A data-driven approach to characterising intron signal in RNA-seq data

Albert Y. Zhang[1,4,✉], **Shian Su**[1,4,✉], **Ashley P. Ng**[2,4,✉], **Aliaksei Z. Holik**[3,4,✉], **Marie-Liesse Asselin-Labat**[3,4,✉], **Matthew E. Ritchie**[1,4,5✉], **and Charity W. Law**[1,4,✉]

[1]Molecular Medicine Division,
[2]Cancer and Haematology Division and
[3]ACRF Stem Cells and Cancer Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia
[4]Department of Medical Biology and
[5]School of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia

RNA-seq datasets can contain millions of intron reads per sequenced library that are typically removed from downstream analysis. Only reads overlapping annotated exons are considered to be informative since mature mRNA is assumed to be the major component sequenced, especially when examining poly(A) RNA samples. By examining multiple datasets, we demonstrate that intron reads are informative and that signal is shared between exon and intron counts. The majority of expressed genes contain reads in both exon and intron regions, where exon and intron log-counts are positively correlated. On a per gene basis intron counts are larger in Total RNA libraries than the same biological sample sequenced under a poly(A) RNA protocol. This is due to 3' coverage bias in poly(A) RNA libraries affecting medium to long intron regions, where a considerable drop in read coverage for introns residing at the 5' end of genes is observed. Looking across thousands of genes simultaneously we characterise coverage profiles of exon and intron regions in poly(A) RNA and Total RNA libraries. Our study provides a general yet comprehensive examination into intron reads that is insightful and applicable to transcriptomics research, especially in the identification of pre-mRNA levels, transcript structure and intron retention detection.

bioinformatics | sequencing | RNA-seq | intron reads | read coverage | data analysis

Correspondence: ***law@wehi.edu.au***

Introduction

Advances in gene profiling technology, such as RNA-sequencing (RNA-seq) have allowed researchers to study transcription in exquisite detail. Previously, quantitative gene expression analyses by microarrays required prior knowledge of the sequences to be interrogated, limiting *de novo* discoveries, understanding into gene transcripts and alternative splicing especially at a high-throughput level. Most research efforts focused on gene-level information and in the comparison of genes that are differentially expressed between two or more groups. Whilst this is still the main focus for RNA-seq, the technology, however, has the ability to examine sub-gene components such as at the transcript-level, exon-level, or even nucleotide base-level without prior sequence knowledge. As a result, there has been increased interest and effort into the study of transcript-level information, alternative gene splicing and gene intron retention at a global level using RNA-seq (1–6).

RNA-seq can be used to characterise and study many RNA types. These include non-coding RNAs that regulate a diverse range of cellular processes (7, 8) but the overwhelming majority of studies focus on messenger RNAs (mRNAs) which encode genes that are translated into protein. The most popular RNA selection protocol is that which captures polyadenylated RNA seeing that it is optimised for mRNA selection – in other words, RNA that has a poly(A) tail at the 3' end of the molecule. In eukaryotes, poly(A) tails are synthesised to aid in the transportation of mature mRNA molecules from the nucleus to the cytoplasm. Total RNA selection is also widely used and often includes a step that depletes ribosomal RNA so that it does not compete with the sequencing of mRNA. RNA expression values are highly correlated between the two RNA selection protocols, with a higher percentage of reads (≈3% more) mapping to protein coding genes in poly(A) RNA samples, and a higher percentage of reads (≈2.5% more) mapping to long non-coding RNAs in Total RNA samples (9).

The general assumption is that for protein coding genes the vast majority of RNA captured are mature mRNA transcripts. This is especially true when it comes to experiments for poly(A) RNA samples. As a result, aligned sequencing reads are usually summarised only for defined exon regions within genes. However, intron reads do in fact account for a significant percentage of sequencing reads (10–12) but it is not common practice to quantify these reads. Importantly, specific differences in RNA processing have identified intron retention as a mechanism of post-translational regulation of protein expression by nonsense mediated decay during normal blood development as well as with certain mutations of spliceosomal genes that have been identified in myelodysplasia (13).

Currently, there is a lack of general consensus about the interpretation of intron reads and the origin of such reads in RNA-seq experiments. Of the studies that have been undertaken, each has focused on specific aspects of RNA biology, thus potentially confounding interpretation of the relative contributions to intron reads from different stages of RNA processing. For instance, Gaidatzis *et al.* (2015) (14) used the presence of intron reads to differentiate between transcriptional and post-transcriptional regulation in both poly(A) RNA and Total RNA datasets. When poly(A)

RNA exon and intron counts were compared to those of chromatin and cytoplasmic fractions from lipid A stimulated mouse bone marrow–derived macrophages, overall changes in exonic counts were similar to those found in cytoplasmic fractions, whilst intronic changes were similar to that of chromatin fractions which represent nascent RNA that had yet to undergo nuclear export. Thus they concluded changes in intron reads were a direct reflection of transcriptional changes that were occuring following macrophage stimulation. Similarly, when IMR90 fibroblasts were stimulated with TNF-α, intron read changes were also shown to be highly correlated with changes using global run-on sequencing (GRO-seq), a method that measures nascent RNA. Here, their exon-intron split method (*EISA*) compared changes in exon counts to the changes in intron counts between two conditions. For genes that were transcriptionally regulated, changes in exon and intron counts are positively and linearly correlated, whilst gene transcripts that were post-transcriptionally regulated demonstrated changes in exon counts while intron counts remain unchanged.

Wong *et al.* (2013) (15) investigated intron reads with respect to granulocyte differentiation and their potential biological role in post-transcriptional gene regulation via intron retention resulting in nonsense mediated RNA decay. Using poly(A) RNA samples, a subset of intron reads were demonstrated to originate from mature mRNA molecules that had undergone nuclear export. Here differentially retained introns were identified by examining intron reads relative to exon-intron boundary reads. Comparing the transcriptome of more primitive promyelocytes to granulocytes, 121 introns in 86 genes were found to be differentially retained by *IRFinder*, a software tool they developed to detect intron retention. Intron presence in mRNA was validated by quantitative reverse transcription PCR (RT-qPCR) for 20 of these transcripts. Sixteen out of 20 transcripts were shown to be enriched in the cytoplasm relative to nuclei of granulocytes and thus represented processed and exported mRNAs. Intron retention in this small subset of genes resulted in a reduction of gene expression via nonsense mediated RNA decay and was unrelated to the level of nascent transcript. Thus intron retention appeared to provide a mechanism of post-transcriptional regulation of gene expression in select genes during granulocyte differentiation. Following their study on granulocyte differentiation, improvements to the *IRFinder* method (16) was made by estimating the ratio of reads supporting a retained event (reads overlapping an exon-intron boundary) relative to a spliced event (reads that are split between multiple exons). Braunschweig *et al.* (2014) (4) also examined exon-intron boundary reads and exon-exon split reads to demonstrate that intron retention is widespread in human and mouse, where intron retention is prevalent in 77% of multi-exonic genes. The majority of studies that incorporate intron reads into their analysis assume that the reads represent biological signal of a specific type. In contrast to these studies, other studies have suggested that intron reads may be a reflection of experimental and transcriptional noise (17), or even that intron reads are unusable in exon and gene quantification (18).

While utilising intron reads were a focus in the aforementioned studies, much of the analytical approach, presentation, and interpretations were driven by a specific aspect of RNA biology, and hence potentially subject to several assumptions. For example, Gaidatzis *et al.* (2015) (14) did not consider the presence of introns as a biological means of post-transcriptional gene regulation but rather a measure of nascent RNA transcription. Middleton *et al.* (2017) (16), on the other hand, assumed intron reads in poly(A) RNA samples could not arise from unprocessed mRNAs but instead represent intron retention. By themselves, each study is supported by their experimental design, relevance of dataset selections and biological validation. These conflicting hypotheses can bias the interpretation and relative significance of intron reads when using assumptions and methods that focus on a specific aspect of RNA biology. The use of intron reads in bioinformatic analysis requires further analyses to better understand the prevalence and pattern of intron reads, the relationship of intron reads to different methods of RNA library preparation, the relationship of intron reads to different biological samples, and how intron reads relate to exon reads at the gene level.

