The Journal of Open Source Software

SneakerNet: A modular quality assurance and quality check workflow for primary genomic and metagenomic read data

Taylor Griswold[1], **Curtis Kapsak**[1, 2], **Jessica C. Chen**[1], **Henk C. den Bakker**[3], **Grant Williams**[1], **Alyssa Kelley**[2, 4], **Eshaw Vidyaprakash**[1, 5], **and Lee S. Katz**[1, 3]

1 Enteric Diseases Laboratory Branch (EDLB), Centers for Disease Control and Prevention, Atlanta, GA, USA **2** Weems Design Studio, Inc., Suwanee, GA, USA **3** Center for Food Safety, University of Georgia, Griffin, GA, USA **4** Waterborne Disease Prevention Branch (WDPB), Centers for Disease Control and Prevention, Atlanta, GA, USA **5** IHRC, Atlanta, GA, USA

DOI: 10.21105/joss.02334

Software

- Review
- Repository
- Archive

Editor: Lorena Pantano

Reviewers:

- @lfaller
- @erinyoung

Submitted: 20 May 2020
Published: 16 June 2020

Summary

Receiving a set of primary data from whole genome sequencing or metagenomics sequencing has become commonplace and perhaps ubiquitous in bioinformatics. However, there is a need to standardize the quality assurance and quality control process (QA/QC).

There are very few published workflows for performing an analysis on primary sequence data that span the breadth of initial standardized QA/QC (e.g., sequence yields, contamination checks, and subtyping). For example, the Pandoo pipeline can be given a set of genomes to run analyses: species inference, 7-gene multilocus sequence typing (MLST), resistance gene profile, plasmid profile, virulence profile, and raw read QC (Schultz, Gonçalves da Silva, Kwong, & Seemann, 2019). The Nullarbor pipeline is similar to Pandoo, but focused on public health datasets (Seemann et al., 2017). Another example is the ASA3P pipeline that runs raw read trimming, assembly, annotation, taxonomic classification, MLST, antibiotic resistance detection, virulence factor detection, reference mapping, and single nucleotide polymorphism (SNP) detection (Schwengers et al., 2019). However, no existing "broad stroke" QA/QC pipelines seem to be focused on a plugins-based architecture for batches of unrelated bacterial sequences or for batches of bacteria from different species. To that end, we have created SneakerNet. The major design principles are centered around the ability to collaboratively design plugins. With the plugins architecture, SneakerNet can dynamically change for current and future needs with input from the bioinformatics and public health community.

Implementation

Plugin design

SneakerNet has a modular plugin design, where the main program calls each plugin in an ordered succession. Each plugin, in turn, reads a set of genomes as input. Each plugin accepts specific flagged parameters, such that the main program can call each plugin in a standardized way. Workflows are thus defined as a specified order of plugins. An example workflow order might be genome assembly, followed by MLST, and finalized with report generation. At the time of this writing, 25 plugins are available. These plugins are listed in the documentation in a summary table, and each plugin has its own documentation page.

Plugin development

The plugin system has drastically lowered the activation energy needed to develop a new step in a SneakerNet workflow. Documentation has been provided on how to develop a new plugin, and 'Hello World' plugins have been published in three different languages: Perl, Python, and Bash. Because plugins are not tied to any specific language, SneakerNet collaborators do not have to be bound by any specific language.

Configuration

SneakerNet is highly configurable as described in the installation documentation. There are many configurations. We would like to highlight some ways that SneakerNet can be configured.

For some genera, SneakerNet comes packaged with some recommended configurations (e.g., *Salmonella* or *Legionella*), and an example genus with all options commented. These options include the minimum coverage needed for a sample to pass QC and even some detailed options to help customize a taxon for a particular plugin such as the antimicrobial resistance plugin. Therefore, a user could easily add a taxon to customize the workflow for his or her instance of SneakerNet. In fact, SneakerNet has been recently configured to accommodate the protist *Cryptosporidium* successfully with input from the CDC WDPB Molecular Epidemiology Laboratory.

Users can also customize the workflow. SneakerNet comes packaged with a default workflow which specifies the order of plugins that are run. However, if a certain analysis is not needed, e.g., 7-gene MLST, then it can be removed from the configuration. Likewise, if a new plugin is needed, it can be added into the workflow.

Acknowledgements

The authors would like to thank the following SneakerNet users in the Enteric Diseases Laboratory Branch for useful feedback: Heather Carleton, Katie Dillon, Blake Dinsmore, Yang Gao, Jessica Halpin, Jasmine Hensley, Monica Im, Justin Kim, Charlotte Lane, Ana Lauer, Rebecca Lindsey, Tori McIntosh, Angela Poates, Zachary Rigney, Katie Roache, Ashley Sabol, Peyton Smith, Cheryl Tarr, Jenny Truong, Maryann Turnsek; Additionally, SneakerNet users from other branches: Dhwani Batra, Shatavia Morrison. The authors would also like to thank Aaron Petkau for useful feedback for staramr.

References

Schultz, M. B., Gonçalves da Silva, A., Kwong, J., & Seemann, T. (2019). *Pandoo*. Retrieved from https://github.com/schultzm/pandoo/releases/tag/0.3.3

Schwengers, O., Hoek, A., Fritzenwanker, M., Falgenhauer, L., Hain, T., Chakraborty, T., & Goesmann, A. (2019). ASA3P: An automatic and scalable pipeline for the assembly, annotation and higher level analysis of closely related bacterial isolates. *bioRxiv*. doi:10.1101/654319

Seemann, T., Goncalves da Silva, A., Bulach, D. M., Schultz, M. B., Kwong, J. C., & Howden, B. P. (2017). *Nullarbor*. Retrieved from https://github.com/tseemann/nullarbor/releases/tag/v1.28