The Journal of Open Source Software

TipToft: detecting plasmids contained in uncorrected long read sequencing data

Andrew J. Page[1] **and Torsten Seemann**[2]

1 Quadram Institute Bioscience, Norwich Research Park, Norwich, UK. **2** Melbourne Bioinformatics, The University of Melbourne, Parkville, Australia.

DOI: 10.21105/joss.01021

Software

- Review
- Repository
- Archive

Submitted: 09 October 2018
Published: 01 March 2019

Summary

With rapidly falling costs, long-read DNA sequencing technology from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), are beginning to be used for outbreak investigations (Faria et al., 2017; J. Quick et al., 2015) and rapid infectious disease clinical diagnostics (Votintseva et al., 2017). ONT instruments can produce data within minutes, and PacBio within hours compared to short-read sequencing technologies which takes hours/days. By reducing the time from swab to an actionable answer, genomics can begin to directly influence clinical decisions, with the potential for a positive impact for patients (Gardy & Loman, 2017). Clinically important genes, like those conferring animicrobial resistance or encoding virulence factors, can be horizontally acquired from plasmids. With the increased speed afforded by long-read sequencing technologies comes increased base errors rates. The high error rates inherent in long-read sequencing reads require specialised tools to correct the reads (Koren et al., 2017), however, these methods require substantial computational requirements, and often take longer to run than the original time to generate the sequencing data, and can result in the loss of small, clinically important plasmids.

We present `TipToft` which uses raw uncorrected reads to predict which plasmids are present in the underlying raw data. This provides an independent method for validating the plasmid content of a *de novo* assembly. It is the only tool which can do this from uncorrected long reads. `TipToft` is fast and can accept streaming input data to provide results in a realtime manner. Plasmids are identified using replicon sequences used for typing from PlasmidFinder (Carattoli et al., 2014). We tested the software on 1975 samples (https://www.sanger.ac.uk/resources/downloads/bacteria/nctc/) sequenced using long read sequencing technologies from PacBio, predicting plasmids from de novo assemblies using abricate (https://github.com/tseemann/abricate). It identified 84 samples containing plasmids with a 100% match to a plasmid sequence, but where no corresponding plasmid was present in the de novo assembly. Taking all the plasmids identified in the assemblies with 100% match, Tiptoft identified 97% (n=326) of these, representing 95% (236) of the samples. A higher depth of read coverage will increase the power to accurately identify plasmid sequences, the level of which depends on the underlying base error rate. For sequence data with 90% base accuracy, an approximate depth of read coverage of 5 is required to identify a plasmid replicon sequence with 99.5% confidence. The software is written in Python 3 and is available under the open source GNU GPLv3 licence from https://github.com/andrewjpage/tiptoft.

Acknowledgements

This work was supported by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1) and by the Wellcome Trust (grant WT 098051).

References

Carattoli, A., Zankari, E., García-Fernández, A., Larsen, M. V., Lund, O., Villa, L., Aarestrup, F. M., et al. (2014). In SilicoDetection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy*, *58*(7), 3895–3903. doi:10.1128/aac.02412-14

Faria, N. R., Quick, J., Claro, I., Thézé, J., Jesus, J. G. de, Giovanetti, M., Kraemer, M. U. G., et al. (2017). Establishment and cryptic transmission of zika virus in brazil and the americas. *Nature*, *546*(7658), 406–410. doi:10.1038/nature22401

Gardy, J. L., & Loman, N. J. (2017). Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics*, *19*(1), 9–20. doi:10.1038/nrg.2017.88

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. doi:10.1101/gr.215087.116

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., Nair, S., et al. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of salmonella. *Genome Biology*, *16*(1). doi:10.1186/s13059-015-0677-2

Votintseva, A. A., Bradley, P., Pankhurst, L., Ojo Elias, C. del, Loose, M., Nilgiriwala, K., Chatterjee, A., et al. (2017). Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. (Y.-W. Tang, Ed.)*Journal of Clinical Microbiology*, *55*(5), 1285–1298. doi:10.1128/jcm.02483-16