Title: Challenges in defining the role of intron retention in normal biology and disease

Authors: Darya P. Vanichkina, Ulf Schmitz, Justin J.-L. Wong, John E.J. Rasko

PII: S1084-9521(17)30293-8
DOI: http://dx.doi.org/doi:10.1016/j.semcdb.2017.07.030
Reference: YSCDB 2293

To appear in: *Seminars in Cell & Developmental Biology*

Received date: 12-6-2017
Revised date: 19-7-2017
Accepted date: 19-7-2017

Please cite this article as: Vanichkina Darya P, Schmitz Ulf, Wong Justin J-L, Rasko John E.J.Challenges in defining the role of intron retention in normal biology and disease.*Seminars in Cell and Developmental Biology* http://dx.doi.org/10.1016/j.semcdb.2017.07.030

Challenges in defining the role of intron retention in normal biology and disease

Darya P. Vanichkina[1,2], Ulf Schmitz[1,2], Justin J.-L. Wong[1,2,3], John E. J. Rasko[1,2,4]

[1] Gene & Stem Cell Therapy Program, Centenary Institute, University of Sydney, Camperdown 2050, Australia
[2] Sydney Medical School, University of Sydney, Camperdown 2050, Australia
[3] Gene Regulation in Cancer Laboratory, Centenary Institute, University of Sydney, Camperdown 2050, Australia
[4] Cell and Molecular Therapies, Royal Prince Alfred Hospital, Camperdown 2050, Australia

Corresponding author: John E.J. Rasko, j.rasko@centenary.org.au

Highlights

- Intron retention is a mechanism of gene expression control in eukaryotes
- Custom computational pipelines are essential for IR detection
- Phylogenetic analyses reveal conserved IR and functional consequences
- Subcellular fractionation helps determine the spatio-functional relationship of IR
- Improvements in mass spectrometry are critical to detect novel IR-derived peptides

Abstract

RNA sequencing has revealed a striking diversity in transcriptomic complexity, to which alternative splicing is a major contributor. Intron retention (IR) is a conserved form of alternative splicing that was originally overlooked in normal mammalian physiology and development, due mostly to difficulties in its detection. IR has recently been revealed as an independent mechanism of controlling and enhancing the complexity of gene expression. IR facilitates rapid responses to biological stimuli, is involved in disease pathogenesis, and can generate novel protein isoforms. Many challenges, however, remain in detecting and quantifying retained introns and in determining their effects on cellular phenotype. In this review, we provide an overview of these challenges, and highlight approaches that can be used to address them.

Keywords

- alternative splicing
- RNA sequencing
- Bioinformatics
- gene expression

- epigenetics
- phylogeny

Abbreviations:
EST – expressed sequence tag
ID – intron detention
IR – intron retention
IRT – intron retaining transcripts
mRNA – messenger RNA
NMD – nonsense mediated decay
PTC – premature termination codon
qRT-PCR – quantitative reverse transcription polymerase chain reaction
SNP – single nucleotide polymorphism

Funding:
DPV is supported by the National Health and Medical Research Council [grant number 1080530] and the Cure the Future Foundation. US is supported by the Sydney Research Excellence Initiative and the Cure the Future Foundation. JJ-LW is supported by the National Health and Medical Research Council [grant numbers 1129901, 1080530, 1128175, 1126306]. JEJR is supported by the National Health and Medical Research Council [grant numbers 1061906, 1129901, 1080530, 1128175]; the Cure the Future Foundation; and an anonymous foundation.

1 Introduction

1.1 Alternative splicing increases mammalian transcriptome diversity

Advances in genome and transcriptome sequencing and gene annotation have revealed the widespread diversification of the proteome via alternative splicing across eukaryotic taxa [1, 2]. For example, exon combinatorics enable neuronal synapse specification via neurexin-neuroligin interactions [3]; maintenance of cell differentiation state as a result of alternative splicing of core pluripotency factors [4]; the immune response in the context of T-cell activation [5, 6]; and a plethora of other disease-related biological processes. Furthermore, introns themselves can contain *cis* and *trans*-acting elements including regulatory non-coding RNAs, such as small nucleolar RNAs [7] and microRNAs [8]. While these RNA transcripts are physically located in the intronic region, they can be transcribed independently of their host gene [9, 10].

1.2 IR is a widespread form of alternative splicing across eukaryotic taxa

Intron retention (IR) is a form of alternative splicing characterised by the inclusion of intronic sequence in a mature transcript (Figure 1). IR is widespread in plants, fungi and unicellular eukaryotes [11-13], but was previously thought to be nearly absent and/or irrelevant in animals [14-16]. Next generation sequencing has propelled the detection and quantitation of RNA originating from introns to an unprecedented extent. Ameur *et al.* [17] observed significant numbers of sequencing reads mapping to introns in total RNA libraries from human brain. Nonetheless, they considered

these intron-retaining transcripts as evidence for co-transcriptional splicing, as immature RNA transcripts, and not *bona fide* functional molecules.

Recent work has revealed that IR is not an artefact of sequencing library preparation, and is actually much more widespread in the animal kingdom than originally thought. Three percent of *Drosophila* introns are nearly completely retained after splicing [18], and IR occurs in 50 - 75% of multiexonic genes across 11 animal species, ranging from chicken and frog - to platypus and human [19]. Our lab has demonstrated that IR affects 80% of coding genes in human, especially those involved in the cell cycle and differentiation [20]. IR has also been reported in approximately 5-6% of expressed genes in the mouse cortex [21].

1.3 The fate of intron retaining transcripts is varied

The fate of intron retaining transcripts (IRTs) in animals can be quite varied (Figure 1). A subset is retained in the nucleus (these are said to be affected by a process called intron detention (ID) [22]), while others are exported to the cytoplasm. When exported into the cytoplasm, IRTs interact with the ribosome, and undergo NMD if premature termination codons (PTCs) are detected in them during the pioneer round of translation [23, 24]. Some IRTs escape NMD to generate new protein isoforms [25], while in other cases, signals in the retained intron serve to specify the subcellular localisation of the transcript or protein [26]. Other functions that have been proposed for IRTs include the regulation of intron-derived microRNA precursor (mirtron) and snoRNA expression, and the modulation of post-transcriptional gene regulation by acting as competing endogenous RNA [27]. However, these functions are not as yet validated experimentally.

1.4 IR is tissue specific, tightly regulated during development and altered in disease

IR is a tissue-specific phenomenon in animals, which further supports its role as a mechanism of gene expression regulation. A higher proportion of introns is retained in neural and immune cell types, whereas IR events are less frequent in embryonic stem and muscle cells [19]. Increased IR in neuronal and immune cells may facilitate rapid respond to external stimuli, within a time frame shorter than that required for de novo transcription and protein synthesis [6, 21]. Specific IR patterns can be characteristic of cell subtypes as well, for example in luminal and myoepithelial breast cells [28]. In a reanalysis of over 2500 mRNA sequencing datasets, over 15000 introns retained in at least one dataset were retained in fewer than 7% of all samples considered, further supporting the tissue-specificity of this process [20]. This specificity can enable tissue-specific sequestration of intron-retaining transcripts, serving to restrict the translation of proteins only to the cells where they are required, while concurrently maintaining transcription from the locus of origin in other tissues. Such a mechanism of action has been demonstrated, for example, in genes encoding critical presynaptic proteins in both neurons and non-neuronal cells: only in neuronal cells is the last intron spliced out, preventing RNA degradation and enabling fully spliced transcripts to be exported from the nucleus [29].

