Briefings in Bioinformatics Advance Access published May 22, 2013

doi:10.1093/bib/bbt010

DRISEE overestimates errors in metagenomic sequencing data

A. Murat Eren, Hilary G. Morrison, Susan M. Huse and Mitchell L. Sogin

Submitted: 7th December 2012; Received (in revised form): 5th February 2013

Abstract

The extremely high error rates reported by Keegan *et al.* in 'A platform-independent method for detecting errors in metagenomic sequencing data: DRISEE' (*PLoS Comput Biol* 2012;**8**:e1002541) for many next-generation sequencing datasets prompted us to re-examine their results. Our analysis reveals that the presence of conserved artificial sequences, e.g. Illumina adapters, and other naturally occurring sequence motifs accounts for most of the reported errors. We conclude that DRISEE reports inflated levels of sequencing error, particularly for Illumina data. Tools offered for evaluating large datasets need scrupulous review before they are implemented.

Keywords: *next-generation sequencing; sequencing error; adapter ligation; PCR; quality score*

INTRODUCTION

Error identification and correction in high-throughput sequencing datasets, especially at the single read level, have been addressed by many investigators [1–18]. Many approaches use platform-dependent quality scores, read consensus or k-mer analysis. Recently, Keegan *et al.* [19] described DRISEE, a method to assess quality of genomic and metagenomic next-generation sequencing runs. The authors analysed numerous publicly available datasets with DRISEE and reported widely variable levels of sequencing errors, generally far higher than other published estimates [1, 20–22].

DRISEE bases its error estimates on variation from a consensus sequence in bins of artificially duplicated reads (ADRs). DRISEE assumes that prior to sequencing, over-amplification from a given start point in the template leads to formation of ADRs, and that sequencing error, not naturally occurring sequence diversity, accounts for sequence variation within an ADR bin. An ADR bin consists of all reads starting with an identical prefix, by default the first 50 nt of the read.

DRISEE as described might provide an improved method for estimating sequencing errors than the platform-based quality scores; however, the authors failed to carefully examine the origins of ADR bins. DRISEE analyses all reads except those that contain ambiguous bases. The authors correctly note, 'Bins can be screened for eukaryotic content, sequences with low complexity, and/or known sequences that may exhibit an unusually high level of biological repetition (16s rRNA-based, sequences with low complexity, eukaryotic sequences etc.). Bins that contain such sequences should be excluded from further consideration'. However, the Supplemental Methods in the DRISEE manuscript reveal that the authors did not exclude such reads.

Widespread Illumina adapter contamination

We obtained from the NCBI Sequence Read Archive (SRA) the 12 metagenomic datasets that were used in the original publication to generate Figure 4b. DRISEE error estimation demonstrated

Corresponding author. Mitchell L. Sogin, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, MBL, Woods Hole, MA 02543-1015, USA. Tel: +1-508-289-7246; Fax: +1-508-457-4727; E-mail: sogin@mbl.edu

A. Murat Eren is a computer scientist and microbial ecologist currently doing a post-doctoral fellowship at the MBL with M.L.S. Eren has extensive experience with the analysis of next-generation sequencing datasets.

Hilary Morrison is a Senior Research Scientist at the MBL developing and using next-generation sequencing techniques for projects ranging from microbial community structures to whole-genome sequencing.

Susan Huse holds a joint appointment with the MBL and the Alpert Medical School at Brown University and studies the importance of microbial communities to human health using next-generation sequencing data.

Mitchell L. Sogin is the director of the Josephine Bay Paul Center for Comparative Molecular Biology and Evolution at the MBL and originated the use of next-generation sequencing technology for studying microbial diversity.

