MICROBIAL GENOMICS

Methods paper template

1   # PlasmidTron: assembling the cause of phenotypes
2   # from NGS data

3   Andrew J. Page*, Alexander Wailan, Yan Shao, Kim Judge, Gordon Dougan, Elizabeth J. Klemm,
4   Nicholas R. Thomson, Jacqueline A. Keane

5

6   Infection Genomics, Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton,
7   Cambridge, UK.

8   *Corresponding author: ap13@sanger.ac.uk

9

10 ## ABSTRACT

11 When defining bacterial populations through whole genome sequencing (WGS) the samples often
12 have detailed associated metadata that relate to disease severity, antimicrobial resistance, or even
13 rare biochemical traits. When comparing these bacterial populations, it is apparent that some of
14 these phenotypes do not follow the phylogeny of the host i.e. they are genetically unlinked to the
15 evolutionary history of the host bacterium. One possible explanation for this phenomenon is that
16 the genes are moving independently between hosts and are likely associated with mobile genetic
17 elements (MGE). However, identifying the element that is associated with these traits can be
18 complex if the starting point is short read WGS data. With the increased use of next generation WGS
19 in routine diagnostics, surveillance and epidemiology a vast amount of short read data is available
20 and these types of associations are relatively unexplored. One way to address this would be to
21 perform assembly *de novo* of the whole genome read data, including its MGEs. However, MGEs are
22 often full of repeats and can lead to fragmented consensus sequences. Deciding which sequence is
23 part of the chromosome, and which is part of a MGE can be ambiguous. We present *PlasmidTron*,
24 which utilises the phenotypic data normally available in bacterial population studies, such as
25 antibiograms, virulence factors, or geographic information, to identify sequences that are likely to
26 represent MGEs linked to the phenotype. Given a set of reads, categorised into cases (showing the
27 phenotype) and controls (phylogenetically related but phenotypically negative), *PlasmidTron* can be
28 used to assemble *de novo* reads from each sample linked by a phenotype. A *k*-mer based analysis is
29 performed to identify reads associated with a phylogenetically unlinked phenotype. These reads are
30 then assembled *de novo* to produce contigs. By utilising *k*-mers and only assembling a fraction of the
31 raw reads, the method is fast and scalable to large datasets. This approach has been tested on
32 plasmids, because of their contribution to important pathogen associated traits, such as AMR, hence
33 the name, but there is no reason why this approach cannot be utilized for any MGE that can move

34 independently through a bacterial population. *PlasmidTron* is written in Python 3 and available
35 under the open source licence GNU GPL3 from https://goo.gl/ot6rT5 .

36

37

DATA SUMMARY

39

40 1. Source code for *PlasmidTron* is available from Github under the open source licence GNU
41   GPL 3; (url – https://goo.gl/ot6rT5)
42

43 2. Simulated raw reads files have been deposited in Figshare; (url –
44   https://doi.org/10.6084/m9.figshare.5406355.v1)
45

46 3. *Salmonella enterica* serovar Weltevreden strain VNS10259 is available from GenBank;
47   accession number GCA_001409135.
48

49 4. *Salmonella enterica* serovar Typhi strain BL60006 is available from GenBank; accession
50   number GCA_900185485.
51

52 5. Accession numbers for all of the Illumina datasets used in this paper are listed in the
53   supplementary tables.
54

55 **I/We confirm all supporting data, code and protocols have been provided within the article or**
56 **through supplementary data files. ⊠**

57

IMPACT STATEMENT

59

60 PlasmidTron utilises the phenotypic data normally available in bacterial population studies, such as
61 antibiograms, virulence factors, or geographic information, to identify sequences that are likely to
62 represent MGEs linked to the phenotype.