IR is tightly regulated during differentiation and development. IR increases in key myeloid-related genes during granulocyte differentiation, leading to reduced RNA and protein levels, critical for the maturation of granulocytes [30]. Differential IR is also a characteristic of other cells of the hematopoietic lineage, including erythroblasts [31], megakaryocyte progenitors [32], and $CD4^+$ T-cells transitioning from an inactive to active state [6]. IR mediates the down-regulation of genes

involved in cell cycle progression and up-regulation of genes with neuron-specific functions during the differentiation of embryonic stem cells into neural progenitors [19], and during reprogramming of mouse embryonic fibroblasts to induced pluripotent stem cells [33]. A subset of intron-containing mature mRNAs are shielded from rapid degradation in embryonic stem cells via their sequestration in the nucleus [22]. IR is also inversely correlated with gene expression levels during the reprogramming of mouse embryonic fibroblasts [33]. Recently, IR in polyadenylated transcripts has been shown to be crucial for modulating mouse cortical mRNA dynamics in response to neuronal activity. Over 200 retained introns were spliced out within 15 minutes in response to depolarisation [21], and a significant proportion of glutamate receptor transcripts are preferentially affected by IR in the cerebellum [34].

IR is widespread across a range of cancer transcriptomes [35]. It has been described as a mechanism of tumour suppressor inactivation [36], and is the predominant form of alternative splicing in hypoxic tumour cells [37]. In hypoxia, IR leads to a reduction in protein levels of the critical cytotoxic response regulator HDAC6 and DNA double strand break pathway member TP53BP1 [37]. Roles for IR in other diseases are currently being investigated [27].

2 There are many bioinformatic challenges in investigating intronic regions

2.1 Intronic regions are rich in repetitive sequences and longer than exonic regions

Introns are rich in repetitive sequences, containing over double the density of these elements as exonic regions (Figure 2). These include Long and Short Interspersed Nuclear Elements (LINEs and SINEs), DNA transposons, tandem and low complexity

repeat sequences. Most of these genomic features are longer than the 75 – 150 nucleotide read length characteristic of current high-throughput RNA sequencing technologies. This presents a unique challenge since RNA-optimised mapping algorithms such as Tophat2 [38], STAR [39] and MapSplice [40] have filters in place to discard reads that map to more loci than a specific threshold. Moreover, counting reads mapping to multiple locations is not straightforward (Figure 3B). Indeed, Bai *et al*. [41], have suggested that considering reads mapping to multiple locations introduces "noise" into IR calling and analysis. Unfortunately, exclusion of multimapping reads can lead to the loss of biologically important information. For example, repetitive sequences have been reported to be ***the*** critical functional component of retained introns in rat neurons [26]. In this case, SINE retrotransposons in retained introns are necessary and sufficient for targeting of cytoplasmic mRNAs to dendrites, while similar repeat elements in the 3' UTR do not alter RNA subcellular localisation. Hence, any bioinformatic tool for IR identification and differential analysis which does not include an approach to deal with multi-mapping reads and repetitive sequences in the genome is likely to miss substantial, functionally relevant biological complexity.

Introns are significantly longer than exons, reaching up to 500 kbp in some mammals. Therefore, it is difficult to evenly "sample" them to achieve adequate read coverage across their entire length. In humans, the average size of exons is 150 nucleotides, and that of introns is 3500 nucleotides [42], meaning that most exons are fully covered by current 150 bp paired-end sequencing approaches, while introns are not. There are two contrasting strategies to deal with this challenge as follows.

2.2 Pitfalls of splice-junction-only approaches to analyse IR from short read sequencing data

The first approach to deal with long introns involves considering only reads mapping across the splice junctions of interest when calculating IR levels. This circumvents the issues described above concerning repetitive sequences within introns, and the issues of adequate normalisation of read coverage across introns of vastly different lengths. However, this approach requires very high read depth, since only ~10 – 25% of sequencing reads in a typical paired-end library span one or more splice junctions. For example, the developers of VAST-tools, which takes such a junction-read-only approach, recommend at least 70 million reads per sample (ideally >150 million reads) for improved detection and quantitation of all splicing types – not IR specifically. They observe improvement in alternative splicing assessment when read depth is increased to 200 – 300 million reads per sample [19, 43] . If insufficient read depth is observed, conclusions about the background distribution and baseline IR levels can be erroneous. Depending on the statistical tests implemented by each tool, this may lead to biased and unreliable analyses. Unfortunately, most publically available datasets and experiments carried out without the specific prior aim of alternative splicing characterisation are usually not as deep as this, meaning that using such tools to analyse pre-existing datasets may not be appropriate.

2.3 Pitfalls of coverage-based approaches to analyse IR from short read sequencing data

The second strategy considers both junction reads and reads mapping within the body of the intron. Tools and protocols that use this approach need to reliably address the mappability issues presented by repetitive elements, and to account for intron length when normalising retention levels between different introns of the same gene or

across genes [19, 20, 30]. Such approaches also require relatively high sequencing depth, since they frequently implement a coverage cut-off, where a certain proportion of the intron, or uniquely mappable region of the intron, must be covered by sequencing reads with a minimum depth (for example, 70% of the intron covered at a depth of 5 reads) [20, 21, 36, 44]. A common caveat with this cut-off is whether to consider the mean number of reads or the median, with most tools using the mean. Unfortunately, this is unlikely to be the best strategy, as the mean is statistically more susceptible to being skewed by outliers than the median. Hence, if there is a small repetitive region not masked by mappability assessment, using the mean coverage will inflate the actual level of reads (Figure 3B). Coverage thresholds are especially critical when considering mammalian protein-coding genes, as many of these harbour independently transcribed small RNAs, such as snoRNAs or microRNAs (Figure 3C). If host transcripts are expressed at high levels, and a coverage cut-off is not implemented or used, it might be erroneously concluded that IR affects the coding gene.

2.4 Differential gene expression can bias IR detection and quantitation

Any assessment of IR and differential IR must also consider the expression levels of the whole gene [41]. For example, if a gene is highly expressed, it is more likely that reads spanning both inclusion and exclusion splice junctions will be detected in the RNA sequencing library; therefore, passing any coverage, splice junction boundary balance or other cut-off. If a differential IR analysis between two tissues is attempted, the differences in gene expression level must be taken into account and subtracted, as these will directly affect detectability of IR (Figure 3A). Finally, the fact that many IRTs are subject to NMD adds an additional layer of complexity, as the gene may be

observed as expressed at a lower level due to the specific degradation of IRTs (Figure 1).

2.5 IR detection can be confounded by antisense transcription

An additional constraint when using next generation sequencing to characterise IR is the abundance of sense and antisense transcription [45]. This means that any analyses of IR must take great care in identifying the strand of origin of the observed sequencing reads (Figure 3D), and use strand-specific sequencing protocols whenever possible. IR detection can be confounded by traces of genomic DNA. Thus, care must be taken to ensure effective DNase treatment of RNA samples and confirmation via quality control testing and filters implemented informatically [20].

2.6 Long read sequencing captures full intron length but is limited by sequencing depth and accuracy

Long read sequencing technologies have been heralded as ground-breaking in improving alternative splicing detection and quantitation [46], as they circumvent many of the length and mappability challenges outlined above. Unfortunately, these technologies are currently limited by their relatively low throughput, high error rates and propensity towards amplification and detection of shorter transcripts [47]. Assessing the extent of IR in species more complex than yeast [48] or beyond a small number of target genes [49] using long read sequencing is currently not readily tractable.

2.7 IR validation rates remain low

Even for dedicated IR studies, validation rates for estimates in the levels of IR remain low. For example, one study reported a correlation coefficient (r) of only 0.63 between IR fold-changes determined using RNA-seq and qRT-PCR [19]. Given the

low number of genes validated using qRT-PCR in most studies to date, the best computational approaches capable of providing the most robust IR predictions remain to be determined.

2.8 Best practices for informatic IR analysis

To circumvent the challenges described above, for both junction-based and coverage-based IR analysis, strand-specific sequencing should be carried out, and followed by a robust assessment of whether the read depth is adequate to sample intronic regions. This can be done by subsampling reads and examining splice junction or intron coverage statistics. If a junction based IR assessment approach is chosen, more sequencing depth will be required, while if a coverage-based approach is selected mappability assessment must be incorporated into the mapping and intron quantitation. Both strategies also require a filtering step to segregate IR in differentially expressed genes from IR in non-differentially expressed genes, with estimates for the latter frequently being more robust and reliable. Finally, new computational techniques need to be developed to incorporate data generated by long-read sequencing with short-read data, and benchmarking studies of best practices for IR detection, normalisation and quantitation similar to those available for differential gene expression analysis [50, 51] need to be carried out.