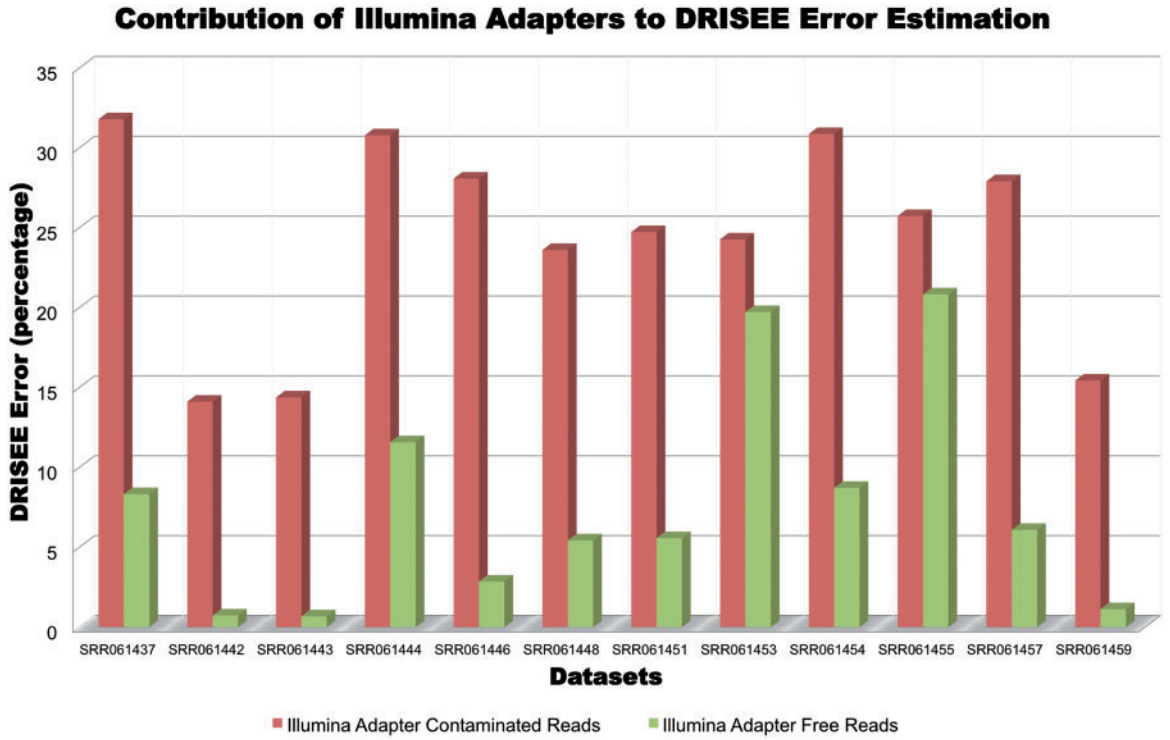

Figure 1: Change in DRISEE error estimation for reads with and without Illumina adapter contamination for all 12 datasets that were used in the original publication to demonstrate how DRISEE error profiles differ markedly from quality scores.

a significant discrepancy from the quality scores reported by the Illumina platform. Our analysis of DRISEE-generated ADR bins with ≥20 reads showed that Illumina adapter sequences drive the formation of these bins. This 65-nt adaptor sequence usually occurs upstream of the sequencing primer-binding site. Unfortunately, Illumina adapter artifacts sometimes contaminate libraries. Unless they are filtered out or trimmed, reads starting with Illumina adapters will present identical 65-nt prefixes at the start of the read and create a spurious ADR bin. DRISEE interprets the actual biological variation that follows the adapter sequences in these bins as extensive sequencing error.

With DRISEE (version 1.2), we re-analysed the 12 datasets, identifying reads as 'adapter contaminated' if they presented at least 15 nt perfect identity to the Illumina adapter sequences in the first 50 nt (see Supplemental Methods). Figure 1 shows the marked difference in error estimation for reads with and without Illumina adapters. Although Keegan *et al.*'s [19] claim that the true error rates are higher than reported in the quality scores may be correct, the exceedingly high error rates presented in Figure 4b from the original publication reflect the presence of untrimmed Illumina adapter sequences and do not support their claims.

Spurious ADR bins caused by adaptor sequences differ markedly from valid bins in the magnitude of errors and their distribution by nucleotide position. Figure 2 shows DRISEE output from individual large bins from dataset SRR061459. The adapter-generated bin exhibits error greater than zero at all positions following the prefix and the average error greatly exceeds that of the valid ADR bin.

