*)
- Total length of contigs in each putative plasmid measured in bp (*pl comp size*)
- Number of long contigs in each putative plasmid (shown in parenthesis) and the number of all contigs in each putative plasmid (*# contigs*)
- Longest contig in each putative plasmid (*max contig*)
- Median coverage of the dataset (*med cov*)
- Coverage ratios for each putative plasmid, i.e. the coverage of each putative plasmid divided by the median coverage (*cov ratio*)
- Confirmation status (*conf*). A 'Y' indicates that the putative plasmids best blast hit to NCBI NT database was to a plasmid. A 'N' indicates that the best blast hit to NT database was identified within a chromosome (for some related species). 'N/A' indicates that there are no significant matches to NCBI NT database. We further analyzed all plasmids annotated as 'N' or 'N/A' and, whenever possible, classified them as 'N/A (plasmid)', 'N/A (phage), 'N/A (rRNA)', 'N (phage)', or 'N (mobile).'

Table 4. Benchmarking plasmidSPAdes on datasets with non-completed assemblies and lacking annotated plasmids

name	plasm comp size	# contigs	Max Contig	med cov	cov ratio	Conf
Aci	66431	(2)20	32433	249	9.9	*Y*
	42116	(1)1	42116		1.8	*N(phage)*
		(0)34	7477		6.3	*N(mobile)*
	30022	(0)1	3964		15.1	*N/A*
But	**3964**	(2)36	63050	130	2.9	*Y*
	146094	(1)1	34095		2.8	*N/A(plasmid)*
	34095	(0)1	5609		4.1	*N/A(plasmid)*
	5609	(0)1	1899		3.5	*N/A(mobile)*
	1899					
Lac	**2296**	(0) 1	2296	188	5.9	N/A (mobile)
Lut	**1445**	(0) 1	1445	227	6.6	N (mobile)
Pre	0	0	0	211	N/A	N/A
Pse	24513	(0) 15	9915	195	5.1	N
Rho	830258	(25)120	112428	127	1.9	*Y*
	138160	(3)3	52367		2.8	*N*
	2583	(0)1	2583		7.5	*Y*
Rum	**9168**	(0) 1	9168	157	3.7	N/A (plasmid)
Sph	0	0	0	171	N/A	N/A
	219971	(7)27	40017		2.3	*N*
	33615	(0)36	6474		5.3	*Y*
Tfi	**25201**	(1)1	25201	388	1.5	*N(rRNA)*
	5150	(0)1	5150		3.7	*N/A*
	1626	(0)1	1626		2.2	*N/A*

We boldfaced the putative plasmid composed of a single circular contig (the edge representing the contig is a loop edge).

In order to validate our results, we ran a *blastn* search of longest contigs from putative plasmids constructed by plasmidSPAdes against the NCBI database of nonredundant nucleotides (NT). The best BLAST hit, defined as the hit with the lowest e-value (ties are broken by the highest bit score) for each putative plasmid component from this search can be used to identify each of the components. If the best hit for a component is to a plasmid sequence in the NT database, then we classify it as a confirmed plasmid and mark by Y in the last column of Table 4. We note that this confirmation approach is unable to confirm still unknown plasmids that have little similarity with known plasmids and so is prone to false negatives.

To further investigate the putative plasmids marked as N/A, we annotated them using RAST server (Aziz *et al.*, 2008) to check if they harbor plasmid-specific proteins. If RAST identified plasmid-specific proteins, we labeled the corresponding putative plasmidic component as 'N/A (plasmid)' to emphasize that it likely represents a previously unknown plasmid. Interestingly, we found that some putative plasmids annotated as 'N/A' or 'N' likely represent phages (labeled as 'N/A (phage)' or 'N (phage).' Also, one of the putative plasmids annotated as 'N/A' harbored an rRNA gene cluster (labeled as 'N/A (rRNA)').

We also conducted further analysis of putative plasmids annotated as 'N' and found that many of them are formed by mobile elements that plasmidSPAdes failed to remove from the assembly graph (labeled as 'N (mobile)').

