Quantitative Biology 2018, 6(3): 267–274
https://doi.org/10.1007/s40484-018-0141-x

METHODOLOGY ARTICLE

ShapeShifter: a novel approach for identifying and quantifying stable lariat intronic species in RNAseq data

Allison J Taggart[1] and William G Fairbrother[1,2,*]

[1] Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, RI 02912, USA
[2] Center for Computational Molecular Biology, Brown University, Providence, RI 02912, USA
* Correspondence: William_Fairbrother@brown.edu

Received November 2, 2017; Revised January 16, 2018; Accepted February 28, 2018

***Background*: Most intronic lariats are rapidly turned over after splicing. However, new research suggests that some introns may have additional post-splicing functions. Current bioinformatics methods used to identify lariats require a sequencing read that traverses the lariat branchpoint. This method provides precise branchpoint sequence and position information, but is limited in its ability to quantify abundance of stabilized lariat species in a given RNAseq sample. Bioinformatic tools are needed to better address these emerging biological questions.**
***Methods*: We used an unsupervised machine learning approach on sequencing reads from publicly available ENCODE data to learn to identify and quantify lariats based on RNAseq read coverage shape.**
***Results*: We developed ShapeShifter, a novel approach for identifying and quantifying stable lariat species in RNAseq datasets. We learned a characteristic "lariat" curve from ENCODE RNAseq data and were able to estimate abundances for introns based on read coverage. Using this method we discovered new stable introns in these samples that were not represented using the older, branchpoint-traversing read method.**
***Conclusions*: ShapeShifter provides a robust approach towards detecting and quantifying stable lariat species.**

Keywords: splicing; RNA; lariat

Author summary: RNA splicing is a post-transcriptional process in which introns are excised and exons are ligated together, forming the mRNA product. While most RNA intron lariats are rapidly degraded after splicing, some introns exhibit unusual stability and persist post-splicing. There is a lack of existing tools to profile these stable RNA species. Here we developed ShapeShifter, an approach to identify and quantify intronic lariats in RNA sequencing data. Using a clustering-based approach over the shape of sequencing read pileup, we defined a characteristic lariat curve. This approach validates known stable introns and also discovers potential new stable introns.

INTRODUCTION

RNA splicing is a post-transcriptional processing step in which long, non-coding sequences (introns) are removed from within a transcript, and the coding sequences (exons) are ligated together to create the mature mRNA transcript. The splicing reaction occurs through two transesterification reactions. In the first reaction, the 2′OH of the branchpoint (BP) nucleotide, typically an adenosine, will attack the 5′ splice site (5′ss). This forms the looped lariat intermediate molecule with an unusual 2′-5′ linkage connecting the BP to the 5′ss. In the second reaction, the free 3′OH of the 5′ exon attacks the 3′ splice site (3′ss), resulting in the ligated exonic product and an excised lariat.

After the splicing reaction, the intronic lariat is typically turned over quickly. DBR1, the RNA debranching enzyme, selectively recognizes and hydrolyzes the 2′-5′ linkage in a rate-limiting manner. The linear debranched intron is then quickly degraded by exonu-

cleases. It has been demonstrated that diminished DBR1 activity causes decreased lariat turnover and subsequent increased lariat accumulation in cells in both *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* [1,2]. There is strong biological imperative to rapidly turn over introns, both to release free nucleotides and sequestered RNA binding proteins for future rounds of transcription and splicing. While the vast majority of research focus has been studying the mRNA product of splicing, it is clear that the cell goes through much energy to both transcribe and splice introns. An average human coding genes contains 9.8 introns, and remarkably, introns make up about half of the human genome [3]. While historically thought of as "junk" nucleic acid, emerging research is suggestive that introns may have functional roles post-splicing.

While most lariats are turned over quickly, there exist cases of well-studied lariats that appear to resist degradation and persist in the cell. One of the first discovered stabilized lariats is derived from an intron from the T cell receptor-beta gene and accumulates in the nucleus of T cells [4]. Additionally, stabilized lariats have also been discovered in viruses. The first discovery of a viral stabilized lariat was a simian virus 40 (SV40) intron accumulating to high levels in infected *Xenopus laevis* oocytes [5]. Other viruses, including herpes simplex virus 1 (HSV1) and human and murine cytomegalovirus were later found to also produce stable a stable lariat [6–10]. While the function of these viral introns remains unknown, viruses are under selective pressure to maintain a small genome size and depend upon host cellular machinery. The presence of these stable viral introns is suggestive of function.

