Multilocus sequence typing by blast from de novo assemblies against PubMLST

Andrew J. Page[1], Ben Taylor[1], and Jacqueline A. Keane[1]

[1]Pathogen Informatics, Wellcome Trust Sanger Institute

3 Nov 2016

Paper DOI: http://dx.doi.org/10.21105/joss.00118
Software Repository: https://github.com/sanger-pathogens/mlst_check
Software Archive: http://dx.doi.org/10.6084/m9.figshare.4285097.v1

Summary

Multilocus sequence typing (MLST) is a standard method for categorising genomes (M. C. Maiden et al. 1998) based on variation in a small set of conserved house keeping genes. It allows for rapid identification of genomes into high level categories and is extremely useful for epidemiological investigations (R. Urwin and Maiden 2003) making it a key tool for public health reference laboratories. Whilst MLST is more commonly associated with classical sequencing methods, it is possible to extract the same information from Next Generation Sequencing data, in particular from de novo assemblies which are generated routinely for bacterial sequencing data (Page et al. 2016).

We provide a scalable command line tool, MLSTcheck, which can take multiple de novo assemblies and output detailed information about the sequence type of the samples. It provides access to 124 MLST databases covering all of the major human disease causing bacterial pathogens. MLSTcheck can search one or more databases at once, is parallelisable, fast and robust. When a sample contains more than one allele, it flags the contaminant since there should only be 1 copy of a house keeping gene in a well designed MLST scheme. A multiple FASTA alignment of the concatenated MLST genes is optionally produced, allowing for the creation of phylogenetic trees. This allows for rapid epidemiological outbreak investigations. Whilst other software applications can perform similar functions [Seeman2016; Jolley2010], this application follows more rigorous software engineering principles, including automated testing, continuous integration, object orientated code, and is installable via CPAN (a Perl package manager). In a large diverse set of 6814 publicly accessible draft assemblies, MLSTcheck was able to assign a sequence type in 99.6% of cases (Page et al. 2016).

References

Maiden, M C, J a Bygraves, E Feil, G Morelli, J E Russell, R Urwin, Q Zhang, et al. 1998. "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms." *Proceedings of the National Academy of Sciences of the United States of America* 95 (6): 3140–5. doi:10.1073/pnas.95.6.3140.

Page, Andrew J., Nishadi De Silva, Martin Hunt, Michael A. Quail, Julian Parkhill, Simon R. Harris, Thomas D. Otto, and Jacqueline A. Keane. 2016. "Robust High-Throughput Prokaryote de Novo Assembly and

Improvement Pipeline for Illumina Data." *Microbial Genomics* 2 (8). doi:10.1099/mgen.0.000083.

Urwin, Rachel, and Martin C.J. Maiden. 2003. "Multi-Locus Sequence Typing: A Tool for Global Epidemiology." *Trends in Microbiology* 11 (10): 479–87. doi:10.1016/j.tim.2003.08.006.