MICROBIAL GENOMICS

RESEARCH ARTICLE

Stohr *et al.*, *Microbial Genomics* 2020 ;6

DOI 10.1099/mgen.0.000400

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

Detection of extended-spectrum beta-lactamase (ESBL) genes and plasmid replicons in *Enterobacteriaceae* using PlasmidSPAdes assembly of short-read sequence data

Joep J.J.M. Stohr[1,2,*], Marjolein F.Q. Kluytmans-van den Bergh[1,3,4], Ronald Wedema[5], Alexander W. Friedrich[6], Jan A.J.W. Kluytmans[1,3,4] and John W.A. Rossen[6]

Abstract

Knowledge of the epidemiology of plasmids is essential for understanding the evolution and spread of antimicrobial resistance. PlasmidSPAdes attempts to reconstruct plasmids using short-read sequence data. Accurate detection of extended-spectrum beta-lactamase (ESBL) genes and plasmid replicon genes is a prerequisite for the use of plasmid assembly tools to investigate the role of plasmids in the spread and evolution of ESBL production in *Enterobacteriaceae*. This study evaluated the performance of PlasmidSPAdes plasmid assembly for *Enterobacteriaceae* in terms of detection of ESBL-encoding genes, plasmid replicons and chromosomal wgMLST genes, and assessed the effect of k-mer size. Short-read sequence data for 59 ESBL-producing *Enterobacteriaceae* were assembled with PlasmidSPAdes using different k-mer sizes (21, 33, 55, 77, 99 and 127). For every k-mer size, the presence of ESBL genes, plasmid replicons and a selection of chromosomal wgMLST genes in the plasmid assembly was determined. Out of 241 plasmid replicons and 66 ESBL genes detected by whole-genome assembly, 213 plasmid replicons [88%; 95% confidence interval (CI): 83.9–91.9] and 43 ESBL genes (65%; 95% CI: 53.1–75.6) were detected in the plasmid assemblies obtained by PlasmidSPAdes. For most ESBL genes (83.3%) and plasmid replicons (72.0%), detection results did not differ between the k-mer sizes used in the plasmid assembly. No optimal k-mer size could be defined for the number of ESBL genes and plasmid replicons detected. For most isolates, the number of chromosomal wgMLST genes detected in the plasmid assemblies decreased with increasing k-mer size. Based on our findings, PlasmidSPAdes is not a suitable plasmid assembly tool for short-read sequence data for ESBL-encoding plasmids of *Enterobacteriaceae*.

DATA SUMMARY

Short-read sequence data for the clinical ESBL-producing *Enterobacteriaceae* isolates included in this study are available from the publicly available European Nucleotide Archive of the European Bioinformatics Institute under study accession number: PRJEB15226. The authors confirm that all supporting data have been provided within the article and through the supplementary data files. Supplementary Material for this article can be found on figshare at http://doi.org/10.6084/m9.figshare.12423875.v1

INTRODUCTION

Plasmids are small DNA molecules that exist naturally within bacterial cells and replicate independently from the chromosome [1]. They can harbour genes involved in virulence and antibiotic resistance and are important vectors

Received 06 December 2019; Accepted 04 June 2020; Published 26 June 2020

Author affiliations: [1]Department of Infection Control, Amphia Hospital, Breda, The Netherlands; [2]Laboratory for Medical Microbiology and Immunology, Elisabeth-TweeSteden Hospital, Tilburg, The Netherlands; [3]Amphia Academy Infectious Disease Foundation, Amphia Hospital, Breda, The Netherlands; [4]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands; [5]Department of Life Science and Technology, Hanze University of Applied Sciences, Groningen, The Netherlands; [6]Department of Medical Microbiology and Infection Prevention, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.

***Correspondence:** Joep J.J.M. Stohr, joep.stohr@gmail.com

Keywords: PlasmidSPAdes; plasmids; ESBL.

