HHS Public Access

Author manuscript

IEEE EMBS Int Conf Biomed Health Inform. Author manuscript; available in PMC 2016 August 14.

Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2016 February ; 2016: 74–77. doi:10.1109/BHI.2016.7455838.

Evaluating the impact of sequencing error correction for RNA-seq data with ERCC RNA spike-in controls

Li Tong [Student Member, IEEE],

Dept. of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA

Cheng Yang [Student Member, IEEE],

Dept. of Biomedical Engineering, Peking University, No.5 Yiheyuan Road Haidian District, Beijing, P.R. China 100871

Po-Yen Wu [Student Member, IEEE], and

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

May D. Wang [Senior Member, IEEE]

Dept. of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA

Li Tong: ltong9@gatech.edu; Cheng Yang: ycheng@gatech.edu; Po-Yen Wu: pwu33@gatech.edu

Abstract

Sequencing errors are a major issue for several next-generation sequencing-based applications such as de novo assembly and single nucleotide polymorphism detection. Several error-correction methods have been developed to improve raw data quality. However, error-correction performance is hard to evaluate because of the lack of a ground truth. In this study, we propose a novel approach which using ERCC RNA spike-in controls as the ground truth to facilitate error-correction performance evaluation. After aligning raw and corrected RNA-seq data, we characterized the quality of reads by three metrics: mismatch patterns (i.e., the substitution rate of A to C) of reads aligned with one mismatch, mismatch patterns of reads aligned with two mismatches and the percentage increase of reads aligned to reference. We observed that the mismatch patterns for reads aligned with one mismatch are significantly correlated between ERCC spike-ins and real RNA samples. Based on such observations, we conclude that ERCC spike-ins can serve as ground truths for error correction beyond their previous applications for validation of dynamic range and fold-change response. Also, the mismatch patterns for ERCC reads aligned with one mismatch can serve as a novel and reliable metric to evaluate the performance of error-correction tools.

INTRODUCTION

Next-generation sequencing (NGS) has enabled unbiased high-throughput analysis for biological systems. The fast development of sequencing techniques has boosted the applications of RNA-seq, a major branch of NGS, to biomedical research and clinical practice [1] [2] [3]. For most RNA-seq applications, the reliability and accuracy of results are influenced by sequencing biases such as coverage variation, batch effect, background noise, and sequencing errors [4]. Among all sequencing biases, sequencing errors have been identified as a major issue for accurate SNP detection and de novo assembly [5].

Many mechanisms have been proposed to explain the occurrence of sequencing errors. Depending on the cause, Meacham et al. classified sequencing errors into three categories: random errors, sequence-specific errors, and systematic errors [6]. Random errors refer to lower sequencing qualities in Illumina sequencers and often occur at later sequencing cycles. Lagging-strand dephasing caused by the incomplete extension of a template ensemble has been suggested as one main reason for this type of errors [5]. Sequence-specific errors refer to errors caused by a continuity of the same nucleotide or any other low complexity regions. Lastly, systematic errors often relate to specific motif patterns (e.g., inverted repeats and GGC sequence) that may lead to dephasing by either inhibiting single-base elongation with folded single-stranded DNA or altering enzyme preference [5]. The sequencing errors can also be broadly classified into four categories, including substitutions, insertions, deletions, and ambiguous calling, depending on their impact on the reading of nucleotide sequences. Substitutions are the most prominent errors in Illumina platform while insertions and deletions are more common in platforms such as 454 and Ion Torrent [7].

A lot of error correction methods have been developed to ameliorate the impact of sequencing errors on downstream RNA-seq applications. As an example, Le et al. have successfully improved the accuracy of de novo transcriptome assembly by applying error correction to sea cucumber RNA-seq data [8]. However, error correction is still seldom applied in real practice. The gap between the development and employment of error-correction tools may result from inadequate evaluation of their performance and advantages. Several studies have used real RNA-seq data for species with the well-characterized genome to evaluate the performance of some error-correction tools [7], [9]. However, error-correction performance should be data-dependent. Thus, results derived from a specific benchmark dataset may not be directly applicable to other datasets. In this study, we propose to use ERCC spike-ins, a set of synthetic RNAs with known sequences, as ground truth to facilitate the evaluation of error-correction performance. We aim to show that error-correction performance of ERCC spike-ins can serve as an indicator for the performance of the entire data. With ERCC spike-ins as the external reference, error-correction methods can be applied with a built-in quality control reference.

