BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 4 2012, pages 593–594
doi:10.1093/bioinformatics/btr708

Sequence analysis

Advance Access publication December 23, 2011

ART: a next-generation sequencing read simulator

Weichun Huang[1,*], Leping Li[1], Jason R. Myers[1,†] and Gabor T. Marth[2,*]

[1]Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709 and [2]Department of Biology, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: ART is a set of simulation tools that generate synthetic next-generation sequencing reads. This functionality is essential for testing and benchmarking tools for next-generation sequencing data analysis including read alignment, *de novo* assembly and genetic variation discovery. ART generates simulated sequencing reads by emulating the sequencing process with built-in, technology-specific read error models and base quality value profiles parameterized empirically in large sequencing datasets. We currently support all three major commercial next-generation sequencing platforms: Roche's 454, Illumina's Solexa and Applied Biosystems' SOLiD. ART also allows the flexibility to use customized read error model parameters and quality profiles.

Availability: Both source and binary software packages are available at http://www.niehs.nih.gov/research/resources/software/art

Contact: weichun.huang@nih.gov; gabor.marth@bc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 3, 2011; revised on December 6, 2011; accepted on December 19, 2011

1 INTRODUCTION

In the past few years, high-throughput next-generation sequencing technologies have effectively replaced earlier data types for genome-wide studies measuring gene expression changes and discovering genomic/epigenetic variations, and many tools were developed for analyzing such datasets. Simulated data is indispensable for guiding tool development and evaluating tool performance, and therefore it is essential to develop simulation software that can produce next-generation sequencing reads that captures the most essential characteristics of real data. Currently available read simulation programs include *Wgsim* from the Samtools package (Li *et al.*, 2009) for generating Illumina sequencing reads, *MetaSim* (Richter *et al.*, 2008) for simulating metagenomic data, *Mason* (http://seqan.de/projects/mason.html) for both Illumina and 454 reads, *SimSeq* (https://github.com/jstjohn/SimSeq) for Illumina reads and *FlowSim* (Balzer *et al.*, 2010) for 454 reads. Although these programs work well in their domain, there is a need for a read simulation program that can deal with all major sequencing platforms, and generate sequence reads with both substitution and insertion–deletion (INDEL) errors, as appropriate for the error modes of each specific platform.

*To whom correspondence should be addressed.

†Present address: Department of Biological Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA.

As a general simulator, our ART software was initially developed for simulation studies helping to design data collection modalities of the 1000 Genomes Project (Durbin *et al.*, 2010). ART has been subsequently used by many users worldwide to facilitate sequencing software development. ART takes a set of DNA sequences (representing e.g. a reference genome), and generates 'synthetic' sequencing reads in a way that mimics the technology-specific sequencing process. ART comes with a set of technology-specific read error profiles, but it can also take user-supplied profiles to generate sequencing data with customized read length and error characteristics. ART can report simulated reads in the standard SAM alignment format and UCSC BED files.

2 FEATURES AND METHODS

ART simulates both single-end and paired-end sequencing reads of the three main commercial next-generation sequencing platforms: 454, Illumina and SOLiD. The built-in read length and read error profiles were derived from large sets of actual real sequencing data (see Supplementary Material). ART supports all three types of common sequencing errors: base substitutions, insertions and deletions.

2.1 Illumina read simulation

Illumina sequencing by synthesis is a base-by-base sequencing technology using a reversible terminator-based method, enabling detection of single bases as they are incorporated into growing DNA strands complementary to the template (Bentley, 2006). Since this technology reads out one base at a time, the main error mode is substitution rather than insertion or deletion. The probability of a substitution error is determined by the base quality score associated with the called base. The distribution of base quality scores is position-dependent: the mean quality score decreases as a function of increasing base position. ART simulates substitution errors according to the empirical, position-dependent distribution of base quality scores, measured in large training datasets. The base quality score does not directly provide information for INDEL errors, and ART simulates insertion and deletion errors directly from empirical distributions from our training data. The current version of ART comes with four empirical read quality score distributions, one for each of four different read lengths: 36, 44, 50 and 75 bp. The built-in insertion and deletion error rates were derived from 35 bp reads aligned with our modified ACANA alignment tool (Huang *et al.*, 2006). For paired-end simulation, ART uses two different quality score distributions and error rates for the first and second reads, each determined empirically.

Table 1. ART simulation speed. Speed measured for generating 10× read coverage of human chromosome 17, for 454, Illumina, and SOLiD technology-specific parameters

Platform	Read length	Running time (s)		Speed (no. of reads/s)	
		Single	Paired	Single	Paired
454	Varied	491	676	7,049	10,490
Illumina	50 bp	290	300	55,997	54,130
SOLiD	33 bp	728	696	33,798	33,870

2.2 454 read simulation

Roche/454 sequencing is a pyrosequencing technology that tests for the presence of each of the four DNA nucleotides (T, A, C, G) in a cyclical fashion. All consecutive bases within a homopolymer run are incorporated within a single cycle, and the read-out is an intensity signal that is proportional with the number of incorporated bases (Margulies *et al.*, 2005). The dominant error mode is base over- or under-call, resulting in INDEL type errors. While sequencing error rate only slightly increases with the number of flow cycles, the error rate increases dramatically with the frequency of long homopolymer runs. Accordingly, ART models the 454 sequencing error profile with homopolymer length-dependent over-call (insertion) and under-call (deletion) error distributions, and models base quality profiles as homopolymer length-dependent first-order Markov chains. ART uses an empirical distribution of 454 read lengths. By default, ART generates 454 reads with built-in distributions derived for the 454 GS FLX sequencer model.

2.3 SOLiD read simulation

Applied Biosystems' SOLiD sequencing technology is based on ligation of oligonucleotides. It uses four fluorescent color dyes to encode the 16 different dinucleotides, each dye encoding four dinucleotides. SOLiD performs double interrogation of each base by combining the four-dye encoding scheme with a sequencing assay that samples every base (http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html). Different from either 454 or Illumina technology, the SOLiD base caller reports nucleotide transition color codes, rather than nucleotide sequences. Accordingly, ART also generates nucleotide transition codes or 'color-space' reads. For paired-end read simulations, a Gaussian distribution is used to model the distribution of DNA fragment sizes. The built-in empirical error profiles of SOLiD reads were derived from the read data generated at Applied Biosystems. ART provides an option to tune sequencing error rates with a linear scaling factor.

2.4 Performance

To test ART's speed, we used human chromosome 17 as reference, and generated reads representing 10× coverage for each of the three sequencing platforms. The test was performed on a desktop computer with Intel Xeon 2.93 GHz CPU, running a Linux operating system. This procedure took <12 min (Table 1), with Illumina reads being the fastest and SOLiD reads the slowest.

ACKNOWLEDGEMENTS

We would like to thank Dr Heather E. Peckham at Applied Biosystems for kindly providing SOLiD read error profiles.

Funding: Intramural Research Program of the National Institutes of Health; National Institute of Environmental Health Sciences (ES101765); National Human Genome Research Institute, National Institutes of Health (HG003698 and HG004719 to G.T.M.) in part.

Conflict of interest: none declared.

REFERENCES

Balzer,S. *et al.* (2010) Characteristics of 454 pyrosequencing data–enabling realistic simulation with flowsim. *Bioinformatics*, **26**, i420–i425.

Bentley,D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, **16**, 545–552.

Durbin,R.M. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Huang,W. *et al.* (2006) Accurate anchoring alignment of divergent sequences. *Bioinformatics*, **22**, 29–34.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Margulies,M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Richter,D.C. *et al.* (2008) MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One*, **3**, e3373.