Abbreviations: EBI, European Bioinformatics Institute; ENA, European Nucleotide Archive; ESBL-E, extended-spectrum beta-lactamase-producing *Enterobacteriaceae*; kb, kilobases; WGA, whole-genome assembly; wgMLST, whole-genome multilocus sequence typing; WGS, whole-genome sequencing.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Six supplementary tables and two supplementary figures are available with the online version of this article.

for horizontal gene transfer [2, 3]. Most plasmids contain specific regions, called replicons, that regulate their replication [4]. These replicons can be used to identify and classify bacterial plasmids into different incompatibility groups [2, 4]. Extended-spectrum beta-lactamases (ESBLs) are an example of an antimicrobial resistance mechanism for which the encoding genes are frequently located on plasmids in *Enterobacteriaceae* [2]. These enzymes confer resistance to the majority of beta-lactam antibiotics, limiting the options for antimicrobial therapy [5]. Knowledge of the epidemiology of plasmids is essential for understanding the evolution and spread of antimicrobial resistance. It remains challenging to study plasmids using short-read whole-genome sequencing (WGS) data [6]. Repeated sequences, often shared between plasmid and chromosomal DNA, hinder the assembly of the bacterial genome from short-read sequence data, often resulting in contigs for which the origin, either plasmid or chromosomal, cannot be resolved. In recent years, several tools to study plasmids using WGS data have been developed, including PLACNET, PlasmidFinder, cBar, HyAsP, MOB-suite, PlasmidTron, Recycler and PlasmidSPAdes [6–13]. Only Recycler and PlasmidSPAdes are fully automated computer programs that aim to *de novo* reconstruct whole plasmid sequences from short-read sequence data [9, 10]. Recent studies benchmarked the PlasmidSPAdes and Recycler algorithm for genomes of Gram-negative bacteria [14, 15]. These studies showed PlasmidSPAdes outperforming Recycler in terms of the plasmid and chromosomal fraction detected in the putative plasmids created by these algorithms [14]. However, data on the performance of PlasmidSPAdes in representative and well-defined clinical data sets for *Enterobacteriaceae* remain limited [13, 15]. PlasmidSPAdes uses the SPAdes *De Bruijn* graph assembler [10], in which k-mer size is an important parameter [16–18] that influences the size, entanglement and location of contigs in the assembly graph, and the accuracy and coverage of the assembly [13]. An algorithm optimizing the k-mer size selection for whole-genome assemblies was previously published [18]. However, no data are available on how this key setting affects the performance of plasmid assemblies using PlasmidSPAdes. The objective of this study was to evaluate the performance of PlasmidSPAdes plasmid assembly of short-read sequence data of ESBL-producing *Enterobacteriaceae* (ESBL-E) in terms of detection of ESBL-encoding genes, plasmid replicons and chromosomal whole-genome multilocus sequence typing (wgMLST) genes, and to assess the effect of k-mer size.

Impact Statement

It remains challenging to study plasmids using short-read whole-genome sequencing data. PlasmidSPAdes is a fully automated computer program that aims to *de novo* reconstruct whole plasmid sequences from short-read sequence data. However, data on the performance of PlasmidSPAdes remains limited. Accurately detecting extended-spectrum beta-lactamase (ESBL) genes and plasmid replicon genes is a prerequisite for the use of plasmid assembly tools to investigate the role of plasmids in the spread and evolution of ESBL production in *Enterobacteriaceae*. This study evaluated the performance of PlasmidSPAdes in terms of detection of ESBL-encoding genes, plasmid replicons and chromosomal whole-genome multilocus sequence typing genes in a large set of genomes of ESBL-producing *Enterobacteriaceae*. Only a limited number of the ESBL-encoding genes and plasmid replicons detected in the whole-genome assemblies were retrieved in the PlasmidSPAdes-derived plasmid assemblies. Based on our findings, PlasmidSPAdes does not seem to be a suitable plasmid assembly tool for short-read sequence data for ESBL-encoding plasmids of *Enterobacteriaceae*. Future studies should be performed to further improve automated short-read plasmid assembly tools.

METHODS

Selection of whole-genome sequencing data

Raw whole-genome sequencing reads for 59 ESBL-E isolates were selected from a well-defined WGS database of clinical ESBL-E isolates collected in the SoM study and deposited under study accession number PRJEB15226 in the publicly available European Nucleotide Archive (ENA) of the European Bioinformatics Institute (EBI) [19]. The selection of genomes was aimed at including genomes containing at least one plasmid belonging to the following plasmid families: A/C, F, L/M, I1, HI2, N. The selection included 21 *Escherichia coli*, 10 *Klebsiella pneumoniae*, 10 *Klebsiella oxytoca*, 10 *Enterobacter cloacae* complex and 8 *Citrobacter* spp. which harboured a variety of plasmid types (Table S1, available in the online version of this article). The sequence data in the database were generated on either an Illumina MiSeq sequencer (Illumina, San Diego, CA, USA) or a HiSeq 2500 sequencer (Illumina, San Diego, CA, USA).

