Covering all your bases: incorporating intron signal from RNA-seq data

Stuart Lee[1,2], Albert Y. Zhang[1], Shian Su[1,3], Ashley P. Ng[3,4], Aliaksei Z. Holik[3,5], Marie-Liesse Asselin-Labat[3,5], Matthew E. Ritchie ⬤[1,3,6] **and Charity W. Law[1,3,*]**

[1]Epigenetics and Development Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia, [2]Department of Econometrics and Business Statistics, Monash University, Clayton, Victoria 3800, Australia, [3]Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia, [4]Blood Cells and Blood Cancer Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia, [5]Personalised Oncology Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia and [6]School of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia

ABSTRACT

RNA-seq datasets can contain millions of intron reads per 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 for poly(A) RNA libraries. In this study, we show that intron reads are informative, and through exploratory data analysis of read coverage that intron signal is representative of both pre-mRNAs and intron retention. We demonstrate how intron reads can be utilized in differential expression analysis using our *index* method where a unique set of differentially expressed genes can be detected using intron counts. In exploring read coverage, we also developed the *superintronic* software that quickly and robustly calculates user-defined summary statistics for exonic and intronic regions. Across multiple datasets, *superintronic* enabled us to identify several genes with distinctly retained introns that had similar coverage levels to that of neighbouring exons. The work and ideas presented in this paper is the first of its kind to consider multiple biological sources for intron reads through exploratory data analysis, minimizing bias in discovery and interpretation of results. Our findings open up possibilities for further methods development for intron reads and RNA-seq data in general.

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 microarray required prior knowledge of the sequences to be interrogated, limiting *de novo* discoveries and understanding of gene transcripts and alternative splicing especially at a high-throughput level. Most research efforts focused on gene-level information and comparison of genes that are differentially expressed (DE) between two or more groups. Whilst this is still the main focus for RNA-seq, the technology 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 (IR) at a global level using RNA-seq (1–3).

RNA-seq can be used to characterize and study many RNA types, including non-coding RNAs that regulate a diverse range of cellular processes (4,5), 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 captures polyadenylated (poly(A)) RNA seeing that it is optimized for mRNA selection. In eukaryotes, poly(A) tails are synthesized to aid transportation of mature mRNA molecules from the nucleus to the cytoplasm, to increase molecule stability and for translation. Total RNA selection is also widely used, often including a step to deplete ribosomal RNA so that it does not compete with 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 (6). The general assumption is that for protein coding genes the vast majority of RNA captured by the experiment are mature mRNA transcripts, as such, aligned sequencing reads are typically summarized only for annotated exons within genes. It is not common practice to quantify reads that overlap uniquely with intronic regions of a gene, perhaps because few reads are expected or due to suggestions that intron reads represent experimental and transcriptional noise (7) and/or are unusable in exon and gene quantification (8). However, intron reads can account for a significant proportion of sequencing reads (9).

In stark contrast, a small subset of studies have highlighted the use of intron reads, showing their correlation with measurements of nascent RNA (10) or by using the reads to study IR (3,11). The inclusion of intron reads into data analysis expands areas in which RNA-seq can be used to interrogate transcriptional biology at a high-throughput level, where for example IR has been shown to play important roles in inactivation of tumour suppressor genes (12) and during neutrophil (11) and erythroblast (13) differentiation. Whilst previous works on intron reads have been ground-breaking, they focus solely on one of the many aspects of transcription based on the biological interests of the study at hand. Gaidatzis *et al.* (10) thoroughly explored nascent transcription using intron reads in both poly(A) and total RNA libraries, but without mention of IR. Wong *et al.* (11) assumed that poly(A) RNA libraries contain only processed mRNAs, and successfully showed that genes with differentially retained introns are enriched in the cytoplasm. It is unclear, however, whether the same methods can or should be applied more generally to separate datasets. Perhaps, this would be determined by an expert with prior and thorough understanding of the underlying biology within a given dataset, but does not really allow an analyst to apply the methods to a randomly selected dataset to see if the results 'make sense'. And yet, there is still little consensus on whether intron reads are informative to begin with.

In this paper, we summarize and explore the general characteristics of intron reads in a data driven manner. The work presented here allows for a novel perspective on technical, as well as multiple biological considerations when using intron reads. Demonstrating that intron reads are informative, we find that coverage profiles within intronic regions of poly(A) RNA libraries differ from that of total RNA libraries for genes that are relatively long. We observe that across most of the genes, their coverage patterns and strong correlation between exon and intron counts is consistent with our understanding of pre-mRNA signal. Amongst the pre-mRNA signal, for a human cell lines dataset we also select a small set of genes that have coverage profiles representative of IR. As a result of our exploratory work on intron reads, we have also made two novel methods available—*index* incorporates intron reads into differential gene expression (DGE) analyses; and *superintronic* is used to summarize read coverage for intronic and exonic regions. We expect that the results presented in this paper will better inform of how RNA-seq intron reads can be applied appropriately for various biological interests and further methods development.

MATERIALS AND METHODS

Datasets

Human cell lines of lung adenocarcinoma HCC827 and NCI-H1975 were cultured on three separate occasions by Holik *et al.* (14) giving three pseudo biological replicates. RNA was extracted from each pseudo biological replicate and split into two and prepared as poly(A) RNA and total RNA libraries. Raw sequencing reads were downloaded from the Gene Expression Omnibus (GEO) (15) under accession number GSE64098. Twelve libraries were examined for this dataset.

Human immune cells were sequenced by Linsley *et al.* (16) using a poly(A) RNA library preparation; GEO accession number GSE60424. RNA samples were taken of whole blood and six immune cell subsets, including pure populations of neutrophils, monocytes, B cells, CD4+ T cells, CD8+ T cells and natural killer (NK) cells. A total of 134 libraries were examined for this dataset.

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.* (17) to obtain poly(A) RNA libraries; GEO accession number GSE63310. Mammary cell populations include mammary stem cell-enriched basal cells, luminal progenitor-enriched (LP) and mature luminal-enriched (ML) cell populations. Nineteen libraries were examined for this dataset.

Megakaryocytes and platelets from mice were sequenced separately, with four and six libraries, respectively. Megakaryocytes were sequenced by Choi *et al.* (18) using a poly(A) RNA protocol; GEO accession number GSE116177. Poly(A) RNA libraries of platelets were sequenced by Chappaz *et al.* (19); GEO accession number GSE141161.

Genomes and gene annotations

FASTQ files containing raw sequencing reads were aligned to the human *hg38* or mouse *mm10* genome using *subjunc* (20) with default parameters in the *Rsubread* software package (21). 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). Using 'gene types' (rather than 'transcript types') from Gencode, the 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). Our Supplementary Materials available at https://github.com/charitylaw/Intron-reads contain the scripts to process annotation files, together with other data analyses and supplementary figures.

Intron and exon counts

Aligned reads were summarized by *featureCounts* (22) using exon annotation and genebody annotation separately

to get gene-level *exon counts* and gene-level *genebody counts* respectively. Gene-level *intron counts* are calculated by subtracting exon counts from genebody counts.

Approximately 15% of genes had exon counts that were greater than genebody counts (by a median value of eight counts). This was due to our conservative approach of excluding reads that overlapped features in multiple genes during the read summarization step by *featureCounts* using the argument allowMultiOverlap=FALSE. Under this strategy, some reads were counted towards the exon count set but not the genebody count set. This happens when a read overlaps the exon in one gene and the intron of another gene—it is counted towards exon counts but not genebody counts due to its overlap of multiple genebodies but not multiple exons.

An alternate count strategy sets allowMultiOverlap=TRUE and does not result in higher exon counts than genebody counts. However, this gives ambiguous assignment of reads to counts via the multi-counting of reads, and can return a larger number of total counts than the original number of sequenced reads. This is not desired for our purpose of quantification and classification of reads into exon and intron sets.

Any gene with a larger exon count than genebody count, had its intron count adjusted to zero. Intron counts represent the gain in information when summarizing reads across the whole genebody relative to exonic regions only. Whilst there are other count strategies, such as counting exon–intron boundary reads separately or towards intron counts, we take this approach since our interest is in assessing whether the intron reads that are not typically used contain additional signal.