METHODS

A. Dataset

The RNA-seq dataset we used in this study is part of the Sequencing Quality Control (SEQC) project (NCBI SRA accession number: SRP025982), which aimed to

comprehensively assess RNA-seq accuracy and reproducibility [10]. The dataset contains two standard human RNA samples, Universal Human Reference RNA (UHRR) and Human Brain Reference RNA (HBRR), each of which was mixed with ERCC RNA spike-in controls. ERCC spike-ins are artificially synthesized RNAs, whose structures are relatively simple. Also, the nucleotide sequences of ERCC are known and publicly available [11]. In previous studies, ERCC spike-ins have been used as the ground truth to assess dynamic range of expression, accuracy of fold-change recovery, and lower limit of detection for several sequencing platforms [12]. With known sequences, we propose that ERCC spike-ins can also be applied to examine sequencing errors that occur in RNA-seq experiments. In this study, we use four replicates of the UHRR sample that have been mixed with ERCC spike-ins. These samples were sequenced by the Illumina HiSeq 2000 instrument at the Beijing Genomics Institute (BGI).

B. Performance Evaluation

The overall pipeline for evaluating error-correction performance for RNA-seq data with ERCC spike-ins is shown in Figure 1. We apply three sequencing error-correction tools to raw RNA-seq data before sequence alignment. Depending on algorithmic design, error-correction tools can be categorized into four types: k-mer spectrum-based tools; suffix tree/array-based tools; multiple sequence alignment-based tools; and hidden Markov model-based tools.

In this study, we choose to use three error-correction tools representing three different types of algorithms: Musket [13] for k-mer spectrum-based tools; Coral [14] for multiple sequence alignment-based tools; and SEECER [8] for hidden Markov model-based tools. These three tools have been compared between each other in previous studies [7] [8] [9]. By evaluating these three tools together in this study, we can compare the evaluation results between previous metrics and the metrics proposed in this paper.

All tools are run with default parameter settings. After error correction, we use TopHat [15] to align both the raw and the corrected sequence reads to ERCC reference sequences as well as the hg38 reference human genome [16].

1) Mismatch Rate—We calculate the number of mismatches between aligned reads and reference sequences as the first evaluation metric. The mismatch rate is defined as the number of aligned reads with a specific mismatch pattern (e.g., A→C, A→G, and G→T) over the total number of aligned reads. We only allow up to two mismatches during the alignment. The previous studies [17] indicates that reads with one mismatch have better alignment results, so we further divide all reads with mismatches into reads aligned with one mismatch and reads aligned with two mismatches and characterize the mismatch patterns respectively.

The mismatch rate can directly capture the true sequencing errors for ERCC samples because the ERCC reference sequences are artificially synthesized RNAs with no junctions. So the mismatches will not be confounded with SNPs or alignment errors. But these mismatches can only be a rough approximation for human RNA samples because the mismatches may result from not only sequencing errors but also SNPs and alignment errors.

However, the overall SNP density is about 5 per 10 kb on exon in human [18], which is lower than the overall error rate of Illumina. Also, the human transcriptome is much more complicated; the reads could be mapped to suboptimal locations. If a read aligns to an incorrect location, it is impossible to differentiate mismatches caused by alignment from true sequencing errors. Therefore, mismatches in reads aligned to the hg38 genome confound with both alignment errors and SNPs, and thus, they can only be a rough approximation of true sequencing errors. But using reads aligned with one mismatch could potentially minimize the influence of aligners.

2) Percentage of Reads Aligned—Besides the mismatch rate, we also compute the percentage of reads aligned to the reference sequences as the third evaluation metric. In sequence alignment, we allow up to two mismatches per read, and thus, reads with more than two errors may not be aligned properly or cannot be aligned. With error correction, we expect to observe more reads aligned to the reference sequences because some reads that previously contained more than two errors may be able to align to the reference sequences after error correction.