De novo whole-genome assembly (WGA)

De novo WGA was performed using CLC Genomics Workbench 7.0.4 (Qiagen, Hilden, Germany) [19]. The k-mer size used in the assembly was based on the maximum N50 value (the largest scaffold length, N, such that 50% of the genome size is made up of scaffolds with a length of at least N) [19].

De novo plasmid assembly

De novo plasmid assembly with raw read error correction was performed using PlasmidSPAdes version 3.9 with default settings. An extensive explanation of the PlasmidSPAdes assembly algorithm can be found in a publication by Antipov *et al.* [10]. Briefly, the PlasmidSPAdes assembly algorithm excludes long edges (>10 kb) from the assembly graph when k-mer coverage of the edge deviates from the median unless this edge belongs to a connected component without 'dead-end' edges with a size smaller than the 'maxComponentSize' (default: 150 kb). Moreover, following

long edge removal, the PlasmidSPAdes algorithm excludes short dead-end edges (<10 kb) and all 'non-plasmidic' connected components from the assembly graph. Incremental plasmid assemblies were performed at different k-mer sizes. Both MiSeq- and HiSeq 2500-derived reads were assembled at k-mer sizes of 21, 33, 55 and 77. The increased read length for MiSeq-derived reads enabled additional assembly of MiSeq-derived reads at k-mer sizes of 99 and 127. The k-mer sizes used were based on the k-mer size options of the automatic k-mer size selection algorithm used by default SPAdes assemblies. For each k-mer size, the plasmid assembly and the assembled genome before chromosomal removal were retained. Quality control statistics were generated for all plasmid assemblies using Quast v.5.0.2 [20]. The median k-mer coverage of all long edges (>10 kb) in the assembly as defined by Antipov *et al.* [10] was calculated for each PlasmidSPAdes assembly using a custom script.

Detection of ESBL genes and plasmid replicons

The CLC-derived WGAs and the PlasmidSPAdes plasmid assemblies and SPAdes-assembled genomes before chromosomal removal were uploaded onto the online bioinformatics tools ResFinder v2.1 and PlasmidFinder v.1.3.1 (Center for Genomic Epidemiology, DTU, Denmark) [7, 21]. ESBL genes were called when at least 60% of the sequence length of the gene in the ResFinder database was covered with a sequence identity of at least 90%. Plasmid replicon genes were called when at least 60% of the sequence length of the replicon gene in the PlasmidFinder database was covered with a sequence identity of at least 80%. ESBL genes and plasmid replicons that were called for the plasmid assemblies were compared with those called for the corresponding WGAs. If genes were detected more than once within the plasmid assembly, only one of the called genes was used for the comparison. ESBL genes and plasmid replicons that were called for the plasmid assembly but not for the WGA were classified as additionally found ESBL genes or plasmid replicons unless the ESBL gene or plasmid replicon was from the same resistance class (e.g. $bla_{\text{CTX-M}}$, bla_{SHV} or bla_{TEM}) or plasmid incompatibility group, but with a lower ambiguity score. The ambiguity score is defined as the percentage of the sequence length that is aligned with the called gene multiplied by the sequence identity of this alignment [22].

Detection of chromosomal DNA

Chromosomal wgMLST genes were blasted against the plasmid assemblies using ABRicate version 0.8.2 (https://github.com/tseemann/abricate) to assess non-specific incorporation of chromosomal DNA in the plasmid assemblies. Chromosomal wgMLST genes were derived from recently developed species-specific wgMLST schemes that excluded plasmid sequences [19]. Similar to the settings used for creating the wgMLST schemes, chromosomal wgMLST genes were called when 100% of the sequence length was covered with an identity of at least 90% [19].