In this paper, we take a data-driven approach to examine intron reads to investigate the general behaviour and pattern of intron reads across multiple datasets and library preparation protocols. By novel application of multi-dimensional scaling (MDS) methods using intron specific read counts, we provide strong evidence that intron reads are not only informative, but their signal is shared with exon counts. We observe that per library, intron read percentages are dependent on library preparation methods where libraries prepared under a Total RNA protocol have 15% more intron reads than poly(A) RNA. We also demonstrate that intron and exon read percentages are highly dependent on cell and tissue types. Using a benchmarking dataset with identical biological samples sequenced under two RNA preparation protocols, we examined differences between poly(A) RNA and Total RNA libraries for intron reads at a more comprehensive level than prior studies. We show that on a per gene basis, intron counts are larger in Total RNA libraries than the same biological sample sequenced under a poly(A) RNA protocol because of a significant drop in intron read coverage at the 5' end of genes in poly(A) selected RNA libraries. This is in contrast to Total RNA libraries which have more uniform intron read coverage. Our study therefore comprehensively examines intron reads using several independent technical and biological datasets. It provides significant insights that are useful for transcriptomics research, including the application of intron reads to intron retention detection, as well as the potential incorporation of intron reads into differential gene transcription analyses that may be of particular importance for single-cell RNA-sequencing applications. R code used in the analyses of datasets are available at http://bioinf.wehi.edu.au/intronSignal.

Materials and Methods

Datasets. *Human cell lines* The human lung adenocarcinoma cell lines HCC827 and NCI-H1975 were cultured on three separate occasions, with RNA extracted and prepared for sequencing on an Illumina HiSeq2500 instrument using a 100-basepair single-end protocol. Within each cell line Replicate 1 (R1), Replicate 2 (R2) and Replicate 3 (R3) samples are considered as biological replicates of each other. RNA from each of the replicates were split into two and prepared using *Illumina's TruSeq RNA v2 kit* and *Illumina's TruSeq Total Stranded RNA kit with Ribozero depletion* to obtain poly(A) RNA and Total RNA respectively. A total of 12 libraries were examined in this paper: poly(A) RNA of HCC827, poly(A) RNA of NCI-H1975, Total RNA of HCC827 and Total RNA of NCI-H1975. Only pure, non-degraded samples were analysed from the larger experiment described in Holik *et al.* (2015) (19). Raw sequencing reads can be downloaded from the Gene Expression Omnibus (GEO) (20, 21) under accession number GSE64098.

Human immune cells Human immune cells were sequenced by Linsley *et al.* (2014) (22) on Illumina's HiScanSQ system for 134 samples from healthy and unhealthy individuals with autoimmune or infectious diseases. RNA samples were taken of whole blood and 6 immune cell subsets and processed using *Illumina's TruSeq preparation kit* to select for poly(A) RNA and sequenced on an Illumina HiScanSQ instrument using a 50-basepair paired-end protocol. Cell subsets include pure populations of neutrophils, monocytes, B cells, CD4+ T cells, CD8+ T cells and natural killer (NK) cells. Raw sequencing reads were downloaded from GEO under accession number GSE60424.

Mouse mammary cells Mouse mammary cells from female virgin mice with additional samples from mammosphere and the CommaD-βGeo (CommaD-bG) cell line were sequenced in a study by Sheridan *et al.* (2015) (23). Mammary cell populations include mammary stem cell-enriched basal cells, luminal progenitor-enriched (LP) and mature luminal-enriched (ML) cell populations. Total RNA was extracted from the samples and prepared for sequencing using *Illumina's TruSeq RNA v2 kit* to obtain poly(A) RNA. A total of 19 libraries were then sequenced using a 100-basepair single-end protocol on Illumina's HiSeq2000 machine.

Gene annotation. FASTQ files containing raw sequencing reads were aligned to the human *hg20* genome or mouse *mm10* genome using *subjunc* (24) in the *Rsubread* software package. Aligned reads were then summarised into tables of counts by GENCODE gene annotation using two counting strategies. GENCODE's main *Comprehensive gene annotation* file in gtf format was downloaded from https://www.gencodegenes.org for human (Release 27) and mouse (Release M12). GENCODE annotation was chosen over other annotations, such as RefSeq or Ensembl, because it provides annotation that is more inclusive and less stringent in terms of gene definitions and exon boundaries. It also includes the annotation of long non-coding RNAs (8). Annotation files were simplified by taking the union of two or more overlapping exons from transcripts of the same gene. The adjustment provides a simplification of genomic positions on each strand, such that each position is classified as belonging to "exon", "intron" or otherwise outside of an annotated gene. Three resultant annotation files were saved in standard annotation format (SAF) – exon annotation, intron annotation (region between exons), and genebody annotation (region spanning first to last exon).

Exon and intron counts. Reads are summarised by *featureCounts* (25) using exon annotation and genebody annotation separately to get gene-level *exon counts* and gene-level *genebody counts* respectively. Reads that overlap features of multiple genes are not counted. Gene-level *intron counts* are calculated by subtracting exon counts from genebody counts. This is a conservative estimate on intron count levels relative to an approach that also includes exon-intron boundary reads into intron quantification. The intron counts represent the gain in information when summarising reads across the whole genebody relative to exon regions only. We take on this approach since our interest is in assessing whether intron counts contain additional signal on top of standard exon counts. Genes can have negative values if genebody counts are smaller than exon counts – this happens when a read overlaps multiple genes when looking across the whole genebody, but does not overlap multiple genes when considering exons only. Negative values are adjusted to zero. Note that exon and intron counts form completely separate read sets, in that reads are only included into either count sets but not both.

Bin coverage estimation. Annotated exons and introns are divided into non-overlapping bins of 30-basepairs in length. The features are defined by the simplified GENCODE annotation (see section on *Gene annotation*). Not all exons or introns divide evenly into 30-basepair bins. As a result, the *last* bin of a feature may be 30-basepairs in length or shorter. The last exon or intron bin in genes on the positive strand will be that which is closest to the 3' end; 5' most on the negative strand.

For both poly(A) RNA and Total RNA libraries in the HCC827 human cell line, aligned reads were summarised into counts using *featureCounts* by passing the binned annotation for exons and introns separately. Reads overlapping multiple bins are counted once for each bin. Resultant counts represent read coverage in the local area. A "last bin" that is smaller in size may result in lower coverage relative to neighboring 30-basepair bins. To be conservative, we do not adjust the coverage of such bins by bin size since this could potentially over-inflate the coverage of very small bins.

Coverage estimation is carried out for 1) genes that have signal (a count of 3 or more) in both exon and intron regions of poly(A) RNA samples; 2) are protein coding genes on reference chromosomes (excludes the mitochondrial chromosome); and 3) to ensure that coverage estimates for a gene are not associated with another annotated gene we include only genes that do not overlap any other annotated gene along its entire genebody. Protein coding genes are as defined by GENCODE. A total of 3,694 genes are examined. Genes

are categorised as having short, regular or long exon regions based on the total number of exon bins in each gene, with roughly 1,231 genes in each category. In a similar fashion, genes were categorised as having short, regular or long intron regions.

Exon and intron coverage patterns. The pattern of coverage for short, regular and long regions were estimated separately for exons and introns. For all bins in a given gene, coverage values were transformed to a $\log_2$-scale using an offset of 1 and then divided by the maximum log-coverage value within the same gene. Adjusted log-coverage values were separated into 20 sections of equal bin numbers based on bin position along the gene. If the number of bins do not divide evenly into 20 sections, extra bins were placed mid-gene in the tenth section. Strand direction of genes was accounted for. The mean of the bins was calculated for each section in each gene to represent the general pattern of coverage for a given gene. The overall pattern of coverage across multiple genes was calculated by taking the mean along all sections across genes.

Results

Intron reads are prevalent across datasets. Taking a conservative approach, we quantify the number of intron reads that map entirely to an intron (or introns) of a gene. Intron counts represent the extra counts one may obtain from within a gene when looking outside of annotated exons. The intron counts presented in this paper are conservative in that they exclude the counting of reads that overlap both annotated exons and introns.

Across poly(A) RNA datasets, the percentage of reads that contribute to gene-level exon counts ranges from 57% to 78%, with a mean of 69% (Figure 1a). A smaller percentage of reads contribute towards gene-level intron counts, 2% to 14% with a mean value of 7% (Figure 1b). Sample-wise read percentages are consistent within biological groups for both exons and introns, with variation in read percentages for different cell types. Despite the relatively small percentage of intron reads, they amount to hundreds of thousands to millions of reads per library under typical sequencing protocols. For a library of size 30 million, the number of intron reads is approximately 2.1 million (using the mean value of 7%).

A higher percentage of intron reads are found in Total RNA libraries in comparison to poly(A) RNA libraries, as noted in prior studies (11, 18). The mean proportion of reads contributing to exon counts and intron counts for Total RNA libraries in human cell lines is 56% and 21%, respectively – a profound difference of roughly 20% fewer exon counts and 15% more intron counts when compared to corresponding poly(A) RNA samples. There are approximately 6.3 million intron reads for a library of size 30 million.