3 Obstacles in phylogenetic IR analyses

We and others have shown that functionally related genes are affected by IR in humans and mice [19, 22, 30]. Phylogenetic IR analyses present the prospect of shedding light on the evolution and functional conservation of IR, but studying the conservation of IR comes with several challenges discussed below.

3.1 Sequencing depth and genome annotation quality vary for different species

IR analysis requires the presence of well-curated genome annotations and adequate depth of coverage relative to genome size across all species considered. Early phylogenetic studies that used expressed sequence tags reported that many retained introns had relatively low coverage, and were less well-represented in assembled transcripts [12]. The first large-scale phylogenetic analyses of alternative splicing in vertebrates based on RNA sequencing data investigated splicing profiles across 7 organs in 11 vertebrate species [52]. They found that alternative splicing complexity increased in species evolutionarily closer to primates, and observed the highest complexity in man [52]. In order to avoid biases in the detection of relative alternative splicing frequencies associated with differences in annotation qualities, the authors generated *de novo* exon-intron structures from the same number of random reads for each sample. In a follow-up study conducted by this group investigating IR in ~40 human and mouse tissues, an evolutionarily conserved IR code was proposed to distinguish retained and constitutively spliced introns [19]. Retained introns were called after filtering based on coverage, depth, and read distribution to avoid consideration of false introns due to mis-annotation, insufficient precision due to lack of coverage, and false IR calling due to neighbouring alternative 5' or 3' splice sites or overlapping genes. Such tissue-specific transcriptome sequencing and refinement of genome assemblies may lead to improved annotations, revealing that regions previously considered to be introns are in fact exons [53].

Finally, the detection of alternative splicing events including IR is strongly dependent on the number of transcripts per gene [54]. This means that a method for transcript number normalization on a gene-by-gene basis needs to be employed in comparative

analysis across taxa. In the case of IR, it may be sensible to also normalize for the average number of introns per gene in a genome, as the likelihood of observing stochastic IR events increases with intron number [54, 55].

3.2 Introns vary in length and number, and their sequences are poorly conserved

Unlike exons, intronic sequences are poorly conserved throughout evolution, apart from the four main splice signals (the 5' and 3' splice sites, the branch site, and the polypyrimidine tract) [56]. Little is known about *cis*-regulatory elements in introns or flanking exons that impact IR. This limits our ability to analyse the effects that base changes such as single nucleotide polymorphisms may have on possible IR-associated functions. Recently, enrichment of particular RNA binding sites has been observed in sequences of frequently retained introns as well as their flanking exons in human [20], however, their conservation has not been examined.

An implicit problem in phylogenetic analyses of IR is the identification of orthologous introns because of their lack of sequence conservation. Phylogenetic analyses of exon-intron structures in orthologous genes have demonstrated that 25-30% of intron positions are shared between at least two out of three lineages of animals, fungi and plants [57]. Two approaches are generally applied to determine orthologous introns: (1) requiring them to occur in the exact same position in orthologous genes (genomic coordinates between species are converted, for example by using the liftOver tool as in [19]), or (2) based on the orthology relationship or conservation of their flanking exons [58]. However, these methods can result in conflicting annotations for the same intron, especially when evolutionarily distant species are considered in the analysis. Additional research needs to be carried out to

ascertain which of these two approaches is appropriate depending on the evolutionary distance between considered organisms.

The number of introns in genes varies considerably among lineages with very few or no introns in protists and fungi, intermediate numbers in plants and large numbers in multicellular animals [59]. Animal introns are on average much longer than exons, while in protists and fungi, where IR is the dominant form of alternative splicing, introns are shorter than exons. McGuire and co-authors suggest that in these species splice junctions are recognised via an intron definition mechanism as opposed to an exon definition mechanism, the primary form of splice site recognition in animals [59]. The authors further suggest that in plants, both mechanisms play a role in splice site recognition due to the great variance in intron lengths. In animals intron size is negatively associated with introns located towards the 3′ end and correlates with genome size [58], and birds and reptiles have shorter introns than mammals. These differences in size and number necessitate normalisation in short read sequencing experiments comparing IR abundance across taxa, because IR events are more likely detected in species with fewer and or shorter introns.

3.3 IR functions may not be conserved across distant lineages

Another challenge in phylogenetic IR analyses is the assessment of conserved functions of IR. Computational tools for the identification of conserved, lineage-specific IR events and their downstream effects have not been developed. Due to the relatively sparse data from the limited number of model organisms described above, broad extrapolations are required to study the evolution of the extent and function of IR. This is distinct from the approach that can be taken for canonical protein-coding genes, for which the underlying assumption is that the most important sequences, critical for the function of the protein, are the most conserved. With respect to IR,

low-conservation of most intron sequences negates this assumption. This has previously led to the conclusion that IR is merely transcriptional noise (e.g. due to errors in splicing) and has no functional implications [60]. A conclusive interpretation of functional consequences mediated by IR can only be achieved when the fate of intron-retaining genes (downregulation, detention in the nucleus, novel protein isoform generation) is clarified, for which currently no prediction algorithm exists. Comparative phylogenetic studies of IR resulting in the development of such algorithms would enable circumventing this challenge.

Enrichment analysis of the list of genes affected by IR can illuminate the molecular functions and biological processes it regulates. Many tools for performing such analyses exist, including GSEA [61], DAVID [62], and PANTHER [63]. These rely on similar statistics, and can be used with predefined whole-genome backgrounds or custom user-specified ones. However, comparing enriched ontology terms or pathways across species may lead to false conclusions, as annotation quality varies between organisms, tending to be more comprehensive in widely studied organisms such as mouse and human. Enrichment analysis of alternative splicing in general is strongly confounded by detectability, to which expression is the biggest contributor [64]. In the case of phylogenetic exploration of IR, it is unclear how a valid background should be constructed: the usual approach is to consider all expressed genes in a sample, but, in a phylogenetic analysis, it might be more appropriate to include all orthologous genes in a comprehensive background gene set.

Ultimately, as IR is likely to play different roles in different taxa, it seems to be more reasonable to study its functional conservation in closely related species with similar physiology, and not necessarily across distant evolutionary lineages.

4 Detecting IRTs undergoing degradation is challenging

Numerous IRTs may be undetectable using RNA sequencing as they are rapidly degraded by NMD [65]. As reviewed elsewhere, NMD allows elimination of mRNAs with PTCs typically positioned more than 50-55 nucleotides upstream of their last exon-exon junction [66, 67]. As such, the majority of intron-retaining transcripts are computationally predicted as NMD targets [19, 30, 36]. However, some PTC-containing IRTs may escape NMD, and are translated. For example, the IR form of the endoplasmic reticulum chaperone GRP78/BiP contains a PTC when intron 1 is retained, but a shorter functional immunologically detectable polypeptide is translated from a downstream initiation codon [68]. This means that although computational predictions of PTC introduction based on sequence may suggest that an IRT should be subject to NMD, a protein with a distinct function may be generated instead. Alternatively, intron-retaining transcripts can be degraded via nuclear degradation mechanism(s) or Staufen-mediated decay [29].

