AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

DATABASES AND SOFTWARE

RRAP: RPKM Recruitment Analysis Pipeline

Conner Y. Kojima,[a] Eric W. Getz,[a] J. Cameron Thrash[a]

[a]Department of Biological Sciences, University of Southern California, Los Angeles, California, USA

ABSTRACT A common method for quantifying microbial abundances *in situ* is through metagenomic read recruitment to genomes and normalizing read counts as reads per kilobase (of genome) per million (bases of recruited sequences) (RPKM). We created RRAP (RPKM Recruitment Analysis Pipeline), a wrapper that automates this process using Bowtie2 and SAMtools.

Quantifying the relative abundance of microorganisms in a sample is a critical component of microbial ecology research. Whole-community metagenomic sequencing can be used to calculate relative abundance after recruiting reads to genomes generated from isolates, metagenomes, or single cells (1–4). Since genomes will have different sizes and each sample will have different numbers of reads, normalizing for these two variables can be accomplished with the RPKM (reads per kilobase [of genome] per million [bases of recruited sequences]) method, which was originally developed to quantify relative transcript abundance (5).

To automate the process of read recruitment and RPKM normalization for use in recruiting hundreds or thousands of samples to similarly large numbers of genomes, we developed RRAP (RPKM Recruitment Analysis Pipeline). RRAP is a wrapper for other established tools that takes paired-end metagenomic sequences and reference genome sequences as the input and generates both read alignment data and RPKM values. The pipeline streamlines the read recruitment process by automatically handling the preprocessing steps of merging contigs, concatenating reference genomes, and indexing reference sequences. RRAP installs the most recent versions of Bowtie2 and SAMtools that are compatible with the other dependencies (6, 7). After performing read recruitment with Bowtie2, the pipeline sorts and indexes sequence alignment data before counting the numbers of mapped and unmapped metagenomic reads per reference sequence with SAMtools. From the output, RRAP calculates both unadjusted and $\log_{10}$-adjusted RPKM values for each reference genome in each metagenomic sample.

Other bioinformatics tools are similar to RRAP but serve different purposes. The Enveomics Collection is a compilation of scripts that analyze metagenomes (8). The scripts BlastTab.catbj.pl and BlastTab.recplot2.R in particular use BLAST results to generate a recruitment plot for visualization purposes. The script anir.rb estimates the average nucleotide identity of reads against a genome using existing alignment data. Anvi'o also provides a metagenomics workflow that assembles reads and maps them to contigs, but this is a much more comprehensive software package than RRAP and serves numerous purposes (9, 10). There are other existing pipelines that perform read recruitment but do not calculate RPKM values. Sunbeam and ngs_backbone are two examples that recruit reads with bwa instead of Bowtie2 to produce alignment data but do not calculate RPKM values (11–13). RRAP is therefore a unique, lightweight, and standalone pipeline for both recruitment and RPKM calculation.

Data availability. The code, detailed instructions for use, and sample data files to install and test run RRAP are available on GitHub (https://github.com/thrash-lab/rrap). Because the pipeline has dependencies, we recommend installation through the Conda package manager (14). Upon installation, RRAP can be accessed from the command line with a single command.

Editor Irene L. G. Newton, Indiana University, Bloomington

Address correspondence to J. Cameron Thrash, thrash@usc.edu.

The authors declare no conflict of interest.

Received 26 June 2022
Accepted 8 August 2022
Published 22 August 2022

Sample metagenomes and reference genomes to allow quick testing of read recruitment and RPKM calculations were obtained from previous studies (15–19).

ACKNOWLEDGMENTS

This work was supported by the Center for Advanced Research Computing (CARC) at the University of Southern California, which provided computing resources that contributed to the research results reported within this publication, as well as a Simons Early Career Investigator in Marine Microbial Ecology and Evolution award and NSF Biological Oceanography Program grants (OCE-1747681 and OCE-1945279) to J.C.T.