Unsupervised analysis of intron reads shows strong evidence of signal, not noise. In differential gene expression analyses, plots of principle components analysis and/or multi-dimensional scaling (MDS) methods are commonly created using counts from exon reads. These plots provide an overview of the experiment demonstrating similarities and differences in transcriptional profiles in an unsupervised manner. MDS plots can be used to confirm whether samples cluster into experimental and biological groups, or alternatively to check for unwanted batch effects present in the data.

MDS plots were created in *limma* (26) for the top 500 most variable genes using exon counts. Samples clustered by experimental and biological groups in each of the datasets as expected (Figure 1c). Using a novel approach of applying MDS methods to intron reads rather than exon reads, samples were also shown to cluster by their respective groups using intron counts (Figure 1d). Importantly, these data demonstrate that intron reads are informative, and contain biological information in RNA-seq data when controlling for the same library preparation method.

Intron MDS plots separate samples by type of library preparation and cell type. By comparing exon and intron MDS plots, we show that plot pairs are very similar for two out of three datasets, with those from mouse mammary cells prepared using identical library methods, almost identical to each other. This provides evidence that sources of variation in RNA-seq data is shared between exon and intron reads, be it of a biological or technical nature. For human cell lines, the plot pairs are similar to each other in that samples cluster distinctly into four clear groups. However the factors of separation are switched between the two – the first dimension for cell line and the second dimension for RNA selection protocol in the exon plot, and vice versa for the intron plot. Notably, the first dimension of separation accounts for a larger proportion of variation in the data than the second dimension, indicating that intron read signal is significantly influenced by RNA selection protocols. The first dimension which corresponds to RNA selection in the intron plot accounts for 48% of the variation in the intron counts, whilst the second dimension corresponding to cell lines accounts for 18% of variation in the data. This is in contrast to the exon plot where RNA selection protocol (second dimension) accounts for 26% of variation in the counts, and cell lines (first dimension) accounts for 37%. Thus, the type of RNA selection protocol used for RNA library preparation has a greater influence on intron than exon reads.

The pair of plots for human immune cells are most different to each other. Here, intron counts result in more distinct group clusters than exon counts although the degree of separation is reduced slightly. In the exon plot, CD4+ T cells completely overlap CD8+ T cells; and three whole blood samples are positioned closer to the cluster of monocytes than they are to their own group. In contrast, CD4+ and CD8+ T cells in the intron plot separate out from one another although with some overlap; while whole blood and monocyte samples form perfectly separated clusters. This suggests there may be intron specific differences between CD4+ and CD8+ T-cell subsets that are less evident when analysing exon data.

It is extraordinary to observe that the scale in intron MDS

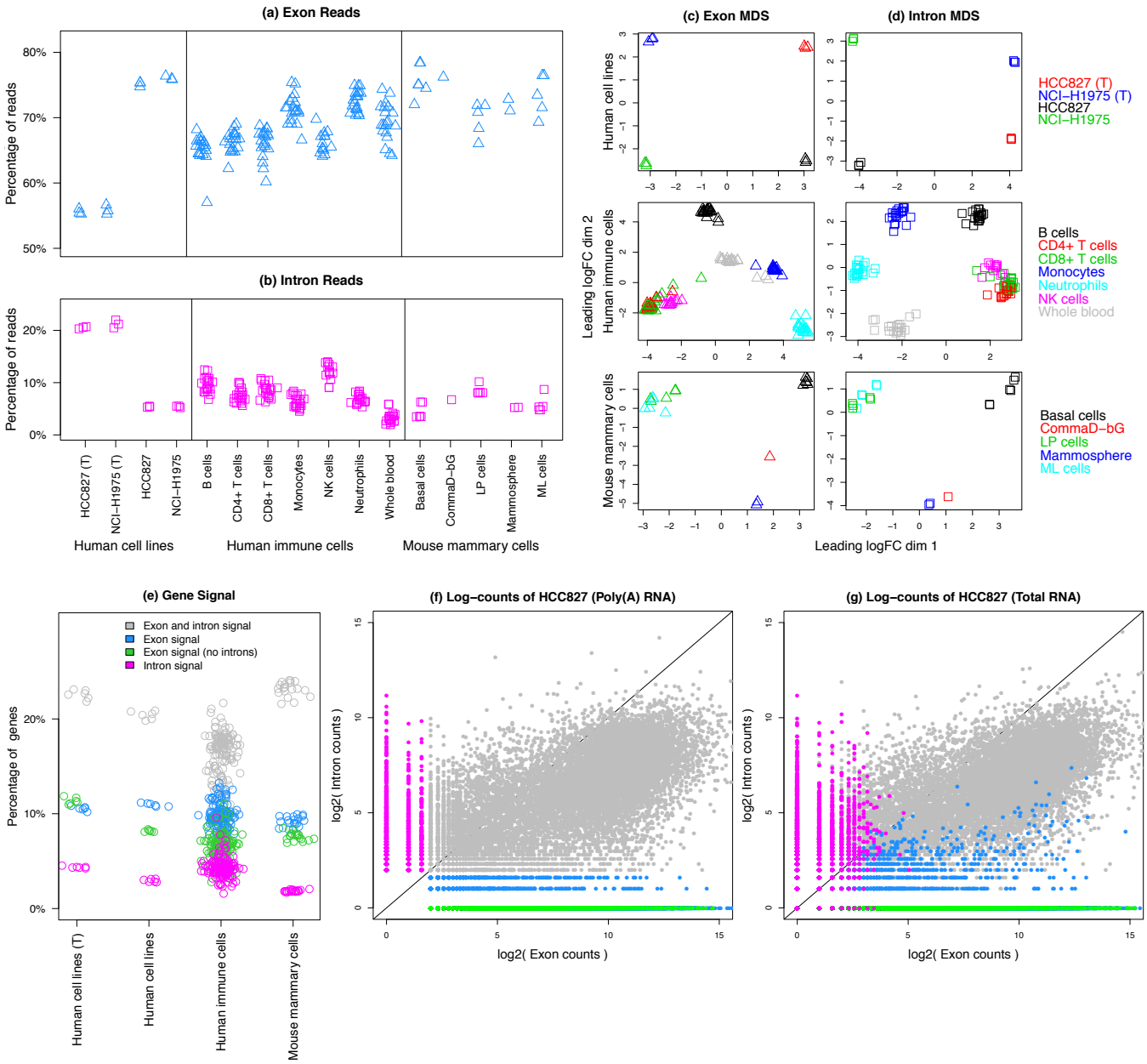


Fig. 1. Intron and exon read characteristics. (a) Percentage of reads assigned to exon (blue triangle) and (b) intron (magenta square) counts across three datasets, with one point per library. Total RNA samples in human cell lines are displayed separately from poly(A) RNA libraries, and have been labelled with a '(T)'. (c) MDS plots of log-counts per million values calculated from gene-level exon counts, and (d) gene-level intron counts. Datasets are separated into different panels. Libraries are coloured by experimental and biological group; triangle and square symbols are used as a visual association with exon and intron reads, respectively. (e) Percentage of genes in each library with exon and intron signal (grey), exon signal (blue) for genes with one or more introns, exon signal with no introns (green) for genes containing a single exon only, and genes with intron signal only (magenta). Each point represents a library. (f) In the poly(A) RNA R1 sample for cell line HCC827, intron log-counts are plotted against exon log-counts. Log-counts are calculated with an offset of 1. Each point represents a gene, where colors differentiate genes with exon and/or intron signal – color legend as in (e). To demonstrate differences in gene-level counts by library preparation protocol, (g) intron versus exon log-counts for the Total RNA R1 HCC8277 sample have genes colored by poly(A) RNA gene signal from (f), rather than the Total RNA count.

plots are quite comparable to exon MDS plots given that there is roughly ten times more exon reads than intron reads. This observation supports the notion that intron reads contain important information. The distance between points on each plot give an indication of the typical $\log_2$-fold change (logFC) between samples in the top most variable genes in each set of counts. In other words, the typical logFC between samples are similar for intron and exon counts.

One hypothesis for the similarity underlying exon and intron based MDS analysis, is that on a gene level, transcriptomes will contain both exon and intron signal, where exon and intron counts are positively correlated. As a result, MDS plot pairs may be comparable in their layout and scale, and such similarities would be expected of intron and exon logFC values.

A majority of expressed genes contain intron signal. We next investigated whether the same genes possess both

exon and intron signal from RNASeq data, or whether genes possess one type of signal and not the other. A gene is considered to contain exon signal if it has an exon count of 3 or more, and intron signal if it has an intron count of 3 or more. Genes are considered to be expressed if they contain exon and/or intron signal. A gene is considered to have no signal if both exon and intron counts are less than 3. Genes consisting of a single exon are non-informative with regards to intron signal – and are therefore separated from genes that have at least one intron.