One successful approach to enhancing the discovery of IR is to inhibit NMD prior to RNA sequencing. This process can be achieved by chemical or molecular means. Several chemical agents including emetine, actinomycin D, cycloheximide and caffeine can inhibit pathways involved in NMD activation such as protein translation and phosphorylation of NMD factors [29-31, 69-71]. Alternatively, RNA interference or gene knockout technologies can be used to ablate or inhibit the core NMD factors including UPF1, UPF2 and SMG6 [72, 73]. Using both chemical and siRNA approaches, we have previously shown that at least 45% of IRTs are subject to NMD in mouse granulocytes [30]. This is in contrast, for example, to work carried out in plants, where only 4% of transcripts containing a PTC were found to be targeted by the NMD machinery [74]. However, these approaches cannot measure transcripts

degraded by pathways other than NMD, including those that take place in the nucleus. We and others have reported that the vast majority of intron-retaining transcripts in erythroblasts and megakaryocytes are not degraded via NMD; although increased IR correlates with reduced gene expression in these cell types [31, 32]. This observation can be explained by the ID in the nucleus [31], where intron-retaining transcripts may be stored for later use. Recent work in mouse cortical neurons has demonstrated that for IRTs that undergo splicing within a 2-hour timeframe following transcriptional inhibition, only 73% undergo degradation, while 9% are instead spliced and go on to contribute to the pool of canonical protein-coding mRNA isoforms available to the cell [21]. A recent study has reported that only cytoplasmic intron-retaining transcripts are engaged by the ribosome [75], and while it is yet unknown whether novel proteins are derived from these transcripts, it is hypothesised that this initial round of translation is a prerequisite for NMD. It is reasonable to speculate that there are many more as yet unidentified functional intron-retaining transcripts, especially among those detained in the nucleus.

5 The extent and function of intron retaining transcripts in the nucleus remain poorly understood

When carrying out RNA sequencing, most studies do not first perform subcellular fractionation. This makes it impossible to assess what proportion of polyadenylated and non-polyadenylated transcripts with intronic reads is sequestered in the nucleus. Introns in these transcripts have been termed “detained” (ID), and these RNA can be degraded or stored in the nucleus, in contrast with classical “intron retaining” transcripts exported to the cytoplasm and accessible to the translation and NMD machineries (Figure 1, [22]).

Investigation of individual intron-detaining mRNA molecules sequestered in the nucleus under various physiological conditions has revealed the diverse roles this can play in reaction to a plethora of biological stimuli. Xu et al. [76] demonstrated a switch-like role of intron 3 retention and nuclear RNA localisation in modulating levels of apolipoprotein E protein in response to excitotoxic challenge. Ninomiya et al. [77] reported that the vast majority of *Clk1* transcripts in mouse brain, spleen, lung, liver and other tissues contained detained introns, and were sequestered in the nucleus until heat shock. Global studies have further extended these single-gene observations. In a study of transcripts detaining the 3' terminal intron, it was observed that the vast majority of these RNAs were not subject to NMD, and instead were detained in the nucleus and eventually degraded there [29]. A similar phenomenon has been observed in neuronal activity [21], heat shock [78] and stress response [22] paradigms. Overall, nuclear detention of IRTs appears to be a tightly controlled process, enabling cells to rapidly respond to stimuli without the need for transcription, and dynamically controlling available mRNA levels. Additional studies involving nuclear/cytoplasmic fractionation need to be carried out to assess the extent and function of IR vs ID, and to elucidate the mechanisms involved in regulating this process for specific RNA molecules.

6 Lack of protocols for isolation of subcellular compartments precludes identification of IR transcripts targeted to them

The presence of global subcellular compartments such as the cytoplasm and nucleus has been recognised since the earliest days of microscopy. However, there is less clarity on how many other spatial compartments and sub-compartments exist within

cells, and what defines the RNA and protein molecules that need to be trafficked there [79].

Several studies have reported how IR can serve as a signal for subcellular targeting. IR was crucial for the correct localization of 33 neuronal mRNAs to dendrites, in turn affecting the localization of the proteins that they encoded [26]. The authors were able to observe this phenomenon because it was well established that neurons have a cell body, partitioned into the cytoplasm, from which dendrites and the axon branch off. Established protocols were available to separate these cellular compartments, and published information was available about which of the mRNAs were localised to the dendrites. Had this confluence of data not pre-existed – which is the case for most cell systems and subcellular compartments– the authors would most likely have erroneously discounted this small number of non-degraded, protein-coding IRTs as being subject to NMD in the cytoplasm.

The localisation of not only RNA but also the proteins translated from them may be altered as a result of IR (Figure 1). Ni et al [68] showed that an intron-retaining form of the endoplasmic reticulum chaperone GRP78 was translated to form a shorter protein localised diffusely throughout the cytoplasm, while the non-IR form was tethered to the endoplasmic reticulum. Intriguingly, the diffusely localised isoform was more prevalent in leukemic patients, and hypothesised to play a pro-survival role during endoplasmic reticulum stress. Furthermore, while most splicing events occur co-transcriptionally in the nucleus, evidence from platelets – which lack a nucleus in their mature state - supports the possibility of cytoplasmic, point-of-action splicing [80]. Experiments carried out in isolated dendrites also demonstrate the possibility of intron removal occurring outside of the nucleus [81]. This has been proposed to enable highly compartmentalised cells to specifically target both mRNA and protein

molecules to key locations. Under such circumstances the intronic sequence forms the critical part of the "sentinel RNA" that contains the informative sequences for determining where in the cell a specific isoform and polypeptide variant will be transported to [82]. Hence, methods for the isolation of novel subcellular compartments is required to increase our understanding of the functions of IRTs as "targeted" molecules, and the alterations such targeting can undergo in disease contexts.

7 Detecting protein-coding intron-retaining transcripts remains challenging

It has long been hypothesised that the pioneer round of translation is a prerequisite for NMD and therefore most IRTs may be translated at least once [83, 84]. However, proteins originating from intron-retaining genes in 9 tissues had significantly lower protein output than those from non intron-retaining genes [20]. Strikingly, when examining ribosome binding data for IR events in a human cell line, there were no reads observed from retained introns, even though they were identified in polyadenylated mRNA [20].

Many individual cases of proteins generated from intron-retaining mRNAs with functions or subcellular localisation distinct from their fully spliced isoforms have been reported. Transcripts of the intermediate filament protein perphirin retaining two introns are upregulated in amyotrophic lateral sclerosis and translated to form a 28kDa protein involved in forming round inclusions – a pathological hallmark of the disease [85]. An intron-retaining isoform of carcinoembryonic antigen-related cell adhesion molecule 6 (Ceacam6) is localised to the interface between Sertoli and germ cells in rat testes, and contains three additional Ig-CAM domains relative to the fully

spliced isoform [86]. In contrast to this, an intron retaining isoform of cyclin D1b is characterised by the introduction of a termination codon which leads to the production of a shorter protein with a higher transforming activity and nuclear vs cytoplasmic localisation [87, 88]. Similarly, a smaller protein isoform of myo-inositol-3-phosphate synthase (*Isyna1*) is generated as a result of IR, and competes with the larger one for NAD+ binding, modulating the activity of this enzymatic complex [89].

Several individual protein-coding intron retaining isoforms have been associated with the development of human disease. IR results in the production of a smaller CYP11B1 protein isoform associated with the development of the steroidogenesis disorder congenital adrenal hyperplasia [90]. IR can also have a protective effect, with the retention of introns 12 and 13 in the calcineurin gene producing an isoform which improves cardiac function and reduces scar formation after myocardial infarction [91].

While it is clear from the above examples that IRTs can encode proteins with functions distinct from their non-intron-retaining counterparts, the full extent of this phenomenon and the functions of all of these IR protein isoforms remain unclear. This is primarily due to the challenge of correlating high-throughput RNA sequencing with proteomics data, as a result of the non-exhaustive, and often non-quantitative, nature of the latter. Indeed, only recently have technologies that purport to profile "full proteomes" been developed [92, 93], but available data support only a small subset of IR events [94]. During neuronal depolarisation, Prabakaran et al. [95] observed quantitative changes in a subset of peptides originating from introns using tandem mass tag labelling. Extrapolating based on canonical, well-known protein-coding transcripts and peptides, proteomics data accounted for only 25% of the intron-retaining transcripts detected by RNA-seq. This indicates that substantial

improvements are needed for an unbiased, global, exhaustive detection of novel peptides and proteins from introns and other "non-canonical" coding regions. This in turn hampers functional studies investigating more than a handful of candidates, and their roles in normal biology and disease.