Recent genomic studies have discovered potential post-splicing functions for introns. One example of a functional lariat is derived from the IgH locus in B cells. When B cells are stimulated to undergo class switching, the IgH locus is spliced, producing an intronic lariat. This lariat is subsequently debranched by DBR1, and then acts as a guide RNA bringing the activation-induced cytidine deaminase (AID) enzyme to the complementary DNA, resulting in class-switching recombination [11]. Another example of a functional intron is derived from the ANKRD52 locus. This lariat intron accumulates near its host gene transcription and interacts with RNA polymerase II to potentially regulate its host gene expression [12]. In genome-wide studies of cellular RNA species, it was recently discovered that stable intronic species accumulated as lariat molecules in *Xenopus* oocytes, suggesting that these lariats may play roles in oogenesis, embryogenesis or development [13,14].

Due to this growing field, new tools are required to better annotate and quantify lariat levels in a cell. When lariats were first discovered, they were characterized by their unusual gel mobility due to their circular structure, and branchpoints were mapped using primer extension strategies [15–19]. It was later discovered that reverse transcriptase (RT) could read through the lariat 2′-5′ linkage and inverted primers were utilized to map branchpoint location in a small number of candidate introns using a RT-PCR based strategy [20,21]. Our group was the first to develop an inverted read alignment strategy to map branchpoints in RNA deep sequencing data [22]. We then applied this strategy to ENCODE publicly available sequencing data to map branchpoints for 16.8% of all human introns [23]. While these strategies allow us to create an extensive branchpoint annotation and learn a great deal about branchpoint chemistry, they are limited in their ability to quantify the levels of individual lariat loci in the cell. An RNA sequencing read that traverses the BP-5′ss junction is incredibly rare due to the low efficiency of RT reading over the 2′-5′ chemistry linkage. In a given sample, the majority of BPs are defined by a single lariat read, which makes it impossible to quantify lariat levels using this technique. Additionally, it is possible that other factors unrelated to the lariat levels, such as secondary structure or BP sequence, can affect RT read-through, which makes quantification from this method challenging.

Here, we present ShapeShifter, a lariat-profiling approach that relies on the shape and density of RNA sequencing reads to identify and quantify stabilized lariats in RNA deep sequencing data. Using an unsupervised machine learning approach, we developed a lariat-calling heuristic from publicly available ENCODE RNA sequencing samples. ShapeShifter can apply this metric to new RNA sequencing datasets to identify stabilized lariats and calculate the intronic read density over the entire intron to quantify its abundance. This abundance value can be compared between samples to identify perturbations that affect lariat levels both in bulk and at individual loci.

RESULTS

In our previous study, we released a branchpoint annotation for 16.8% of human introns using an inverted alignment strategy [23]. This inverted alignment strategy requires an RNA sequencing read that reads through the 2′-5′ lariat linkage and traverses the BP-5′ss junction (Figure 1A). These lariat reads are rare due to both biological and technical reasons. Biologically speaking, introns are typically rapidly turned over post-splicing to recycle both nucleotides and sequestered RNA binding proteins. Technically, lariat reads are difficult to capture both because it has a strict positional requirement, but also due to the chemistry of the linkage. While RT is capable of reading through the lariat 2′-5′ linkage, it is a low

efficiency read-through and the RT will often pause, misincorporate nucleotides, or fall off of the molecule. Due to these reasons, the majority of recovered lariats have low numbers of validating reads. When examining on a per-intron basis, about 80% of introns measured have only one unique branchpoint-traversing read in a given RNAseq sample (Figure 1B). While this provides valuable information about the branchpoint sequence and position, it is impossible to use this data to deduce information about steady-state lariat levels in the cell.