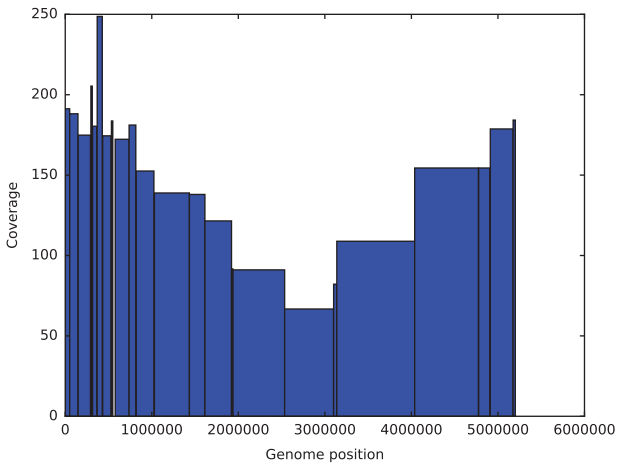

Fig. 4. Variations in coverage of long contigs in *Bacillus anthracis* (along the genome). The coverage of long contigs arranged in the order of their positions along the *B. anthracis* A1144 genome

Appendix 'Evaluating plasmidSPAdes on unannotated plasmids' provides additional information about this benchmarking.

3.2.3 Analyzing sequencing datasets with non-uniform read coverage

To investigate why some datasets (like *B. anthracis* A1144 dataset) have a rather non-uniform coverage, we ordered contigs along the *B. anthracis* A1144 genome and represented each contig as a bar of height equal to the coverage of this contigs. The resulting histogram (Fig. 4) reveals a characteristic shape with the minimum around position 2.5 Mb. The shape of the histogram in Figure 4 is similar to the shape of the *skew diagrams* that are used for identifying the origin of replication in bacterial genomes (Compeau and Pevzner, 2015). This similarity between the histogram of coverage and the skew diagram (albeit with a few outliers) suggests that the *B. anthracis* A1144 culture was sequenced in the growth phase when some cells have been replicating. The abundance of replicating cells leads to the increased coverage near the origin of replication (around position 0 Mb) and the decreased coverage near the origin of termination (around position 2.5 Mb) in *B. anthracis* A1144. See Korem *et al.* (2015) for the link between uneven coverage, rate of bacterial growth and the origin of replication. Since the coverage of the *B. anthracis* A1144 dataset is non-uniform, plasmidSPAdes (run with the default parameter *maxDeviation* = 0.3) is unable to remove a significant fraction of chromosomal edges in the assembly graph. Since the median coverage of the *B. anthracis* A1144 dataset is 130, plasmidSPAdes only removes the edges with coverage exceeding 169 thus retaining a large number of chromosomal edges (Fig. 4). While increasing the *maxDeviation* parameter to 0.4 or even higher removes nearly all chromosomal edges, it also removes some plasmidic edges (*B. anthracis* A1144 has two plasmids with coverage 135 and 183, respectively). This example illustrates additional challenges in reconstructing plasmids from sequencing datasets with highly non-uniform coverage.

4 Discussion

We described a novel plasmidSPAdes algorithm for assembling plasmids from whole genome sequencing data. Since plasmidSPAdes does not require any specialized sample preparation to isolate plasmids, it has a potential to increase the throughput of plasmid discovery. It thus complements a recently published approach mainly aimed at analyzing plasmids after plasmid isolation (Carattoli *et al.*, 2014). As Table 3 illustrates, plasmidSPAdes identifies eight plasmids in ten randomly selected SRA dataset (three of them are not similar to any previously identified plasmids). We thus expect that 1000s of new plasmids will be identified when plasmidSPAdes runs on all bacterial and achaeal SRA genome datasets.

plasmidSPAdes uses coverage to remove chromosomal contigs from the assembly while retaining the plasmid contigs. We have demonstrated that in many cases it successfully removes over 99% of the chromosomal contigs while retaining over 99% of the plasmid contigs. However, plasmidSPAdes has the following limitations:

- plasmidSPAdes may misclassify contigs from plasmids with copy numbers close to one as being chromosomal. However, in certain cases, plasmidSPAdes correctly assembles circular plasmids even when they have similar coverage to the chromosome. For example, an isolated cycle in the assembly graph is classified as a putative plasmid irrespectively of its coverage.
- The filtering procedure in plasmidSPAdes relies on estimating a median chromosomal coverage. As long as the chromosomes are significantly larger than the total length of plasmids, plasmidSPAdes reliably separates edges in the assembly graph into chromosomal and plasmidic. However, when chromosomes and plasmids are comparable in size (like in the case of megaplasmids (Zheng *et al.*, 2013), filtering based on the median coverage may fail.
- While the vast majority of plasmids are circular, some bacteria harbour difficult-to-detect linear plasmids. The performance of plasmidSPAdes deteriorates in the case of linear plasmids. For example, plasmidSPAdes assembled only one out of ten linear pasmids in *Borrelia burgdorferi* B31 (SRA accession number SRR1772332) into a single contig.
- Since coverage of some short edges in the assembly graph significantly differ from the median coverage, plasmidSPAdes may misclassify short chromosmal edges in the assembly graph as plasmidic.
- Since chromosomes are typically much longer than plasmids, if even a small portions of a chromosome is not filtered out, the small percent remaining can result in some false positive putative plasmidic contigs.