Coverage patterns

Read coverage of intronic and exonic regions were calculated for poly(A) RNA and total RNA HCC827 human cell lines using our *superintronic* package, via the *Rsamtools* package (23). Genes of interest were restricted to protein coding genes on reference chromosomes, and we removed any genes that overlapped another to simplify the analysis and reduce coverage ambiguity. Genes were then further filtered if they were not expressed in the poly(A) RNA protocol (requiring at least three reads overlapping intronic and exonic regions). A total of 3262 genes were examined and categorized as short, regular or long (roughly 1087 genes in each category) by splitting the length of each gene into three bins by tertiles.

Using associated BAM files and GENCODE v27 annotation GTF, *superintronic* summarized the number of bases covered at a given coverage score for each gene and sample. For exonic regions and intronic regions, coverage scores were transformed to $\log_2$-scale using an offset of 0.5, and then normalized by dividing by each gene and each sample's maximum log-coverage score. Normalized log-coverage scores, or *relative log-coverage* (relative to each gene's maximum coverage), were divided into 20 windows along the length of each gene using the *GenomicRanges* package (24). To summarize coverage patterns across genes, *plyranges* (25) was applied to relative log-coverage scores by intersecting it with the positional windows. Each window's mean coverage score (mean relative log-coverage) was calculated for each gene. Based on the position of the windows, they were further summarized across genes by taking its mean. The summary values were calculated separately for genes categorized as short, regular and long to represent general coverage trends along the gene body.

RESULTS

Intron reads are prevalent across datasets

Taking a conservative approach, we quantify the number of intron reads that map entirely to an intron of a gene, excluding those that overlap an exon-intron boundary. Gene-level intron counts represent the extra counts one may obtain from within a gene when looking outside of annotated exons. The proportion of reads contributing to gene-level intron counts ranges from 2 to 14% with a mean value of 7% for poly(A) RNA libraries across three datasets examined (Figure 1A). A greater proportion of reads contribute towards gene-level exon counts, ranging from 57 to 78% with a mean of 69% (Figure 1B). Despite the relatively small proportion 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 ∼2.1 million (using the mean value of 7%).

A higher proportion of intron reads are found in total RNA libraries in comparison to poly(A) RNA libraries, as noted in prior studies (8,9). The mean proportion of reads contributing to intron counts and exon counts for total RNA libraries in human cell lines is 21 and 56%, respectively—a profound difference of roughly 15% more intron counts and 20% fewer exon counts when compared to corresponding poly(A) RNA samples. This equates to ∼6.3 million intron reads for a library of size 30 million. Intron and exon read proportions are fairly consistent for libraries within the same biological and experimental groups (such as within cell lines, cell types, tissues and RNA library protocol). Some variation in read proportions can be observed for different biological (e.g. neutrophils versus whole blood) and experimental (e.g. poly(A) RNA versus total RNA HCC827 cell line) groups. Within libraries, exon log-counts are positively correlated with intron log-counts (Figure 1C).

Intron reads are informative and contain biological signal

In DGE analyses, plots of principle components analysis and multi-dimensional scaling (MDS) methods are commonly created from exon counts to provide an overview of the similarities and differences in transcriptional profiles in an unsupervised manner. To determine whether intron reads contain any biological signal, we applied MDS methods to intron counts instead. Samples cluster by experimental and biological groups in intron MDS plots across all datasets (Figure 1D) indicating that intron reads are informative, rather than a result of sequencing noise. As expected, samples also cluster by experimental and biological groups in exon MDS plots (Figure 1E).

Strikingly, the scales observed in the first and second dimensions of the intron MDS plots are comparable to that of the exon MDS plots even though there are roughly ten times
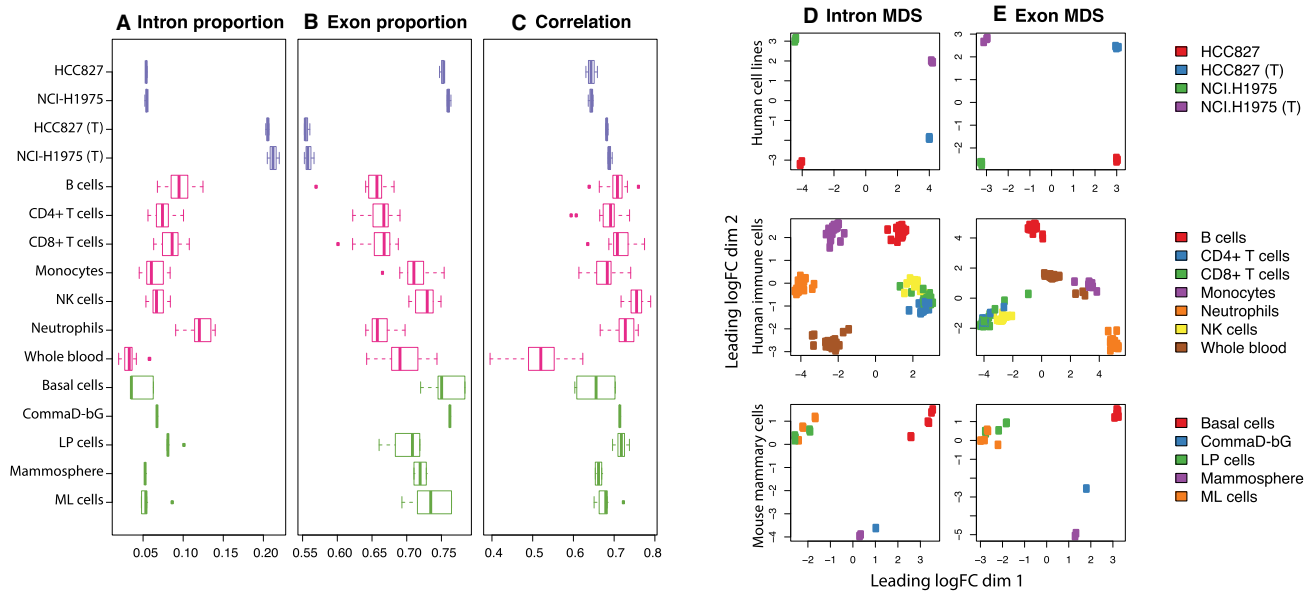
Figure 1. With libraries separated by biological and experimental groups, various statistics are summarized as boxplots across the datasets (distinguished by colour—purple for human cell lines, pink for human immune cells and green for mouse mammary cells) and total RNA samples labelled with a '(T)'. (**A**) Proportion of reads assigned to intron and (**B**) exon counts. (**C**) Pearson correlation of gene-level exon $\log_2$-counts (log-counts) and gene-level intron log-counts. Log-counts are calculated for genes expressed (count of three or more) in both intron and exon count sets, using an offset of 1. (**D**) MDS plots of $\log_2$-counts-per-million (log-CPM) values calculated using an offset of two for gene-level intron counts and (**E**) gene-level exon counts for each of the three datasets. MDS plots were created using *limma*'s (32) *plotMDS* function based on the top 500 most variable genes.

fewer intron counts than exon counts. The distance between points on each plot give an indication of the typical $\log_2$-fold change (logFC) between samples in the top 500 most variable genes in each set of counts. In other words, the typical logFC between samples are similar for intron and exon counts.

Note that the first dimension of separation accounts for a larger proportion of variation in the data than the second dimension. The MDS plots for human cell lines indicate that intron read signal is significantly influenced by RNA selection protocols. The first dimension in the intron plot separates samples based on RNA selection method and accounts for 48% of the variation in the intron counts, whilst the second dimension relates to cell line identity and accounts for 18% of variation in the data. This is in contrast to the exon plot where the RNA selection protocol (second dimension) accounts for 26% of the variation in counts, and cell lines (first dimension) accounts for 37%. The type of RNA selection protocol used in library preparation has a greater influence on intron reads than exon reads.

Comparing counts from poly(A) and total RNA libraries

The human cell line dataset allows us to further explore count differences between library preparation methods. Gene-level exon log-CPM values are similar between poly(A) RNA and total RNA libraries and have a very strong positive correlation (Supplementary Figure S1). Gene-level intron log-CPM values are also positively correlated but counts tend to be greater in total RNA than poly(A) RNA libraries (Supplementary Figure S1). Log-CPM values were calculated using an offset of two and by setting the library size as the sum of counts from exons and

introns (log-RPKM values are calculated in the same way). This allows adjustment of intron and exon counts by the same sequencing depth per library, rather than an intron- or exon-specific proportion of the original sequencing depth.