63

INTRODUCTION

65   When defining bacterial populations through whole genome sequencing (WGS) the samples often
66   have detailed associated metadata that relate to disease severity, antimicrobial resistance, or even
67   rare biochemical traits. When comparing these bacterial populations, it is apparent that some of
68   these phenotypes do not follow the phylogeny of the host i.e. they are genetically unlinked to the
69   evolutionary history of the host bacterium. One possible explanation for this phenomenon is that
70   the genes are moving independently between hosts and are likely associated with mobile genetic
71   elements (MGE). However, identifying the element that is associated with these traits can be
72   complex if the starting point is short read WGS data. With the increased use of next generation WGS
73   in routine diagnostics, surveillance and epidemiology a vast amount of short read data is available
74   and these types of associations are relatively unexplored. One way to address this would be to
75   perform assembly *de novo* of the whole genome read data, including its MGEs. However, MGEs are
76   often full of repeats and can lead to fragmented consensus sequences. Deciding which sequence is
77   part of the chromosome, and which is part of a MGE can be ambiguous (1).

78   A number of recent methods have been developed to address the problem of assembling some of
79   these MGEs, from NGS data (1). *plasmidSPAdes* (2) detects plasmids by analysing the coverage of
80   assembled contigs to separate out chromosomes from plasmid like sequences. By filtering the
81   dataset, a higher quality assembly is possible. However, if the copy number of the plasmids are
82   similar to the chromosome, it is not possible to separate out plasmids. *Unicycler* (3) is a hybrid
83   assembler which can combine short and long read data to produce fully circularised chromosomes
84   and plasmids. It essentially fixes many of the deficiencies of *SPAdes* (4) and fine tunes it for
85   assembling bacteria. *Recycler* (5) takes an assembly graph and aligned reads to search for cycles in
86   the graph which may correspond to plasmids. The method is only partially implemented with
87   substantial work required on the researcher's part to generate input files in the correct formats. It is
88   shown to work well on small simple plasmids, however it does not scale to larger more complex
89   plasmids. All of these software applications utilise *SPAdes* within their methods, work on a single
90   sample at a time, and require no *a priori* knowledge about the samples themselves.

91   We present *PlasmidTron*, which utilises the phenotypic data normally available in bacterial
92   population studies, such as antibiograms, virulence factors, or geographic information, to identify
93   sequences that are likely to represent MGEs linked to the phenotype. Given a set of reads,
94   categorised into cases (showing the phenotype) and controls (phylogenetically related but
95   phenotypically negative), *PlasmidTron* can be used to assemble *de novo* reads from each sample
96   linked by a phenotype. A *k*-mer based analysis is performed to identify reads associated with a
97   phylogenetically unlinked phenotype. These reads are then assembled *de novo* to produce contigs.
98   By utilising *k*-mers and only assembling a fraction of the raw reads, the method is fast and scalable
99   to large datasets. This approach has been tested on plasmids, because of their contribution to
100   important pathogen associated traits, such as AMR, hence the name, but there is no reason why this
101   approach cannot be utilized for any MGE that can move independently through a bacterial
102   population. The method is tested on simulated and real datasets, compared to other methods, and
103   the results are validated with long read sequencing. *PlasmidTron* is a command-line tool, is written
104   in Python 3 and is available under the open source licence GNU GPL3 from https://goo.gl/ot6rT5 .

105

106

METHOD

108 *PlasmidTron* takes two spreadsheets as input, one containing paired ended reads in FASTQ format
109 for samples displaying the phenotype (cases), the other containing FASTA or FASTQ files for samples
110 not displaying the phenotype (controls). The full method is shown in Figure 1. A *k*-mer analysis of
111 each of the samples is performed using KMC (syntax versions v2.3.0 or v3.0.0) (6,7) to produce
112 databases of *k*-mer counts. *k*-mers occurring less than 5 times are excluded by default since
113 assembly is more error prone below this level of coverage. A union is taken of the cases *k*-mer
114 databases to produce a new database of all *k*-mers ever seen in any of the trait samples, and
115 similarly for the controls. The two sets are then subtracted from each other, leaving only *k*-mers
116 uniquely present in the cases dataset. The raw reads, plus their mates, which match these unique *k*-
117 mers are extracted from each sample where each read must be covered by a defined percentage of
118 *k*-mers. Each set of reads is assembled *de novo* with SPAdes. The assembly contigs are filtered to
119 remove small contigs (default 300 bases), and low coverage contigs (below 10X). This is because a
120 single erroneous *k*-mer can draw in reads on either side equating to approximately the fragment size
121 of the library. The resulting sequences can be fragmented so a second scaffolding step is
122 undertaken. A k-mer database is generated for each assembly and the raw reads, plus their mates,
123 are extracted for a second assembly with *SPAdes*. This allows for gaps of up to twice the fragment
124 size to be closed. A final filtering step of the assembled sequences is performed, as previously
125 described. An assembly in FASTA format is created for each of the trait samples, along with a plot of
126 the shared *k*-mers in each sample, indicating the level of identity between samples. Parallelisation
127 support is provided by GNU parallel (8).

