Genome Biology

RESEARCH

Open Access

Retained introns in long RNA-seq reads are not reliably detected in sample-matched short reads

Julianne K. David[1,2,3†], Sean K. Maden[1,2,4†], Mary A. Wood[1,5,6], Reid F. Thompson[1,2,7,8,9*] and Abhinav Nellore[1,2,10*]

†Julianne K. David and Sean K. Maden contributed equally to this work.

*Correspondence: thompsre@ohsu.edu; anellore@gmail.com

[2] Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA
[9] Department of Radiation Medicine, Oregon Health & Science University, Portland, OR, USA
Full list of author information is available at the end of the article

Abstract

Background: There is growing interest in retained introns in a variety of disease contexts including cancer and aging. Many software tools have been developed to detect retained introns from short RNA-seq reads, but reliable detection is complicated by overlapping genes and transcripts as well as the presence of unprocessed or partially processed RNAs.

Results: We compared introns detected by 8 tools using short RNA-seq reads with introns observed in long RNA-seq reads from the same biological specimens. We found significant disagreement among tools (Fleiss' $\kappa = 0.113$) such that 47.7% of all detected intron retentions were not called by more than one tool. We also observed poor performance of all tools, with none achieving an F1-score greater than 0.26, and qualitatively different behaviors between general-purpose alternative splicing detection tools and tools confined to retained intron detection.

Conclusions: Short-read tools detect intron retention with poor recall and precision, calling into question the completeness and validity of a large percentage of putatively retained introns called by commonly used methods.

Keywords: RNA-seq, Splicing, Intron retention

Background

During RNA transcription, multiple spliceosomes may act on the same transcript in parallel to remove segments of sequence called introns and splice together flanking exons [1]. Most splicing occurs stochastically [2] during transcription [3–5], although up to 20% of splicing may occur after transcription and polyadenylation [5, 6] (Additional file 1: Fig. S1). Introns are spliced by U2 and U12 spliceosomes [7], primarily in the nucleus [8], though studies suggest that cytoplasmic splicing may also occur [9–12].

Intron retention (IR) is a form of alternative splicing where an anticipated intron remains after transcript processing is complete. IR occurs in up to 80% of protein-coding genes in humans [13] and may affect gene expression regulation [14–20] as well as