REFERENCES

1. Henson MW, Lanclos VC, Faircloth BC, Thrash JC. 2018. Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. ISME J 12:1846–1860. https://doi.org/10.1038/s41396-018-0092-2.
2. Savoie ER, Lanclos VC, Henson MW, Cheng C, Getz EW, Barnes SJ, LaRow DE, Rappé MS, Thrash JC. 2021. Ecophysiology of the cosmopolitan OM252 bacterioplankton (Gammaproteobacteria). mSystems 6(3):e00276-21. https://doi.org/10.1128/mSystems.00276-21.
3. Krüger K, Chafee M, Ben Francis T, Glavina del Rio T, Becher D, Schweder T, Amann RI, Teeling H. 2019. In marine Bacteroidetes the bulk of glycan degradation during algae blooms is mediated by few clades using a restricted set of genes. ISME J 13:2800–2816. https://doi.org/10.1038/s41396-019-0476-y.
4. Tully BJ. 2019. Metabolic diversity within the globally abundant marine group II Euryarchaea offers insight into ecological patterns. Nat Commun 10:271. https://doi.org/10.1038/s41467-018-07840-4.
5. Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628. https://doi.org/10.1038/nmeth.1226.
6. Langmead B, Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
7. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. Gigascience 10:giab008. https://doi.org/10.1093/gigascience/giab008.
8. Rodriguez-R LM, Konstantinidis KT. 2016. The Enveomics Collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ PrePrints 4:e1900v1. https://peerj.com/preprints/1900/
9. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, Fink I, Pan JN, Yousef M, Fogarty EC, Trigodet F, Watson AR, Esen OC, Moore RM, Clayssen Q, Lee MD, Kivenson V, Graham ED, Merrill BD, Karkman A, Blankenberg D, Eppley JM, Sjödin A, Scott JJ, Vázquez-Campos X, McKay LJ, McDaniel EA, Stevens SLR, Anderson RE, Fuessel J, Fernandez-Guerra A, Maignien L, Delmont TO, Willis AD. 2021. Community-led, integrated, reproducible multi-omics with anvi'o. Nat Microbiol 6:3–6. https://doi.org/10.1038/s41564-020-00834-3.
10. Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28:2520–2522. https://doi.org/10.1093/bioinformatics/bts480.
11. Clarke EL, Taylor LJ, Zhao C, Connell A, Lee J-J, Fett B, Bushman FD, Bittinger K. 2019. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. Microbiome 7:46. https://doi.org/10.1186/s40168-019-0658-x.
12. Blanca JM, Pascual L, Ziarsolo P, Nuez F, Cañizares J. 2011. ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using next generation sequence. BMC Genomics 12:285. https://doi.org/10.1186/1471-2164-12-285.
13. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
14. Anaconda Inc. 2022. Anaconda software distribution (4.13.0). Anaconda Inc, Austin, TX. https://www.anaconda.com. Retrieved 22 September 2021.
15. Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, Rappé MS. 2012. Streamlining and core genome conservation among highly divergent members of the SAR11 clade. mBio 3(5):e00252-12. https://doi.org/10.1128/mBio.00252-12.
16. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur EJ. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science 309:1242–1245. https://doi.org/10.1126/science.1114057.
17. Fortunato CS, Crump BC. 2015. Microbial gene abundance and expression patterns across a river to ocean salinity gradient. PLoS One 10:e0140578. https://doi.org/10.1371/journal.pone.0140578.
18. Sakowski EG, Arora-Williams K, Tian F, Zayed AA, Zablocki O, Sullivan MB, Preheim SP. 2021. Interaction dynamics and virus-host range for estuarine actinophages captured by epicPCR. Nat Microbiol 6:630–642. https://doi.org/10.1038/s41564-021-00873-4.
19. Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11:e0163962. https://doi.org/10.1371/journal.pone.0163962.