128

RESULTS

130 To evaluate the effectiveness of *PlasmidTron* three datasets were used including: 1) simulated reads
131 to show the impact of copy number variation in identifying plasmids, 2) the effectiveness of different
132 methods in recalling plasmid type sequences on real world data, and 3) identification of a novel AMR
133 plasmid with subsequent validation using long read sequencing. All experiments were performed
134 using the Wellcome Trust Sanger Institute compute infrastructure, running Ubuntu 12.04.

135

IMPACT OF COPY NUMBER VARIATION

137 Simulated reads were generated to show the impact of copy number variation compared to other
138 methods. A trivial set of simulated perfect reads was generated. A reference genome, which was
139 sequenced using the PacBio RSII for *Salmonella enterica* serovar Weltevreden *(S.* Weltevreden)
140 (accession number GCA_001409135), was shredded using FASTAQ (v3.15.0)
141 (https://github.com/sanger-pathogens/fastaq) to generate perfect paired-ended reads with a read
142 length of 125 bases and a mean fragment size of 400 bases. The reference contains a single
143 chromosome (5,062,936 bases) and single plasmid (98,756 bases), where the chromosome depth of
144 coverage was fixed at 30X, and the plasmid depth of coverage was varied from 1 to 60X in steps of 2.
145 The break point for the plasmid was varied, in steps of 500 bases, to simulate a circular genome.

146

147   The results of *PlasmidTron* (v0.3.5) were compared to 4 other methods, *recycler* (v0.6), *Unicycler*
148   (v0.4.0), *SPAdes* (v3.10.0), and *plasmidSPAdes* (v3.10.0). *SPAdes* (v3.10.0) was used as the assembler
149   for each of these methods. *recycler* required pre-processing steps using *bwa* (v0.7.12) (9) and
150   *samtools* (v0.1.19) (10). *SPAdes* and *Unicycler* are not dedicated plasmid assemblies and are
151   agnostic to the underlying structures being sequenced, however they provide a good baseline for
152   what is possible, though the final plasmid sequences are contained in a large collection of
153   chromosome sequences. *plasmidSPAdes* and *PlasmidTron* are dedicated plasmid assemblers, and
154   *recycler* is post assembly plasmid analysis tool, with each employing a fundamentally different
155   analysis strategy.

156

157   Each resulting assembly was measured based on the percentage of plasmid assembled, how
158   fragmented the plasmid was, and the proportion of non-plasmid bases to plasmid bases (signal to
159   noise ratio). The assemblies are blasted (v.2.6.0) (11) against the expected plasmid sequence, with
160   an e-value of 0.0001. Blast hits of less than 200 bases long or less than 90% identity were excluded.
161   *Recycler* identified no plasmids on the real or simulated data, which appears to be due to the large
162   complex size of the plasmid.

163

164   Figure 2 shows that as the copy number of the plasmid in the input reads changes, the percentage of
165   the plasmid recovered changes. *plasmidSPAdes* only begins to start identifying plasmid sequences at
166   40X, recovering the full plasmid sequence. Below this level the plasmid copy number is too similar to
167   the chromosome coverage so the algorithm filters it out. The *SPAdes* and *Unicycler* assemblers
168   identify all of the plasmid sequence with less than 10X coverage, however the plasmid sequences
169   are fragmented and makes up only ~1.9% of the final assembly as show in Figure 3. *PlasmidTron*
170   requires slightly more coverage (16X) to generate an assembly which covers the full plasmid
171   sequence. At 16X more than 90% of the resulting assembly contains plasmid sequences, increasing
172   to 100% at 40X.

