Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw286
Advance Access Publication Date: 10 May 2016
Applications Note

OXFORD

Sequence analysis

SimLoRD: Simulation of Long Read Data

Bianca K. Stöcker[1], Johannes Köster[2,3] and Sven Rahmann[1,*]

[1]Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, Essen, 45147, Germany, [2]Life Sciences, Centrum Wiskunde & Informatica (CWI), Amsterdam 1098 XG, The Netherlands and [3]Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02215, USA

*To whom correspondence should be addressed.

Received on January 26, 2016; revised on April 11, 2016; accepted on April 25, 2016

Abstract

Motivation: Third generation sequencing methods provide longer reads than second generation methods and have distinct error characteristics. While there exist many read simulators for second generation data, there is a very limited choice for third generation data.
Results: We analyzed public data from Pacific Biosciences (PacBio) SMRT sequencing, developed an error model and implemented it in a new read simulator called SimLoRD. It offers options to choose the read length distribution and to model error probabilities depending on the number of passes through the sequencer. The new error model makes SimLoRD the most realistic SMRT read simulator available.
Availability and Implementation: SimLoRD is available open source at http://bitbucket.org/genomeinformatics/simlord/ and installable via Bioconda (http://bioconda.github.io).
Contact: Bianca.Stoecker@uni-due.de or Sven.Rahmann@uni-due.de.
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Third generation sequencing technologies like SMRT (single molecule real time) sequencing are increasingly used because they yield considerably longer reads than second generation methods. The error characteristics of SMRT are fundamentally different from previous technologies: The basic error rates are higher (10–15%), but errors are considered unbiased and uniformly distributed (Eid *et al.*, 2009), which means that they can be reduced by sequencing a molecule several times. As more bioinformatics applications are developed for sequence analysis tasks from SMRT data or hybrid data, e.g. genome assembly, SNP calling, structural variant discovery, authors of such tools will benefit from read simulators that take into account the specifics of the SMRT technology. The existing manifold simulators for second generation technologies, such as ART for 454, Illumina and SOLiD reads (Huang *et al.*, 2012), do not do this.

There are few simulators designed for SMRT reads, e.g. PBSIM (Ono *et al.*, 2013), FASTQSim (Shcherbina, 2014) and Alchemy from the BLASR package (Chaisson and Tesler, 2012), the latter now being deprecated along with the .bas.h5 format. PBSIM's defaults are based on now outdated chemistry and cannot be completely re-configured. Even if the read length is adapted, the conditional read quality distribution does not match well existing data (Fig. 1 and Supplement). Also, PBSIM does not provide SAM-formatted alignments between the reference and the simulated reads. FASTQSim is a general-purpose tool for both read analysis and simulation. In particular, it provides pre-set parameters for SMRT simulation, but also allows to analyze an existing dataset with regard to its properties and simulate accordingly. However, it is unable to provide mapping information or alignments of simulated reads and it simulates reads rather slowly (8700 *N. crassa* reads with 30 cores took 90 min). The simulated length/quality distributions do not agree well with data (Fig. 1 and Supplement), and it is difficult to change parameters directly.

To improve upon the existing solutions, we developed a new read simulator called 'SimLoRD – Simulation of Long Read Data' that is convenient to use and easily re-configured when technical specifications change. The default values provide realistic simulation results according to the current state of the SMRT technology (March 2016); see Figure 1.

2 Methods

Because the sequenced DNA fragments in a SMRT library are circular with adapter sequences between forward and backward strand, a

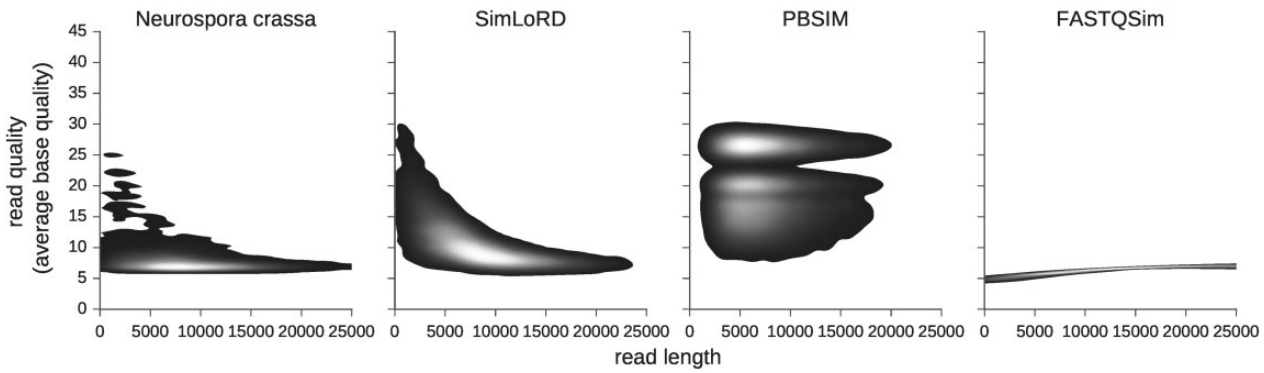

Fig. 1. Joint distribution of read length and average base quality per read on a real dataset (D1 in Table 1), in a SimLoRD simulation, in a PBSIM simulation (--length-mean =7000 and --length-sd =3000) and in a FASTQSim simulation (parameters estimated from D1)