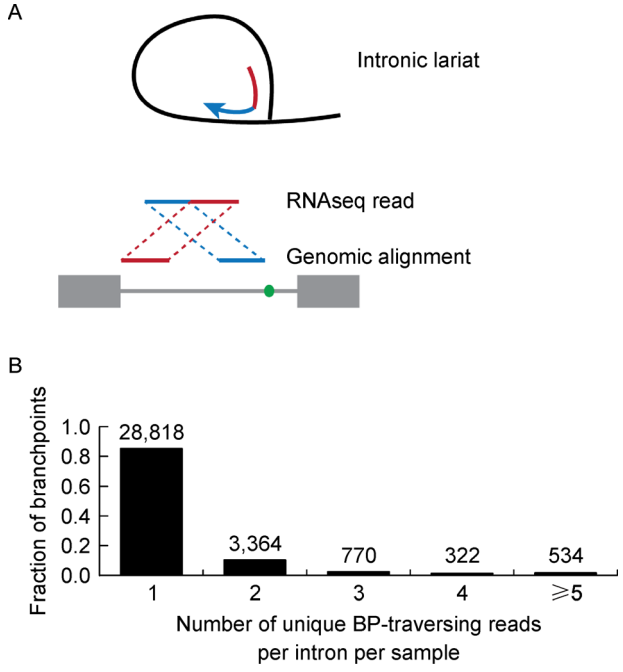

Figure 1. Number of unique lariat branchpoint-traversing reads per intron. (A) Schematic of the inverted read alignment strategy to identify branchpoint traversing reads. (B) Histogram of the number of unique branchpoint-traversing reads per intron in each sample.

In order to determine a better way to quantitate lariat levels, we visualized the RNAseq read coverage of raw read pileup over several introns with high lariat read counts in a given RNAseq sample. Five examples of different introns in different cell lines are pictured in Figure 2. In these cases, we observe a characteristic lariat shape of the read density across the splice sites. High read coverage in the center of the intron is consistent with high levels of the intron in the cell. Additionally, we observe a drop-off of sequencing read coverage at the splice sites, indicative of a separate, circular molecule (as opposed to intron retention). By using all of the intron-aligning reads, instead of just branchpoint-traversing reads, we have much more power to estimate individual lariat levels.

To study this phenomenon on a genome-wide scale, we used an unsupervised machine learning approach to learn the "lariat" shape in RNAseq read coverage data. To do this, 44 total RNA, whole cell RNAseq datasets from various cell lines from the ENCODE project were analyzed. Normalized read coverage curves were calculated for each annotated *hg19* intron in each of the samples in a window downstream of the 5′ss (Figure 3A). The 5′ end of the lariat curve was used because 5′ss positions are much more exhaustively annotated than branchpoint positions. These curves were then subjected to K-means clustering (Figure 3B). In order to identify which cluster contained stable lariats, the set of introns that was assigned to each cluster was overlapped with sets of introns with increasing levels of branchpoint-traversing read evidence. We find that of the introns that have at least five branchpoint-traversing reads there is a clear enrichment for cluster 6 (Figure 3C). Consistent with our hypothesis from the introns shown in Figure 2, cluster 6 has high read density at the end of the curve, but a steep drop off of read coverage at the 5′ end of the intron. This sloped drop-off suggests that these intronic reads arise not from intron retention, but from a separate, circular molecule.

We present the output of our clustering approach as ShapeShifter, an approach that can be used to identify stable lariats by the shape of intronic read pileup and quantify abundance levels by calculating read coverage over the intron. Using this approach we discovered other stable lariats in the ENCODE datasets with no previous lariat read evidence by selecting other high-coverage introns that cluster into cluster 6. Read density plots of 10 stable lariat exemplars with no previous lariat read evidence are depicted in Figure 4.

DISCUSSION

Here we describe an alternate method for profiling lariat abundance in RNAseq data. Previous lariat profiling methods describe precise branch point sequence and position, but provide limited quantitative information. Using these methods does not provide an adequate tool for comparing lariat levels across samples due to low read counts. The ShapeShifter approach, on the other hand, can be used to estimate lariat abundance in a given sample by calculating the read coverage over the entire intron and normalizing to intron length and sequencing depth. This is of particular use in comparative or perturbation future work, in which a particular intronic locus can be quantitatively compared across samples.