8 The mechanisms regulating IR are complex

One of the major questions in understanding the roles of IR in normal biology and disease is why some introns are retained while others are not. Several factors are known to be involved in IR regulation including the expression levels of splicing factors, RNA polymerase II occupancies and epigenetic changes [19, 20, 28, 30, 96, 97]. Our group has previously reported the reduced expression of exon-defining splicing factors in granulocytes that harbour higher levels of IR than their progenitors [30]. In a comparative analysis of the 1000 most frequently retained introns versus an identical number of rarely retained introns in over 2500 tissue or cell types, we recently identified an enrichment of SR family protein binding sites in retained introns [20]. Knockdown of SR family proteins results in a dramatic increase in IR levels, indicating that most IR events can be modulated via common splicing regulatory mechanisms involving these proteins. However, the specificity of why one intron is retained whereas a nearby intron within the same gene is spliced out remains largely unexplained.

We and others have also discovered the enrichment of RNA pol II stalling across retained compared to constitutively spliced introns, indicating that a slower transcription elongation rate is associated with IR [19, 97]. We have further demonstrated the association between enriched RNA Pol II occupancy at retained introns and reduced splicing factor recruitment to splice junctions flanking retained

introns. We have shown that this mechanism is linked to epigenetic changes. Consequent to reduced levels of DNA methylation, the occupancy of the methylated DNA-binding protein, MeCP2, decreases. Given that MeCP2 can act as an adaptor to recruit splicing factors, its de-enrichment near splice junctions results in reduced splicing factor recruitment, leading to IR. Reduced RNA Pol II occupancy may result as a consequence of inefficient splicing factor recruitment as previously reported [19].

IR can also be regulated via the epigenetic mark H3.3 lysine 36 tri-methylation (H3.3K36me3) [96]. This chromatin mark recruits a reader protein BS69, which physically interacts with a component of the U5 snRNP complex, EFTUD2 [96]. Knockdown of BS69 *in vitro* led to an increased steady-state of processed mRNA in the cytoplasm, indicating that the binding of BS69 to EFTUD2 antagonises its activity and supresses splicing [96].

Collectively, there is dynamic cooperation between epigenetic and splicing machineries in the regulation of IR. However, each of the mechanisms described above individually accounted for less than 20% of IR events observed in the investigated cell types [96, 97], indicating that other factors regulating IR remain to be determined. The dynamic regulation of IR by multiple distinct cellular processes is intriguing in itself, and indicates that our understanding of IR remains somewhat limited.

9 Conclusions

IR is an important form of alternative splicing that is crucial for normal human, animal and plant biology. However, many challenges remain in identifying IRTs and

understanding their functions. These challenges include the complexity of genome structure and organisation in higher organisms, and the plethora of algorithmic and bioinformatic difficulties in identifying and quantitating retained introns. This could be addressed by comparative studies of informatics methodologies, which have a long history in the differential gene expression analysis space [50, 98, 99], but are lacking for IR analysis. There are substantial pitfalls when adapting existing techniques developed for protein-coding genes for carrying out comparative phylogenetic analysis, as the roles of IR appear to be conserved within closely related taxa, but not across distant phyla. Finally, there is a lack of tools and approaches to investigate the mechanisms of IR regulation and its functional consequences in terms of protein production, RNA-based regulation, subcellular localisation or possible decay pathways. Overcoming these challenges will facilitate a more complete understanding of why cells retain certain introns and not others, and the contribution of this form of alternative splicing to the structure and function of gene transcripts in normal biology and disease.

Figure 1: The diverse fates of intron retaining transcripts.

(A) Splicing results in the excision of all intronic sequence from a pre-mRNA molecule. (B) The mRNA is then exported to the cytoplasm, where it is translated into the canonical protein isoform. (C) Intron retention (IR) occurs when sequence corresponding to one or more introns is not spliced out of the pre-mRNA molecule, giving rise to an intron retaining transcript (IRT). (D) IRTs can be exported into the cytoplasm, where they can (E) interact with the ribosome and (F) be subject to NMD if a premature termination codon (PTC) is encountered, or (G) be translated to give rise to alternative protein isoforms. (H) Cytoplasmic IRTs can be targeted to specific subcellular compartments, where they are stored or translated. (I) Alternatively, an IRT can be sequestered in the nucleus – in which case it is termed an intron detaining transcript (IDT). This IDT can be subject to (J) stimulus-dependent splicing or (K) degradation.

Figure 2: Introns are more enriched in repetitive sequences than exons

In both the human and mouse genomes, intronic regions contain proportionally more repetitive sequence than exonic regions. RepeatMasker [100] annotations were intersected with collapsed human and mouse Gencode [101, 102] exonic and intronic regions using BEDTools [103], and the proportion of bases covered by each repeat type was calculated and visualised using tidyverse and ggplot2 [104].

Figure 3: Informatic pitfalls of IR detection

A schematic of a gene affected by differential IR between two cellular states. (A) Differential expression alters the normalised density of reads across the entire gene body, resulting in better coverage in cell state 1 than 2. Not taking these differences

into account can prevent or bias IR detection and quantification. (B) Unless coverage filters and adequate multimapping read handling are implemented, expression of the repetitive elements in introns 1 and 2 can result in spurious intronic and junction read counts. (C) These filters must also prevent the differential expression of the miRNA hosted in intron 3 from contributing to assessment of IR for this intron. (D) The antisense transcript hosted in introns 3 - 4, which is expressed in cell state 1 and not 2, could also confound analysis, since only stranded RNA-seq with subsequent coverage filtering could permit identification of this phenomenon. (E) If the sensitivity and specificity of the employed informatics pipeline is adequate, only intron 5 in cell state 2 should be reported as differentially retained.

10 References

[1] Q. Pan, O. Shai, L.J. Lee, B.J. Frey, B.J. Blencowe, Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, Nat. Genet. 40(12) (2008) 1413-1415.
[2] B.-B. Wang, V. Brendel, Genomewide comparative analysis of alternative splicing in plants, Proc. Natl. Acad. Sci. U. S. A. 103(18) (2006) 7175-7180.
[3] C. Reissner, F. Runkel, M. Missler, Neurexins, Genome Biol 14(9) (2013) 213.
[4] A. Kalsotra, T.A. Cooper, Functional consequences of developmentally regulated alternative splicing, Nat. Rev. Genet. 12(10) (2011) 715-729.
[5] V. Cho, Y. Mei, A. Sanny, S. Chan, A. Enders, E.M. Bertram, A. Tan, C.C. Goodnow, T.D. Andrews, The RNA-binding protein hnRNPLL induces a T cell alternative splicing program delineated by differential intron retention in polyadenylated RNA, Genome Biol. 15(1) (2014) R26.
[6] T. Ni, W. Yang, M. Han, Y. Zhang, T. Shen, H. Nie, Z. Zhou, Y. Dai, Y. Yang, P. Liu, K. Cui, Z. Zeng, Y. Tian, B. Zhou, G. Wei, K. Zhao, W. Peng, J. Zhu, Global intron retention mediated gene regulation during CD4+ T cell activation, Nucleic Acids Res 44(14) (2016) 6817-29.
[7] T. Hirose, J.A. Steitz, Position within the host intron is critical for efficient processing of box C/D snoRNAs in mammalian cells, Proc. Natl. Acad. Sci. U. S. A. 98(23) (2001) 12914-12919.
[8] Y.-K. Kim, V.N. Kim, Processing of intronic microRNAs, EMBO J. 26(3) (2007) 775-783.
[9] D. Lutter, C. Marr, J. Krumsiek, E.W. Lang, F.J. Theis, Intronic microRNAs support their host genes by mediating synergistic and antagonistic regulatory effects, BMC Genomics 11 (2010) 224.
[10] X. Gao, Y. Qiao, D. Han, Y. Zhang, N. Ma, Enemy or partner: relationship between intronic micrornas and their host genes, IUBMB Life 64(10) (2012) 835-840.
[11] H. Ner-Gaon, R. Halachmi, S. Savaldi-Goldstein, E. Rubin, R. Ophir, R. Fluhr, Intron retention is a major phenomenon in alternative splicing in Arabidopsis, Plant J. 39(6) (2004) 877-885.
[12] Y. Marquez, J.W.S. Brown, C. Simpson, A. Barta, M. Kalyna, Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis, Genome Res. 22(6) (2012) 1184-1195.
[13] A. Sebé-Pedrós, M. Irimia, J. Del Campo, H. Parra-Acero, C. Russ, C. Nusbaum, B.J. Blencowe, I. Ruiz-Trillo, Regulated aggregative multicellularity in a close unicellular relative of metazoa, Elife 2 (2013) e01287.
[14] P.A. Galante, N.J. Sakabe, N. Kirschbaum-Slager, S.J. de Souza, Detection and evaluation of intron retention events in the human transcriptome, RNA 10(5) (2004) 757-65.
[15] N.J. Sakabe, S.J. de Souza, Sequence features responsible for intron retention in human, BMC Genomics 8 (2007) 59.