Assessment of coverage deviation

For all ESBL gene- or plasmid replicon-containing contigs in the SPAdes-derived assembly before chromosome removal, the k-mer coverage was compared with the median k-mer coverage of all long edges (>10 kb) in the assembly. Only contigs containing an ESBL gene and/or plasmid replicon that was also detected in the corresponding CLC-derived WGA were included in the analysis. A contig k-mer coverage below 0.7 or above 1.3 times the median k-mer coverage of all long edges (>10 kb) in the assembly was defined as deviating. Subsequently, all contigs containing an ESBL gene or plasmid replicon with and without a deviating coverage were checked for their presence in the plasmid assemblies. For each assembly that included a contig with a non-deviating k-mer coverage, the assembly graph before chromosomal removal was inspected using BANDAGE [23]. As a sensitivity analysis, the total number of long contigs and the ESBL gene- and plasmid replicon-containing contigs with a k-mer coverage below 0.8 or above 1.2 the median k-mer coverage was calculated to assess the effect of reducing the maxDeviation parameter from 0.3 to 0.2.

RESULTS

PlasmidSPAdes assembly characteristics before and after chromosome removal were dependent on k-mer size (Table S2). An increasing k-mer size was associated with a decrease in the median k-mer coverage and the number of contigs, but with an increase in the N50 and the maximum contig size. The size of the plasmid assemblies was largest for the smallest k-mer size, and decreased with increasing k-mer size, but did not further decrease at k-mer sizes higher than 55. For every k-mer size used, the median size of the plasmid assembly was larger than the maxComponentSize parameter of 150 kb.

For 59 isolates, 241 plasmid replicons and 66 ESBL genes were detected in the WGA. Of those, 213 plasmid replicons [88%; 95% confidence interval (CI): 83.7–91.9] and 43 ESBL genes (65%; 95% CI: 53.1–75.6) were detected in at least 1 of the plasmid assemblies (Table 1). Eight plasmid replicons and one ESBL gene that were detected in the plasmid assembly were not identified in the WGA, despite their presence in the raw reads (Table S3).

The detection of ESBL-genes and plasmid replicons was not dependent on k-mer size. Concordant results for all k-mer size plasmid assemblies were found for 173 (72%) of 241 plasmid replicons and 55 (83%) of 66 ESBL genes. For individual k-mer sizes, the percentage of plasmid replicons detected ranged from 72% (k-mer size 55) up to 81% (k-mer size 127) and the percentage of ESBL genes detected ranged from 53% (k-mer size 21) up to 61% (k-mer sizes 55 and 77), with no statistically significant differences (Table S4). Thirteen plasmid replicons and two ESBL genes were detected at one k-mer size only, although at varying k-mer sizes.

The retrieval of plasmid replicons and ESBL genes differed between bacterial species. For plasmid replicons, retrieval ranged from 67% (14/21) in *Citrobacter* spp. to 96% (44/46) in

Table 1. Detection of ESBL genes and plasmid replicons in the plasmid assembly as compared to the WGA

	Gene/replicon detected in WGA *	Gene/replicon detected in at least one of the plasmid assemblies	
	n	n	%
ESBL gene			
$bla_{CTX-M-1}$	4	4	100
$bla_{CTX-M-3}$	1	1	100
$bla_{CTX-M-9}$	16	4	25
$bla_{CTX-M-14}$	4	4	100
$bla_{CTX-M-14b}$	1	1	100
$bla_{CTX-M-15}$	17	14	82
$bla_{CTX-M-27}$	1	1	100
$bla_{CTX-M-30}$	2	2	100
$bla_{CTX-M-55}$	1	1	100
bla_{GES-1}	1	0	0
bla_{SHV-12}	13	8	62
bla_{SHV-28}	1	0	0
bla_{TEM-15}†	1	0	0
$bla_{TEM-52B}$	3	3	100
Total	66	43	65
Plasmid replicon family			
Col	61	56	92
IncA/C2	2	2	100
IncB/O/K/Z	4	4	100
IncFIA	12	9	75
IncFIB	42	38	90
IncFII	45	37	82
IncHI2	34	27	79
IncI1	10	10	100
IncI2	1	1	100
IncL/M	1	1	100
IncN	2	2	100
IncN3	2	2	100
IncP1†	1	0	0
IncQ1	2	2	100
IncQ2	1	1	100
IncR	5	5	100
IncX1	8	7	88
IncX3	4	3	75
IncX4	3	3	100
IncY	4	3	75
Total	241	213	88

*Based on WGAs constructed using CLC Genomics Workbench.
†Not detected in SPAdes WGA (all other genes were also detected in the SPAdes WGA).