While most introns are rapidly turned over, some introns appear to be selectively stable. Interestingly, several viruses, which are under selective pressure to keep their genome size small, code for introns that result in stable lariats. Recent research has suggested functional

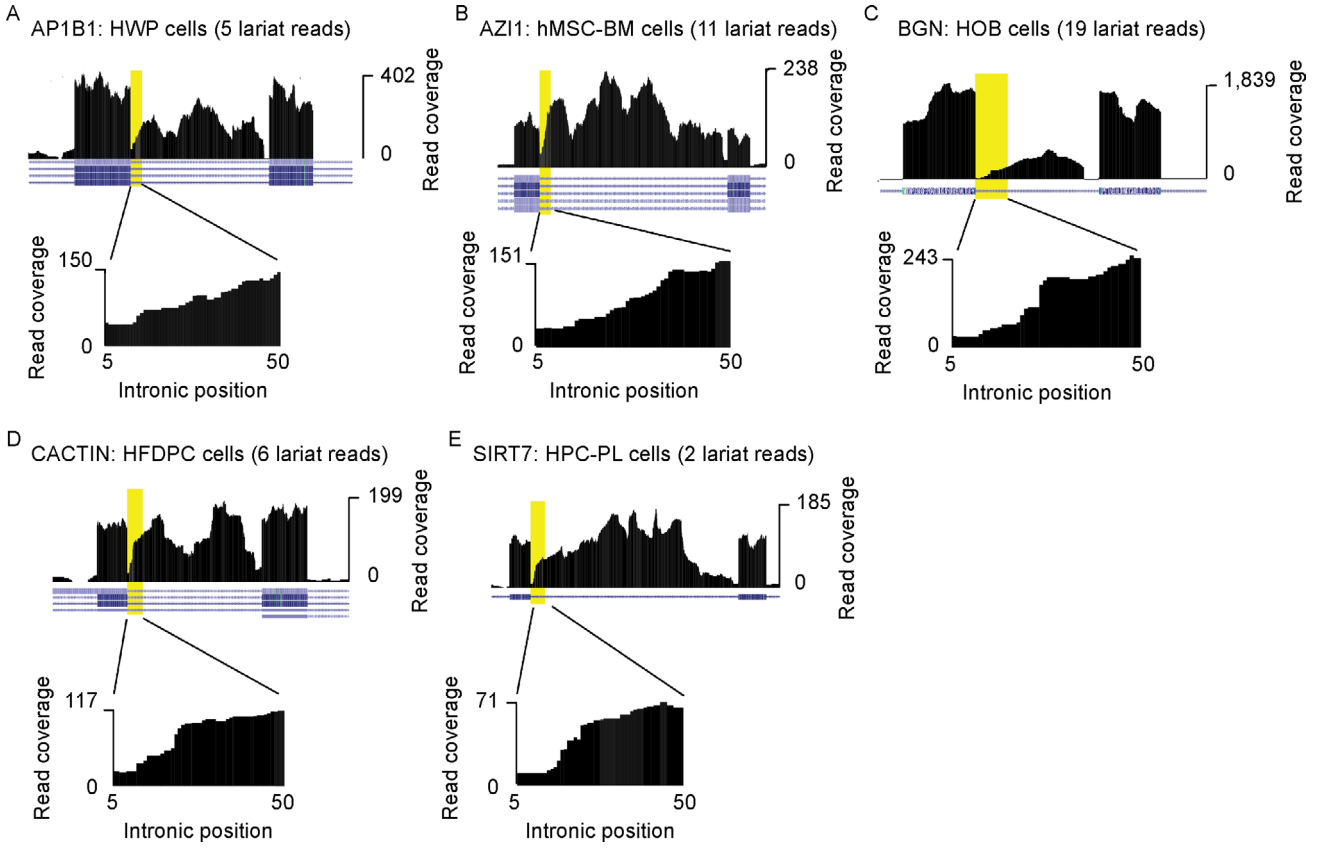

Figure 2. RNAseq read coverage plots over five different introns in five different cell lines with multiple unique branchpoint-traversing reads. Yellow highlight and over intronic window 0 to 50 nucleotides downstream of the 5′ splice site. (A) AP1B1 intron in HWP cells. (B) AZI1 intron in hMSC-BM cells. (C) BGN intron in HOB cells. (D) CACTIN intron in HFDPC cells. (E) SIRT7 intron in HPC-PL cells.

roles for specific stable intron species. The role of stabilized introns after splicing is an emerging field, and better bioinformatics algorithms are required to better annotate and characterize these species.