Within libraries, we found that the majority (56% on average) of expressed, multi-exonic genes contain both intron and exon signal simultaneously. This was calculated by looking for the percentage of multi-exonic genes with counts of three or more in both intron and exon count sets, out of genes that are expressed. Expressed genes were defined as having a count of three or more in exon and/or intron counts. 32% of expressed, multi-exonic genes were expressed in exon regions only (exon count $\geq$3, intron count $\leq$2) and 13% of genes were expressed in intron regions only (exon count $\leq$2, intron count $\geq$3) on average.

To understand the nature of intron counts in relation to exon counts, we focus on the set of genes that are expressed in both regions, noting that total intron length of genes tend to become disproportionately large relative to total exon length (Supplementary Figure S2). Within poly(A) RNA libraries, intron and exon log-CPM and log-RPKM values are positively correlated (Supplementary Figure S3). Log-CPMs provide a reflection of the size of counts used as inputs to many analysis methods, and log-RPKMs are adjusted for length differences and provide a representation of read coverage levels. Intron coverage tends to be lower than exon coverage within the same gene, but the relative difference is quite stable across the genes (Supplementary Figure S3). The median difference between genewise exon log-RPKM and intron log-RPKM values is $\approx$5.1 across all poly(A) RNA HCC827 and NCI-H1975 cell line libraries, such that gene-wise exon coverage is roughly 34 times greater than intron coverage on average. Average

intron coverage is affected by total length of intron regions in genes, such that genes with longer intron regions tend to have lower log-RPKM values (Supplementary Figure S3). Also, relative coverage of exons over introns increases as the total length of intron regions increases, and as exon log-RPKM increases (Supplementary Figure S3).

Similar trends are observed in total RNA samples (Supplementary Figure S3), though it is worth noting that for total RNA libraries the intron and exon counts have similar count size and dynamic range, and have stronger correlation of log-CPM and log-RPKM values between introns and exons. For total RNA libraries, exon coverage is roughly 10 times greater than intron coverage on average (median difference between gene-wise exon log-RPKM and intron log-RPKM values is ~3.3 across all total RNA HCC827 and NCI-H1975 cell line libraries).

Intron reads are predominantly from pre-mRNA

The notion that intron reads originate from pre-mRNA molecules rather than genes with retained introns is supported by the observation that gene-level exon counts tend to have a strong positive correlation with intron counts across all genes (Supplementary Figure S3), and that a large proportion of expressed, multi-exonic genes express intron and exon signal simultaneously. Assuming that IR is generally not widespread across all genes, a weak positive correlation is expected between intron and exon counts if intron reads were predominantly coming from genes with retained introns.

To verify this, we examined intron and exon read counts from nucleated megakaryocytes and compared this to their anucleate platelet progeny. We observe overwhelmingly that intron reads are detected in megakaryocytes but are not detected in platelets (Figure 2A), leading us to conclude that the majority of intron reads correspond to premRNA. Exon reads, on the other hand, are detected in both megakaryocytes and platelets.

Coverage patterns across the genebody of multi-exonic genes provide further evidence that intron signal is predominantly from pre-mRNAs with unspliced or partially spliced introns, where intron reads tend to be uniformly distributed in total RNA libraries and increasing gradually towards the 3′ end of genes in poly(A) RNA libraries (Figure 2B). Total gene length appears to play a part in intron coverage patterns. Genes under examination were catergorized by total gene length (length from first base in first exon to last base in last exon) such that a third were considered to be *short*, *regular* and *long* genes each. Coverage patterns were similar between poly(A) RNA and total RNA libraries for short genes, whilst patterns differed substantially for regular and long genes. Exon coverage patterns were similar for poly(A) RNA and total RNA libraries.

The coverage patterns were calculated by dividing per base log-coverage values in exonic regions by the maximum exonic log-coverage value in each gene. We refer to these values as *relative log-coverage* values in exonic regions (see 'Materials and Methods'). The same is carried out for intronic regions. To summarize relative log-coverage over multiple genes, the values were averaged within windows
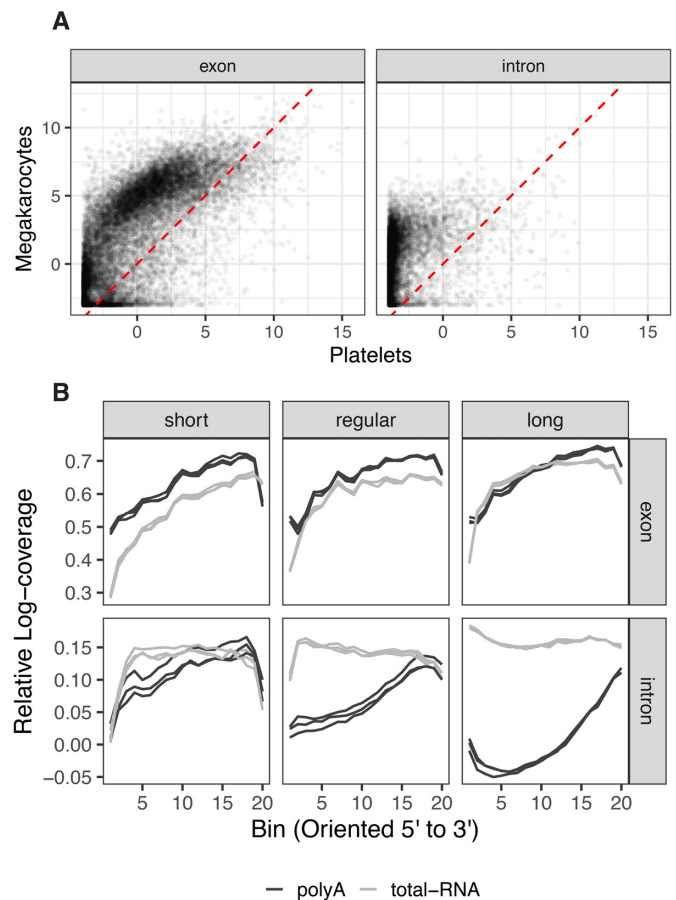

Figure 2. (**A**) Log-CPM values for exon counts in mouse megakaryocytes versus platelets; and for intron counts. Log-CPM values are calculated using the combined library size of intron and exon count sets, and the average value across biological replicates is plotted. Only multi-exonic genes and those that have a mean raw count of three or more in at least one count set are included. There are a total of four count sets—exon and intron counts for megakaryocytes, and exon and intron counts for platelets. (**B**) Coverage of intron and exon regions across the genebody. Top: Coverage patterns of exon regions across the genebody, separated into groups based on gene length. Coverage is represented by the mean of mean relative log-coverage, where relative log-coverage is calculated as local log-coverage divided by maximum log-coverage in a given gene. HCC827 cell line samples are depicted with black lines representing poly(A) RNA libraries R1, R2 and R3, and grey lines representing the corresponding total RNA libraries. Bottom: Coverage patterns of intron regions across the genebody.

in each gene before taking the average of windows across genes.