K. oxytoca, and for ESBL genes from 17% (2/12) in *E. cloacae* to 91% (10/11) in *K. pneumoniae*) (Table S4). The retrieval of ESBL genes differed between plasmid families. For isolates with an IncHI2 replicon, for example, only 5 of 22 (23%) WGA-detected ESBL genes were detected in at least 1 of the plasmid assemblies as compared to 32 of 42 (76%) WGA-detected ESBL genes in isolates with an IncFIB replicon. Only the isolate containing an IncQ2 plasmid replicon had a lower retrieval rate for ESBL genes (Table S5).

In general, the percentage of chromosomal wgMLST genes detected in the plasmid assemblies decreased with increasing k-mer size (Fig. 1). In several *K. oxytoca* isolates and one *E. coli* isolate, however, the percentage of chromosomal wgMLST genes detected increased when assembling with a larger k-mer size.

A total number of 312 ESBL-gene-containing contigs and 989 plasmid-replicon-containing contigs were present in any of the SPAdes assembly graphs before removal of the bacterial chromosome. Of these, 128 (41%) of the ESBL gene-containing contigs and 192 (19%) the plasmid-containing contigs were not present in the plasmid assemblies. The main reason for exclusion from the assembly graph was non-deviating k-mer coverage in 119 (38%) of the ESBL-E-containing contigs and 113 (11%) of the plasmid replicon-containing contigs. A contig size smaller than 10 kb explained the exclusion of 8 (3%) ESBL-containing contigs and 76 (8%) plasmid-containing contigs with deviating k-mer coverage, which probably resulted in the removal of a 'dead-end edge' or 'non-plasmidic component'. Four contigs containing an ESBL gene or plasmid replicon were not included in the plasmid assembly despite being longer than 10 kb and having a coverage that deviated from the median k-mer coverage of the assembly.

Five ESBL genes and 16 plasmid replicon contigs that were included in the plasmid assembly but not located on a contig with a coverage that deviated from the median were part of a component with no dead-end edges smaller than MaxComponentSize (<150 kb) (Table 2, Fig. S2).

The percentage of contigs encoding an ESBL gene or plasmid replicon that were excluded from the assembly despite having

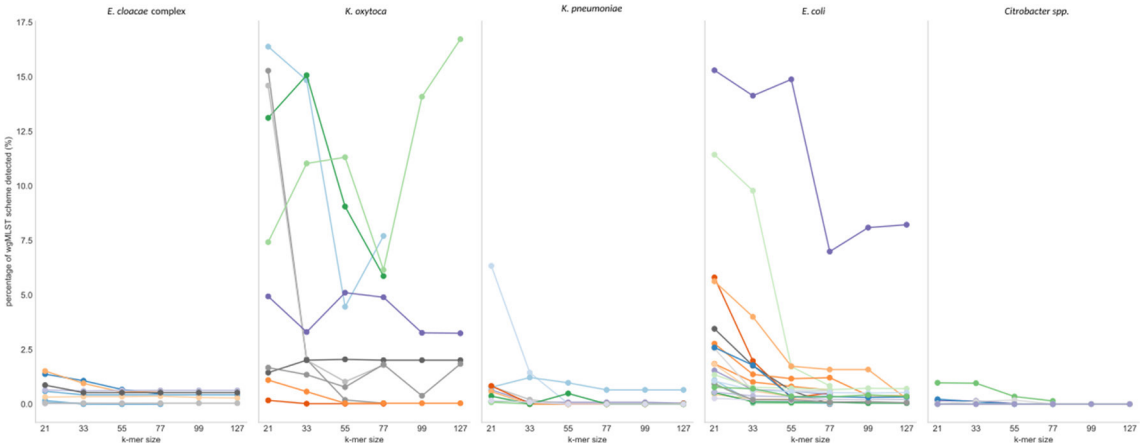

Fig. 1. The percentage of chromosomal wgMLST genes that were detected in the plasmid assemblies per species. Every line represents a plasmid assembly of an isolate at different k-mer sizes.

deviating coverage differed per k-mer size used in the plasmid assembly (Table S6).