MATERIALS AND METHODS

ENCODE RNAseq data was obtained from GEO accession GSE30567. Whole cell, total RNA samples were used for this analysis (SRR534289, SRR534290, SRR534317, SRR534318, SRR534325, SRR534326, SRR534327, SRR534334, SRR534335, SRR545685, SRR545686, SRR545689, SRR545690, SRR545691, SRR545692, SRR545693, SRR545694, SRR545701, SRR545702, SRR545703, SRR545704, SRR545705, SRR545706, SRR545707, SRR545708, SRR545709, SRR545710, SRR545711, SRR545712, SRR545713, SRR545714, SRR545715, SRR545716, SRR545717, SRR545718, SRR545719, SRR545720, SRR545721, SRR545722, SRR545723, SRR768411, SRR768412, SRR768413, SRR768414).

Branchpoint traversing reads were mapped as previously published [23] (http://fairbrother.biomed.brown.edu/data/Lariat2016/). Briefly, using the Bowtie aligner, reads were aligned to the *hg19* genome, allowing up to three mismatches. Ungapped, complete forward alignments were discarded. The remaining reads were iteratively split into head and tail regions, and aligned to the genome. Reads with a gapped, inverted alignment with one portion mapping directly to the 5′ splice site, and the other portion mapping downstream intronically, were called as branchpoint-traversing reads. The branchpoint is determined as the last nucleotide of the downstream intronic read portion. Splice site annotation is using *hg19* UCSC genes.

In the ShapeShifter approach, reads were aligned to *hg19* genome using STAR [24]. Alignments were outputted from STAR in bedgraph format, using uniquely mapping reads. Using in-house perl scripts, bedgraphs were converted into tables of read coverage values over each nucleotide of all annotated introns. Intronic annotation was obtained from UCSC *hg19* knowngenes. Windowed read coverage tables were obtained by extracting windows downstream of the 5′ splice site of