Low-complexity and conserved gene reads

Analysis of all 10 Illumina genomic datasets, as well as 10 randomly chosen Illumina metagenomic runs from Keegan *et al.*'s Figure 3 [19], detected significant Illumina adapter contamination and a high proportion of low-complexity reads in all datasets, both of which generated spurious bins that inflated DRISEE error drastically. Table 1 demonstrates the inflation of DRISEE error for one of these datasets chosen randomly (SRA accession SRR061488). Genes with conserved regions followed by biological variation that commonly occurs in both bacterial and eukaryotic genomes can create bins large enough

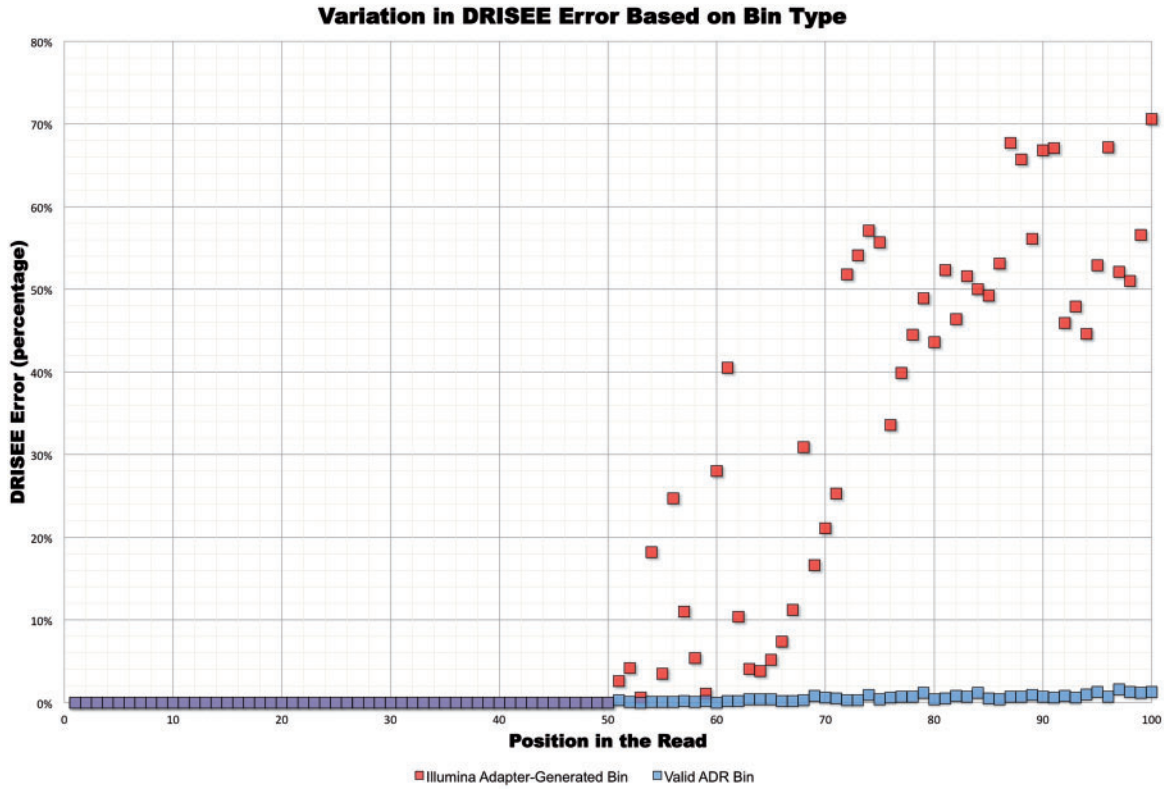

Figure 2: DRISEE error by position. The largest bin contained 15 264 reads and the prefix appeared to be a true ADR (bacterial genomic sequence). The per cent error at each position is plotted on the *y*-axis (light blue). Scores for an adapter-generated bin with 8177 reads are shown for comparison (dark red).

```
99 TTGTATTGTTATTAACTCTTTCTTCAAATCGTAGTCCTTAAGAACAGTAT
68 ACTTATAGTGATTACATAACACAATCATCTTCTATCAGTTGTATGACTAC
64 TTATAGTGATTACATAACACAATCATCTTCTATCAGTTGTATGACTACAA
62 ATAGAGACTTATAGTGATTACATAACACAATCATCTTCTATCAGTTGTAT
58 TTGTGTTATGTAATCACTATAAGTCTCTATCGTAGAGTATAGAAGATTGA
57 GTGATTACATAACACAATCATCTTCTATCAGTTGTATGACTACAACTGAT
56 CGATAGAGACTTATAGTGATTACATAACACAATCATCTTCTATCAGTTGT
54 GTGTTATGTAATCACTATAAGTCTCTATCGTAGAGTATAGAAGATTGAGT
53 ATAGTGATTACATAACACAATCATCTTCTATCAGTTGTATGACTACAACT
51 TATAGTGATTACATAACACAATCATCTTCTATCAGTTGTATGACTACAAC
48 TACGATAGAGACTTATAGTGATTACATAACACAATCATCTTCTATCAGTT
48 ATCTTCTATCAGTTGTATGACTACAACTGATTTCTTTTGGATACCCAAAA
46 GATAGAGACTTATAGTGATTACATAACACAATCATCTTCTATCAGTTGTA
46 ATTGTATTGTTATTAACTCTTTCTTCAAATCGTAGTCCTTAAGAACAGTA
46 AGACTTATAGTGATTACATAACACAATCATCTTCTATCAGTTGTATGACT
45 TTGTGATGTTATTAACTCTTTCTTCAAATCGTAGTCCTTAAGAACAGTAT
44 ATTGTGTTATGTAATCACTATAAGTCTCTATCGTAGAGTATAGAAGATTG
44 AGAGACTTATAGTGATTACATAACACAATCATCTTCTATCAGTTGTATGA
42 TGATTACATAACACAATCATCTTCTATCAGTTGTATGACTACAACTGATT
42 CTTATAGTGATTACATAACACAATCATCTTCTATCAGTTGTATGACTACA
```

Figure 3: Some of the motifs that generated invalid bins for dataset 4441625.3. The first 20 largest bins are shown. The first column is bin size and the second is the 50-nt prefix. Similar motifs are shown using the same font colour.