Confirming the same patterns for individual genes, coverage profiles of two short genes F3 and MYC are observed to be similar between RNA library preparation (Figure 3A and B). In contrast, the coverage profiles of two long genes TSC22D2 and FAM3C differ at the 5′ end where poly(A) RNA libraries are observed to have deflated intron coverage relative to the 3′ end, as well as relative to total RNA libraries (Figure 3C and D). Reduction in read coverage at 5′ introns for poly(A) RNA libraries explain why poly(A) libraries are observed to have relatively low proportions of intron reads (Figure 1A) and dynamic range in log-CPM values relative to total RNA libraries (Supplementary Figure S3).
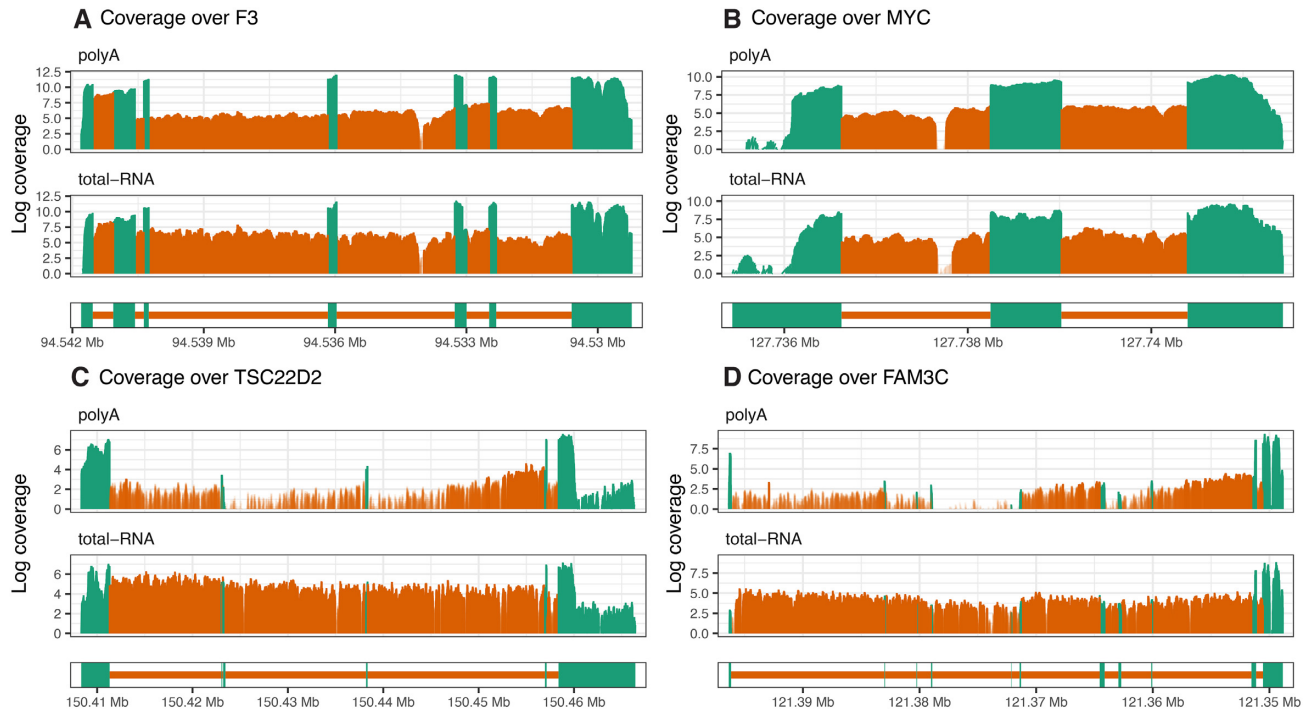
Figure 3. Log-coverage in exon regions (green) and intron regions (orange) in HCC827 poly(A) RNA R1 library (top) and HCC827 total RNA R1 library (bottom) are displayed for two short genes, (**A**) F3 and (**B**) MYC, and two long genes, (**C**) TSC22D2 and (**D**) FAM3C. Genes are oriented from 5′ to 3′, from left to right. These genes were selected based on having a high median intron log-coverage. Intron regions with high expression (log-coverage greater than three) are highlighted by a shade of darker orange.

Relative gene-level contribution by pre-mRNAs is low compared to mature mRNAs for the majority of genes, as reflected in low intron log-RPKM relative to exon log-RPKM values (Supplementary Figure S3). If intron and exon log-RPKM values provide an estimate of the relative proportions of pre-mRNA and mRNA molecules captured, then on average roughly 1 pre-mRNA molecule is captured for every 10 molecules in a sequencing experiment (since exon coverage is roughly 10 times greater than intron coverage in total RNA libraries, and total RNA libraries have uniform intron coverage). Unless the sequencing experiment is carried at very high depths, intron signal may not be detected for genes with relatively short intron regions since pre-mRNA levels are low, whilst genes with long intron regions have greater ability to accumulate sequencing reads over the gene.

DGE analyses of transcriptional activity using intron and exon counts

Classical DGE analyses are performed on gene-level exon counts, where in light of results from the previous section we have an understanding that the associated reads originate from mRNA as well as pre-mRNA molecules. Previously, signal from microarrays designed with exonic and intronic probe sets were used to study transcriptional dynamics of pre-mRNAs and mRNAs (26). For RNA-seq data, we propose a method that complements the classical DGE analysis and includes intron counts to measure changes in early transcriptional activity. We call our method

index, **in**tron **d**ifferences to **ex**on, a DGE method categorizing genes by significance and directional changes in intron and exon counts. Relative to a classical DGE analysis which requires gene-level exon counts and some information about the experimental design and comparisons of interest, *index* simply requires an addition of gene-level intron counts. *Index* is an R package which is available to download and install at https://github.com/Shians/index.

The *index* workflow (Figure 4A) is carried out on genes that are expressed in both intron and exon regions. Firstly, sufficiently large intron and exon counts are selected by the *filterByExpr* function in *edgeR* (27,28). Trimmed mean of M-values (TMM) normalization (29) is then carried out on intron and exon count sets separately using the combined library size for samples (sum of both pre-filtered intron and exon counts). This is a variation on the standard library size calculation which only sums counts within a single count set. Intron and exon counts evaluated based on standard library sizes will be affected by intron and exon read proportions (Figure 1A and B) which vary between samples, groups and experiments and gives a poor estimate of original sequencing depth. The combined library size is used also for downstream calculations, such as in obtaining log-CPM values by *voom* (30).

DE genes are obtained for intron and exon counts separately following a standard *limma-voom* pipeline (31) where log-CPM values and variables of interest are modelled on a Normal distribution with precision weights calculated by *voom*. Moderated *t*-statistics are calculated for each gene using Empirical Bayes' methods (32,33) and *P*-values are
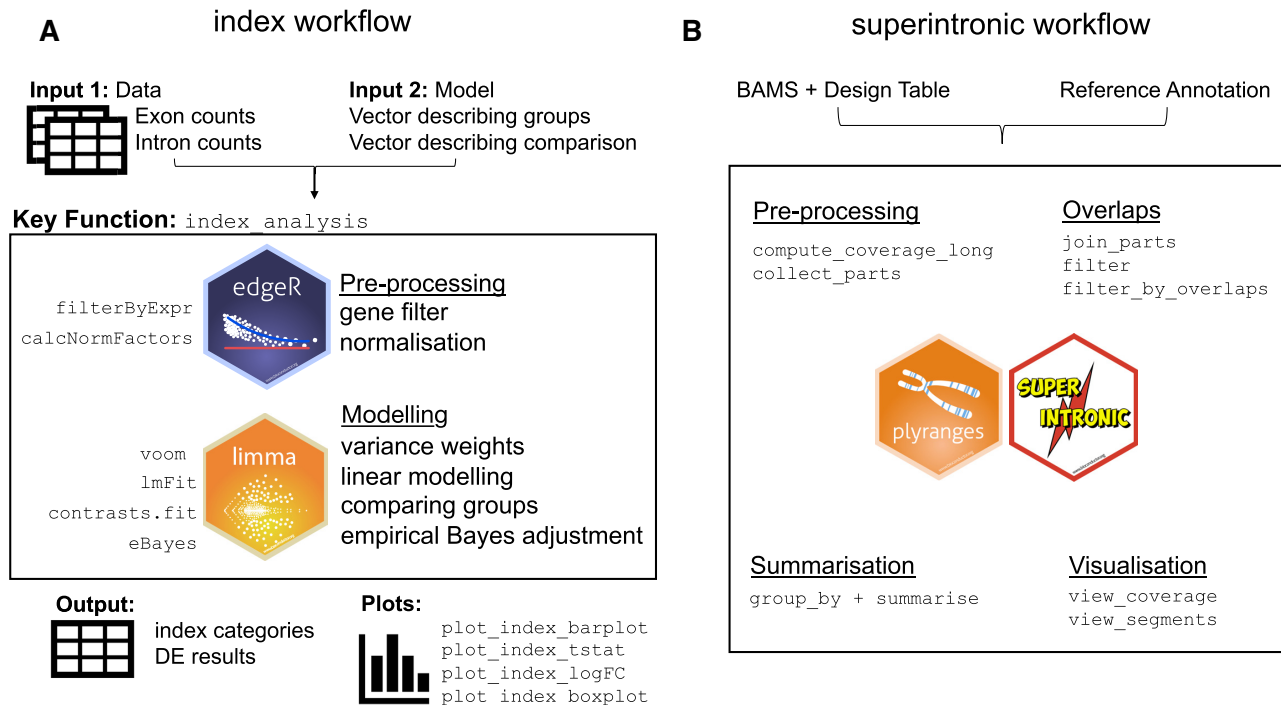
A index workflow