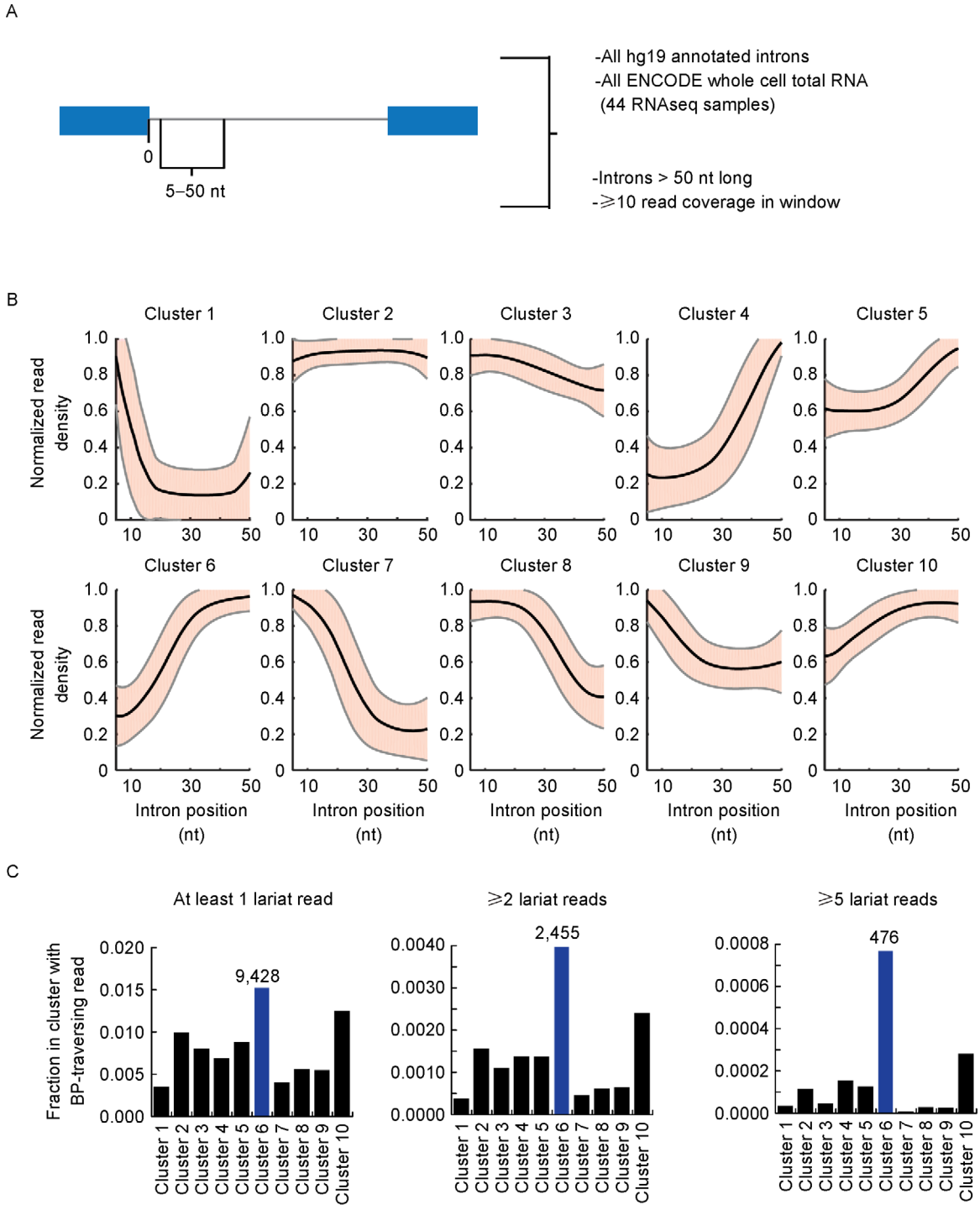

Figure 3. Identifying lariat shape in ENCODE RNAseq data. (A) Schematic showing intronic windows used for clustering analysis. (B) K-means clustering output. Cluster centroid plotted with shaded red area as one standard deviation from centroid. (C) Introns with branchpoint-traversing reads are enriched in cluster 6.

each intron in intronic positions 5–50 nt. Introns that were shorter than 50 nucleotides were discarded. A minimum read height of 10 reads in this window was required for clustering (introns with less than 10 reads in this window were discarded). All remaining intronic windows were normalized for read pileup height [0,1] within this window. K-means clustering approach was applied to the normalized read coverage curves over this window