Table 1. Datasets; see References for URLs

ID	type	organism	CCSs	subreads	URL
D1	DNA	*Neurospora crassa*	103 Mbp	982 Mbp	[a]
D2	RNA	*Homo sapiens*	481 Mbp	6 Gbp	[b]
D3	RNA	*Homo sapiens*, MCF-7 line	1.9 Gbp	15 Gbp	[c]
D4	DNA	*Caenorhabditis elegans*	350 Mbp	5 Gbp	[d]

[a]https://github.com/PacificBiosciences/DevNet/wiki/Neurospora-Crassa-(Fungus)-Genome,-Epigenome,-and-Transcriptome

[b]http://blog.pacificbiosciences.com/2014/10/data-release-whole-human-transcriptome.html

[c]http://blog.pacificbiosciences.com/2013/12/data-release-human-mcf-7-transcriptome.html

[d]https://github.com/PacificBiosciences/DevNet/wiki/C.-elegans-data-set

fragment may be sequenced multiple times in a single run. For a single pass through the sequence (*subread*), the error rate is high, but it is possible to calculate a consensus after multiple passes (*circular consensus sequence read*, CCS). Thus the error rate of CCSs decreases with the number of passes.

Each CCS is simulated as follows from a given (or randomly generated) reference genome. First, a random chromosome of the reference and a random start position are chosen. Next, a read length is chosen according to the user-specified model, usually a log-normal distribution for genomic data and an empirical distribution corresponding to library size selection for RNA-seq data. If the reference contains Ns in the relevant part, those Ns are replaced randomly in the read. This yields an error-free simulated read. To determine base qualities and error probabilities, we first draw the (fractional) number of passes over the fragment, depending on the read length, from a χ^2 distribution (details below). For example, 2.37 passes means that the whole read is read at least twice and a part (0.37 of the read) is read three times. The number of passes is used to determine final error probabilities for each base, starting from given baseline error probabilities (different for substitutions, insertions and deletion) for subreads. The read is traversed, and changes are applied for each base according to the final error probabilities. In the process, the true alignment to the reference is tracked. With probability 1/2, the completed read is reverse-complemented.

To determine appropriate distributions and parameters for the simulation, we analyzed the properties of two freely available datasets from Pacific Biosciences (D1, D2; Table 1). The identified models and parameters were then validated with two different datasets (D3, D4).

We found that the length of CCS reads has a log-normal distribution with certain parameters that are now the defaults in SimLoRD, while RNA reads are usually size selected, so their lengths should be drawn from a given empirical distribution. We also found that the number of passes p, given the read length ℓ, can be modeled by a scaled chi-squared distribution with parameters $n(\ell)$ (degrees of freedom) and scale parameter $s(\ell)$, both of which depend on the read length ℓ. The exact dependency of n and s on ℓ is documented in the Supplement. If $f_n(x) := 1/(2^{n/2}\Gamma(n/2)) \cdot x^{n/2-1}e^{-x/2}$ is the chi-squared density with n degrees of freedom, then p has scaled density $g_{n,s}(p) := f_n(p/s)/s$. With increasing p, basepair error probabilities decrease. We found that this dependency can be modeled by a noisy square root function: When ε is a basepair error probability in a subread, in a CCS it becomes becomes $\varepsilon^{\tau(p)}$ with $\tau(p) = \sqrt{p+a} - b + N$ with parameters a, b and normally distributed noise N (with additional parameters; see Supplement).

3 The SimLoRD tool

SimLoRD is a command line tool implemented in Python 3 that uses the observations above to simulate SMRT CCS reads. The only required positional argument is the path prefix of the simulated reads. The parameter -n determines the number of simulated reads. The true alignments of the simulated reads to the reference are stored in SAM format (using .sam instead of .fastq as file extension; this can be customized). The reference can be either read from a FASTA file (`-rr PATH`) or randomly generated (`-gr GC LEN`) with given GC content and length and stored.

Many parameters controlling the properties of the generated reads exist (see the Supplement for details). For choosing the read length distribution, there are four possibilities: (i) providing parameters for a log-normal distribution (`-ln SIGMA LOC SCALE`); (ii) setting a fixed read length (`-fl LEN`); (iii) sampling the read length from an existing FASTQ file (`-sf PATH`); (iv) sampling the read length from a file containing one integer per line (`-st PATH`). The baseline error probabilities for subreads can be specified individually for substitutions (`-ps`), insertions (`-pi`) and deletions (`-pd`). Consider the following example, where 10 000 reads are simulated, sampled from random positions of the reference `ref.fa` and written to `reads.fastq`. Error probabilities for subreads are 1, 12 and 2% for substitutions, insertions and deletions, respectively, on average (15% total error probability). Alignments are written to `reads.sam`. With the *Neurospora crassa* reference, this example takes 2:10 min.

```
simlord -n 10000 -rr ref.fa -pi .12 -pd .02 -ps .01 reads
```

To conclude, we have presented a Python-based read simulator (SimLoRD) whose error model corresponds to third-generation SMRT error characteristics, with default parameters based on public datasets. Relevant parameters are easily adjustable via command line arguments, so the simulator can be quickly adapted to new chemistries as they are developed. SimLoRD is convenient to install with standard Python tools and runs on all standard platforms. We hope that many researchers will benefit from the ability to generate simulated SMRT data when developing novel analysis applications.

Conflict of Interest: none declared.

References

Chaisson,M.J. and Tesler,G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**, 238.

Eid,J. *et al*. (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Huang,W. *et al*. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Ono,Y. *et al*. (2013) PBSIM: PacBio reads simulator – toward accurate genome assembly. *Bioinformatics*, **29**, 119–121.

Shcherbina,A. (2014) FASTQSim: platform-independent data characterization and in silico read generation for NGS datasets. *BMC Res Notes*, **7**, 533.