173

174   **RECOVERY OF TYPING SEQUENCE**

175   A real dataset of 114 isolates of *S.* Weltevreden, was sequencing using Illumina as described in (12).
176   The samples are clonal, with most sharing a similar plasmid, although the payload of the plasmid
177   itself varies greatly. To get a baseline for what plasmids are present in the input dataset, all of the
178   samples were compared to the PlasmidFinder (13) database (retrieved 2017-07-25) using Ariba
179   (v2.10.0) (14), providing the Incompatibility group. PlasmidFinder identifed one plasmid group,
180   IncFII$_S$, as present in 89.5% of samples. *plasmidSPAdes, Unicycler, SPAdes, Recycler* and *PlasmidTron*
181   were provided with the dataset and the results were searched for the IncFII$_S$ sequence using blastn,
182   with full details listed in Supplementary Table 1. In 4 cases PlasmidFinder failed to identify the
183   sequence, when it was found by 2 or more other applications. *SPAdes* and *Unicycler* identify the
184   sequence in 88.6% and 87.7% of samples. *plasmidSPAdes* identifies the plasmid sequence in just

185 8.8% of samples. Recycler failed to identify the plasmid sequence in any sample. *PlasmidTron*
186 identified the plasmid sequence in 87.7% of cases where the chromosome sequence of *S.*
187 Weltevreden strain VNS10259 was used as the control, giving identical results to *Unicycler*. The
188 benefit though over *Unicycler* is that the majority of the assembled sequences correspond only to
189 the plasmid.

190

191 **OUTBREAK AMR**

192 *PlasmidTron* was used to analyse an outbreak of 87 *Salmonella enterica* serovar Typhi (*S.* Typhi)
193 samples with a resistance profile which had not been previously observed in the haplotype (H58,
194 4.3.1)(15). Further analysis using PlasmidFinder, as previously described, indicated that the antibiotic
195 resistance may reside on an IncY plasmid, a plasmid type which had not been associated before with
196 this haplotype. The chromosomes of 6 complete reference genomes for *S.* Typhi were used as the
197 controls (accessions GCA_000195995, GCA_000007545, GCA_001157245, GCA_000245535,
198 GCA_001302605, GCA_000385905) for *PlasmidTron*, and 87 Illumina sequenced outbreak samples
199 were used as cases (Supplementary Table 2). For each outbreak sample, *PlasmidTron* identified
200 similar sequences, split over 4-5 contigs. One contig carried the IncY sequence and a second carried
201 AMR genes. Subsequent resequencing of 1 sample (ERS1670682) using long read technology (Oxford
202 Nanopore MinION), revealed that these 4 sequences comprised a single plasmid (accession number
203 GCA_900185485.1), which was identical in all of the outbreak strains. The sequences generated by
204 *PlasmidTron* recovered an average of 96% of the plasmid sequence. The fragmentation (mean 4.6) of
205 the plasmid in the Illumina sequenced samples was due to repeats which could not be resolved with
206 short read sequencing. Overall 65% of the sequences in the resulting assemblies were part of the
207 plasmid sequence, with the remainder resulting from a phage recombination in the main
208 chromosome. This indicates the power of *PlasmidTron* to rapidly, accurately and cost effectively
209 extract sequences of clinical importance from short reads alone.

210

211 _____

212 **CONCLUSION**

213 We can utilise the wealth of phenotypic data usually generated for bacterial population studies, be it
214 routine diagnostics, surveillance or outbreak investigation, to reconstruct plasmids responsible for a
215 particular phenotype. Rather than just identifying that an AMR or virulence gene exists in a sample,
216 *PlasmidTron* can reconstruct all of the sequences of the plasmid it is carried on, providing more
217 insight into the underlying mechanisms. We demonstrated with simulated and real sequences that
218 *PlasmidTron* more accurately reconstructs large plasmids compared to other methods. We present
219 the results of a real outbreak of *S.* Typhi where *PlasmidTron* was used to identify the plasmid
220 sequence carrying a novel AMR resistance profile, not previously described in *S.* Typhi H58/4.3.1,
221 and validated the results using long read sequencing. Whilst plasmid assembly remains difficult with
222 short reads, *PlasmidTron* allows for phenotypic data to be utilised to greatly reduce the complexity
223 of the challenge.