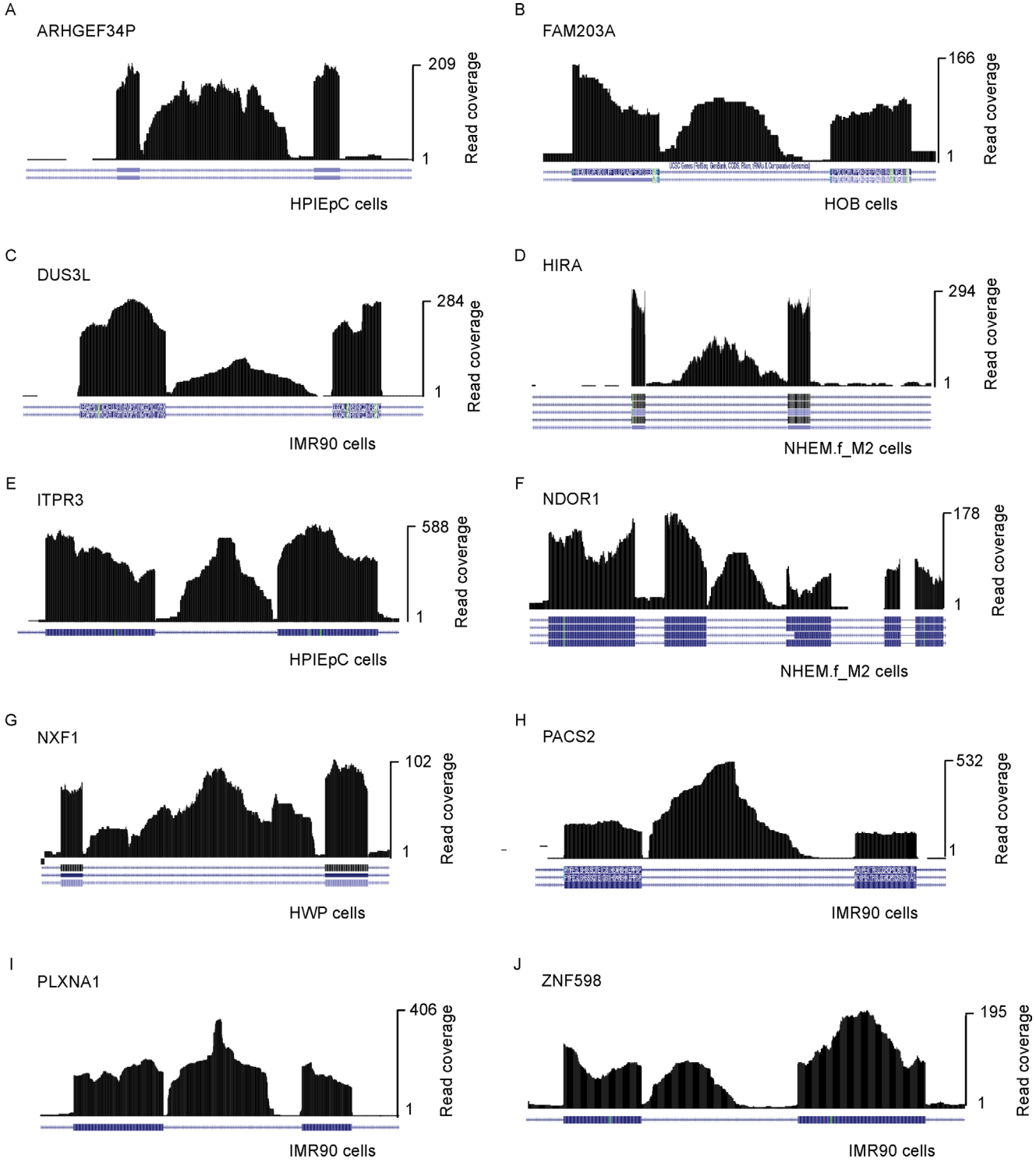

Figure 4. RNAseq read coverage plots in stable introns discovered by ShapeShifter, without any prior branchpoint-traversing read evidence. (A) ARHGEF34P intron in HPIEpC cells. (B) FAM203A intron in HOB cells. (C) DUS3L intron in IMR90 cells. (D) HIRA intron in NHEM.f_M2 cells. (E) ITPR3 intron in HPIEpC cells. (F) NDOR1 intron in NHEM.f_M2 cells. (G) NXF1 intron in HWP cells. (H) PACS2 intron in IMR90 cells. (I) PLXNA1 intron in IMR90 cells. (J) ZNF598 intron in IMR90 cells.

using the MATLAB K-means tool. Clustering was performed using 10 clusters.

For visualization, bigwig files of read coverage were created using a combination of samtools [25], bedtools [26], and the utility bedGraphToBigWig [27]. Bigwig files were then uploaded to the UCSC genome browser to generate read coverage plots (http://genome.ucsc.edu) [28].

ABBREVIATIONS

BP,	branchpoint
5′ss,	5′splice site
3′ss,	3′splice site
HSV,	herpes simplex virus 1
SV40,	simian virus 40