[16] E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, C.B. Burge, Alternative isoform regulation in human tissue transcriptomes, Nature 456(7221) (2008) 470-476.
[17] A. Ameur, A. Zaghlool, J. Halvardson, A. Wetterbom, U. Gyllensten, L. Cavelier, L. Feuk, Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain, Nat. Struct. Mol. Biol. 18(12) (2011) 1435-1440.
[18] Y.L. Khodor, J. Rodriguez, K.C. Abruzzi, C.-H.A. Tang, M.T. Marr, 2nd, M. Rosbash, Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila, Genes Dev. 25(23) (2011) 2502-2512.
[19] U. Braunschweig, N.L. Barbosa-Morais, Q. Pan, E.N. Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia, B.J. Blencowe, Widespread intron retention in mammals functionally tunes transcriptomes, Genome Res 24(11) (2014) 1774-86.
[20] R. Middleton, D. Gao, A. Thomas, B. Singh, A. Au, J.J.L. Wong, A. Bomane, B. Cosson, E. Eyras, J.E.J. Rasko, W. Ritchie, IRFinder: assessing the impact of intron retention on mammalian gene expression, Genome Biol. 18(1) (2017) 51.
[21] O. Mauger, F. Lemoine, P. Scheiffele, Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity, Neuron 92(6) (2016) 1266-1278.
[22] P.L. Boutz, A. Bhutkar, P.A. Sharp, Detained introns are a novel, widespread class of post-transcriptionally spliced introns, Genes Dev 29(1) (2015) 63-80.
[23] O. Jaillon, K. Bouhouche, J.-F. Gout, J.-M. Aury, B. Noel, B. Saudemont, M. Nowacki, V. Serrano, B.M. Porcel, B. Ségurens, A. Le Mouël, G. Lepère, V. Schächter, M. Bétermier, J. Cohen, P. Wincker, L. Sperling, L. Duret, E. Meyer, Translational control of intron splicing in eukaryotes, Nature 451(7176) (2008) 359-362.
[24] R.K. Gudipati, Z. Xu, A. Lebreton, B. Séraphin, L.M. Steinmetz, A. Jacquier, D. Libri, Extensive degradation of RNA precursors by the exosome in wild-type cells, Mol. Cell 48(3) (2012) 409-421.
[25] A.M. Gontijo, V. Miguela, M.F. Whiting, R.C. Woodruff, M. Dominguez, Intron retention in the Drosophila melanogaster Rieske Iron Sulphur Protein gene generated a new protein, Nat. Commun. 2 (2011) 323.
[26] P.T. Buckley, M.T. Lee, J.-Y. Sul, K.Y. Miyashiro, T.J. Bell, S.A. Fisher, J. Kim, J. Eberwine, Cytoplasmic intron sequence-retaining transcripts can be dendritically targeted via ID element retrotransposons, Neuron 69(5) (2011) 877-884.
[27] J.J.L. Wong, A.Y.M. Au, W. Ritchie, J.E.J. Rasko, Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology, Bioessays 38(1) (2016) 41-49.
[28] P. Gascard, M. Bilenky, M. Sigaroudinia, J. Zhao, L. Li, A. Carles, A. Delaney, A. Tam, B. Kamoh, S. Cho, M. Griffith, A. Chu, G. Robertson, D. Cheung, I. Li, A. Heravi-Moussavi, M. Moksa, M. Mingay, A. Hussainkhel, B. Davis, R.P. Nagarajan, C. Hong, L. Echipare, H. O'Geen, M.J. Hangauer, J.B. Cheng, D. Neel, D. Hu, M.T. McManus, R. Moore, A. Mungall, Y. Ma, P. Plettner, E. Ziv, T. Wang, P.J. Farnham, S.J.M. Jones, M.A. Marra, T.D. Tlsty, J.F. Costello, M. Hirst, Epigenetic and transcriptional determinants of the human breast, Nat. Commun. 6 (2015) 6351.

[29] K. Yap, Z.Q. Lim, P. Khandelia, B. Friedman, E.V. Makeyev, Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention, Genes Dev 26(11) (2012) 1209-1223.
[30] Justin J.L. Wong, W. Ritchie, Olivia A. Ebner, M. Selbach, Jason W.H. Wong, Y. Huang, D. Gao, N. Pinello, M. Gonzalez, K. Baidya, A. Thoeng, T.-L. Khoo, Charles G. Bailey, J. Holst, John E.J. Rasko, Orchestrated Intron Retention Regulates Normal Granulocyte Differentiation, Cell 154(3) (2013
) 583-595.
[31] H. Pimentel, M. Parra, S.L. Gee, N. Mohandas, L. Pachter, J.G. Conboy, A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis, Nucleic Acids Res 44(2) (2016) 838-851.
[32] C.R. Edwards, W. Ritchie, J.J.L. Wong, U. Schmitz, R. Middleton, X. An, N. Mohandas, J.E.J. Rasko, G.A. Blobel, A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages, Blood 127 (2016) e24-e34
.
[33] S.M.I. Hussein, M.C. Puri, P.D. Tonge, M. Benevento, A.J. Corso, J.L. Clancy, R. Mosbergen, M. Li, D.-S. Lee, N. Cloonan, D.L.A. Wood, J. Munoz, R. Middleton, O. Korn, H.R. Patel, C.A. White, J.-Y. Shin, M.E. Gauthier, K.-A. Lê Cao, J.-I. Kim, J.C. Mar, N. Shakiba, W. Ritchie, J.E.J. Rasko, S.M. Grimmond, P.W. Zandstra, C.A. Wells, T. Preiss, J.-S. Seo, A.J.R. Heck, I.M. Rogers, A. Nagy, Genome-wide characterization of the routes to pluripotency, Nature 516(7530) (2014) 198-206.
[34] S. Martin, N. Bellora, J. González-Vallinas, M. Irimia, K. Chebli, M. de Toledo, M. Raabe, E. Eyras, H. Urlaub, B.J. Blencowe, J. Tazi, Preferential binding of a stable G3BP ribonucleoprotein complex to intron-retaining transcripts in mouse brain and modulation of their expression in the cerebellum, J. Neurochem. 139(3) (2016) 349-368.
[35] H. Dvinge, R.K. Bradley, Widespread intron retention diversifies most cancer transcriptomes, Genome Med. 7(1) (2015) 45.
[36] H. Jung, D. Lee, J. Lee, D. Park, Y.J. Kim, W.-Y. Park, D. Hong, P.J. Park, E. Lee, Intron retention is a widespread mechanism of tumor-suppressor inactivation, Nat. Genet. 47(11) (2015) 1242-1248.
[37] D. Memon, K. Dawson, C.S.F. Smowton, W. Xing, C. Dive, C.J. Miller, Hypoxia-driven splicing into noncoding isoforms regulates the DNA damage response, npj Genomic Medicine 1 (2016) 16020.
[38] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S.L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, Genome Biol 14(4) (2013) R36.
[39] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29(1) (2013) 15-21.
[40] K. Wang, D. Singh, Z. Zeng, S.J. Coleman, Y. Huang, G.L. Savich, X. He, P. Mieczkowski, S.A. Grimm, C.M. Perou, J.N. MacLeod, D.Y. Chiang, J.F. Prins, J. Liu, MapSplice: accurate mapping of RNA-seq reads for splice junction discovery, Nucleic Acids Res. 38(18) (2010) e178.
[41] Y. Bai, S. Ji, Y. Wang, IRcall and IRclassifier: two methods for flexible detection of intron retention events from RNA-Seq data, BMC Genomics 16 Suppl 2 (2015) S9.