224

AUTHOR STATEMENTS

229

230

ABBREVIATIONS

232 MGE: mobile genetic element

233 WGS: whole genome sequencing

234 AMR: anti-microbial resistance

235

236

REFERENCES

238

239 1. Arredondo-Alonso S, Schaik W van, Willems RJ, Schurch AC. On the (im)possibility of
240   reconstructing plasmids from whole-genome short-read sequencing data. Microb Genomics.
241   2017 Aug 18;

242 2. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. plasmidSPAdes: assembling
243   plasmids from whole genome sequencing data. Bioinformatics. 2016 Nov 15;32(22):3380–7.

244 3. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from
245   short and long sequencing reads. PLOS Comput Biol. 2017 Jun 8;13(6):e1005595.

246 4. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New
247   Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Vol. 19. 2012. 455–
248   477 p.

249 5. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E, Mizrahi I, et al. Recycler: an algorithm
250   for detecting plasmids from de novo assembly graphs. Bioinformatics. 2017 Feb 15;33(4):475–
251   82.

252 6. Deorowicz S, Kokot M, Grabowski S, Debudaj-Grabysz A. KMC 2: fast and resource-frugal k-mer
253   counting. Bioinformatics. 2015 May 15;31(10):1569–76.

254   7. Kokot M, Długosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics.
255   Bioinformatics. 2017 Sep 1;33(17):2759–61.

256   8. Tange O. GNU Parallel - The Command-Line Power Tool. Login USENIX Mag. 2011 Feb;36(1):42–
257   47.

258   9. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
259   ArXiv13033997 Q-Bio [Internet]. 2013 Mar 16 [cited 2017 Jul 26]; Available from:
260   http://arxiv.org/abs/1303.3997

261   10. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map
262   format and SAMtools. Bioinformatics. 2009;25(16):2078–2079.

263   11. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture
264   and applications. BMC Bioinformatics. 2009;10:421.

265   12. Makendi C, Page AJ, Wren BW, Le Thi Phuong T, Clare S, Hale C, et al. A Phylogenetic and
266   Phenotypic Analysis of Salmonella enterica Serovar Weltevreden, an Emerging Agent of
267   Diarrheal Disease in Tropical Regions. PLoS Negl Trop Dis. 2016 Feb;10(2):e0004446.

268   13. Carattoli A, Zankari E, Garcia-Fernandez A, Larsen MV, Lund O, Villa L, et al. In Silico detection
269   and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob
270   Agents Chemother. 2014;58(7):3895–3903.

271   14. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, et al. ARIBA: rapid
272   antimicrobial resistance genotyping directly from sequencing reads. Microb Genomics. 2017 Sep
273   4;

274   15. Wong VK, Baker S, Connor TR, Pickard D, Page AJ, Dave J, et al. An extended genotyping
275   framework for Salmonella enterica serovar Typhi, the cause of human typhoid. Nat Commun.
276   2016 Oct;7:12827.

277

278

279   # DATA BIBLIOGRAPHY

280   Parkhill J et al, *Salmonella enterica subsp. enterica* serovar Typhi str. CT18, 2001, EMBL accession
281   number GCA_000195995.

282

283   Deng W et al, *Salmonella enterica subsp. enterica* serovar Typhi str. Ty2, 2006, EMBL accession
284   number GCA_000007545.

285

286   Ong SY, *Salmonella enterica subsp. enterica* serovar Typhi str. P-stx-12, 2012, EMBL accession
287   number GCA_000245535.

288

289   Xu D et al, *Salmonella enterica subsp. enterica* serovar Typhi str. Ty21a, 2013, EMBL accession
290   number GCA_000385905.