COMPLIANCE WITH ETHICS GUIDELINES

The authors Allison J Taggart and William G Fairbrother declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Nam, K., Lee, G., Trambley, J., Devine, S. E. and Boeke, J. D. (1997) Severe growth defect in a *Schizosaccharomyces pombe* mutant defective in intron lariat degradation. Mol. Cell. Biol., 17, 809–818
2. Kim, J. W., Kim, H. C., Kim, G. M., Yang, J. M., Boeke, J. D. and Nam, K. (2000) Human RNA lariat debranching enzyme cDNA complements the phenotypes of *Saccharomyces cerevisiae dbr1* and *Schizosaccharomyces pombe dbr1* mutants. Nucleic Acids Res., 28, 3666–3673
3. Hubé, F. and Francastel, C. (2015) Mammalian introns: when the junk generates molecular diversity. Int. J. Mol. Sci., 16, 4429–4452
4. Qian, L., Vu, M. N., Carter, M. and Wilkinson, M. F. (1992) A spliced intron accumulates as a lariat in the nucleus of T cells. Nucleic Acids Res., 20, 5345–5350
5. Michaeli, T., Pan, Z. Q. and Prives, C. (1988) An excised SV40 intron accumulates and is stable in *Xenopus laevis* oocytes. Genes Dev., 2, 1012–1020
6. Farrell, M. J., Dobson, A. T. and Feldman, L. T. (1991) Herpes simplex virus latency-associated transcript is a stable intron. Proc. Natl. Acad. Sci. USA, 88, 790–794
7. Zabolotny, J. M., Krummenacher, C. and Fraser, N. W. (1997) The herpes simplex virus type 1 2.0-kilobase latency-associated transcript is a stable intron which branches at a guanosine. J. Virol., 71, 4199–4208
8. Kulesza, C. A. and Shenk, T. (2004) Human cytomegalovirus 5-kilobase immediate-early RNA is a stable intron. J. Virol., 78, 13182–13189
9. Kulesza, C. A. and Shenk, T. (2006) Murine cytomegalovirus encodes a stable intron that facilitates persistent replication in the mouse. Proc. Natl. Acad. Sci. USA, 103, 18302–18307
10. Schwarz, T. M. and Kulesza, C. A. (2014) Stability determinants of murine cytomegalovirus long noncoding RNA7.2. J. Virol., 88, 11630–11633
11. Zheng, S., Vuong, B. Q., Vaidyanathan, B., Lin, J. Y., Huang, F. T. and Chaudhuri, J. (2015) Non-coding RNA generated following lariat debranching mediates targeting of AID to DNA. Cell, 161, 762–773
12. Zhang, Y., Zhang, X. O., Chen, T., Xiang, J. F., Yin, Q. F., Xing, Y. H., Zhu, S., Yang, L. and Chen, L. L. (2013) Circular intronic long noncoding RNAs. Mol. Cell, 51, 792–806
13. Gardner, E. J., Nizami, Z. F., Talbot, C. C. Jr and Gall, J. G. (2012) Stable intronic sequence RNA (sisRNA), a new class of noncoding RNA from the oocyte nucleus of *Xenopus tropicalis*. Genes Dev., 26, 2550–2559
14. Talhouarne, G. J. and Gall, J. G. (2014) Lariat intronic RNAs in the cytoplasm of *Xenopus tropicalis* oocytes. RNA, 20, 1476–1487
15. Domdey, H., Apostol, B., Lin, R. J., Newman, A., Brody, E. and Abelson, J. (1984) Lariat structures are *in vivo* intermediates in yeast pre-mRNA splicing. Cell, 39, 611–621
16. Rodriguez, J. R., Pikielny, C. W. and Rosbash, M. (1984) *In vivo* characterization of yeast mRNA processing intermediates. Cell, 39, 603–610
17. Zeitlin, S. and Efstratiadis, A. (1984) *In vivo* splicing products of the rabbit β-globin pre-mRNA. Cell, 39, 589–602
18. Padgett, R. A., Konarska, M. M., Grabowski, P. J., Hardy, S. F. and Sharp, P. A. (1984) Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. Science, 225, 898–903
19. Ruskin, B., Krainer, A. R., Maniatis, T. and Green, M. R. (1984) Excision of an intact intron as a novel lariat structure during pre-mRNA splicing *in vivo*. Cell, 38, 317–331
20. Gao, K., Masuda, A., Matsuura, T. and Ohno, K. (2008) Human branch point consensus sequence is yUnAy. Nucleic Acids Res., 36, 2257–2267
21. Vogel, J., Hess, W. R. and Börner, T. (1997) Precise branch point mapping and quantification of splicing intermediates. Nucleic Acids Res., 25, 2030–2031
22. Taggart, A. J., DeSimone, A. M., Shih, J. S., Filloux, M. E. and Fairbrother, W. G. (2012) Large-scale mapping of branchpoints in human pre-mRNA transcripts *in vivo*. Nat. Struct. Mol. Biol., 19, 719–721
23. Taggart, A. J., Lin, C. L., Shrestha, B., Heintzelman, C., Kim, S. and Fairbrother, W. G. (2017) Large-scale analysis of branchpoint usage across species and cell lines. Genome Res., 27, 639–649
24. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15–21
25. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., and the 1000 Genome Project Data Processing Subgroup. (2009) The sequence alignment/map format and SAMtools. Bioinformatics, 25, 2078–2079
26. Quinlan, A. R. and Hall, I. M. (2010) BEDTools: a flexible suite of

utilities for comparing genomic features. Bioinformatics, 26, 841–842

27. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. and Karolchik, D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics, 26, 2204–2207
28. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002) The human genome browser at UCSC. Genome Res., 12, 996–1006