Of genes that are expressed, most genes contained both exon and intron signal (Figure 1e). This observation is the most likely explanation as to why signal is shared between exon and intron counts and the similarities between MDS plot pairs. Only a small percentage of genes contained intron only reads that would provide information independent of exon counts. Gene signal percentages were consistent across all libraries in all datasets, and in both RNA selection protocols. Log-counts were positively correlated for genes with both exon and intron signal in poly(A) RNA and Total RNA samples (Supplementary Figure 1). For libraries across the three datasets, Pearson correlation between intron and exon log-counts ranged between 60% and 80%, with the exception of whole blood libraries in the human immune cell dataset which has smaller and more variable correlation values that range between 40% to 60%. The level of correlation between intron and exon log-counts is observed to be consistent within cell type and library preparation type. Consistent with our previous analysis, gene-level exon counts tended to be higher than that of gene-level intron counts in poly(A) RNA samples (Figure 1f). Whereas the two count types are similar in magnitude in Total RNA samples (Figure 1g). In general, genes with both exon and intron signal under poly(A) selection also have exon and intron signal under Total RNA selection (Figure 1g). Genes with exon signal only under poly(A) selection tend to have much lower exon read coverage (Supplementary Figure 2) than that of genes with exon and intron signal (Figure 2a and b). Under the Total RNA protocol, some of these genes gain intron signal and are classified as genes with exon and intron signal (Figure 1g). Although intron counts may be similar in magnitude to exon counts, they are generally summarised over a larger interval of bases relative to exons and thus coverage in these regions are lower on average.

Short genes have high intron and exon read coverage. Consistency of intron read properties were demonstrated across multiple datasets in the previous sections, providing general insight into the prevalence of such reads across entire datasets. Next, we delved further into the HCC827 human cell line data for both poly(A) RNA and Total RNA samples which were each available in triplicate. Note that this experiment was designed such that R1 samples were considered to be the same biological sample rather than biological replicates, as was the case for R2 and R3 samples. The only difference between poly(A) RNA R1 and Total RNA R1 samples was in the RNA selection protocol.

Initially we sought to quantify the number of introns that were "expressed" (unspliced, retained, or otherwise) within genes and their positions. However, this analysis demonstrated that both raw and length-adjusted intron-level summarised counts were highly sensitive to the threshold at which "expression" was called. This was largely due to low but highly variable coverage profiles in intron regions. High coverage variability has been demonstrated previously in exons of transcripts. Studying both poly(A) RNA and ribo-depleted Total RNA, Lahens *et al.* (2014) (27) observed that the difference between lowest and highest coverage points is greater than 2-fold in over 50% of transcripts when uniform coverage was expected.

To overcome high coverage variability within and between introns, we took a novel approach, examining local coverage by counting all overlapping reads in 30-basepair bins across the body of genes. We chose bins of 30-basepairs in length to ensure that individual exons and intron were separated into several subregions per feature, where the median length of a single exon in the simplified human GENCODE annotation was 160-basepairs in length with a mean of 414-basepairs. Coverage values were naturally standardised for length since bins are of equal length. Local, or binned coverage is measured for protein coding genes. To discourage coverage unambiguity, only genes that do not overlap another annotated gene were included. We examined specifically the set of genes that had signal in both exon and intron regions in poly(A) RNA samples as these genes also tended to have both exon and intron signal in Total RNA samples (Figure 1g).

By examining the log-coverage of bins we demonstrated that read coverage in exon regions is greater than that of intron regions, and that coverage of intron regions is higher by Total RNA selection relative to poly(A) RNA selection (Figure 2a) – both results are consistent with previous analyses. For each gene, we calculated the average log-coverage of exon bins to represent overall exon expression levels and similarly for intron bins (Figure 2b). Surprisingly, the average exon log-coverage tended to be higher for genes with short exon regions (fewer number of exon bins) than genes with long exon regions in both poly(A) RNA and Total RNA capture protocols. The same was also true of introns – short regions have higher average coverage and long regions have lower average coverage. Overall, this indicates that short genes have a tendency to have higher read coverage and gene-level reads-per-kilobase-per-million (rpkm) values than that of long genes. The trend is subtle for exons in both RNA selection protocols and introns in Total RNA samples, but it is more prominent in introns of poly(A) RNA samples.

Intron coverage patterns depend on library preparation method. Coverage patterns along exons were examined by joining consecutive exon bins from the same gene together. For each gene, exon log-coverage values are adjusted by dividing by its the maximum exon log-coverage. The same was separately performed for intron bins. Coverage patterns in exons and introns of genes were summarised across multiple genes for short, regular and long regions, where gene strandedness is taken into account. For exons, a short region has a median length of 2,490-basepairs, a regular region has median length of 4,680-basepairs, and 7,950-