In the future, we plan to improve and extend plasmidSPAdes by adding the following features:

- When two plasmids share highly similar sequences, they may assemble into the same connected component in the assembly graph. However, If these plasmids have significantly different copy numbers, it may be possible to separate them from each other by analyzing their coverage using methods similar to the one described in Rozov *et al.* (2015). For example, plasmidSPAdes merged five annotated plasmids in *Rsp* genomes into a single component in the plasmid graph (Table 2). However, four of them feature significantly different coverages, which would permit the identification of individual plasmids from this connected component.
- Identification of plasmidic contigs can be improved by analyzing plasmid-diagnostic sequence features using methods similar to Lanza *et al.* (2014). For example, because plasmids are often self-regulated, they contain certain plasmid-specific genes to control their regulation. Similarly to PLACNET, these genes could be used as the basis for plasmid classification.
- Identification of linear plasmids can be improved by using a less aggressive approach to removing tips in the plasmid graph.
- The plasmidSPAdes algorithm can be extended to sequencing organellar genomes using the coverage arguments that were recently applied to sequencing the spruce organelles (Jackman *et al.*, 2016) by combining Illumina and Pacific Biosciences reads.

Acknowledgements

We are grateful to Anton Korobeynikov and the SPAdes development team for many thoughtful discussions that helped to improve the paper.
The sequence data for *Acinetobacter* sp. UNC434CL69Tsu2S25, *Butyrivibrio* sp. INlla16, *Lachnospiraceae bacterium* NK3A20 and *Prevotellaceae bacterium* HUN156 were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/in collaboration with the user community.
Funding: This work was supported by St. Petersburg State University, St. Petersburg, Russia [grant number 15.61.951.2015].

Conflict of Interest: none declared.

References

Anda,M. *et al.* (2015) Bacterial clade with the ribosomal RNA operon on a small plasmid rather than the chromosome. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 14343–14347.

Aziz,R.K. *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.

Bankevich,A. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.

Carattoli,A. *et al.* (2014) In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.*, **58**, 3895–3903.

Compeau,P. and Pevzner,P. (2015) *Bioinformatics Algorithms: An Active Learning Approach*. Active Learning Publishers, La Jolla.

Conlan,S. *et al.* (2014) Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing enterobacteriaceae. *Sci. Transl. Med.*, **6**, 254ra126–254ra126.

de Toro,M. *et al.* (2014) Plasmid diversity and adaptation analyzed by massive sequencing of *Escherichia coli* plasmids. *Microbiol. Spectrum*, **2**,

Gunge,N. *et al.* (1982) Transformation of *Saccharomyces cerevisiae* with linear DNA killer plasmids from *Kluyveromyces lactis*. *J. Bacteriol.*, **151**, 462–464.

Gurevich,A. *et al.* (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Jackman,S.D. *et al.* (2016) Organellar genomes of white spruce (Picea glauca): assembly and annotation. *Genome Biol. Evol.*, **8**, 29–41.

Kav,A.B. *et al.* (2012) Insights into the bovine rumen plasmidome. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 5452–5457.

Korem,T. *et al.* (2015) Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science*, **349**, 1101–1106.

Koren,S. *et al.* (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.*, **14**, R101.

Lanza,V.F. *et al.* (2014) Plasmid flux in Escherichia coli ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. *PLoS Genet.*, **10**, e1004766.

Marçais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

Pevzner,P. *et al.* (2004) De novo repeat classification and fragment assembly. *Genome Res.*, **14**, 1786–1796.

Prjibelski,A.D. *et al.* (2014) Exspander: a universal repeat resolver for DNA fragment assembly. *Bioinformatics*, **30**, i293–i301.

Rozov,R. *et al.* (2015) Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *bioRxiv*, 029926.

Shintani,M. *et al.* (2015) Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Front. Microbiol.*, **6**, 242.

Williams,L.E. *et al.* (2006) Facile recovery of individual high-molecular-weight, low-copy-number natural plasmids for genomic sequencing. *Appl. Environ. Microbiol.*, **72**, 4899–4906.

Zheng,J. *et al.* (2013) Evolution and dynamics of megaplasmids with genome sizes larger than 100 kb in the bacillus cereus group. *BMC Evolut. Biol.*, **13**, 1.