291

292   Muhamad Harish S et al, *Salmonella enterica subsp. enterica* serovar Typhi str. PM016/13, 2015,
293   EMBL accession number GCA_001302605.

294

295   Page AJ, *Salmonella enterica* serovar Weltevreden str. VNS10259, 2016, EMBL accession number
296   GCA_001409135.

297

298   Page AJ, *Salmonella enterica* serovar Typhi str. BL60006, 2017, EMBL accession number
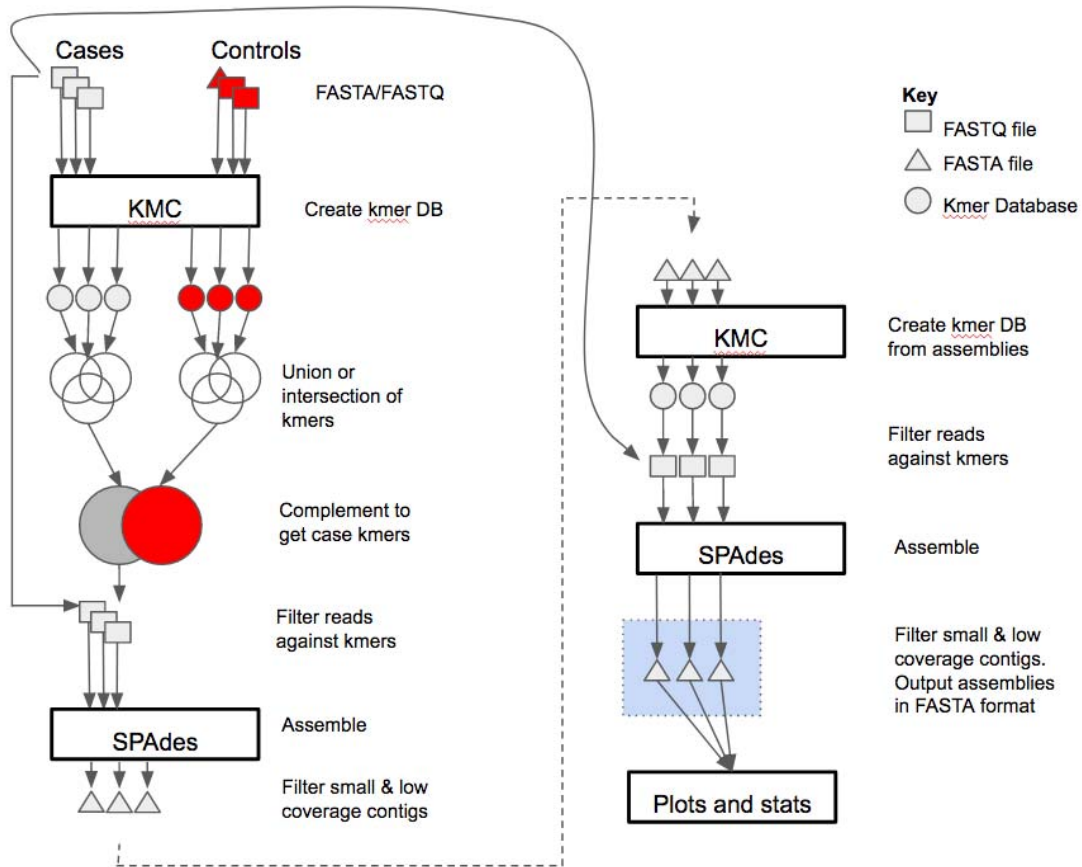299   GCA_900185485.