Input 1: Data
Exon counts
Intron counts

Input 2: Model
Vector describing groups
Vector describing comparison

Key Function: `index_analysis`

`filterByExpr`
`calcNormFactors`

edgeR

Pre-processing
gene filter
normalisation

`voom`
`lmFit`
`contrasts.fit`
`eBayes`

limma

Modelling
variance weights
linear modelling
comparing groups
empirical Bayes adjustment

Output:
index categories
DE results

Plots:
`plot_index_barplot`
`plot_index_tstat`
`plot_index_logFC`
`plot_index_boxplot`

B superintronic workflow

BAMS + Design Table

Reference Annotation

Pre-processing
`compute_coverage_long`
`collect_parts`

Overlaps
`join_parts`
`filter`
`filter_by_overlaps`

plyranges

SUPER INTRONIC

Summarisation
`group_by + summarise`

Visualisation
`view_coverage`
`view_segments`

Figure 4. Overview of the *index* and *superintronic* workflows. (**A**) Beginning with matrices containing exon and intron counts by gene and vectors describing an experimental design and contrasts of interest, *index* uses *edgeR* for normalization and gene filtering. Then using *voom*, DE results are computed and are categorized according to significance in exon and intron counts. (**B**) The modular workflow for exploratory genomics data analysis using *superintronic* and *plyranges*. The software requires BAM files, gene annotation and information on experimental design as input prior to any computation. Each stage can be performed independently depending on the end goal of the analysis. The pre-processing steps compute coverage as a long form *GRanges* object in parallel and with respect to an experimental design. The annotation can also be used to construct the exonic and intronic parts of a gene. Region-based overlaps or filters can be performed to either zoom in on a gene of interest or to split coverage over intron and exon parts. Coverage can then summarized over regions with respect to a design or a sample using a statistical summary such as the mean, sum or standard deviation or any suitable R function. A suite of visualization functions are provided to look at coverage in the context of gene annotations or for finding interesting regions of coverage.

adjusted for multiple testing by controlling the FDR (34). Genes with an adjusted *P*-value of less than a nominal cutoff are considered to be DE.

Index categories are formed based on significance in intron and exon counts: **+** for genes upregulated in both intron and exon counts, **-** for downregulated in intron and exon counts, **exon+** and **exon-** for up- and downregulated in exon counts only, **intron+** and **intron-** for up- and downregulated in intron counts only, **mixed+-** for upregulated in exons and downregulated in intron counts and **mixed-+** when in the opposite direction and **0** for no significant difference in either exon or intron counts.

The *index* software performs analysis on intron and exon DGEList objects (a native object of *edgeR*) to classify genes into the respective *index* categories. *Index* outputs categories assigned to each gene, *limma*-style tables of DGE results for introns and exons, and other data used by the software to create plots. This allows the *index* analysis to be easily performed on any dataset where intron and exon counts can be obtained separately.

Index analysis of human cell lines and immune cells

An *index* analysis comparing NCI-H1975 versus HCC827 cell lines in total RNA libraries reveals that the majority of genes are DE in the same direction between intron and

exon counts—2406 genes upregulated (+) in NCI-H1975 and 2464 downregulated (−) using an adjusted *P*-value cutoff of 0.01 (Figure 5A–C). Genes DE by exon counts only form the second biggest group, with 989 genes upregulated (exon+) and 914 genes downregulated (exon-) in exon counts. Interestingly, these genes tend to have short intron regions (Figure 5D). There are 547 genes upregulated (intron+) and 459 genes downregulated for intron counts only (intron-), where genes tend to have relatively long intron regions. Similarly, genes DE in opposite directions also have relatively long intron regions—a small group of 25 genes upregulated in exon counts but downregulated in intron counts, and 29 genes upregulated in intron counts but downregulated in exon counts. The analysis was carried out on 11608 genes after lowly expressed genes were filtered out.

An identical analysis comparing monocytes versus neutrophils in the immune cells dataset similarly reveals that the majority of genes are DE in the same direction between intron and exon counts—3491 genes are upregulated (+) in neutrophils relative to monocytes and 2628 downregulated (−) using an adjusted *P*-value cutoff of 0.01 (Figure 5E–G). Again, genes DE by exon counts only form the second biggest group, with 1073 genes upregulated (exon+) and 969 genes downregulated (exon−) for neutrophils in exon counts. These genes again tend to have shorter intron regions, but the difference in length between *index* cater-
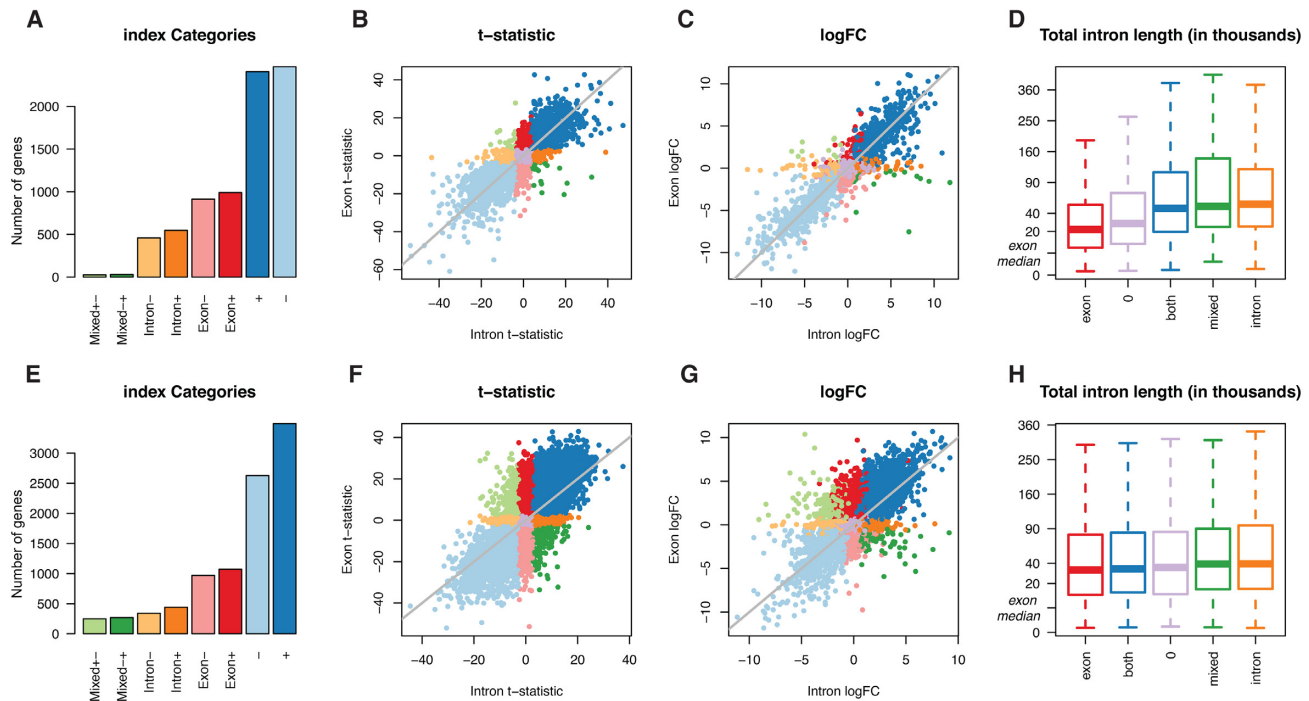
Figure 5. Results for DGE analysis of intron and exon counts using *index*. (**A**) Number of significant genes in *index* categories for NCI-H1975 versus HCC827 cell lines in total RNA libraries, with (**B**) t-statistics from exon counts plotted against those from intron counts and (**C**) logFC values from exon counts plotted against those from intron counts, where colours are associated with different *index* categories. (**D**) Distribution of total intron length ordered by median total intron length of *index* categories. The categories are combined here, such that genes that are upregulated or downregulated for both intron and exon counts are reassigned to *both*; *exon* for genes up- or downregulated for exon counts only, similarly for *intron* and *mixed*. A square-root scale is used along the vertical axis, with the median total exon length (5126 bases) marked as a reference. (**E–H**) Similar plots are displayed for a comparison of monocytes and neutrophils in the immune cells dataset.

gories is subtle compared to that of cell lines (Figure 5H). There are 441 genes upregulated (intron+) and 340 genes downregulated (intron-) for intron counts only, where genes again tend to have longer intron regions. The number of genes DE in opposite directions between intron and exon counts is larger for this comparison than in cell lines, with 249 genes upregulated in exon counts but downregulated in intron counts (mixed+−) and 269 genes downregulated in exon counts but upregulated in intron counts (mixed−+). The analysis was performed on 10789 genes after lowly expressed genes were filtered out.