Table I: Change in DRISEE error estimation for SRR061488 after removing adapter-contaminated and low-complexity bins from the analysis

Category	Number[a]	DRISEE error (%)
All bins	4766	39.9
Adapter-contaminated bins	1645	45.3
Low-complexity bins	2718	34.8
Remaining bins	403	6.6

[a]Number of bins containing ≥20 reads and no ambiguous bases in their prefixes.

to be considered by DRISEE and inflate overall error estimations. For instance, 74 of 403 bins in SRR061488 derive from the 16S rRNA gene.

Platform-specific error

Keegan *et al.* [19] also report a striking difference in error rates between 454 and Illumina datasets. As we have shown, contaminating adapter sequences account for much of the DRISEE error in Illumina datasets. We next analysed 55 of the 65 Roche/454 metagenomic datasets used to generate Figure 3 in the DRISEE manuscript (the other 10 datasets were not available in MG-RAST or SRA).

Our analysis showed that while adapter contamination is rare in 454 data, the 50-nt prefixes from 34 of the datasets were dominated by similar sequence motifs from sources we could not identify (see Supplemental Methods). Figure 3 exemplifies some of these motifs in one dataset (MG-RAST ID 4441625.3). Identical motifs in multiple datasets from the same research project suggest a library preparation artifact. Bins from another eight datasets had low-complexity, repetitive sequence prefixes. Whole genome amplification provided material for at least six of these libraries. Other datasets derived from metatranscriptomic material and contained a high proportion of rRNA-templated reads. The majority of the datasets used to compare the error rates of sequencing platforms in Figure 3 from the original publication violate underlying assumptions of DRISEE and led to publication of misleading results.

Improving DRISEE

Not all reads that share the same first 50 bases represent artificial duplication. Meaningful results from DRISEE require understanding the source and distribution of sequence sets with identical prefixes. Suspicious bins must be excluded. However, this adds a layer of complexity and might result in too few bins to reach a robust error estimate. The minimum number of bins necessary to reach a reliable estimate and the impact of the sub-sampling necessary to complete the analysis in a reasonable time were not adequately addressed by the authors.

