The Journal of Open Source Software

Rasusa: Randomly subsample sequencing reads to a specified coverage

Michael B. Hall[1]

1 European Molecular Biology Laboratory, European Bioinformatics Institute EMBL-EBI, Hinxton, UK

DOI: 10.21105/joss.03941

Software

- Review
- Repository
- Archive

Editor: Mikkel Meyer Andersen

Reviewers:

- @k3yavi
- @holtgrewe

Submitted: 18 October 2021
Published: 22 November 2021

Summary

A fundamental requirement for many applications in genomics is the sequencing of genetic material (DNA/RNA). Different sequencing technologies exist, but all aim to accurately reproduce the sequence of nucleotides (the individual units of DNA and RNA) in the genetic material under investigation. The result of such efforts is a text file containing the individual fragments of genetic material - termed "reads" - represented as strings of letters (A, C, G, and T/U).

The amount of data in one of these read files depends on how much genetic material was present and how long the sequencing device was operated. Read depth (coverage) is a measure of the volume of genetic data contained in a read file. For example, coverage of 5x indicates that, on average, each nucleotide in the original genetic material is represented five times in the read file.

Many of the computational methods employed in genomics are affected by coverage; counterintuitively, more is not always better. For example, because sequencing devices are not perfect, reads inevitably contain errors. As such, higher coverage increases the number of errors and potentially makes them look like alternative sequences. Furthermore, for some applications, too much coverage can cause a degradation in computational performance via increased runtimes or memory usage.

We present Rasusa, a software program that randomly subsamples a given read file to a specified coverage. Rasusa is written in the Rust programming language and is much faster than current solutions for subsampling read files. In addition, it provides an ergonomic command-line interface and allows users to specify a desired coverage or a target number of nucleotides.

Statement of need

Read subsampling is a useful mechanism for creating artificial datasets, allowing exploration of a computational method's performance as data becomes more scarce. In addition, the coverage of a sample can have a significant impact on a variety of computational methods, such as RNA-seq (Baccarella et al., 2018), taxonomic classification (Gweon et al., 2019), antimicrobial resistance detection (Gweon et al., 2019), and genome assembly (Maio et al., 2019) - to name a few.

There is limited available software for subsampling read files. Assumably, most researchers use custom scripts for this purpose. However, two existing programs for subsampling are Filtlong (Wick, 2021) and Seqtk (Li, 2018). Unfortunately, neither of these tools provides subsampling to a specified coverage "out of the box."

Filtlong is technically a filtering tool, not a subsampling one. It scores each read based on its length and quality and outputs the highest-scoring subset. Additionally, minimum and maximum read lengths can be specified, along with the size of the subset required. Ultimately, the subset produced by Filtlong is not necessarily representative of the original reads but is biased towards those with the greatest length or quality. While this may sound like a good thing, in some applications, such as genome assembly, it has been shown that a random subsample produces superior results to a filtered subset (Maio et al., 2019).

Seqtk does do random subsampling via the sample subcommand. However, the only option available is to specify the number of reads required. Thus, it is up to the user to determine the number of reads required to reach the desired coverage. While this serves for Illumina sequencing data, which generally have uniform(ish) read lengths, it does not work for other modalities like PacBio and Nanopore, where read lengths vary significantly.

Rasusa provides a random subsample of a read file (FASTA or FASTQ), with two ways of specifying the size of the subset. One method takes a genome size and the desired coverage, while the other takes a target number of bases (nucleotides). In the genome size and coverage option, we multiply the genome size by the coverage to obtain the target number of bases for the subset. As such, the resulting read file will have, on average, the amount of coverage requested. In addition, Rasusa allows setting a random seed to allow reproducible subsampling. Other features include user control over whether the output is compressed and specifying the compression algorithm and level.

Rasusa is 21 and 1.2 times faster than Filtlong and Seqtk, respectively.

Availability

Rasusa is open-source and available under an MIT license at https://github.com/mbhall88/rasusa.

Acknowledgements

We acknowledge contributions from Pierre Marijon and suggestions from Zamin Iqbal. In addition, MBH is funded by the EMBL International PhD Programme.

References

Baccarella, A., Williams, C. R., Parrish, J. Z., & Kim, C. C. (2018). Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance. *BMC Bioinformatics*, *19*(1), 423. https://doi.org/10.1186/s12859-018-2445-2

Gweon, H. S., Shaw, L. P., Swann, J., Maio, N. D., AbuOun, M., Niehus, R., Hubbard, A. T. M., Bowes, M. J., Bailey, M. J., Peto, T. E. A., Hoosdally, S. J., Walker, A. S., Sebra, R. P., Crook, D. W., Anjum, M. F., Read, D. S., Stoesser, N., Abuoun, M., Anjum, M., … Woodford, N. (2019). The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. *Environmental Microbiome*, *14*(1), 7. https://doi.org/10.1186/s40793-019-0347-1

Li, H. (2018). Seqtk: Toolkit for processing sequences in FASTA/q formats. In *GitHub repository*. GitHub. https://github.com/lh3/seqtk

Maio, N. D., Shaw, L. P., Hubbard, A., George, S., Sanderson, N. D., Swann, J., Wick, R., AbuOun, M., Stubberfield, E., Hoosdally, S. J., Crook, D. W., Peto, T. E. A., Sheppard,

A. E., Bailey, M. J., Read, D. S., Anjum, M. F., Walker, A. S., Stoesser, N., & Consortium, O. B. O. T. R. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics*, *5*(9). https://doi.org/10.1099/mgen.0.000294

Wick, R. (2021). Filtlong: Quality filtering tool for long reads. In *GitHub repository*. GitHub. https://github.com/rrwick/Filtlong