Index analysis detects additional DE genes

The *index* DGE analyses demonstrate that transcriptional changes detected by exon counts are similar to those detected by intron counts. This is expected since exon counts represent mRNA and pre-mRNA levels, whilst intron counts largely represent pre-mRNA levels. For most genes, similarity between intron and exon logFCs (Figure 5C and G) indicate that pre-mRNA and mRNA levels are simultaneously up- or downregulated at similar proportions between groups.

We hypothesize that genes categorized as intron+ or intron− mostly contain changes in pre-mRNA levels only. To verify this, we examine read coverage profiles for genes with the largest expression differences in cell line samples (largest absolute *t*-statistics in intron+ and intron− categories). We observe that genes in the intron+ and intron−

categories have coverage profiles that are consistent with what we would expect of changes in pre-mRNA levels, such that reads are covering most of or all of the genebody for one cell line and higher than that of the other cell line (see 'Analyses' page in Supplementary Materials).

Assignment of genes into different *index* categories is associated with total intron length of a gene (Figure 5D and H), such that genes DE for exon counts only tend to have relatively short intron regions. Naturally, these genes are unlikely to accumulate high intron counts due to low coverage and short region lengths, thus lacking power during statistical testing. On the other hand, genes DE for intron counts only tend to have relatively long intron regions; due to their length they are able to accumulate high intron counts even if coverage levels are low, giving them a power advantage when testing for differential expression.

In other words, exon+ and exon− genes may also contain changes in intron regions even though they remain undetected. Alternative explanations for observing significant changes in exon counts only are less likely, for example, that there are no pre-mRNAs observed, or that pre-mRNA levels are consistent between groups. The former is contradicted by Supplementary Figure S3 which shows high intron coverage for genes with short intron regions, and the latter is unlikely to be a trait specific to genes with short intron regions. Similarly, intron+ and intron− genes may also contain changes in exon regions even though they remain undetected.

If intron counts represent pre-mRNA levels, then any change observed between groups in intron counts should also be reflected in exon counts. However, exons are unlikely to accumulate high counts over its relatively short regions if pre-mRNA (and mRNA) levels are very low. If genes have retained introns or differentially retained introns in one group versus another, it is also possible for genes to be detected as DE in intron counts. Intron+ and intron− genes can be compared against a list of genes detected with retained introns (see *superintronic* in next section). Given that significance is influenced by total intron length, it is possible that exon+, exon−, intron+ and intron− genes may be reclassified into + and − *index* categories if sequencing was performed at greater depths.

Mixed+− and mixed−+ genes form a relatively small set of genes relative to other *index* categories. Biologically, a simultaneous increase in pre-mRNA levels and decrease in mRNA levels between two groups can induce changes in opposing directions between intron and exon counts. For example, this may occur during nonsense mediated decay (35), where post-transcriptional regulation may lead to a decrease in mature transcripts.

An *index* DGE analysis adds an extra layer of information by overlapping intron and exon results, where additional DE genes are detected that are not observed in a classic DGE analysis alone. The *index* method has increased power in genes with long intron regions, where high counts can be detected for low coverage genes. A classic DGE analysis, by *limma-voom* or like methods, followed by an *index* DGE analysis allows researchers to make use of a larger proportion of reads that are already sequenced and available to them to detect additional DE genes.

Superintronic: an exploratory approach to detecting genes with IR

Considering IR also as another possible source of intron reads we propose a new method using our *superintronic* software to explore intron signal directly from aligned sequencing data with the assumption that most intron reads do not point to IR but pre-mRNA instead. *Superintronic* is an R package that is available to download and install at https://github.com/sa-lee/superintronic. It extends the *plyranges* Bioconductor package (25) for genomics data analysis to develop a simple and modular interface for performing exploratory genomics data analysis via coverage estimation. Each aspect of the *superintronic* data analysis workflow as it has been applied for exploring intron signal is outlined in Figure 4B.

Our software records the per base coverage over intron and exon regions of each gene, with the option of storing these per sample or summarized over variables in the experimental design such as by biological group or by RNA library preparation. Coverage scores are normalized using a $\log_2$-transformation with an offset of 0.5 to get log-coverage values for which intron and exon summary statistics are constructed for each gene (described below). Within *superintronic*, a suite of visualization tools to construct coverage plots for genes with intron and exon structures and scatter plots are provided.

Superintronic finds genes with IR-like coverage profiles in human cell lines

Using *superintronic*, poly(A) RNA HCC827 cell lines were examined for genes with IR after selecting genes in the hg38 reference that were protein coding, did not overlap any other gene and were placed on the main contigs—a total of 6606 genes. These genes were then split into intron and exon regions and intersected with the coverage of each sample. Per gene intron and exon summary statistics, mean and standard deviation, were computed on log-coverage values. We selected genes enriched for IR-like coverage profiles by looking for 'expressed' genes, where for a substantial number of intron bases its coverage is much higher than other intron features within the same gene whilst having similar expression levels to the exon features. To do this we used the following thresholds—genes had an average exon log-coverage of greater than two (corresponding to the mean of average exon log-coverage values across all genes), a standard deviation of intron log-coverage >1.5 (corresponding to the mean of intron standard deviation values across all genes), and genes with a large number of intron bases with log-coverage greater than two (top 1% of genes). The thresholds were chosen after examining distributions of the summary statistics (Supplementary Figure S4). Forty-three genes met these criteria, where a manual check of coverage profiles revealed that 36 genes indeed appear IR-like (see 'Analyses' page in Supplementary Materials). We highlight three of these genes in Figure 6. The coverage of remaining seven genes appear to be more pre-mRNA-like, with large variation in intron coverage, where at its peak it is expressed similarly to exon features.

We use poly(A) RNA samples in our analysis for consistency with previous studies on IR. In theory, total RNA samples may be more appropriate for this exercise since it is less biased towards the 3′ end, where high 3′ intron coverage as a result of 3′ bias in poly(A) RNA libraries can be mistaken as a 3′ retained intron. For this reason, we review both the coverage of poly(A) RNA and total RNA samples in our Supplementary Materials for all selected genes to ensure that retained introns found in poly(A) RNA samples are not an artifact of RNA library preparation.

If differentially retained introns are of interest, one could simply run *superintronic* on two conditions separately and compare lists of IR-like genes between groups. We found that differences between the lists of IR-like genes were concordant with *index* results for genes DE by intron counts, as expected (see 'Analyses' page in Supplementary Materials). For example, 14 genes were uniquely selected as IR-like in total RNA HCC827 cell line when compared to total RNA NCI-H1975 cell line using *superintronic*, where 13 of those genes were also found to be DE in intron counts by *index*, and in the expected direction. Similarly, 19 genes were uniquely detected as IR-like in total RNA NCI-H1975 by *superintronic*, 16 of which were also detected as DE in intron counts by *index*. Additionally, 12 genes which were found to be IR-like in both cell lines appear to have varying intron expression according to *index* since it is found to be DE in intron counts.