The percentage of ESBL gene- and plasmid replicon-containing contigs with a deviating k-mer coverage increased when the maxDeviation parameter was reduced from 0.3 to 0.2 (Table 3). At the same time, the total number of long contigs (>10 kb) with deviating k-mer coverage also increased, which may reduce the specificity of the plasmidome.

DISCUSSION

PlasmidSPAdes assembly was not useful to reliably detect the presence of ESBL genes and plasmid replicons in short-read sequence data of ESBL-E, independent of k-mer size. Non-deviating k-mer coverage was the main reason for the exclusion of ESBL gene- or plasmid replicon-containing contigs from the plasmid assembly. The presence of chromosomal wgMLST genes in the plasmid assembly varied between bacterial species and was not consistently related to k-mer size.

The retrieval rates for ESBL genes and plasmid replicons are similar to the retrieval rate for plasmid DNA by plasmidSPAdes in earlier studies [12–14], in which long-read sequence data were used as a reference for plasmid DNA detection. However, these studies included *Enterobacteriaceae* sequences belonging to different undefined datasets, used only the default setting of the plasmidSPAdes algorithm, and did not evaluate which steps in the plasmidSPAdes algorithm were critical in the plasmid assembly.

Although for most isolates the presence of chromosomal wgMLST genes decreased with increasing k-mer size, in four *K. oxytoca* isolates and one *E. coli* isolate, several chromosomal contigs were not excluded from the plasmid assemblies

Table 2. The effect of deviating k-mer coverage on the presence of WGA-detected ESBL gene-containing contigs and plasmid replicon-containing contigs in the plasmid assembly

Deviating coverage*	ESBL gene detected in WGA		Plasmid replicon detected in WGA	
	Detected in plasmid assembly	Not detected in plasmid assembly	Detected in plasmid assembly	Not detected in plasmid assembly
	n (%)	*n* (%)	*n* (%)	*n* (%)
Yes	179 (95)	9 (5)	780 (91)	79 (9)
No	5 (4)	119 (96)	17 (13)	113 (87)

*Deviating coverage k-mer coverage of contig/median coverage of all long edges (>10 kb)>1.3 or k-mer coverage of contig/ median coverage of all long edges (>10 kb)<0.7.

Table 3. The effect of deviating k-mer coverage cut-off alteration on the number of long contigs (>10 kb), ESBL gene- and plasmid replicon-containing contigs with a k-mer coverage that deviated from the median k-mer coverage of all long contigs (>10 kb)

Deviating coverage cutoff	Deviating coverage*	Long contigs (>10 kb)	WGA-detected plasmid replicon-containing contigs	WGA-detected ESBL gene-containing contigs
		n (%)	*n* (%)	*n* (%)
0.3	Yes	2439 (7)	859 (87)	188 (60)
	No	30270 (93)	130 (13)	124 (40)
0.2	Yes	4785 (15)	905 (92)	213 (68)
	No	27924 (85)	84 (8)	99 (32)

*Deviating coverage, k-mer coverage of contig/median coverage of all long edges (>10 kb)>median coverage×(1+deviation cutoff) or k-mer coverage of contig/ median coverage of all long edges (>10 kb)<median coverage×(1-deviation cutoff).

at large k-mer sizes. Large chromosomal contigs with a non-deviating k-mer coverage were formed at these large k-mer sizes, resulting in the erroneous assignment of these contigs as plasmidic.

The low detection rate for both plasmid replicons and ESBL genes in *Enterobacter* and *Citrobacter* isolates when compared to the other genera investigated coincides with a low detection rate of the IncHI2a and $bla_{\text{CTX-M-9}}$ gene when compared to the other plasmid replicons and ESBL genes. This may be because, in contrast to the isolates of the other investigated genera, half or more than half of the *Citrobacter* and *Enterobacter* isolates contained a $bla_{\text{CTX-M-9}}$ and IncHI2a replicon in our selection. The co-existence of these IncHI2 plasmid replicons and $bla_{\text{CTX-M-9}}$ genes in *E. cloacae* complex and *Citrobacter* spp. was also seen in other studies [2, 24–26].

