BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 8 2013, pages 1076–1077
doi:10.1093/bioinformatics/btt074

Sequence analysis

Advance Access publication February 14, 2013

Wessim: a whole-exome sequencing simulator based on *in silico* exome capture

Sangwoo Kim[1,*], Kyowon Jeong[2] and Vineet Bafna[1,*]

[1]Department of Computer Science and Engineering and [2]Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093, USA

Associate Editor: Michael Brudno

ABSTRACT

Summary: We propose a targeted re-sequencing simulator Wessim that generates synthetic exome sequencing reads from a given sample genome. Wessim emulates conventional exome capture technologies, including Agilent's SureSelect and NimbleGen's SeqCap, to generate DNA fragments from genomic target regions. The target regions can be either specified by genomic coordinates or inferred from *in silico* probe hybridization. Coupled with existing next-generation sequencing simulators, Wessim generates a realistic artificial exome sequencing data, which is essential for developing and evaluating exome-targeted variant callers.

Availability: Source code and the packaged version of Wessim with manuals are available at http://sak042.github.com/Wessim/.

Contact: sak042@cs.ucsd.edu or vbafna@cs.ucsd.edu

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

Received on December 13, 2012; revised on February 3, 2013; accepted on February 7, 2013

1 INTRODUCTION

Generation of simulated next-generation sequencing (NGS) reads serves the essential gold standard in developing and evaluating variant detection algorithms. Many tools, including *wgsim* (Li *et al*., 2009), have been developed for artificial NGS read generation. Recent studies, such as *ART* (Huang *et al*., 2012), *pIRS* (Hu *et al*., 2012) and *GemSim* (McElroy *et al*., 2012), have accomplished more realistic simulation by reproducing known biases coming from sequence context and empirical platform-dependent errors.

Although the current simulators mainly focus on realistic NGS read generation, we address another important problem on specifying target regions. Whole-exome sequencing (WES) is currently being regarded as a superior option for disease variant finding studies; it is cost-effective, much smaller in size and easier to interpret results. As typical statistics of whole-exome sequencing data (e.g. coverage, read distribution and bias) are distinct from that of whole-genome sequencing data, variant detection tools are usually required to calibrate their algorithms for practical use of exome data (Krumm *et al*., 2012; Sathirapongsasuti *et al*., 2011). This naturally demands more realistic simulation of exome sequencing to promote accuracy of variant calling tools by providing a statistical base for performance evaluation.

Wessim emulates the exome capture procedure that forms the basis of major commercial solutions (such as Agilent's SureSelect and Roche/NimbleGen's SeqCap) by implementing (i) DNA shearing to generate random fragments, (ii) probe capture by hybridization and (iii) single- or paired-end sequencing of the selected fragments. Other important features including fragment length and GC-content were rigorously considered to reproduce more accurate coverage biases. We compared our synthetic data with real WES data to confirm the similarity of major statistics. The exome capture process is highly optimized so that the overall running time is only bounded by the NGS read generation step.

2 METHODS

Wessim takes (i) a FASTA file of sample genome sequence such as the human reference assembly and (ii) genomic target regions. Wessim first generates random DNA fragments from designated target regions, which is specified either by a BED file that contains target regions' coordinates or a set of probe sequences that are used for capturing fragments. Each fragment is further filtered by length and GC-content to reproduce potential biases. Finally, NGS reads are generated from selected DNA fragments using major emulated platforms.

2.1 Generation of DNA fragments

We define a DNA fragment $f = (c_f, s_f, e_f)$ of length $L(f)$ and sequence $S(f)$, where c_f, s_f and e_f are the chromosome, start and end position of f's genomic origin, respectively. Two distinct approaches for the fragment generation are described later in the text.

Ideal target approach: Each exome capture platform manifests its own *ideal* target regions. For example, Agilent's SureSelect All Exon V4 targets ~186 kb exonic regions. All BED files are freely available online from vendor's websites. In this approach, each DNA fragment f is generated *within* a randomly selected target region $R = (c, s, e)$, where c is the chromosome, s and e are start and end position, respectively. The probability of selecting each target region R is proportional to its length $e - s + 1$. A slack boundary variable ξ allows a fragment to be generated from an extended interval $(c, s - \xi, e + \xi)$.

Probe hybridization approach: This approach implements a probe-level capture procedure for more realistic DNA fragment generation. Agilent's probe sequences are available at the SureDesign website (https://earray.chem.agilent.com/suredesign/); probe sequence of NimbleGen is not publicly available at this time.

Given an oligonucleotide probe p of sequence $S(p)$, we first retrieve p's *hybridizable regions* $h_i^p \in H^p$, where each sequence $S(h_i^p)$ matches $S(p)$ with a *good* score (e.g. sequence identity $\geq 95\%$). The probability of

*To whom correspondence should be addressed.

