The Journal of Open Source Software

WGS2NCBI - Toolkit for preparing genomes for submission to NCBI

Rutger Aldo Vos[1], **Nnadi Nnaemeka Emmanuel**[2], **and John Chinyere Aguiyi**[3, 4]

1 Research Group 'Understanding Evolution', Naturalis Biodiversity Center, Leiden, The Netherlands **2** Department of Microbiology, Faculty of Natural and Applied Science, Plateau State University, Bokkos, Plateau State, Nigeria **3** African Centre of Excellence on Phytomedicine Research and Development (ACEPRD), University of Jos, Jos, Plateau State, Nigeria **4** Department of Pharmacology, Faculty of Pharmaceutical Sciences, University of Jos, Jos, Plateau State, Nigeria

DOI: 10.21105/joss.01364

Software

- Review
- Repository
- Archive

Submitted: 28 March 2019
Published: 19 April 2019

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

The process of submitting annotated genomes to NCBI GenBank (Benson et al., 2015–1AD) - and having them pass review - is a labour intensive, iterative process due to the stringent quality requirements that NCBI imposes (Pirovano, Boetzer, Derks, & Smit, 2015). These requirements cannot typically be met on the first iteration because they involve checks based on the entirety of NCBI's presently held data (e.g. contamination checks), checks for the continuously moving target of vendor-specific adaptor sequences, and checks for the validity of gene product names. In many genome annotation projects, these product names are copied over from homologous sequences in related model organisms. This may introduce terminology that is not, or no longer, permitted by NCBI, such as molecular weights and protein structure, organism names, database identifiers, and so on.

wgs2ncbi is a standalone Perl package for preparing the results of whole genome sequencing (WGS) and annotation projects to NCBI GenBank. The purpose of the package is to automate responding to NCBI's reviews by allowing batch corrections to detected problems. The functionality consists of a command line program that takes sub-commands for the various steps of the process:

1. Preparing the input data for rapid processing downstream (sub-command `prepare`).
2. Generating valid scaffolds and feature tables from the sequence data and genome annotation (`process`).
3. Converting the processing results to ASN.1/seqin files for upload to NCBI (`convert`).
4. Compressing the converted files to archives acceptable to NCBI's upload service (`compress`).

In step 2., the toolkit allows for easy masking of detected contaminations and adaptors, a generalized mapping between invalid product names (as detected by NCBI) and valid alternatives, and automatic conversion of putative-but-invalid genes (e.g. those with introns that are 'too short') to pseudogenes. This functionality has helped make the submission of the genome of the King Cobra (Vonk et al., 2013) and (AZIM00000000.1), and that of the Velvet Bean (QJKJ00000000.1).

References

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2015–1AD). GenBank. *Nucleic acids research*, *43*(Database issue), D30–D35. doi:10.1093/nar/gku1216

Pirovano, W., Boetzer, M., Derks, M. F. L., & Smit, S. (2015). NCBI-compliant genome submissions: tips and tricks to save time and money. *Briefings in Bioinformatics*, *18*(2), 179–182. doi:10.1093/bib/bbv104

Vonk, F. J., Casewell, N. R., Henkel, C. V., Heimberg, A. M., Jansen, H. J., McCleary, R. J. R., Kerkkamp, H. M. E., et al. (2013). The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proceedings of the National Academy of Sciences*, *110*(51), 20651–20656. doi:10.1073/pnas.1314702110