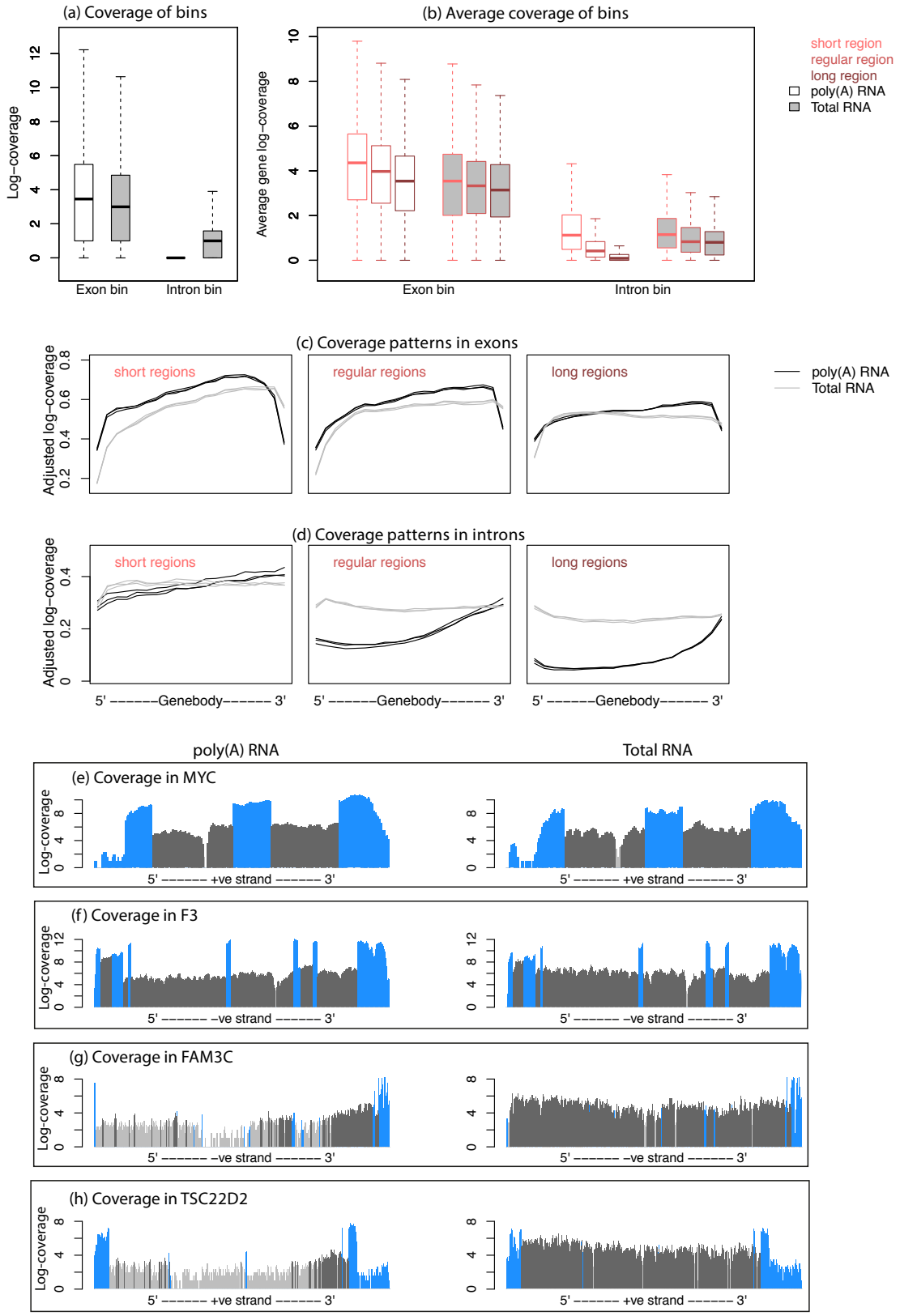


Fig. 2. Coverage of bins. (a) Distribution of log-coverage of exon (left) and intron bins (right), separating poly(A) RNA (white) and Total RNA samples (grey). (b) Distribution of average bin log-coverage of each gene, separating genes into those with short, regular and long exon and intron regions. The average gene log-coverage is calculated as the trimmed mean $\log_2$-coverage, removing the top and bottom 10% of values from each end. Boxplots in (a) and (b) are displayed for R1 samples only and omit outliers. Plots for R2 and R3 samples are similar and not shown. (c) Coverage patterns across the body of genes represented by average adjusted log-coverage. Each line represents a sequencing library. Genes are separated into those with short exon regions (left), genes with regular exon regions (center) and long exon regions (right). (d) Similarly for genes with short intron regions (left), regular intron regions (center) and long intron regions (right). Log-coverage of exon bins (blue) and intron bins (grey) in R1 poly(A) RNA sample (left) and R1 Total RNA sample (right) are displayed for two genes with short intron regions, (e) MYC and (f) F3, and two genes with long intron regions, (g) FAM3C and (h) TSC22D2. Intron bins with high expression (log-coverage greater than 3) are highlighted in dark grey; other intron bins are in light grey.

basepairs for long regions. Exon regions were almost always shorter than intron regions. The top end of short intron regions were as long as some long exon regions. For intron regions, a short region had median length of 5,760-basepairs, regular region had a median length of 20,760-basepairs, and 64,770-basepairs for long regions.

Using our binned coverage approach, it was observed that coverage across exons in poly(A) RNA samples tended to increase from 5' to 3' end, with a drop in coverage at both ends, which is typical of coverage at the terminal ends of a gene (Figure 2c). The gradient of the upward trend decreased as length of the exon region increased. The observed trend is consistent with the results from Lahens *et al.* (2014) (27), where they show that transcripts from poly(A) libraries exhibit 3' coverage bias in their examination of technical biases introduced during the generation of sequencing libraries. In introns, the increasing trend was also observed in poly(A) RNA samples with a steady increase in relative coverage in genes with short intron regions and an exponential increase observed for genes with regular and long intron regions (Figure 2d).

Exon coverage patterns in Total RNA libraries were similar to those of poly(A) RNA libraries, although 3' coverage bias in Total RNA samples was relatively mild. In comparison, intron coverage patterns were drastically different between Total RNA and poly(A) samples, with the exception of short intron regions. In genes with long intron regions, read coverage in Total RNA samples was maintained at a relatively uniform level, whereas poly(A) RNA samples suffered from a decrease in intron coverage at the 5' end. Differences in intron read percentages between poly(A) RNA and Total RNA (Figure 1b) appeared to be a direct result of differences in intron read coverage patterns between the two library preparation methods. Total RNA libraries have higher per library and per gene intron counts since read coverage at the 5' end was greater than when the same biological sample is sequenced under a poly(A) RNA protocol.

Coverage profiles of individual genes were examined to see if the same coverage patterns can be recovered. Two genes with short intron regions, MYC and F3 (Figure 2e,f), and two genes with long intron regions, FAM3C and TSC22D2 (Figure 2g,h), were selected based on having a high average intron log-coverage. Relative to low coverage genes, coverage profiles were easier to observe for genes with higher coverage, and are also of greater interest for the purpose of this study. Consistent with coverage patterns, coverage profiles were similar between poly(A) RNA and Total RNA samples for genes with short intron regions, as shown in the genes individually for MYC and F3. Coverage profiles for genes with long intron regions were distinct from one another, where intron coverage is lower at the 5' end for poly(A) RNA samples and relatively uniform for Total RNA. This was demonstrated by the coverage of FAM3C and TSC22D2 individually.

Genes with the appearance of intron retention. The previous section highlighted common patterns of coverage observed in genes. These analyses showed that coverage was often uniform or changes gradually across the body of a gene. There were, however, exceptions to this observation. Some introns demonstrated coverage that was much higher than other introns within the same gene. An example of this was seen in the genes PSMB7, EIF2S3 and ARGLU1. These three genes resulted from filtering genes for having at least 50 high coverage intron bins (log-coverage greater than 6) and a low average intron bin log-coverage (trimmed mean of 2 or less, removing 10% of values from each end) in the poly(A) RNA R1 HCC827 sample (Figure 3, left).

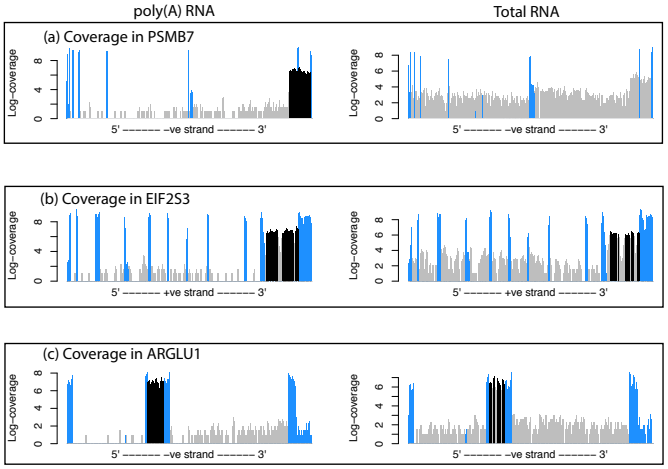


Fig. 3. Distinct and high coverage intron bins (black) for genes (a) PSMB7, (b) EIF2S3, and (c) ARGLU1 within R1 poly(A) RNA sample (left) and R1 Total RNA sample (right).

The coverage profiles of the genes suggested that these may be genes with retained introns. As the cells from which RNA was extracted were not fractionated into nuclear and cytoplasmic components, PSMB7 and EIF2S3, however, could also represent pre-mRNAs with unspliced 3' introns. Of the three genes, ARGLU1 may represent a gene with intron retention since the high coverage intron is positioned in the middle of the gene and has coverage levels similar to its flanking exons. Figure 3 also displays coverage profiles of the same genes in the Total RNA R1 sample (right). We ruled out the possibility that the high coverage region may represent an unannotated exon because the length of region (50 bins or 1500-basepairs) is much larger than a typical exon.

Searching for intronic split reads. Sequencing reads for mRNA would typically contain split reads, or reads that start in one exon and end in another. These reads span the boundary where two exons are joined together after intron splicing, and thus alignment is discontinuous (or split) at the splice site. In this way, split reads are informative of splice events and transcript identification.

Split reads were summarised for HCC827 poly(A) RNA samples for exon and intron annotation separately by setting the *splitOnly* argument in *featureCounts* to TRUE. We allowed reads to be assigned to multiple features. Raw counts of split reads were transformed to $\log_2$-counts (or simply log-counts) using an offset of 1.

Log-counts of split reads for individual exons had a median value of 5.2 in the R1 poly(A) RNA sample. As expected, log-counts of split reads were much lower for individual in-

trons – median value of 0. Forty-two introns in 38 unique genes had high log-counts for split reads (log-count of greater than 5.2) and may provide evidence of unannotated exon boundaries. For genomes of organisms that are less studied and poorly defined, such as chicken or shark, a greater number of introns with high log-counts for split reads are expected and can be used to improve existing annotations.

Discussion

Observing genes with exon and intron signal. Previously, we showed that the majority of expressed genes contain both exon and intron signal. There is therefore great potential in incorporating intron reads into data analysis methods to enhance the amount of information that may be harvested from the sequencing experiment. However, appropriate use of such reads requires an understanding of their properties and where they originated from. In our study, we explored intron read properties in a data-driven manner. Here we considered three possible reasons for observing genes with exon and intron signal and the origin of these reads.

In the first scenario, we considered that pre-mRNA is sequenced (Figure 4a and b). The process of RNA transcription, from DNA to mRNA, is complex and much remains unknown in terms of the exact timing and order of events within the process. The processes of transcription from DNA to primary RNA transcript, polyadenylation of the primary transcript, and the removal of introns through the process of splicing are critical aspects of RNA processing. Polyadenylation occurs upon transcription termination to assist transportation of mRNA into the cytoplasm and for this reason it is assumed that polyadenylation is a marker of the end stages of RNA transcript processing and possibly an identifier of mRNA itself. The assumption that poly(A)+ RNA is (almost) equivalent to mRNA is not unreasonable under co-transcriptional splicing which posits that splicing begins during transcription – an area of study that has been of increasing research interest (11, 28, 29). In fact, Ameur *et al.* (2011) (11) demonstrated that introns are already spliced prior to termination of transcription in humans; Khodor *et al.* (2011) (30) in drosophila and Oesterreich *et al.* (2016) (31) in yeast. However, co-transcriptional splicing has only been demonstrated in approximately 50% of transcripts (32). Moreover, splicing of 5' introns may begin during transcription but due to the longer processing time required relative to transcription, 3' introns may remain unspliced upon transcription termination. Seeing that 3' cleavage of the primary transcript followed by polyadenylation is a relatively fast process, taking up to 30 seconds to complete, this means that co-transcriptionally spliced transcripts can produce poly(A)+ pre-mRNA with 3' unspliced introns.

