BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 11 2012, pages 1533–1535
doi:10.1093/bioinformatics/bts187

Sequence analysis

Advance Access publication April 15, 2012

pIRS: Profile-based Illumina pair-end reads simulator

Xuesong Hu[1,2,†], Jianying Yuan[1,†], Yujian Shi[1,†], Jianliang Lu[1,†], Binghang Liu[1], Zhenyu Li[1], Yanxiang Chen[1], Desheng Mu[1], Hao Zhang[1], Nan Li[1], Zhen Yue[1], Fan Bai[2], Heng Li[3] and Wei Fan[1,2,*]

[1]BGI-Shenzhen, Shenzhen 518083, [2]Biodynamic Optical Imaging Center, Peking University, Beijing, 100871 China and [3]Medical Population Genetics Program, Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The next-generation high-throughput sequencing technologies, especially from Illumina, have been widely used in re-sequencing and *de novo* assembly studies. However, there is no existing software that can simulate Illumina reads with real error and quality distributions and coverage bias yet, which is very useful in relevant software development and study designing of sequencing projects.

Results: We provide a software package, pIRS (profile-based Illumina pair-end reads simulator), which simulates Illumina reads with empirical Base-Calling and GC%-depth profiles trained from real re-sequencing data. The error and quality distributions as well as coverage bias patterns of simulated reads using pIRS fit the properties of real sequencing data better than existing simulators. In addition, pIRS also comes with a tool to simulate the heterozygous diploid genomes.

Availability: pIRS is written in C++ and Perl, and is freely available at ftp://ftp.genomics.org.cn/pub/pIRS/.

Contact: fanweisz09@gmail.com

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

Received on January 9, 2012; revised on April 3, 2012; accepted on April 10, 2012

1 INTRODUCTION

In recent years, the next-generation high-throughput sequencing technologies (NGS) have been widely used for re-sequencing and *de novo* assembly, such as the 1000 human genomes project (http://www.1000genomes.org) and the 10 000 vertebrate genomes project (http://genome10k.soe.ucsc.edu). According to information from GenomeWeb (http://www.genomeweb.com), Illumina has become dominant in the NGS market by the end of 2010.

It has been reported that the Illumina technology rarely contains insertion and deletion errors, but it may produce systematic (Dohm *et al*., 2008) and sequence-specific substitution errors (Nakamura *et al*., 2011) and coverage bias problems (Aird *et al*., 2011) during library construction and sequencing process. It is thus important to be aware of these characteristics for method development. Although, evaluating software performance on real sequencing data is preferred, not all evaluations can be conducted on real data due to the lack of a ground truth. It is frequently more convenient to test the software with simulated data from a known genome, which helps us monitoring every step.

There are already some simulators that can produce Illumina reads, such as MAQ (Li *et al*., 2008), wgsim from SAMTOOLS (Li *et al*., 2009), MetaSim (Richter *et al*., 2008) and ART (Huang *et al*., 2012). Of all the existing softwares, wgsim only simulates reads with uniform substitution errors and dummy quality values; MetaSim is designed for simulating metagenomic data, and it supports empirical error profile to generate reads without quality values; ART simulates Illumina reads with empirical quality profile, which determines quality value first and then uses the quality-derived error rate to randomly generate substitution error; MAQ adopts a first-order Markov chain to model quality distribution of each cycle, which generates quality values based on the transition probabilities and then uses the quality-derived error rate to randomly generate substitution error.

In this article, we present a new Illumina pair-end (PE) reads simulator pIRS (profile-based Illumina pair-end reads simulator), which adopts an empirical Base-Calling profile that is more like the real Illumina data. Moreover, it can use a GC content-coverage depth (GC%-depth) profile to simulate reads with coverage bias along the genome. We also provide a tool to introduce changes (substitution, insertion, deletion and other variations) to the genome, to facilitate simulation of heterozygous data.

2 DESCRIPTION

The overall workflow of pIRS is shown in Figure 1. The empirical Base-Calling profiles are generated by analyzing the alignment results of SOAP2 (Li *et al*., 2009) or SAM/BAM files (Li and Durbin, 2010) from re-sequencing data of known genomes. Only the uniquely mapped reads with full length matches are used to avoid the influence of mis-alignments. An SNP set can be optimally provided to eliminate non-error substitutions caused by true sequence variations. As the determination of a base-calling (base and quality) is highly related with the current cycle number and reference base on reads, we use a 4D distribution matrix (Dist matrix) to store the overall distribution information. The dimensions are read cycle, reference base, called base and called quality in sequential.