300

301   Page AJ, *Salmonella enterica* serovar Typhi str. ERL12148, 2017, EMBL accession number
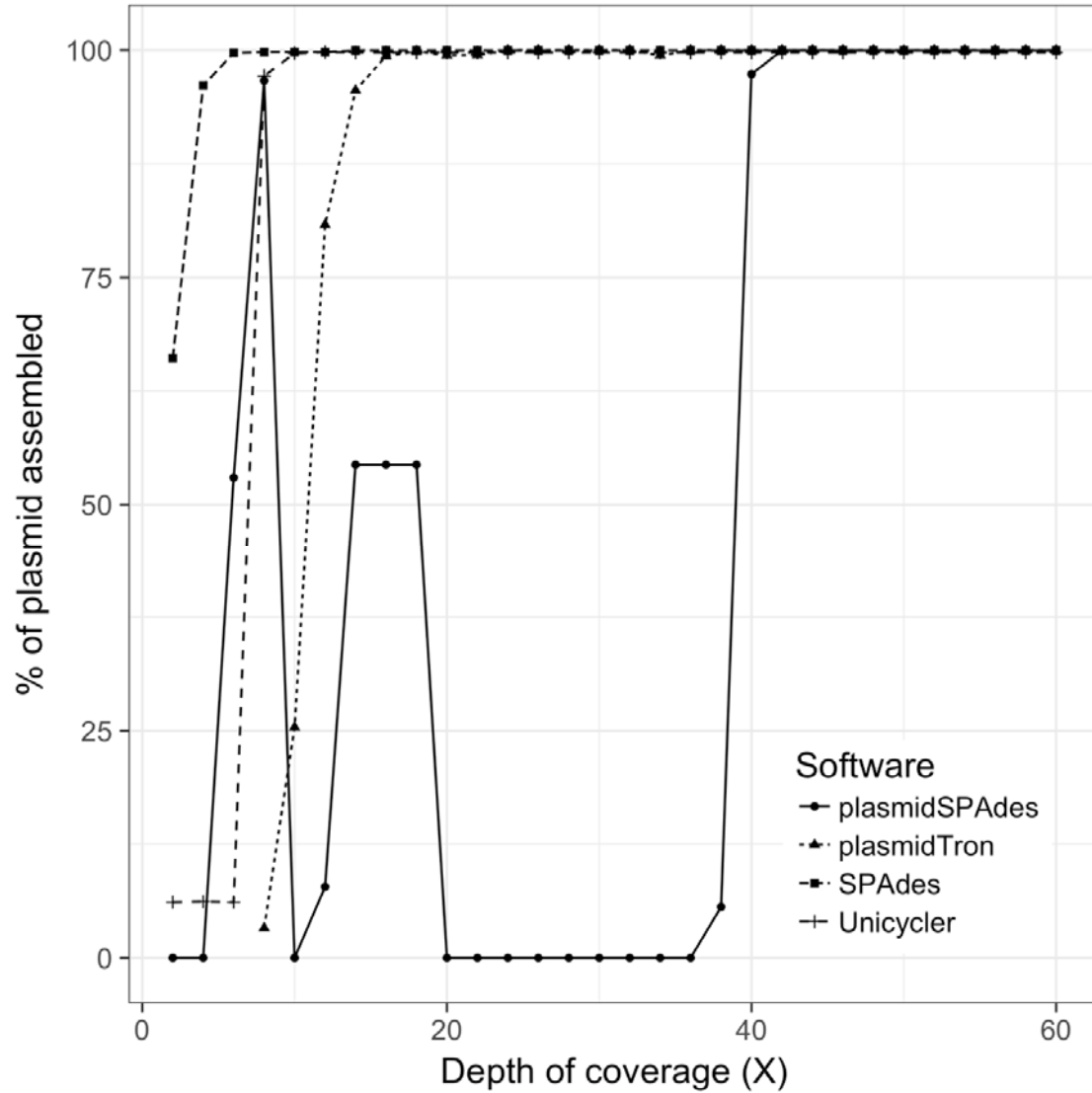302   GCA_001157245.