The poor retrieval rates for ESBL genes and plasmid replicons in plasmidSPAdes-derived plasmid assemblies confirms that ESBL-carrying plasmids in *Enterobacteriaceae* frequently belong to plasmid families with copy numbers that resemble the chromosome (1, 2, 27). The higher retrieval rate for plasmid replicons compared to ESBL genes may be explained by the presence of other (non-ESBL gene-containing) plasmids with a higher copy number or a higher copy number of the plasmid replicon itself, increasing the k-mer coverage of the contig on which the plasmid replicon is located.

Our findings suggest that lowering of the maxDeviation parameter may increase the retrieval rate for plasmid replicon- and ESBL gene-containing contigs, but may, on the other hand, reduce the specificity of the plasmid assembly. Since the maxDeviation parameter could not be adjusted in the command line of the PlasmidSPAdes version used in this study, the actual effect of the alteration of this parameter on the plasmid assemblies in our dataset remains unknown. A study by *Page et al.* also reported that read coverage can be a major determinant in PlasmidSPAdes performance [11]. However, this study only included one genome, and read coverage alterations were manipulated *in silico* [11]. Despite coverage information being the primary determinant for inclusion in the plasmid assemblies, several genes encoded on a contig with deviating k-mer coverage were not incorporated. Most of these genes were located on contigs smaller than 10 kb, suggesting that either the short dead-end edge removal step or the non-plasmidic component removal step might falsely exclude these contigs with deviating k-mer coverage from the plasmid assembly. Only a limited number of ESBL genes and plasmid replicons were incorporated in the plasmid assemblies when not present on a contig with deviating k-mer coverage, as defined by PlasmidSPAdes. Given that the median plasmid assembly size in our isolates was larger than 150 kb for all the k-mer sizes used, and given the possible entanglement of the various plasmids in one component, as observed in previous studies [14], increasing the maxComponentSize could include more genes through this 'escape route' without incorporating more chromosomal DNA in the plasmid assemblies.

A strength of our study is that we studied a broad spectrum of ESBL-producing *Enterobacteriaceae* isolates, including 5 different genera of various sequence types, harbouring 21 different plasmid families. All isolates belonged to a well-defined collection of ESBL-producing isolates that were collected, cultured and whole-genome sequenced using the same methods.

On the other hand, the use of plasmid replicon, ESBL gene and wgMLST gene detection instead of a complete genome as a reference to evaluate the plasmid assembly algorithm may have limited the resolution at which the algorithm could be evaluated. However, accurately detecting ESBL genes and plasmid replicon genes is a prerequisite for the use of plasmid assembly tools to investigate the role of plasmids in the spread and evolution of ESBL production in *Enterobacteriaceae*. Moreover, the ESBL genes and plasmid replicons detected in assemblies of short-read sequence data using SPAdes corresponded with the ESBL genes and plasmid replicons identified in assemblies of long-read or long- and short-read sequence data [28, 29]. Some studies have revealed that additional loci of the same ESBL gene or plasmid replicon can be detected at different locations in the bacterial genome when combining short- and long-read sequence data rather than using short-read data only [28, 30]. However, in the current study, a qualitative comparison, i.e. gene present or absent, was made between plasmidSPAdes and CLC assemblies. A quantitative comparison, i.e. the number of loci present, was not performed. Although the vast majority of ESBL genes are still believed to be located on plasmids, some studies have revealed that ESBL genes can also be found on the chromosome of *Enterobacteriaceae* [30], possibly leading to an overestimation of falsely undetected ESBL genes by plasmidSPAdes in the current study.

In conclusion, based on our data, plasmidSPAdes is not a suitable plasmid assembly tool for short-read sequence data for ESBL-encoding plasmids of *Enterobacteriaceae*.

Funding information
The SoM study was financially supported by The Netherlands Organisation for Health Research and Development (ZonMw, project number 205100010).

Acknowledgement
We are grateful to the members of the SoM Study Group for their contribution to the collection, culturing and whole-genome sequencing of ESBL-E isolates.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Data Bibliography
Short-read sequence data included in this study are available from the publicly available European Nucleotide Archive of the European Bioinformatics Institute under study accession number: PRJEB15226. Accession numbers and plasmid replicon content (based on CLC genomics workbench WGA) of the 59 isolates used for the current paper are listed in Table S1.