Whilst genes that are uniquely IR-like were concordant with directional changes in intron counts, we did not find that IR-like genes played a large part in *index* gene cat-
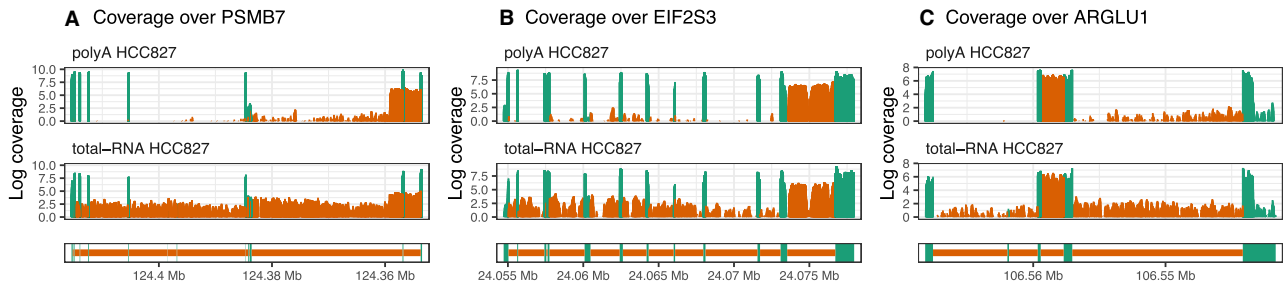
Figure 6. *Superintronic* selects genes with IR-like coverage profiles in poly(A) RNA HCC827 cell lines. Genes (**A**) PSMB7, (**B**) EIF2S3 and (**C**) ARGLU1 are highlighted out of 43 genes selected. Whilst the analysis was carried out on poly(A) RNA libraries, coverage is shown for both poly(A) RNA HCC827 (top) and total RNA HCC827 (bottom) samples to ensure that results are not an artifact of RNA library preparation. Coverage is oriented from 5′ to 3′, with exon regions coloured green and intron regions coloured orange.

egorization. This was mostly expected since thousands of genes were detected as significantly DE by *index* and a relatively small number of genes were detected as IR-like by *superintronic*. Specifically, IR-like genes did not overlap with one particular *index* category (see 'Analyses' page in Supplementary Materials). Of genes that are classified as intron+ or intron−, only 3 out of 1006 genes were also detected by *superintronic* as IR-like. In the 'mixed' category, 2 out of 54 genes were also detected as IR-like.

DISCUSSION

The work presented in this paper provides a broad view of the characteristics and expression patterns associated with intron reads in model organisms, namely human and mice. Incomplete gene annotations may result in misclassification of reads as intron reads. We looked into this by examining split reads as indicative of splice events, and found the effect of unannotated exons in our analyses of human cell line data to be minimal. For example, HCC827 cell line samples have a median of zero split reads within introns (third-quartile of 0.33), as compared to a median number of 38 split reads in exons (third-quartile of 121.33). This demonstrates that splice events are rare within annotated intronic regions, and suggests that unannotated exons play a very minor role in the results we presented overall. For example, only 2% of genes in any of the *index* categories contain introns with 10 or more split reads (see 'Analyses' page in Supplementary Materials).

Interestingly, 41% of *superintronic*'s IR-like genes contain introns with 10 or more split reads. However, we do not believe that this indicates that the genes were detected because of unannotated exons since our thresholds ensure the selection of high coverage regions that are much larger than the typical exon. For the understanding of complex splice events, further biological validation of the IR-like genes is of interest but beyond the scope of this paper.

For organisms that are poorly annotated, the number of reads misclassified as intron reads may be significantly inflated. In such cases, 'intron reads' that are split reads can be useful for identifying new exons. DGE analyses using genebody counts or by our *index* method would naturally incorporate the transcriptional changes within unannotated exons of annotated genes.

Using gold standard differential expression methods, *index* selects and categorizes genes of interest based on *P*-values from moderated *t*-statistics that are adjusted for multiple testing and the direction of change. This is statistically more sophisticated than the *EISA* method which uses intron and exon logFCs alone and provides no prioritization of genes of interest—this is, however, sufficient for their purpose of categorizing genes as sets, and determining whether the biological system under study is driven transcriptionally or post-transcriptionally as a whole. In contrast, *index* looks for DE genes using evidence from intron and exon counts. *Index*'s logFC plot (Figure 5C and G) is analogous to *EISA*'s main result and logFC plot.

The presence of pre-mRNA in poly(A) RNA libraries may be somewhat surprising since the RNA library preparation is optimized for mRNA selection. However, premRNA can be captured by both poly(A) RNA and total RNA protocols since transcription from DNA to a primary RNA transcript, 3′ cleavage of the RNA molecule and polyadenylation can be completed before splicing is complete at the 3′ end because the splicing mechanism requires a relatively long processing time (36) – this is regardless of whether genes are co-transcriptionally (9,37–38) or post-transcriptionally spliced. Evidence supporting this includes the presence of poly(A)-positive molecules in the nucleus that are larger than final mRNAs in the cytoplasm (36).

Handling of RNA in preparation for sequencing results in a degree of fragmentation of the original molecule regardless of the level of care taken during this process. This has minimal downstream effects on total RNA libraries since 3′ and 5′ fragments are selected randomly. However, the selection of poly(A)-positive RNA molecules in poly(A) RNA libraries bias fragments at the 3′ end whilst 5′ fragments are lost in the process. This results in 3′ coverage bias in poly(A) RNA libraries (Figure 2B)—demonstrated also by Lahens *et al.* (39) in their study on technical biases introduced during generation of sequencing libraries. Shorter genes with fewer and/or shorter introns are less affected by fragmentation than genes with long RNA molecules, thus coverage profiles are more similar between total RNA and poly(A) RNA libraries for these genes (Figure 2B). Read coverage in total RNA libraries may provide a more accurate representation of the original RNA molecule than in

poly(A) RNA libraries, especially in long genes. In poly(A) RNA libraries, the inflated 3′ exon expression and 3′ 'background' intron signal may negatively impact on methods for *de novo* transcriptome assembly and transcript quantification. total RNA libraries which have uniform 'background' intron signal is easier to model in theory and should be better suited to such applications. Moreover, existing IR detection methods applied to poly(A) RNA sequencing libraries will have increased difficulty in interrogating introns residing towards the 5′ end of genes.

We have explored intron signal arising in this context via coverage estimation performed using *superintronic* and found that we can visualize IR-like coverage profiles by using simple summary statistics generated from genomic overlaps. Our method differs from existing methods such as *IRFinder* (40) and *IsoformSwitchAnalyzeR* (41) in that it can detect IR-like coverage profiles in individual conditions, rather than differences between two conditions. Naturally, the detection of differentially retained introns within a gene should firstly include a retained intron in at least one condition, and secondly contain differences in the expression of the intron. Since detection of either of these steps are non-trivial, we believe that it is more important to focus on detecting retained introns directly via visualization of 'interesting' coverage profiles. However, a downside of our approach is that does not perform any statistical inference on a given coverage profile to say whether a region is truly IR-like. In this way we see the use of *superintronic* as both complimentary to *index* and useful in its own right for flexibly summarizing reads with respect to an experimental design, it can be used to perform quality control on *index* results interrogate exon/intron count data further by viewing coverage profiles. We have also found that estimation of coverage profiles can provide a visual check of differential IR results from other methods; for example we have observed significant results in poly(A) RNA libraries tending towards the 3′ end of genes for both *IRFinder* and *IsoformSwitchAnalyzeR* (Supplementary Figure S5). Although the results are not directly comparable, we note that 23% of genes uniquely detected as IR-like in either of the cell lines using *superintronic* overlap with differential IR results using *IRFinder*'s generalized linear models method. Whilst this is not a big overlap, it shows some level of consistency between the methods. On the other hand, none of the uniquely IR-like genes are in common with results from *IsoformSwitchAnalyzeR* (see 'Analyses' page in Supplementary Materials).

Compared to bulk RNA-seq, single-cell RNA-seq (scRNA-seq) data have much smaller library sizes and relatively high proportions of intron reads leading to much interest in the incorporation of intron reads in scRNA-seq data analyses (42). Though yet to be tested, DGE analysis by *index* should theoretically perform well on scRNA-seq data since it increases the number of testable genes and libraries by increasing the amount of information used.

The work presented in this paper explores multiple origins of intron reads and signal in RNA-seq data. We demonstrate the usefulness of applying intron reads to study multiple aspects of transcriptional biology, and provide tools to interrogate changes in pre-mRNA and mRNA levels, as well as genes with IR-like coverage profiles. Biological validation of our results was not carried out and is beyond the

scope of this paper, since the intention was to make conclusions using a data-driven approach. However, further work includes closer examination into pre-mRNA-specific and IR-specific signal, such as by using full length transcripts by long-read sequencing by Pacific Biosciences (43) or Oxford Nanopore Technologies (44). It is also of interest to examine intron reads in datasets with RNA from cytoplasmic and nuclear fractions versus whole cell.