For genes that are not co-transcriptionally spliced but post-transcriptionally spliced, splicing may be completed after polyadenylation (32). Evidence to support this process includes the presence of poly(A)+ molecules in the nucleus that are much larger than final mRNAs in the cytoplasm. Regardless of whether genes are co- or post-transcriptionally spliced, the proportion at which genes are co-transcriptionally spliced relative to post-trancriptionally spliced and the efficiency of intron splicing, it is possible to capture polyadenylated pre-mRNA with unspliced introns especially at the 3' end.

In the second scenario, we considered the possibility of observing genes with exon and intron signal due to incomplete gene annotation. An unannotated exon or part of an exon can result in reads classified as intron reads (Figure 4d). The missing feature may belong to an annotated gene or a new and unannotated gene. Split reads beginning in an annotated exon and ending within an intron region can be particularly informative since they can provide further information on gene identification and transcript structure.

In the third scenario, we considered the possibility of intron reads originating from processed transcripts with retained introns (Figure 4e). This model has been used as explanation for observing intron reads as studied by Wong *et al.* (2013) (15) and Braunschweig *et al.* (2014) (4). Indeed a retained intron in an expressed gene could result in exon and intron signal observed from the same gene.

DNA contamination was not considered as a possible source of intron reads since an underlying base signal for all genes is expected in the presence of DNA contamination. Instead, more than half of all annotated genes have no reads across all datasets, such that it is likely that contamination levels were so low that it was not observable at the sequencing depths used for the datasets.

Differences in 3' coverage bias. At first glance, intron coverage profiles for genes in the poly(A) RNA sample, such as that of FAMC3C and TSC22D2 in Figure 2g and h, may be explained by the capture of poly(A)+ pre-mRNAs with 3' unspliced introns; but the corresponding Total RNA coverage profiles suggested a different story. With the exception of histone genes for which mRNAs do not include the classical poly(A)-tail (33), we can loosely assume that the two library preparation protocols select for protein coding genes of the same RNA type: poly(A)+ pre-mRNAs and poly(A)+ mRNAs. Whilst an obvious distinguishing feature between Total RNA and poly(A) RNA libraries is in the selection of poly(A)- RNAs by the former, we do not expect considerable contribution by poly(A)- pre-mRNAs in Total RNA libraries given that 3' cleavage and polyadenylation of a new primary transcript occurs quickly, taking only up to 30 seconds for both processes together. As a result, we conclude that the observed differences in coverage profiles is mostly due to library preparation protocol, rather than the sequencing of different RNA types.

Coverage patterns in exon regions (Figure 2c) indicate that both Total RNA and poly(A) RNA libraries suffer from 3' coverage bias. The difference in intron coverage patterns (Figure 2c, g and h) between the two protocols indicates that coverage bias is more prominent in the poly(A) selected libraries, where region length dictates the degree at which 5' coverage is affected. This means that coverage of short intron regions in poly(A) RNA libraries remain mostly unaffected and appear relatively uniform, like that of Total RNA libraries. On the other hand, long intron regions sequenced

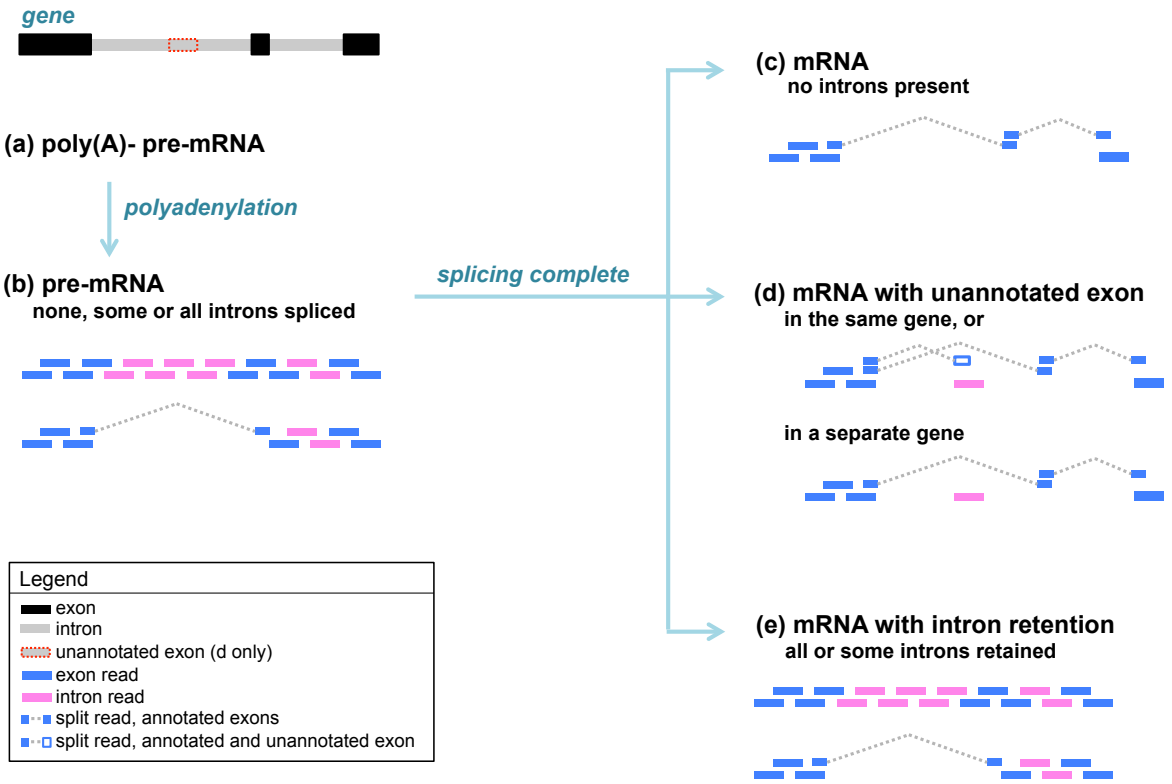


Fig. 4. Reads for mRNA and non-mRNA. (a) DNA is transcribed to a primary RNA transcript, or pre-mRNA, and cleaved at the 3' end upon termination of transcription. (b) A poly(A) tail is synthesised at the 3' end of the transcript. Intron splicing can occur co-transcriptionally or post-transcriptionally, such that poly(A)+ pre-mRNA molecules may be partially or completely spliced. (c) An mRNA molecule is formed upon completion of intron splicing. No introns or intron signal should be present, except in the case of (d) unannotated exons or exon boundaries, or (e) introns are retained. Note that RNA molecules are polyadenylated for (b) to (e); and the unannotated exon is only relevant to (d).

under poly(A) RNA selection suffer from decreasing 5' coverage. The pronounced drop in 5' coverage of long intron regions in poly(A) RNA libraries also explains why gene-wise averages of intron coverage is much lower for long intron regions than they are for short regions (Figure 2b).

Detection and estimation bias in libraries. As a result of 3' coverage bias in introns of poly(A) RNA libraries, our findings have a direct impact on existing intron retention detection methods applied to poly(A) RNA sequencing libraries. There is increased difficulty in interrogating introns residing towards the 5' end of genes due to the decrease in intron information at the 5' end. This may explain why five out of six genes from Figure 1C of the poly(A) RNA experiment by Wong *et al.* (2013) (15) have retained introns in the 3' half of the gene (LMNB1, LBR, PYGL, FXYD5 and NRM); three of which are retained in the 3' most intron. Supplementary Figure 3 shows that coverage profiles of LMNB1, LBR and FXYD5 in our HCC827 cell line data also have high 3' intron coverage in the poly(A) RNA R1 sample, but intron coverage is relatively uniform in the corresponding Total RNA sample. (Annotation of PYGL and NRM overlaps another annotated gene and was omitted from our coverage estimation). When applied to poly(A) RNA libraries, current methods that do not account for intron coverage bias will be bias in detecting 3' intron features.