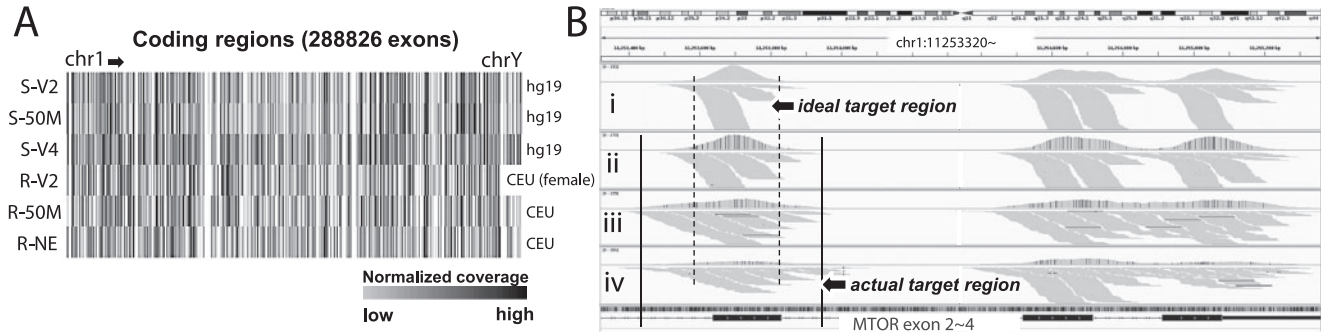

Fig. 1. Simulated exome sequencing by Wessim. (**A**) Read coverage of simulated (upper 3 rows) and real (lower 3 rows) exome sequencing data over 288 826 consensus coding exons are depicted. Wessim's probe-based simulation reproduced the coverage distribution of real data with high correlation (Pearson's $r = 0.44 \sim 0.88$). S: simulated, R: real data; V2, 50M, V4: SureSelect Human All Exon V2, 50M and V4; NE: NimbleGen SeqCap 44M. (**B**) Exon-level read distribution of three simulated (*i*, *ii* and *iii*) and one real data (*iv*). Ideal target approach (*i*) generates more tightly bounded fragments to the designated regions, whereas probe-based approach (*ii*: single-end and *iii*: paired-end) allows more relaxed fragment ranges like in the real data

selecting h_i^p is inversely proportional to the number of mismatches between $S(h_i^p)$ and $S(p)$. To generate a fragment, we first choose a random probe p_x from the entire probe set and select a random hybridizable region h from H^{p_x}. A fragment f can only be generated when more than a certain fraction of f overlaps the selected hybridizable region, which we defined as a minimum overlap ratio b_0. The probability of generating a fragment f from h can be calculated by:

$$P(f) = \max\left(P(h|p_x)P(p_x)\frac{b - b_0}{1 - b_0}, 0\right), \quad (1)$$

where $P(p_x)$ is the probability of selecting a probe p_x, $P(h|p_x)$ is the conditional probability of selecting h from H^{p_x} given p_x and b is the fraction of hybridizable region in f.

2.2 Reproducing bias

To emulate the fragment bias on GC-content and fragment length, we implemented a filter as follows. Denote the fractional GC-content as $G(f)$ and the joint distribution of $L(f)$ and $G(f)$ as $d_{L,G}$. Given f, such that $L(f) = \ell$ and $G(f) = g$, we retain f with the probability $\frac{\gamma(\ell,g)}{\max_{i,j}\gamma(i,j)}$ where $\gamma(\ell,g) := \frac{d'_{L,G}(\ell,g)}{d_{L,G}(\ell,g)}$ and $d'_{L,G}$ is the observed distribution as computed by Benjamini and Speed (2012).

2.3 Sequencing fragments

Generating and sequencing DNA fragments are two separate processes. This ideally enables various existing NGS simulators to be incorporated with Wessim as an independent module. Here, we used an advanced NGS simulator GemSim, which widely supports benchmarked empirical error models of major sequencing platforms (e.g. Solexa, SOLiD and 454). We downloaded and modified the source code of GemSim to convert input unit from genome to fragment while maintaining its core error models.

3 RESULT ANALYSIS

We generated simulated exome data using three different platforms (SureSelect V2, 50M and V4) in two different modes (ideal target and probe hybridization) to compare with real data (Fig. 1A). The read coverage of our simulated data over 288 826 consensus coding exons showed a highly correlated $r = 0.44 \sim 0.88$) pattern with that of the real data. We also confirmed that the probe hybridization approach can reproduce a realistic *per exon* read distribution, whereas ideal target approach did not (refer Fig. 1B and Supplementary Figs S5 and S6 for full results). In a performance test, Wessim could generate 226∼316 kb of reads per second including fragment generation and filtering in an 4 core Intel i7-2600K system. This corresponds to ∼5.45 h of running time for generating 66 million reads (∼two GAIIx lanes).

ACKNOWLEDGEMENT

The authors thank Dr Terry Speed and Yuval Benjamini for providing their GC-content bias data.

Funding: National Institute of Health (U54-HL108460, 5R01-H6004962); National Science Foundation (CCF-1115206); National Institute of Child Health and Human Development (1P01HD070494-01).

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, e72.

Hu,X. *et al.* (2012) pIRS: profile-based illumina pair-end reads simulator. *Bioinformatics*, **28**, 1533–1535.

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

Krumm,N. *et al.* (2012) Copy number variation detection and genotyping from exome sequence data. *Genome Res.*, **22**, 1525–1532.

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

McElroy,K. *et al.* (2012) GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*, **13**, 74.

Sathirapongsasuti,J.F. *et al.* (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, **27**, 2648–2654.