303

304   _____

305   **FIGURES AND TABLES**

306

Figure 1: The PlasmidTron algorithm. FASTQ files are denoted as squares, FASTA files as triangles and *k*-mer databases as circles.
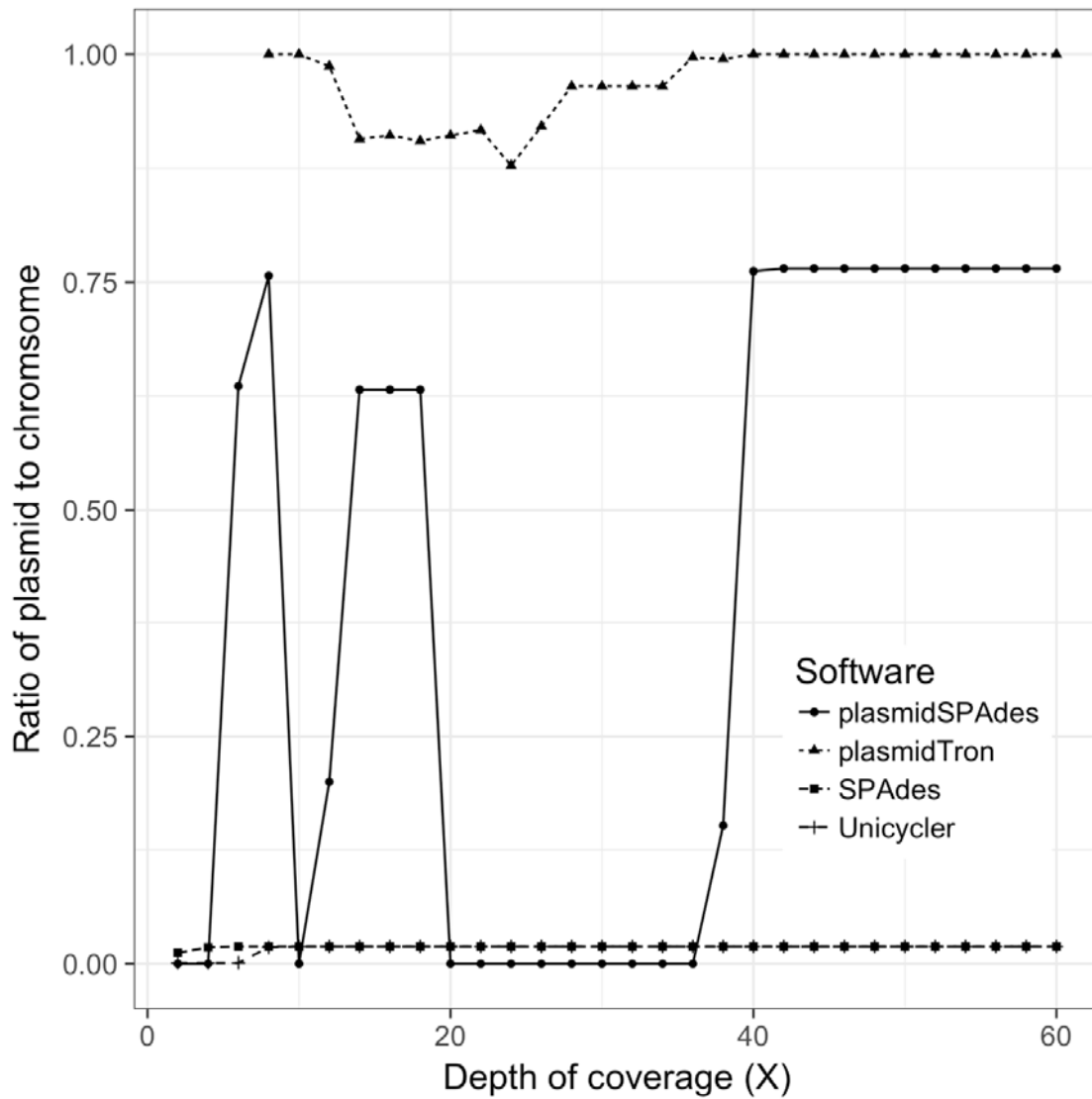
307
308

309

310

311 Figure 2: The percentage of the plasmid sequence which was assembled with different
312 software applications as the depth of coverage of a plasmid increases in the raw data.

313

314

315 Figure 3: The ratio of the plasmid sequence to the chromosome sequence in the final
316 assembly produced by each software application as the depth of coverage of the plasmid
317 increases in the raw reads. This is akin to the signal to noise ratio.

318