RESULTS

A. Mismatch Rates

For all scenarios, error correction effectively reduced the mismatch rates (Figures 2 and 3) with SEECER always achieved the lowest mismatch rates, followed by Musket and Coral. In fact, Coral corrected much less sequencing errors than SEECER and Musket. Note that we used default parameter settings for these error-correction tools, so the performance of these tools may be further improved after optimizing parameters for each tool.

B. Percentage of Reads Aligned

Error correction also increased the percentage of reads aligned to the reference sequences, which held true for both ERCC and hg38 cases (Figure 4). SEECER and Musket showed a significant increase in the percentage of reads aligned with SEECER outperformed Musket in the hg38 case, and Musket outperformed SEECER in the ERCC case. Coral only slightly increased the percentage of reads aligned to the ERCC reference but had a relatively better performance for the hg38 case.

C. Correlated Performance Between ERCC and Hg38

For reads aligned to ERCC reference and reads aligned to hg38 reference, the mismatch patterns of the previous two metrics are similar. We further examine the linear correlation between mismatch patterns in Figure 2 and Figure 3 for reads aligned with one mismatch and reads aligned with two mismatches respectively. We found that a significant positive correlation (correlation coefficients $r > 0.65$, $p < 0.05$) exists between the mismatch patterns in reads aligned with one mismatch to the ERCC reference and those to the hg38 reference (Table I). Such positive correlation demonstrated that error-correction performance for ERCC spike-ins may be the proxy of the error-correction performance for the entire dataset.

For reads with two mismatches, a slight correlation exists but not significant (Table I). We hypothesized that the two-mismatch alignment is more likely to be suboptimal alignment for the reads. The errors introduced by alignment can be further confounded with the sequencing errors. Therefore, mismatch rates in aligned reads with one mismatch would be a better choice for evaluating error-correction performance based on our experiments.

CONCLUSIONS

In this study, we used the external reference ERCC spike-ins as the ground truth for evaluating error-correction performance. Based on the significant linear correlation between mismatch patterns of reads aligned with one mismatch before and after error correction, we conclude that the mismatch patterns for reads aligned with one mismatch can serve as a good indicator for evaluating error-correction performance. Based on this metric, among the three tools we evaluated in this study, SEECER has the best performance, which is consistent with previous studies [7] [8] [9].

There are multiple advantages of this metric. First, ERCC spike-ins have been widely used as the ground truth for assessing the accuracy of fold-change recovery. We have demonstrated another potential application of ERCC spike-ins as the indicator for error-correction performance. Second, ERCC sequences are simple and short, which enables fast and efficient evaluation of error-correction performance for the whole dataset.

DISCUSSIONS

For the four replicates of the UHRR samples we analyzed, we found that the mismatch rates are highly correlated between each pair of replicates. One reasonable hypothesis is that these errors are systematic errors. Thus, miscalls may happen more often for certain bases than the others. Similar observations have been reported in previous studies and have been identified as systematic errors caused by sequence-specific variances. Meacham *et al.* reported that T-to-G was the most prominent mismatch pattern while A-to-G was the most prominent pattern we discovered [6]. Further investigation is needed to identify the cause of these specific patterns so as to improve error-correction performance by integrating knowledge of sequence-specific variances. With these premises, one future direction of this study is to distinguish random errors from systematic errors. We plan to include different samples sequenced at different sequencing sites, which may enable the comparison between systematic errors and sequence-specific errors. By identifying systematic errors, we may be able to target them during error correction, and thus, further improve the error-correction performance.

In this study, we mainly used the external reference ERCC spike-ins as the ground truth for evaluating error-correction performance. Because of SNPs and the complexity of the human transcriptome, we cannot simply take the human reference genome as a ready-to-use ground truth for evaluating error-correction performance. Although we used the hg38 reference genome as the pseudo truth to estimate sequencing errors, the results are not exactly the same as what we observed in ERCC spike-ins. One way to possibly resolve this issue is to establish a reliable internal reference using low SNP density regions in the human

transcriptome to minimize the influence of SNPs. In this way, we can realize error-correction performance evaluation for samples without ERCC spike-ins. Also, the more complicated structures on low SNP density regions could be more representative of the error-correction performance in the whole sequencing data.