References

1. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile genetic elements associated with antimicrobial resistance. *Clin Microbiol Rev* 2018;31:1–61.
2. Carattoli A. Resistance plasmid families in *Enterobacteriaceae*. *Antimicrob Agents Chemother* 2009;53:2227–2238.
3. Carattoli A. Plasmids and the spread of resistance. *Int J Med Microbiol* 2013;303:298–304.
4. Couturier M, Bex F, Bergquist PL, Maas WK. Identification and classification of bacterial plasmids. *Microbiol Rev* 1988;52:375–395.
5. Rodriguez-Baño J, Gutiérrez-Gutiérrez B, Machuca I, Pascual A. Treatment of infections caused by extended-spectrum-beta-. *Clin Microbiol Rev* 2018;31:1–42.
6. Lanza VF, de Toro M, Garcillán-Barcia MP, Mora A, Blanco J *et al*. 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* 2014;10:e1004766.
7. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O *et al*. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 2014;58:3895–3903.
8. Zhou F, Xu Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* 2010;26:2051–2052.
9. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E *et al*. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* 2017;33:475–482.
10. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A *et al*. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics* 2016;32:btw493–3387.
11. Page AJ, Wailan A, Shao Y, Judge K, Dougan G *et al*. PlasmidTron: assembling the cause of phenotypes and genotypes from NGS data. *Microb Genom* 2018;4:1–6.
12. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* 2018;4
13. Müller R, Chauve C. HyAsP, a greedy tool for plasmids identification. *Bioinformatics* 2019;35:4436–4439.
14. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom* 2017;3:e000128.
15. Laczny CC, Galata V, Plum A, Posch AE, Keller A. Assessing the heterogeneity of in silico plasmid predictions based on whole-genome-sequenced clinical isolates. *Brief Bioinform* 2019;20:857–865.
16. Zerbino DR, Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* 2008;18:821–829.
17. Nurk S, Bankevich A, Antipov D *et al*. Assembling genomes and Mini-metagenomes from highly chimeric reads. *In* 2013:158–170.
18. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 2014;30:31–37.
19. Kluytmans-van den Bergh MFQ, Rossen JWA, Bruijning-Verhagen PCJ, Bonten MJM, Friedrich AW *et al*. Whole-Genome multilocus sequence typing of extended-spectrum-beta-lactamase-producing Enterobacteriaceae. *J Clin Microbiol* 2016;54:2919–2927.
20. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 2018;34:i142–i150.
21. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S *et al*. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 2012;67:2640–2644.
22. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
23. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 2015;31:3350–3352.
24. García A, Navarro F, Miró E, Villa L, Mirelis B *et al*. Acquisition and diffusion of bla CTX-M-9 gene by R478-IncHI2 derivative plasmids. *FEMS Microbiol Lett* 2007;271:71–77.
25. Miró E, Segura C, Navarro F, Sorlí L, Coll P *et al*. Spread of plasmids containing the bla(VIM-1) and bla(CTX-M) genes and the qnr determinant in Enterobacter cloacae, Klebsiella pneumoniae and Klebsiella oxytoca isolates. *J Antimicrob Chemother* 2010;65:661–665.
26. Nilsen E, Haldorsen BC, Sundsfjord A, Simonsen GS, Ingebretsen A *et al*. Large IncHI2-plasmids encode extended-spectrum β-lactamases (ESBLs) in Enterobacter spp. bloodstream isolates, and support ESBL-transfer to *Escherichia coli*. *Clin Microbiol Infect* 2013;19:E516–E518.
27. Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B *et al*. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *J Antimicrob Chemother* 2018;73:1121–1137.
28. Stohr JJJM, Verweij JJ, Buiting AGM, Rossen JWA, Kluytmans JAJW. Within-patient plasmid dynamics in Klebsiella pneumoniae during an outbreak of a carbapenemase-producing Klebsiella pneumoniae. *PLoS One* 2020;15:e0233313.
29. Lemon JK, Khil PP, Frank KM, Dekker JP. Rapid nanopore sequencing of plasmids and resistance gene detection in clinical isolates. *J Clin Microbiol* 2017;55:3530–3543.
30. Decano AG, Ludden C, Feltwell T, Judge K, Parkhill J *et al*. Complete Assembly of *Escherichia coli* Sequence Type 131 genomes using long reads demonstrates antibiotic resistance gene variation within diverse plasmid and chromosomal contexts. *mSphere* 2019;4:1–12.