In the Dist matrix, base-calling is only determined by information from the current cycle. However, the called quality is often seriously affected by the quality of previous cycle. To better simulates the real

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

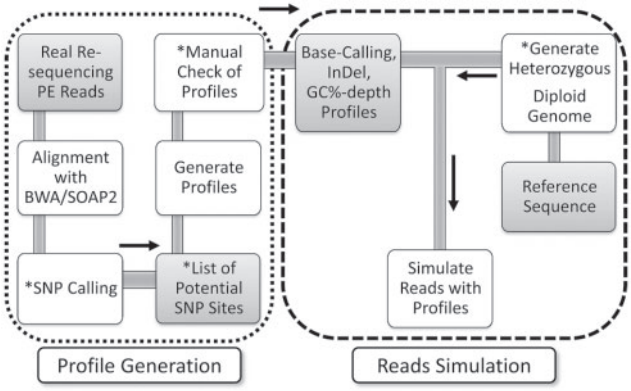

Fig. 1. The overall workflow of pIRS, which can be generally divided into two parts: profile generation and reads simulation. Optional steps are marked with *

data, we introduce a quality-transition (Qtrans) matrix based on the first-order Markov Chain similar to MAQ, in which the distribution of leading and following quality-pairs are recorded for each read cycle. The Qtrans matrix can be co-used with the Dist matrix after the first cycle, in which the called quality is determined first by the Qtrans matrix, and then the called base is determined from a subset of the Dist matrix with specified called quality. The combination of Dist and Qtrans matrics forms our Base-Calling profile.

The relationship between GC content and coverage depth is analyzed using tiling windows along the whole genome. In each window, the GC content and average base depth are calculated, respectively, then the average base depth values are grouped by GC content intervals (1%), and finally the mean depth value is calculated for each GC content interval. This is defined as the GC%-depth profile in this article, which reflects the mean coverage depth in sequence regions with similar GC content. The GC%-depth profile can then be chosen to simulate reads with coverage bias, in which the sampling probability of a read with specific GC content is proportional to the mean coverage depth within that CG content interval.

To simulate InDel errors, we provide a practical empirical InDel profile, which consists of three dimensions: read cycle, type (insertion and deletion) and length of continuous bases (1–3 bp). Our InDel profile is more comprehensive than that of other simulators, which generally use the simple uniform distributions of single-base InDel errors. Considering the fact that InDel error is quite rare in Illumina, this function is often negligible in many real applications.

Our reads simulator pIRS generates pair-end (PE) reads in the FASTQ format. For each read pair, the insert size is randomly drawn from the normal distribution with given mean and SD values. The Dist and Qtrans matrics are used to generate the called base and quality, the Indel profile is used to generate InDel errors and the GC%-depth profile is used to generate coverage bias. Note that the Qtrans, InDel and GC%-depth profiles can be optionally chosen from the pIRS parameter settings.

Since heterozygosity is an important factor for many applications, such as variation detection and *de novo* assembly, we provide a tool to randomly introduce variations including single-nucleotide polymorphism (SNP), small insertion and deletion (InDel) and structural variation (large insertion, deletion and inversion) into the given reference genome, and output the resulting haploid genome in the FASTA format. The combination of two haploid genomes form a heterozygous diploid genome and our simulator can take these two genome files together to simulate heterozygous reads data.

With pIRS, we have generated a set of empirical Base-Calling, GC%-depth and InDel profiles from our testing data, and common users can use these profiles directly to simulate Illumina reads. Only those users with special purpose need to generate profiles with new sequencing data by themselves. Note that the amount of training data should be enough to capture all the properties of real sequencing data. We suggest using ~30 G bases that is equivalent to 10 × human data.

The core programs in pIRS is written in C++ language and optimized for efficiency. The reference genomes are parsed one chromosome at a time to save memory. Both compressed (gzip) input and output files are supported to save disk space. Other programs are organized into perl pipelines to facilitate usage. The detailed methods and results for profile generation, heterozygous genome generation, reads simulation, as well as performance comparison to other existing simulators and the testing results of simulated reads in SNP calling are shown in the Supplementary Material.

3 DISCUSSION

Illumina has introduced EAMSS filtering in its CASAVA package, which masks low-quality G-rich regions at the end of each read with fixed quality value 'B'. It is mainly designed for easier data trimming in general re-sequencing analyses, but this will slightly influence our profile generation. We suggest all quality-aware analysis to run CASAVA with '--no-eamss' to disable the EAMSS filtering. The current Base-Calling and InDel profiles in pIRS focus on the per-read-cycle properties but do not take the sequence-structure-specific errors into consideration. In Illumina sequencing, coverage bias is a complex problem and GC-content is a well-studied one of all the influencing factors. We will introduce more types of coverage biases once they are well-studied.

4 CONCLUSION

We designed and implemented an effective Illumina reads simulator pIRS using empirical profiles, to reproduce reads that are more like real Illumina data compared with existing simulators, which is likely to be very helpful for development of NGS software such as *de novo* assembly, SNP calling and structural variation detection. The independent profile generating module in pIRS grants great freedom to users, who can generate new profiles with their own sequencing data when machine or reagent updates. Moreover, the additional tool for simulating heterozygous genome is especially useful for applications that need heterozygous data.

Funding: Basic Research Program of Shenzhen City (grants JC2010526019); Key Laboratory Project of Shenzhen City (grants CXB200903110066A and CXB201108250096A); Shenzhen Key Laboratory of Gene Bank for National Life Science.

Conflict of Interest: none declared.

REFERENCES

Aird,D. *et al*. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.

Dohm,J.C. *et al*. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.

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

Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**, 589–595.

Li,H.*et al*. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Li,H. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,R. *et al*. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Nakamura,K. *et al*. (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, **39**, e90.

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