The Journal of Open Source Software

Badread: simulation of error-prone long reads

Ryan R Wick[1]

1 Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria 3004, Australia

DOI: 10.21105/joss.01316

Software

- Review
- Repository
- Archive

Submitted: 01 March 2019
Published: 04 April 2019

Background

DNA sequencing platforms aim to measure the sequence of nucleotides (A, C, G and T) in a sample of DNA. Sequencers made by Illumina have been the dominant technology for much of the past decade, but their platforms generate fragments of sequence ('reads') that are relatively small (~100–300 nucleotides in length). In contrast, Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) produce 'long-read' sequencers that can generate sequence fragments with tens of thousands of nucleotides or more (Eisenstein, 2017). Long reads from these platforms can be very beneficial for genome assembly and other bioinformatic analyses (Koren, Walenz, Berlin, Miller, & Phillippy, 2017; Phillippy, 2017). ONT and PacBio sequencers achieve their long read lengths because they detect nucleotides in individual molecules of DNA, a.k.a. single-molecule sequencing (Heather & Chain, 2016). However, the stochastic nature of measuring at the single-molecule scale means that ONT and PacBio reads are 'noisy' – they contain a significant amount of errors.

Since sequencing reads from ONT and PacBio platforms are qualitatively different from Illumina reads (long and noisy vs short and accurate), they often require novel methods of analysis. The last few years have seen much research in this space, and one useful technique for evaluating new methods is read simulation: generating fake sequencing reads from a reference nucleotide sequence (Huang, Li, Myers, & Marth, 2012). This approach has some key advantages over using real sequencing data: it can be faster, more affordable and allow for a greater number of tests. Additionally, when using simulated reads, the reference nucleotide sequence provides a confident ground truth which may not be otherwise available.

Summary

Here we introduce Badread, a software tool for *in silico* simulation of long reads. Its primary aim is to generate simulated read sets for the purpose of evaluating tools or methods that take long reads as input. Badread differs from existing tools (e.g. PBSIM (Ono, Asai, & Hamada, 2013), LongISLND (Mu et al., 2016) and NanoSim (Yang, Chu, Warren, & Birol, 2017)) in two key ways. First, it can simulate types of read errors that other tools cannot. While other long read simulation tools focus on modelling read length and sequencing errors, Badread can additionally include chimeras (when a single read which consists of two or more non-contiguous sequences), adapters (additional sequences from the library preparation at the start or end of a read), glitches (localised regions of low accuracy) and junk reads (low-complexity repetitive sequences).

The second way Badread differs from existing tools is that it prioritises control over realism. Using read length as an example, other long read simulation tools may sample

read lengths from a real read set, so their simulated reads follow a realistic distribution. Badread instead uses a gamma distribution for read lengths where the user specifies the mean and standard deviation – less realistic but highly tuneable. Users can therefore generate many read sets which quantitatively vary, e.g. mean lengths of 1000, 2000, 3000, etc. Other characteristics of the read set (read accuracy, chimera rate, glitch rate, etc.) can be similarly tuned in Badread, allowing users to systematically evaluate how they affect the performance of a tool or method.

Availability

Badread is open-source and available via the GPLv3 license at github.com/rrwick/Badread.

References

Eisenstein, M. (2017). An ace in the hole for DNA sequencing. *Nature*, *550*(7675), 285–288. doi:10.1038/550285a

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. doi:10.1016/j.ygeno.2015.11.003

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593–594. doi:10.1093/bioinformatics/btr708

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. doi:10.1101/071282

Mu, J. C., Mohiyuddin, M., Dallett, C., Lau, B., Bani Asadi, N., Fang, L. T., & Lam, H. Y. K. (2016). LongISLND: in silico sequencing of lengthy and noisy datatypes. *Bioinformatics*, *32*(24), 3829–3832. doi:10.1093/bioinformatics/btw602

Ono, Y., Asai, K., & Hamada, M. (2013). PBSIM: PacBio reads simulator—toward accurate genome assembly. *Bioinformatics*, *29*(1), 119–121. doi:10.1093/bioinformatics/bts649

Phillippy, A. M. (2017). New advances in sequence assembly. *Genome Research*, *27*(5), xi–xiii. doi:10.1101/gr.223057.117

Yang, C., Chu, J., Warren, R. L., & Birol, I. (2017). NanoSim: Nanopore sequence read simulator based on statistical characterization. *GigaScience*, *6*(4). doi:10.1093/gigascience/gix010