CONCLUSION

Intron reads are prevalent at small to moderate proportions in RNA-seq datasets, however, they provide signal that can distinguish between biological and experimental groups. Harvesting these extra reads as pre-mRNA signal, DGE analysis can be carried out more thoroughly with the addition of intron counts into *index*. The extra layer of analysis enables distinction between changes in pre-mRNA and mRNA signal enhancing the understanding of transcriptional changes and dynamics under study. IR remains an important mechanism in biology and can be explored through the use of *superintronic*, which can discover genes with IR-like coverage profiles.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

REFERENCES

1. Sultan,M., Schulz,M.H., Richard,H., Magen,A., Klingenhoff,A., Scherf,M., Seifert,M., Borodina,T., Soldatov,A., Parkhomchuk,D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.
2. Trapnell,C., Hendrickson,D.G., Sauvageau,M., Goff,L., Rinn,J.L. and Pachter,L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, **31**, 46–53.
3. Braunschweig,U., Barbosa-Morais,N.L., Pan,Q., Nachman,E.N., Alipanahi,B., Gonatopoulos-Pournatzis,T., Frey,B., Irimia,M. and Blencowe,B.J. (2014) Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.*, **11**, 1774–1786.
4. Esteller,M. (2011) Non-coding RNAs in human disease. *Nat. Rev. Genet.*, **12**, 861–874.

5. Mattick,J. and Rinn,J. (2015) Discovery and annotation of long noncoding RNAs. *Nat. Struct. Mol. Biol.*, **22**, 5–7.

6. Guo,Y., Zhao,S., Sheng,Q., Guo,M., Lehmann,B., Pietenpol,J., Samuels,D.C. and Shyr,Y. (2015) RNAseq by total RNA library identifies additional RNAs compared to Poly(A) RNA library. *Biomed. Res. Int.*, **2015**, 862130.

7. Harati,S., Phan,J.H. and Wang,M.D. (2014) Investigation of factors affecting RNA-seq gene expression calls. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, **2014**, 5232–5235.

8. Zhao,S., Zhang,Y., Gamini,R., Zhang,B. and von Schack,D. (2018) Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci. Rep.*, **8**, 4781.

9. Ameur,A., Zaghlool,A., Halvardson,J., Wetterbom,A., Gyllensten,U., Cavelier,L. and Feuk,L. (2011) Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.*, **18**, 1435–1441.

10. Gaidatzis,D., Burger,L., Florescu,M. and Stadler,M.B. (2015) Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat. Biotechnol.*, **33**, 722–729.

11. Wong,J.J., Ritchie,W., Ebner,O.A., Selbach,M., Wong,J.W.H., Huang,Y., Gao,D., Pinello,N., Gonzalez,M., Baidya,K. *et al.* (2013) Orchestrated intron retention regulates normal granulocyte differentiation. *Cell*, **154**, 583–595.

12. Jung,H., Lee,D., Lee,J., Park,D., Kim,Y., Park,W., Hong,D., Park,P. and Lee,E. (2015) Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.*, **47**, 1242–1248.

13. Pimentel,H., Parra,M., Gee,S., Mohandas,N., Pachter,L. and Conboy,J. (2015) A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res.*, **44**, 838–851.

14. Holik,A.Z., Law,C.W., Liu,R., Wang,Z., Wang,W., Ahn,J., Asselin-Labat,M., Smyth,G.K. and Ritchie,M.E. (2017) RNA-seq mixology: designing realistic control experiments to compare protocols and analysis methods. *Nucleic Acids Res.*, **45**, e30.

15. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.

16. Linsley,P.S., Speake,C., Whalen,E. and Chaussabel,D. (2014) Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. *PLoS One*, **9**, e109760.

17. Sheridan,J.M., Ritchie,M.E., Best,S.A., Jiang,K., Beck,T.J., Vaillant,F., Liu,K., Dickins,R.A., Smyth,G.K., Lindeman,G.J. *et al.* (2015) A pooled shRNA screen for regulators of primary mammary stem and progenitor cells identifies roles for Asap1 and Prox1. *BMC Cancer*, **15**, 221.

18. Choi,J., Baldwin,T.M., Wong,M., Bolden,J.E., Fairfax,K.A., Lucas,E.C., Cole,R., Biben,C., Morgan,C., Ramsay,K.A. *et al.* (2019) Haemopedia RNA-seq: a database of gene expression during haematopoiesis in mice and humans. *Nucleic Acids Res.*, **47**, D780–D785.

19. Chappaz,S., Law,C.W., Dowling,M.R., Carey,K.T., Lane,R.M., Ngo,L.H., Wickramasinghe,V.O., Smyth,G.K., Ritchie,M.E. and Kile,B.T. (2020) Germline heterozygous mutations in Nxf1 perturb RNA metabolism and trigger thrombocytopenia and lymphopenia in mice. *Blood Adv.*, **4**, 1270–1283.

20. Liao,Y., Smyth,G.K. and Shi,W. (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.*, **41**, e108.

21. Liao,Y., Smyth,G.K. and Shi,W. (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.*, **47**, e47.

22. Liao,Y., Smyth,G.K. and Shi,W. (2014) featurecounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

23. Morgan,M., Pagès,H., Obenchain,V. and Hayden,N. (2019) Rsamtools: binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R package version 2.1.2, doi:10.18129/B9.bioc.Rsamtools.

24. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M. and Carey,V. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.

25. Lee,S., Cook,D. and Lawrence,M. (2019) plyranges: a grammar of genomic data transformation. *Genome Biol.*, **20**, 4.

26. Zeisel,A., Köstler,W.J., Molotski,N., Tsai,J.M., Krauthgamer,R., Jacob-Hirsch,J., Rechavi,G., Soen,Y., Jung,S., Yarden,Y. and et,al. (2011) Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Mol. Syst. Biol.*, **7**, 529.

27. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

28. McCarthy,D.J., Chen,Y. and Smyth,G.K. (2012) Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.

29. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.

30. Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.

31. Law,C.W., Alhamdoosh,M., Su,S., Dong,X., Tian,L., Smyth,G.K. and Ritchie,M.E. (2016) RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000 Res.*, **5**, 1408.

32. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

33. Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 3.

34. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.

35. Colombo,M., Karousis,E.D., Bourquin,J., Bruggmann,R. and Mühlemann,O. (2017) Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6- and SMG7-mediated degradation pathways. *RNA*, **23**, 189–201.

36. Darnell,J.E. (2013) Reflections on the history of pre-mRNA processing and highlights of current knowledge: A unified picture. *RNA*, **19**, 443–460.

37. Merkhofer,E.C., Hu,P. and Johnson,T.L. (2014) Introduction to cotranscriptional RNA splicing. *Methods Mol. Biol.*, **1126**, 83–96.

38. Zlotorynski,E. (2016) RNA metabolism: co-transcriptional splicing at nucleotide resolution. *Nat. Rev. Mol. Cell. Biol.*, **17**, 264–265.

39. Lahens,N.F., Kavakli,I.H., Zhang,R., Hayer,K., Black,M.B., Dueck,H., Pizarro,A., Kim,J., Irizarry,R., Thomas,R.S. *et al.* (2014) IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol.*, **15**, R86.

40. Middleton,R., Gao,D., Thomas,A., Singh,B., Au,A., Wong,J.J., Bomane,A., Cosson,B., Eyras,E., Rasko,J. *et al.* (2017) IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol.*, **18**, 51.

41. Vitting-Seerup,K. and Sandelin,A. (2019) IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics*, **35**, 4469–4471.

42. La Manno,G., Soldatov,R., Zeisel,A., Braun,E., Hochgerner,H., Petukhov,V., Lidschreiber,K., Kastriti,M.E., Lonnerberg,P., Furlan,A. *et al.* (2018) RNA velocity in single cells. *Nature*, **560**, 494–498.

43. Rhoads,A. and Au,K.F. (2015) PacBio sequencing and its applications. *Genomics Proteomics Bioinform.*, **13**, 278–289.

44. Lu,H., Giordano,F. and Ning,Z. (2016) Oxford nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinform.*, **14**, 265–279.