If intron reads were instead used for the estimation of pre-mRNA abundance, results of pre-mRNA versus mature mRNA levels may varying depending on the library preparation protocol that was carried out and whether estimates are calculated from reads across the whole gene, from reads at the 5' end or 3' end of the gene. In addition, intron signal in RNA-seq libraries are likely to affect *de novo* assembly of transcripts (34) where high intron read coverage in Total RNA libraries and non-uniform intron coverage in poly(A) RNA libraries would complicate transcript detection.

Intron reads in single-cell RNA-seq. The incorporation of intron reads into the analysis of single-cell RNA-seq (scRNA-seq) data has been of much recent interest, where they have been utilised in the processing of raw reads into count tables (35) and estimation of time-derived RNA abundance (BioRxiv: https://doi.org/10.1101/206052). Compared to bulk RNA-seq, scRNA-seq data have much smaller library sizes and relatively high percentages of intron reads. This means that incorporation of intron reads into data analysis may be of signficant importance for single-cell technology. Observing that intron and exon read percentages (Figure 1a and b) and intron-exon log-count correlations (Supplementary Figure 1) are consistent for cell types in bulk RNA-seq data, incorporating these values in modeling and clustering analyses of scRNA-seq data may improve the resolution of unknown cell types.

Given that library preparation in most scRNA-seq protcols

capture only the 3' end of molecules, its sample-wise intron and exon read percentages can be approximately represented by the coverage observed at 3' ends (3' most exon and intron) of bulk poly(A) RNA and Total RNA libraries (Figure 2e-h). Due to the common selection of 3' features, characteristics of scRNA-seq should be more comparable to that of bulk poly(A) RNA data more than it is to Total RNA data. However, because bulk poly(A) RNA libraries have an under-representation of reads for 5' introns and thus an under-representation of intron reads in general, scRNA-seq data is often observed to contain high percentages of intron reads relative to bulk poly(A) RNA libraries. Intron read percentages in scRNA-seq data may be dependent on the length at which 3' fragments are captured relative to the length of 3' exons.

Further work. Much of the difficulty in estimating intron information is by and large due to low but highly variable read coverage in addition to library-preparation-induced coverage bias, resulting in calculation of intron-level estimates that are rather *ad hoc* and/or sensitive to threshold-selection. Improvements to intron summary values, in terms of accuracy and robustness, may be achieved through the development and use of more sophisticated models that adjust for technical sources of variation. Examination of full length transcripts by long-read sequencing, such as that of Pacific Biosciences (36) or Oxford Nanopore Technologies (37), may lead to further insight into the classification of contributing factors for intron reads. If used correctly, future analyses may benefit from the inclusion intron reads per sequenced library.

Conclusions

For methods tailored to the discovery of intron retention and alternative splicing, or even pre-mRNA estimation, it is important to understand the patterns and traits of intron reads that are common to most datasets such that technical artifacts may be adjusted. This in turn would enhance the interpretation of results in carefully designed experiments and allow one to distinguish the results that are unique to the study at hand.

In this paper we examined sequencing reads that are routinely removed from analyses. We show that intron reads are prevalent across multiple datasets and contain signal that can differentiate samples into their biological groups, much to the likes of exon reads. The majority of expressed genes had reads mapping to both exon and intron regions, where log-counts of exons and introns are positively correlated. Coverage profiles of genes tended to be similar between poly(A) RNA and Total RNA samples for genes with short exon and intron lengths, but were different for genes with long introns. Due to 3' coverage bias in poly(A) RNA libraries, genes have reduced intron coverage at the 5' end, whilst coverage is relatively uniform for Total RNA samples. The work presented here therefore provides a broad view of intron coverage patterns and intron-level characteristics across multiple datasets and will help inform future work relating to the study and use of intron reads.

Funding

This work was supported by the National Health and Medical Research Council (NHMRC), Australia (Fellowship 1104924 and Project Grant 1124812 to MER, Project Grant 1060179 to APN); Victorian State Government Operational Infrastructure Support; and Australian Government NHMRC Independent Research Institute Infrastructure Support Scheme (IRIISS).

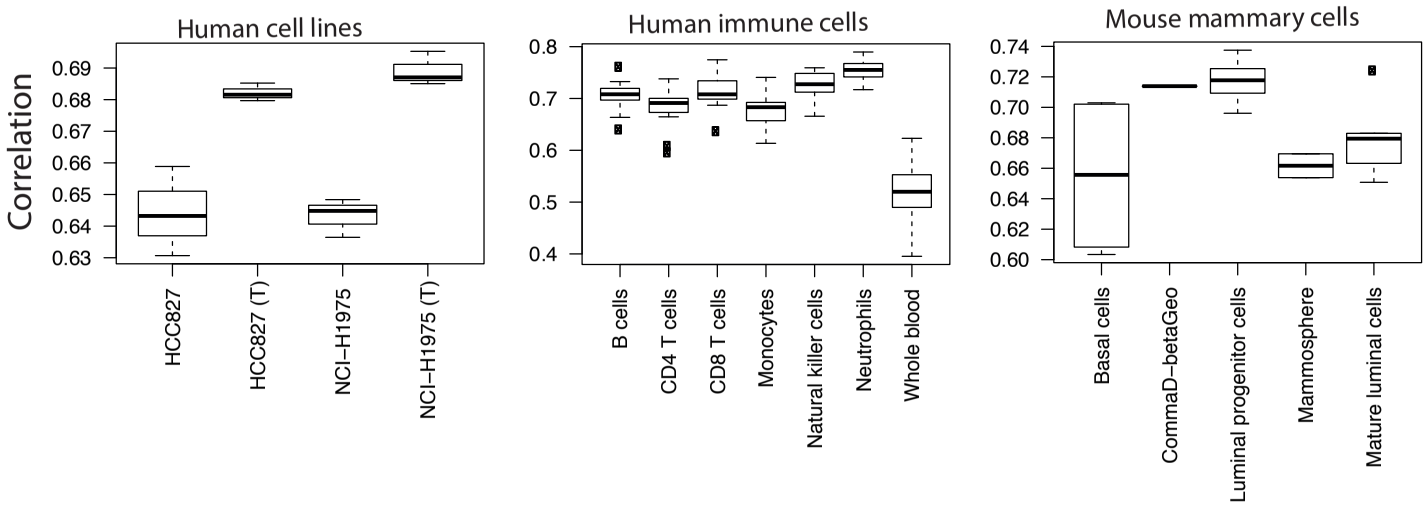
ACKNOWLEDGEMENTS

The authors would like to thank Dr Stephane Chappaz, Dr Quentin Gouil and Dr Clare Morgan for their helpful discussions and suggestions that have enhanced the work presented in this paper.