Although DRISEE may eventually have the potential to identify problematic datasets and assess the sequencing quality of next-generation sequencing runs based on ADRs, the current version of the software is inadequate and its results are unrealistic.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key points

- DRISEE is proposed as a method for detecting errors in metagenomic sequencing data by binning reads that contain the same prefix and investigating their divergence.
- DRISEE does not eliminate bins created by adapter contamination or that arise from closely related or low-complexity sequences, which results in inflated error estimates.
- DRISEE in its current implementation is inaccurate, and error rates reported in the DRISEE publication regarding Illumina and 454 technologies are misleading.

FUNDING

National Institutes of Health [1UH2DK083993 to M.L.S.]; National Science Foundation [BDI-096026 to S.M.H.].

References

1. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol* 2011;**12**:R112.
2. Salmela L. Correction of sequencing errors in a mixed set of reads. *Bioinformatics* 2010;**26**:1284–90.
3. Meacham F, Boffelli D, Dhahbi J, *et al*. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* 2011;**12**:451.
4. Yang X, Aluru S, Dorman KS. Repeat-aware modeling and correction of short read errors. *BMC Bioinformatics* 2011; **12(Suppl 1)**:S52.
5. Yang X, Dorman KS, Aluru S. Reptile: representative tiling for short read error correction. *Bioinformatics* 2010;**26**: 2526–33.
6. Kao WC, Chan AH, Song YS. ECHO: a reference-free short-read error correction algorithm. *Genome Res* 2011; **21**:1181–92.

7. Schroder J, Schroder H, Puglisi SJ, *et al*. SHREC: a short-read error correction method. *Bioinformatics* 2009;**25**: 2157–63.
8. Shi H, Schmidt B, Liu W, Müller-Wittig W. A parallel algorithm for error correction in high-throughput short-read data on CUDA-enabled graphics hardware. *J Comput Biol* 2010;**17**:603–15.
9. Zhang T, Luo Y, Liu K, *et al*. BIGpre: a quality assessment package for next-generation sequencing data. *Genomics Proteomics Bioinformatics* 2011;**9**:238–44.
10. Zhao X, Palmer LE, Bolanos R, *et al*. EDAR: an efficient error detection and removal algorithm for next generation sequencing data. *J Comput Biol* 2010;**17**:1549–60.
11. Qu W, Hashimoto S, Morishita S. Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing. *Genome Res* 2009;**19**:1309–15.
12. Robinson T, Killcoyne S, Bressler R, Boyle J. SAMQA: error classification and validation of high-throughput sequenced read data. *BMC Genomics* 2011;**12**:419.
13. Smeds L, Kunstner A. ConDeTri—a content dependent read trimmer for Illumina data. *PLoS One* 2011;**6**:e26314.
14. Wang VX, Blades N, Ding J, *et al*. Estimation of sequencing error rates in short reads. *BMC Bioinformatics* 2012;**13**:185.
15. Schroder J, Bailey J, Conway T, Zobel J. Reference-free validation of short read data. *PLoS One* 2010;**5**:e12681.
16. Medvedev P, Scott E, Kakaradov B, Pevzner P. Error correction of high-throughput sequencing datasets with non-uniform coverage. *Bioinformatics* 2011;**27**:i137–41.
17. Bokulich NA, Subramanian S, Faith JJ, *et al*. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* 2013;**10**(1):57–9.
18. Rosen MJ, Callahan BJ, Fisher DS, Holmes SP. Denoising PCR-amplified metagenome data. *BMC Bioinformatics* 2012;**13**:283.
19. Keegan KP, Trimble WL, Wilkening J, *et al*. A platform-independent method for detecting errors in metagenomic sequencing data: DRISEE. *PLoS Comput Biol* 2012;**8**: e1002541.
20. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 2008; **36**:e105.
21. Huse SM, Huber JA, Morrison HG, *et al*. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 2007;**8**:R143.
22. Quinlan AR, Stewart DA, Stromberg MP, Marth GT. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* 2008;**5**:179–81.