[42] J.Y. Wu, J.A. Potashkin, Alternative Splicing in the Nervous System, in: L.R. Squire (Ed.), Encyclopedia of Neuroscience, Academic Press, Oxford, 2009, pp. 245-251.
[43] S. Gueroussov, T. Gonatopoulos-Pournatzis, M. Irimia, B. Raj, Z.Y. Lin, A.C. Gingras, B.J. Blencowe, An alternative splicing event amplifies evolutionary differences between vertebrates, Science 349(6250) (2015) 868-73.
[44] V. Madan, D. Kanojia, J. Li, R. Okamoto, A. Sato-Otsubo, A. Kohlmann, M. Sanada, V. Grossmann, J. Sundaresan, Y. Shiraishi, S. Miyano, F. Thol, A. Ganser, H. Yang, T. Haferlach, S. Ogawa, H.P. Koeffler, Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome, Nat. Commun. 6 (2015) 6042.
[45] T.R. Mercer, M.B. Clark, J. Crawford, M.E. Brunck, D.J. Gerhardt, R.J. Taft, L.K. Nielsen, M.E. Dinger, J.S. Mattick, Targeted sequencing for gene discovery and quantification using RNA CaptureSeq, Nat. Protoc. 9(5) (2014) 989-1009.
[46] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M.W. Szcześniak, D.J. Gaffney, L.L. Elo, X. Zhang, A. Mortazavi, A survey of best practices for RNA-seq data analysis, Genome Biology 17(1) (2016) 13.
[47] H. Tilgner, F. Jahanbani, T. Blauwkamp, A. Moshrefi, E. Jaeger, F. Chen, I. Harel, C.D. Bustamante, M. Rasmussen, M.P. Snyder, Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events, Nat. Biotechnol. 33(7) (2015) 736-742.
[48] Z. Kuang, J.D. Boeke, S. Canzar, The dynamic landscape of fission yeast meiosis alternative-splice isoforms, Genome Res. 27(1) (2017) 145-156.
[49] M.T. Bolisetty, G. Rajadinakaran, B.R. Graveley, Determining exon connectivity in complex mRNAs by nanopore sequencing, Genome Biol. 16(1) (2015) 204.
[50] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézic, C. French StatOmique, A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, Brief. Bioinform. 14(6) (2013) 671-683.
[51] C. Everaert, M. Luypaert, J.L.V. Maag, Q.X. Cheng, M.E. Dinger, J. Hellemans, P. Mestdagh, Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data, Sci Rep 7(1) (2017) 1559.
[52] N.L. Barbosa-Morais, M. Irimia, Q. Pan, H.Y. Xiong, S. Gueroussov, L.J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C.M. Misquitta-Ali, M.D. Wilson, P.M. Kim, D.T. Odom, B.J. Frey, B.J. Blencowe, The evolutionary landscape of alternative splicing in vertebrate species, Science 338(6114) (2012) 1587-93.
[53] W.B. Barbazuk, Y. Fu, K.M. McGinnis, Genome-wide analyses of alternative splicing in plants: opportunities and challenges, Genome Res 18(9) (2008) 1381-92.
[54] L. Chen, S.J. Bush, J.M. Tovar-Corona, A. Castillo-Morales, A.O. Urrutia, Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity, Mol Biol Evol 31(6) (2014) 1402-13.
[55] E. Kim, A. Magen, G. Ast, Different levels of alternative splicing among eukaryotes, Nucleic Acids Res 35(1) (2007) 125-31.

[56] E. Kim, A. Goren, G. Ast, Alternative splicing: current perspectives, Bioessays 30(1) (2008) 38-47.
[57] I.B. Rogozin, Y.I. Wolf, A.V. Sorokin, B.G. Mirkin, E.V. Koonin, Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution, Curr Biol 13(17) (2003) 1512-7.
[58] Q. Zhang, S.V. Edwards, The evolution of intron size in amniotes: a role for powered flight?, Genome Biol Evol 4(10) (2012) 1033-43.
[59] A.M. McGuire, M.D. Pearson, D.E. Neafsey, J.E. Galagan, Cross-kingdom patterns of alternative splicing and splice recognition, Genome Biol 9(3) (2008) R50.
[60] J.T. Mendell, N.A. Sharifi, J.L. Meyers, F. Martinez-Murillo, H.C. Dietz, Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise, Nat Genet 36(10) (2004) 1073-8.
[61] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci U S A 102(43) (2005) 15545-50.
[62] B.T. Sherman, W. Huang da, Q. Tan, Y. Guo, S. Bour, D. Liu, R. Stephens, M.W. Baseler, H.C. Lane, R.A. Lempicki, DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis, BMC Bioinformatics 8 (2007) 426.
[63] H. Mi, A. Muruganujan, P.D. Thomas, PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees, Nucleic Acids Res 41(Database issue) (2013) D377-86.
[64] J.A. Timmons, K.J. Szkop, I.J. Gallagher, Multiple sources of bias confound functional enrichment analysis of global -omics data, Genome Biol 16 (2015) 186.
[65] T. Trcek, H. Sato, R.H. Singer, L.E. Maquat, Temporal and spatial characterization of nonsense-mediated mRNA decay, Genes Dev 27(5) (2013) 541-551.
[66] S. Lykke-Andersen, Y. Chen, B.R. Ardal, B. Lilje, J. Waage, A. Sandelin, T.H. Jensen, Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes, Genes Dev. 28(22) (2014) 2498-2517.
[67] S. Lykke-Andersen, T.H. Jensen, Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes, Nat Rev Mol Cell Biol 16(11) (2015) 665-77.
[68] M. Ni, H. Zhou, S. Wey, P. Baumeister, A.S. Lee, Regulation of PERK signaling and leukemic cell survival by a novel cytosolic isoform of the UPR regulator GRP78/BiP, PLoS One 4(8) (2009) e6868.
[69] Y. Ionov, N. Nowak, M. Perucho, S. Markowitz, J.K. Cowell, Manipulation of nonsense mediated decay identifies gene mutations in colon cancer Cells with microsatellite instability, Oncogene 23(3) (0000) 639-645.
[70] I. Ivanov, K.C. Lo, L. Hawthorn, J.K. Cowell, Y. Ionov, Identifying candidate colon cancer tumor suppressor genes using inhibition of nonsense-mediated mRNA decay in colon cancer cells, Oncogene 26(20) (2006) 2873-2884.
[71] F. Usuki, A. Yamashita, I. Higuchi, T. Ohnishi, T. Shiraishi, M. Osame, S. Ohno, Inhibition of nonsense-mediated mRNA decay rescues the phenotype in Ullrich's disease, Ann Neurol 55(5) (2004) 740-744.