References

1. Oszolak F, Milos PM. Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*. 2011; 12(2):87–98.
2. Wang Z, Gerstein M, Snyder M. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009; 10(1):57–63.
3. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*. 2008; 5(7):621–628. [PubMed: 18516045]
4. Taub MA, Corrada Bravo H, Irizarry RA. Overcoming bias and systematic errors in next generation sequencing data. *Genome Med*. 2010; 2(12):87. [PubMed: 21144010]
5. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, et al. Sequence-specific error profile of illumina sequencers. *Nucleic acids research*. 2011:gkr344.
6. Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L. Identification and correction of systematic error in highthroughput sequence data. *BMC bioinformatics*. 2011; 12(1):451. [PubMed: 22099972]
7. Yang X, Chockalingam SP, Aluru S. A survey of error correction methods for next-generation sequencing. *Briefings in bioinformatics*. 2013; 14(1):56–66. [PubMed: 22492192]
8. Le H-S, Schulz MH, McCauley BM, Hinman VF, Bar- Joseph Z. Probabilistic error correction for rna sequencing. *Nucleic acids research*. 2013:gkt215.
9. Molnar M, Ilie L. Correcting illumina data. *Briefings in bioinformatics*. 2014:bbu029.
10. S.-I. Consortium. et al. A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature biotechnology*. 2014; 32(9):903–914.
11. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, Causton H, Conley MP, Elespuru R, Fero M, et al. The external rna controls consortium: a progress report. *Nature methods*. 2005; 2(10):731–734. [PubMed: 16179916]
12. Munro SA, Lund SP, Pine PS, Binder H, Clevert D-A, Conesa A, Dopazo J, Fasold M, Hochreiter S, Hong H, et al. Assessing technical performance in differential gene expression experiments with external spike-in rna control ratio mixtures. *Nature communications*. 2014; 5
13. Liu Y, Schröder J, Schmidt B. Musket: a multistage k-mer spectrum-based error corrector for illumina sequence data. *Bioinformatics*. 2013; 29(3):308–315. [PubMed: 23202746]
14. Salmela L, Schröder J. Correcting errors in short reads by multiple alignments. *Bioinformatics*. 2011; 27(11):1455–1461. [PubMed: 21471014]
15. Trapnell C, Pachter L, Salzberg SL. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*. 2009; 25(9):1105–1111. [PubMed: 19289445]
16. Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. Centromere reference models for human chromosomes x and y satellite arrays. *Genome research*. 2014; 24(4):697–707. [PubMed: 24501022]
17. Yang, C.; Wu, P-Y.; Tong, L.; Phan, J.; Wang, M. Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics. ACM; 2015. The impact of rna-seq aligners on gene expression estimation; p. 462-471.
18. Zhao Z, Fu Y-X, Hewett-Emmett D, Boerwinkle E. Investigating single nucleotide polymorphism (snp) density in the human genome and its implications for molecular evolution. *Gene*. 2003; 312:207–213. [PubMed: 12909357]

Fig 1.
The pipeline for evaluating error-correction performance with ERCC spike-ins.

Fig 2.

The distribution of mismatch rates for reads with one mismatch (panel a) or two mismatches (panel b) after aligning the raw and the three corrected sequencing data to the ERCC reference sequences.

Fig 3.

The distribution of mismatch rates for reads with one mismatch (panel a) or two mismatches (panel b) after aligning the raw and the three corrected sequencing data to the hg38 reference human genome.

Fig 4.

The increase rate of reads aligned to ERCC reference (a) and hg38 human genome reference (b) respectively before and after correction with three different tools.

Table I

Correlation of mismatch patterns

One Mismatch	Correlation(r)	P-value(p)
SEECER	0.6758	0.0021
Musket	0.7748	0.0031
Coral	0.7822	0.0026
Two Mismatches	Correlation Coefficient (r)	P-value (p)
SEECER	0.1244	0.7001
Musket	0.0563	0.8619
Coral	0.3174	0.3147