DRAFT

Vol. 00 no. 00 2016
Pages 1–2

plasmidSPAdes: Assembling Plasmids from Whole Genome Sequencing Data

Extended abstract

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, USA
[3]Bioinformatics and Systems Biology Program, Massachusetts Institute of Technology, USA

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/
Keywords: genome assembly, plasmid detection, plasmid assembly.
Contact: d.antipov@spbu.ru

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 *plasmid sequencing* (Williams *et al.*, 2006; Kav *et al.*, 2012). 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 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. 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 plasmidSPAdes tool aimed at sequencing of 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 that represent the lion's share of bacterial sequencing projects.

*to whom correspondence should be addressed.
Dmitry Antipov, email: d.antipov@spbu.ru

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 open. 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 *C. freundii CFNIH1* genome with well-annotated plasmids and identifying a new previously overlooked plasmid in this genome as well as discovering 7 new plasmids in ten randomly chosen bacterail datasets in the Short Reads Archive. We further provide the first analysis of accuracy of a plasmid sequencing tool across a wide variety of diverse bacterial genomes.

ACKNOWLEDGMENTS

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 study was funded by the Russian Science Foundation [grant 14-50-00069]

REFERENCES

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

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, **19**(5), 455–477.

Carattoli, A., Zankari, E., García-Fernández, A., Larsen, M. V., Lund, O., Villa, L., Aarestrup, F. M., and Hasman, H. (2014). In silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrobial agents and chemotherapy*, **58**(7), 3895–3903.

Conlan, S., Thomas, P. J., Deming, C., Park, M., Lau, A. F., Dekker, J. P., Snitkin, E. S., Clark, T. A., Luong, K., Song, Y., *et al.* (2014). Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing enterobacteriaceae. *Science Translational Medicine*, **6**(254), 254ra126–254ra126.

de Toro, M., Garcilláon-Barcia, M. P., and De La Cruz, F. (2014). Plasmid diversity and adaptation analyzed by massive sequencing of escherichia coli plasmids. *Microbiology Spectrum*, **2**(6).

Gunge, N., Murata, K., and Sakaguchi, K. (1982). Transformation of saccharomyces cerevisiae with linear dna killer plasmids from kluyveromyces lactis. *Journal of bacteriology*, **151**(1), 462–464.

Kav, A. B., Sasson, G., Jami, E., Doron-Faigenboim, A., Benhar, I., and Mizrahi, I. (2012). Insights into the bovine rumen plasmidome. *Proceedings of the National Academy of Sciences*, **109**(14), 5452–5457.

Lanza, V. F., de Toro, M., Garcillán-Barcia, M. P., Mora, A., Blanco, J., Coque, T. M., and de la Cruz, F. (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 Genetics*, **10**(12), e1004766.

Rozov, R., Brown, A. K., Bogumil, D., Halperin, E., Mizrahi, I., and Shamir, R. (2015). Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *bioRxiv*, page 029926.

Williams, L. E., Detter, C., Barry, K., Lapidus, A., and Summers, A. O. (2006). Facile recovery of individual high-molecular-weight, low-copy-number natural plasmids for genomic sequencing. *Applied and environmental microbiology*, **72**(7), 4899–4906.