[72] C. Gong, Y.K. Kim, C.F. Woeller, Y. Tang, L.E. Maquat, SMD and NMD are competitive pathways that contribute to myogenesis: effects on PAX3 and myogenin mRNAs, Genes & Development 23(1) (2009) 54-66.
[73] S.A. Schmidt, P.L. Foley, D.-H. Jeong, L.A. Rymarquis, F. Doyle, S.A. Tenenbaum, J.G. Belasco, P.J. Green, Identification of SMG6 cleavage sites and a preferred RNA cleavage motif by global analysis of endogenous NMD targets in human cells, Nucleic Acids Res 43(1) (2015) 309-323.
[74] S. Li, M. Yamada, X. Han, U. Ohler, P.N. Benfey, High-Resolution Expression Map of the Arabidopsis Root Reveals Alternative Splicing and lincRNA Regulation, Dev. Cell (2016).
[75] R.J. Weatheritt, T. Sterne-Weiler, B.J. Blencowe, The ribosome-engaged landscape of alternative splicing, Nat Struct Mol Biol 23(12) (2016) 1117-1123.
[76] Q. Xu, D. Walker, A. Bernardo, J. Brodbeck, M.E. Balestra, Y. Huang, Intron-3 retention/splicing controls neuronal expression of apolipoprotein E in the CNS, J. Neurosci. 28(6) (2008) 1452-1459.
[77] K. Ninomiya, N. Kataoka, M. Hagiwara, Stress-responsive maturation of Clk1/4 pre-mRNAs promotes phosphorylation of SR splicing factor, J. Cell Biol. 195(1) (2011) 27-40.
[78] R. Shalgi, J.A. Hurt, S. Lindquist, C.B. Burge, Widespread inhibition of posttranscriptional splicing shapes the cellular transcriptome following heat shock, Cell Rep. 7(5) (2014) 1362-1370.
[79] P. van Bergeijk, C.C. Hoogenraad, L.C. Kapitein, Right Time, Right Place: Probing the Functions of Organelle Positioning, Trends Cell Biol. 26(2) (2016) 121-134.
[80] M.M. Denis, N.D. Tolley, M. Bunting, H. Schwertz, H. Jiang, S. Lindemann, C.C. Yost, F.J. Rubner, K.H. Albertine, K.J. Swoboda, C.M. Fratto, E. Tolley, L.W. Kraiss, T.M. McIntyre, G.A. Zimmerman, A.S. Weyrich, Escaping the nuclear confines: signal-dependent pre-mRNA splicing in anucleate platelets, Cell 122(3) (2005) 379-91.
[81] J. Glanzer, K.Y. Miyashiro, J.Y. Sul, L. Barrett, B. Belt, P. Haydon, J. Eberwine, RNA splicing capability of live neuronal dendrites, Proc. Natl. Acad. Sci. U. S. A. 102(46) (2005) 16859-16864.
[82] P.T. Buckley, M. Khaladkar, J. Kim, J. Eberwine, Cytoplasmic intron retention, function, splicing, and the sentinel RNA hypothesis, Wiley Interdiscip. Rev. RNA 5(2) (2014) 223-230.
[83] O. Muhlemann, A.B. Eberle, L. Stalder, R. Zamudio Orozco, Recognition and elimination of nonsense mRNA, Biochim Biophys Acta 1779(9) (2008) 538-49.
[84] O. Muhlemann, J. Lykke-Andersen, How and where are nonsense mRNAs degraded in mammalian cells?, RNA Biol 7(1) (2010) 28-32.
[85] S. Xiao, S. Tjostheim, T. Sanelli, J.R. McLean, P. Horne, Y. Fan, J. Ravits, M.J. Strong, J. Robertson, An aggregate-inducing peripherin isoform generated through intron retention is upregulated in amyotrophic lateral sclerosis and associated with disease pathology, J. Neurosci. 28(8) (2008) 1833-1840.
[86] H. Kurio, E. Murayama, T. Kaneko, Y. Shibata, T. Inai, H. Iida, Intron retention generates a novel isoform of CEACAM6 that may act as an adhesion molecule in the ectoplasmic specialization structures between spermatids and sertoli cells in rat testis, Biol. Reprod. 79(6) (2008) 1062-1073.
[87] F. Lu, A.B. Gladden, J.A. Diehl, An alternatively spliced cyclin D1 isoform, cyclin D1b, is a nuclear oncogene, Cancer Res. 63(21) (2003) 7056-7061.

[88] D.A. Solomon, Y. Wang, S.R. Fox, T.C. Lambeck, S. Giesting, Z. Lan, A.M. Senderowicz, E.S. Knudsen, Cyclin D1 Splice Variants: DIFFERENTIAL EFFECTS ON LOCALIZATION, RB PHOSPHORYLATION, AND CELLULAR TRANSFORMATION, J. Biol. Chem. 278(32) (2003) 30339-30347.
[89] R.S. Seelan, J. Lakshmanan, M.F. Casanova, R.N. Parthasarathy, Identification of myo-Inositol-3-phosphate Synthase Isoforms: CHARACTERIZATION, EXPRESSION, AND PUTATIVE ROLE OF A 16-kDa γc ISOFORM, J. Biol. Chem. 284(14) (2009) 9443-9457.
[90] H.H. Nguyen, A. Eiden-Plach, F. Hannemann, E.M. Malunowicz, M.F. Hartmann, S.A. Wudy, R. Bernhardt, Phenotypic, metabolic, and molecular genetic characterization of six patients with congenital adrenal hyperplasia caused by novel mutations in the CYP11B1 gene, J Steroid Biochem Mol Biol 155(Pt A) (2016) 126-34.
[91] L.E. Felkin, T. Narita, R. Germack, Y. Shintani, K. Takahashi, P. Sarathchandra, M.M. López-Olañeta, J.M. Gómez-Salinero, K. Suzuki, P.J.R. Barton, N. Rosenthal, E. Lara-Pezzi, Calcineurin splicing variant calcineurin Aβ1 improves cardiac function after myocardial infarction without inducing hypertrophy, Circulation 123(24) (2011) 2838-2847.
[92] M. Beck, A. Schmidt, J. Malmstroem, M. Claassen, A. Ori, A. Szymborska, F. Herzog, O. Rinner, J. Ellenberg, R. Aebersold, The quantitative proteome of a human cell line, Mol. Syst. Biol. 7 (2011) 549.
[93] N. Nagaraj, J.R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Pääbo, M. Mann, Deep proteome and transcriptome mapping of a human cancer cell line, Mol. Syst. Biol. 7 (2011) 548.
[94] J.E. Kroll, S.J. de Souza, G.A. de Souza, Identification of rare alternative splicing events in MS/MS data reveals a significant fraction of alternative translation initiation sites, PeerJ 2 (2014) e673.
[95] S. Prabakaran, M. Hemberg, R. Chauhan, D. Winter, R.Y. Tweedie-Cullen, C. Dittrich, E. Hong, J. Gunawardena, H. Steen, G. Kreiman, J.A. Steen, Quantitative profiling of peptides from RNAs classified as noncoding, Nat. Commun. 5 (2014) 5429.
[96] R. Guo, L. Zheng, Juw W. Park, R. Lv, H. Chen, F. Jiao, W. Xu, S. Mu, H. Wen, J. Qiu, Z. Wang, P. Yang, F. Wu, J. Hui, X. Fu, X. Shi, Yujiang G. Shi, Y. Xing, F. Lan, Y. Shi, BS69/ZMYND11 Reads and Connects Histone H3.3 Lysine 36 Trimethylation-Decorated Chromatin to Regulated Pre-mRNA Processing, Mol Cell 56(2) (2014) 298-310.
[97] G.D. Wong J.J.-L., Nguyen T.V., Kwok C.-T., van Geldermalsen M., Middleton R., Pinello N., Thoeng A., Nagarajah R., Holst J., Ritchie W., Rasko J.E.J., Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment. , Nature Commun. (in press) (2017).
[98] J.H. Bullard, E. Purdom, K.D. Hansen, S. Dudoit, Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, BMC Bioinformatics 11 (2010) 94.
[99] B. Ding, L. Zheng, W. Wang, Assessment of single cell RNA-seq normalization methods, bioRxiv (2016).
[100] A. Smit, R. Hubley, P. Green, RepeatMasker Open-3.0, 1996-2010. http://www.repeatmasker.org/.
[101] J. Harrow, A. Frankish, J.M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B.L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V.

Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J.M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, T.J. Hubbard, GENCODE: the reference human genome annotation for The ENCODE Project, Genome Res 22(9) (2012) 1760-74.
[102] J.M. Mudge, J. Harrow, Creating reference gene annotation for the mouse C57BL6/J genome assembly, Mamm Genome 26(9-10) (2015) 366-78.
[103] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26(6) (2010) 841-842.
[104] H. Wickham, ggplot2: elegant graphics for data analysis, Springer Science & Business Media2009.