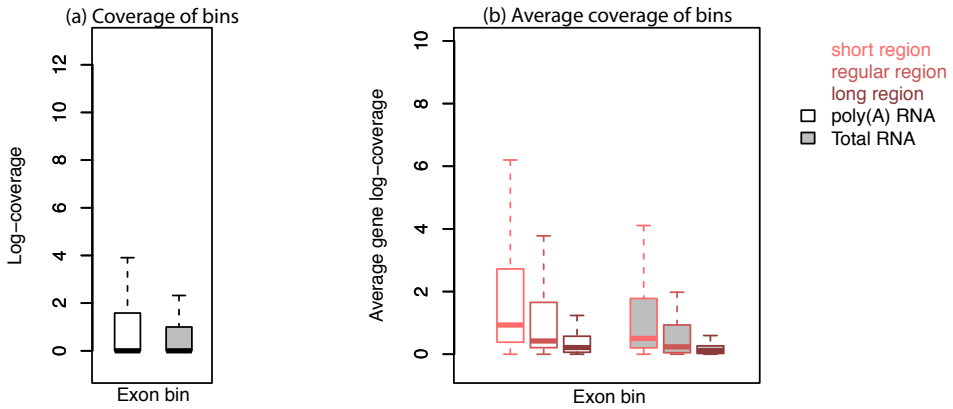
Bibliography

1. M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321:956–960, 2008.
2. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature*, 5(7), 2008.
3. C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol*, 31:46–53, 2013.
4. U. Braunschweig, N. L. Barbosa-Morais, Q. Pan, E. N. Nachman, B. Alipanahi, T Gonatopoulos-Pournatzis, B. Frey, M. Irimia, and B. J. Blencowe. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res*, 11:1774–1786, 2014.
5. N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*, 34:525–527, 2016.
6. R. Patro, G Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat Methods*, 14 (4):417–419, 2017.
7. M. Esteller. Non-coding RNAs in human disease. *Nat Rev Genet*, 12:861–874, 2011.
8. J.S. Mattick and J.L. Rinn. Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol*, 22:5–7, 2015.
9. Y. Guo, S. Zhao, Q. Sheng, M. Guo, B. Lehmann, J. Pietenpol, D. C. Samuels, and Y. Shyr. RNAseq by Total RNA library identifies additional RNAs compared to Poly(A) RNA library. *Biomed Res Int*, 2015(862130), 2015.
10. P. Kapranov, G. St Laurent, T. Raz, F. Ozsolak, C. P. Reynolds, P. H. B. Sorensen, G. Reaman, P. Milos, R. J. Arceci, J. F. Thompson, and T. J. Triche. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' unannotated RNA. *BMC Biology*, 8(149), 2010.
11. A. Ameur, A. Zaghlool, J. Halvardson, A. Wetterbom, U. Gyllensten, L. Cavelier, and L. Feuk. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol*, 18(12):1435–1441, 2011.
12. G. St Laurent, D. Shtokalo, M. R. Tackett, Z. Yang, T. Eremina, C. Wahlestedt, S. Urcuqui-Inchima, B. Seilheimer, T. A. McCaffrey, and P. Kapranov. Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. *BMC Genomics*, 13(504), 2012.
13. D. Inoue, R.K. Bradley, and O. Abdel-Wahab. Spliceosomal gene mutations in myelodysplasia: molecular links to clonal abnormalities of hematopoiesis. *Gene Dev*, 30:989–1001, 2016.
14. D. Gaidatzis, L. Burger, M. Florescu, and M. B. Stadler. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat Biotechnol*, 33(7), 2015.
15. J. J.-L. Wong, W. Ritchie, O. A. Ebner, M. Selbach, J. W. H. Wong, Y. Huang, D. Gao, N. Pinello, M. Gonzalez, K. Baidya, A. Thoeng, T.-L. Khoo, C. G. Bailey, J. Holst, and J. E. J. Rasko. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell*, 154: 583–595, 2013.
16. R. Middleton, D. Gao, A. Thomas, B. Singh, A. Au, J. J.-L. Wong, A. Bomane, B. Cosson, E. Eyras, J. E. J. Rasko, and W. Ritchie. IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol*, 18(51), 2017.
17. S. Harati, J. H. Phan, and M. D. Wang. Investigation of factors affecting RNA-seq gene expression calls. *Conf Proc IEEE Eng Med Biol Soc*, 2014:5232–5235, 2014.
18. S. Zhao, Y. Zhang, R. Gamini, B. Zhang, and D. von Schack. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep*, 8(4781), 2018.
19. A. Z. Holik, C. W. Law, R. Liu, Z. Wang, W. Wang, J. Ahn, M.-L. Asselin-Labat, G. K. Smyth, and M. E. Ritchie. RNA-seq mixology: designing realistic control experiments to compare protocols and analysis methods. *Nucleic Acids Res*, 45(5):e30, 2016.
20. R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–210, 2002.
21. T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*, 41:D991–995, 2013.
22. P. S. Linsley, C. Speake, E. Whalen, and D. Chaussabel. Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. *PLoS One*, 9(10):e109760, 2014.

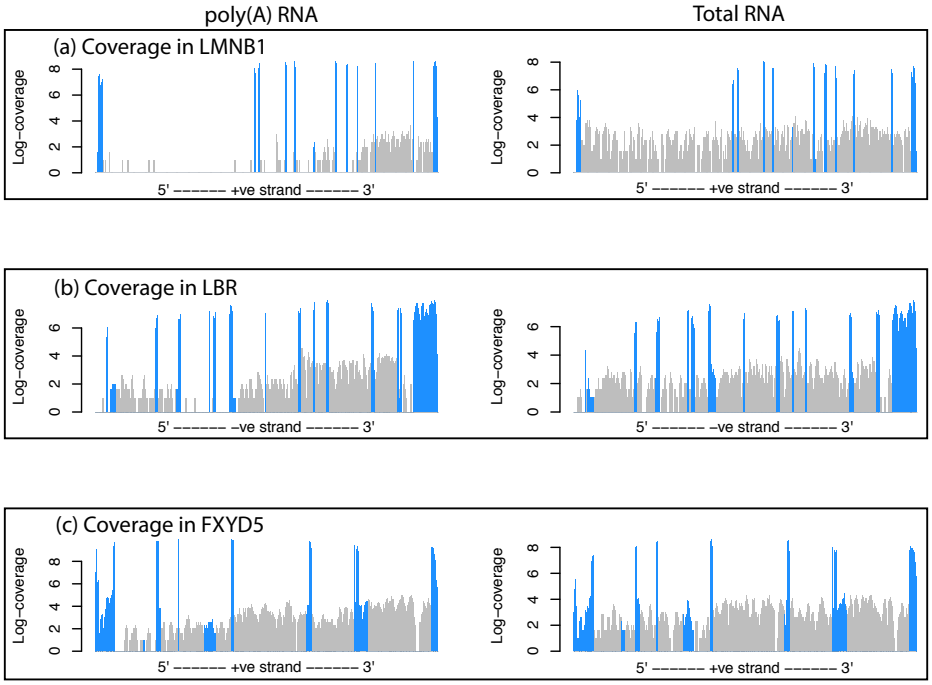
23. J. M. Sheridan, M. E. Ritchie, S. A. Best, K. Jiang, T. J. Beck, F. Vaillant, K. Liu, R. A. Dickins, G. K. Smyth, G. J. Lindeman, and J. E. Visvader. A pooled shRNA screen for regulators of primary mammary stem and progenitor cells identifies roles for Asap1 and Prox1. *BMC Cancer*, 15(1):221, 2015.
24. Y. Liao, G. K. Smyth, and W. Shi. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res*, 41:e108, 2013.
25. Y. Liao, G. K. Smyth, and W. Shi. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–30, 2014.
26. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, 43(7):e47, 2015.
27. N. F. Lahens, I. H. Kavakli, R. Zhang1, K. Hayer, M. B. Black, H. Dueck, A. Pizarro, J. Kim, R. Irizarry, R. S. Thomas, Grant G. R, and J. B. Hogenesch. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol*, 15(6):R86, 2014.
28. E. C. Merkhofer, P. Hu, and T. L. Johnson. Introduction to cotranscriptional RNA splicing. *Methods Mol Biol*, 1126:83–96, 2014.
29. E. Zlotorynski. RNA metabolism: co-transcriptional splicing at nucleotide resolution. *Nat Rev Mol Cell Biol*, 17:264–265, 2016.
30. Y. L. Khodor, J. Rodriguez, K. C. Abruzzi, C. H. Tang, M. T. Marr, and M. Rosbash. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes Dev*, 25(23):2502–2512, 2011.
31. F. C. Oesterreich, L. Herzel, K. Straube, K. Hujer, J. Howard, and K. M. Neugebauer. Splicing of nascent RNA coincides with intron exit from RNA polymerase II. *Cell*, 165(2):372–381, 2016.
32. J. E. Darnell. Reflections on the history of pre-mRNA processing and highlights of current knowledge: A unified picture. *RNA*, 19:443–460, 2013.
33. L. Yang, M. O. Duff, B. R. Graveley, G. G. Carmichael, and L.-L. Chen. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol*, 12(R16), 2011.
34. C. Trapnell, B. A. Williams, G. Pertea, G. Mortazavi, Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28 (5), 2010.
35. S. Parekh, C. Ziegenhain, B. Vieth, W. Enard, and I. Hellmann. zUMIs - a fast and flexible pipeline to process RNA sequencing data with UMIs. *GigaScience*, giy059, 2018.
36. A. Rhoads and K. F. Au. PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289, 2015.
37. H. Lu, F. Giordano, and Z. Ning. Oxford Nanopore MinION sequencing and genome assembly. *Genomics, Proteomics & Bioinformatics*, 14(5):265 – 279, 2016.



Supplementary Fig. 1. Distributions of Pearson correlations calculated between exon and intron log-counts per library. Outliers are marked by black squares.



Supplementary Fig. 2. (a) Distributions of log-coverage of exon bins, separating poly(A) RNA (white) and Total RNA samples (grey). (b) Distribution of average bin log-coverage of each gene, separating genes into those with short, regular and log exon regions. The average gene log-coverage is calculated as the trimmed mean $\log_2$-coverage, removing the top and bottom 10% of values from each end. Boxplots are displayed for R1 HCC827 samples only. Plots for R2 and R3 samples are similar and not shown.



Supplementary Fig. 3. Log-coverage of exon bins (blue) and intron bins (grey) in R1 poly(A) RNA sample (left) and R1 Total RNA sample (right) for genes detected to contain retained introns by Wong *et al.* (2013).