The Journal of Open Source Education

Genome-centric metagenomics, from experimental design to metagenomic Hi-C

Matthew Z DeMaere[1], **Daniela Gaio**[1], **Kay Anantanawat**[1], **and Aaron E Darling**[1]

1 University of Technology Sydney, The ithree institute

DOI: 10.21105/jose.00079

Software

- Review
- Repository
- Archive

Submitted: 08 January 2020
Published: 28 February 2020

Summary

This short manuscript describes an open educational resource for teaching applied metagenomic data analysis called The Gene School: Metagenomics. The material assumes that trainees have a basic knowledge of metagenomics and computing, or that such knowledge is supplied separately via lecture material. Text documents provided herein are licensed as CC-BY-SA and all third-party software used is available under an OSI-approved license.

Statement of need

The importance and pervasiveness of microbial communities in biological systems is now widely appreciated, and high throughput DNA sequencing has provided a means to characterize and quantify such microbial communities, in particular via metagenome sequencing. Successful application of metagenome sequencing benefits from careful experimental design and planning, not just for the sampling process but also for the sequencing and data analysis components. A wholistic view of the entire process from sampling to data analysis is required. The educational materials we present in this open module are designed to give students hands-on experience with several aspects of experimental design and data analysis for metagenomics. The material introduces several state of the art (as of 2019) data analysis methods for metagenome quality control, genome assembly, and genome binning. The module also covers the application of emerging techniques such as metagenomic Hi-C for which open source analysis software has only recently been introduced and no previous educational material has been developed.

Learning objectives

Students trained with this module are expected to develop the following skills:

- Work in a Jupyter bash environment
- Install software via conda
- Run containerized software via docker and singularity
- Carry out QC for metagenome data samples
  - Sequencing QC with FastQC and MultiQC (Ewels, Magnusson, Lundin, & Käller, 2016)

    - Taxonomic profiling as QC using metaphlan and kraken2 (Segata et al., 2012; Wood, Lu, & Langmead, 2019)
- Clean sequencing data using fastp (Chen, Zhou, Chen, & Gu, 2018)
- Understand the limitations of reference databases for metagenome analysis
- Learn how a pilot study can inform experimental design
    - Learn how to estimate metagenome sequencing depth requirements
    - How do host-associated samples affect sequencing depth requirements
    - Designs that maximize experimental interpretation via genome-centric metagenomics
        * Time-series
        * Spatial or population transects
- Learn how to assemble metagenomes
    - Use the megahit assembler (Li, Liu, Luo, Sadakane, & Lam, 2015)
    - Understand whether or not to co-assemble multiple samples
- Binning genomes from metagenomes with MetaBAT2 (Kang et al., 2019)
- Visualization of bins and bin refinement with anvi'o (**???**)
- Using Hi-C for genome binning
    - Carry out QC for a metagenomic Hi-C library with qc3C
    - Extract Metagenome-assembled genomes (MAGs) with Hi-C data with bin3C (DeMaere & Darling, 2019)

Content

This open educational resource comprises a collection of markdown-formatted workshop pages, a collection of publicly available data sets which we have generated, and instructions to set up a virtual machine image preloaded with data & results so that compute-intensive steps can be skipped during course delivery. The precomputed results are also provided via Zenodo DOI 10.5281/zenodo.3585993 for ease of use in other computing environments.

Instructional design

The material is designed for use in a hybrid lecture and active learning format. Lecture material that introduces the concepts of microbial communities, their genetics, and metagenome sequencing provides appropriate context for the hands on computational steps in this module. The active learning material has been designed for use in Jupyter, enabling students to gain some basic familiarity with reproducible research environments as they learn about metagenome analysis.

Experience of use

This course material was first delivered at The Gene School 2019, held at Kasetsart University, Bangkok, Thailand. For that workshop a budget from registration fees was used to spawn a fleet of virtual machines in the Amazon Web Services cloud, with one VM per student. Each student received a unique URL at the start of the workshop to access the preconfigured Jupyter server on their own VM. This approach worked very well, providing a stable and homogenous computing environment for the learning. An interactive online discussion system was used during the workshop, and the student questions posed in that

system as well as verbally, highlighted deficiencies in the first revision of the workshop material. The material was delivered a second time at a workshop held at Western Sydney University, Sydney, Australia. When combined with lecture material, this module requires approximately 6 hours to deliver.

Acknowledgements

This work was funded in part via the Australian Research Council's Discovery scheme, under ARC Discovery project DP180101506. We thank Passorn Wonnapinij, Arinthip Thamchaipenet, Alexie Papanicolaou, and Thomas Jeffries for their roles in organizing the workshop at which this material was first delivered. That workshop was supported in part by the Australian Academy of Sciences Regional Collaborations Programme and the Genetics Society of Thailand.

References

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. doi:10.1093/bioinformatics/bty560

DeMaere, M. Z., & Darling, A. E. (2019). Bin3C: Exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome Biology*, *20*(1), 46. doi:10.1186/s13059-019-1643-1

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. doi:10.1093/bioinformatics/btw354

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *7*, e7359. doi:10.7717/peerj.7359

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)*, *31*(10), 1674–1676. doi:10.1093/bioinformatics/btv033

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, *9*(8), 811–814. doi:10.1038/nmeth.2066

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. doi:10.1186/s13059-019-1891-0