GFF3toEMBL: Preparing annotated assemblies for submission to EMBL

Andrew J. Page[1], Sascha Steinbiss[1], Ben Taylor[1], Torsten Seemann[2], and Jacqueline A. Keane[1]

[1]Pathogen Informatics, Wellcome Trust Sanger Institute
[2]University of Melbourne

19 Sept 2016

Paper DOI: http://dx.doi.org/10.21105/joss.00080
Software Repository: https://github.com/sanger-pathogens/gff3toembl.git
Software Archive: https://dx.doi.org/10.6084/m9.figshare.3988815.v1

Summary

An essential part of open reproducible research in genomics is the deposition of annotated de novo assembled genomes in public archives such as EMBL/GenBank (Blaxter et al. 2016). The interfaces provided by the major archives do not allow for data to be easily submitted on a large scale without substantial prior knowledge on the part of the submitter. This has lead to a situation where less than 15% of all sequenced bacteria have corresponding public assemblies. We address this by providing GFF3toEMBL, which converts the output of the most commonly used automatic annotation tool, Prokka (Seemann 2014), and converts it to a format suitable for submission to EMBL. Built on the GenomeTools annotation processing library (Gremme, Steinbiss, and Kurtz 2013), GFF3toEMBL is robust, fast, memory efficient and well tested, and has been used to submit more than 30% of all annotated genomes in EMBL/GenBank (Page et al. 2016). It is a small, but essential missing step in making genomic research more open and reproducible.

References

Blaxter, Mark, Antoine Danchin, Babis Savakis, Kaoru Fukami-Kobayashi, Ken Kurokawa, Sumio Sugano, Richard J. Roberts, Steven L. Salzberg, and Chung-I Wu. 2016. "Reminder to Deposit DNA Sequences." *Science*. American Association for the Advancement of Science. doi:10.1126/science.aaf7672.

Gremme, Gordon, Sascha Steinbiss, and Stefan Kurtz. 2013. "*GenomeTools*: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10 (3). Los Alamitos, CA, USA: IEEE Computer Society: 645–56. doi:10.1109/TCBB.2013.68.

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). http://mgen.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000083.

Seemann, Torsten. 2014. "Prokka: rapid prokaryotic genome annotation." *Bioinformatics (Oxford, England)*

30 (14): 2068–9. doi:10.1093/bioinformatics/btu153.