Computational methods for intron retention identification and quantification

Author and Co-author Contact Information
Ulf Schmitz
Computational BioMedicine Laboratory Centenary Institute, The University of Sydney, Australia
Gene & Stem Cell Therapy Program Centenary Institute, The University of Sydney, Australia
Sydney Medical School, The University of Sydney, Australia
u.schmitz@centenary.org.au
T: +61 2 9565 6209

Geoffray Monteuuis
Gene & Stem Cell Therapy Program Centenary Institute, The University of Sydney, Australia
g.monteuuis@centenary.org.au
T: +61 2 9565 6162

Veronika Petrova
Computational BioMedicine Laboratory Centenary Institute, The University of Sydney, Australia
Gene & Stem Cell Therapy Program Centenary Institute, The University of Sydney, Australia
v.petrova@centenary.org.au
T: +61 2 9565 6209

Jaynish S. Shah
Gene & Stem Cell Therapy Program Centenary Institute, The University of Sydney, Australia
Sydney Medical School, The University of Sydney, Australia
j.shah@centenary.org.au
T: +61 2 9565 6289

John E.J. Rasko
Gene & Stem Cell Therapy Program Centenary Institute, The University of Sydney, Australia;
Sydney Medical School, The University of Sydney, Australia;
Cell and Molecular Therapies, Royal Prince Alfred Hospital, Camperdown, Australia
j.rasko@centenary.org.au
T: +61 2 9565 6156

Keywords
Aberrant splicing; Alternative splicing; Epigenetics; Gene isoforms; Gene regulation; Intron retention; Isoform detection; RNA sequencing; Transcriptomics; Transcriptomic complexity; Transcript quantification

Abstract (200-250 words)

Alternative splicing is a ubiquitous process that increases transcriptomic and proteomic complexity across the animal kingdom. Intron retention (IR) is a particular form of alternative splicing that is different from the other forms as it only increases transcriptomic complexity

but rarely directly affects the proteome. IR has long been neglected as it was considered a mis-splicing event and was referred to as transcriptional noise. However, recent reports have attributed a pivotal role to IR in normal physiology and diseases.
Studying IR comes with specific technical and analytical requirements, that enable a robust detection and quantification of this phenomenon. Advances in sequencing technologies and the development of IR calling and quantification software have facilitated numerous novel insights into the complex life of introns.
In this chapter, we describe computational methods for the analysis of IR events, their characteristics and conservation, the regulation of IR, and downstream consequences. We also introduce experimental approaches that are used in IR research.

Introduction

Intron retention and the mammalian transcriptome

With the advent of next-generation sequencing technologies, in particular RNA sequencing, we were able to study cellular transcriptomes at great detail. Recent landmark studies suggest that more than 95% of human multi-exonic genes are subject to alternative splicing and thereby give rise to at least two alternative isoforms (Merkin et al. 2012; Barbosa-Morais et al. 2012; Nilsen and Graveley 2010).
A striking transcriptomic diversity, enabled by alternative splicing, was revealed across many species. A major contributor to this diversity is IR, the only form of alternative splicing that does not affect proteomic complexity (Wong et al. 2015). Moreover, IR was found to be a new form of post-transcriptional gene regulation that is important, for example, in the differentiation of hematopoietic lineages (Wong et al. 2013; Edwards et al. 2016; Ni et al. 2016; Pimentel et al. 2016). It is known that introns contain *cis*-regulatory elements, such as regulatory motifs, but they can also accommodate *trans*-acting elements, such as small nucleolar RNAs and microRNAs (Hirose and Steitz 2001; Kim and Kim 2007). Thus, IR has novel gene regulatory implications that can, for example, facilitate stem cell differentiation (Naro et al. 2017), rapid responses to biological stimuli (Mauger, Lemoine, and Scheiffele 2016; Ni et al. 2016), as well as disease pathogenesis and progression (Dvinge and Bradley 2015; Jung et al. 2015; Wong, Rasko, and Wong 2018).

IR is a widespread form of post-transcriptional gene regulation

Formally, IR occurs when the splicing machinery fails to excise an intron from a pre-mRNA transcript so that the introns remains part of the mature mRNA. While most mRNA transcripts are transported to the cytoplasm, where they function as a blueprint for protein synthesis, many intron-retaining transcripts remain in the nucleus (Boutz, Bhutkar, and Sharp 2015). Others, that are transported into the cytoplasm are often subjected to nonsense-mediated decay, a process initiated by the cellular surveillance machinery that detects premature termination codons. Retained introns are enriched in premature termination codons (Lareau et al. 2007), hence they are considered a mediator of nonsense-mediated decay and IR is seen as a distinct form of post-transcriptional gene regulation (Wong et al. 2015).
While it has been shown that IR affects ~80 % of protein coding genes in human (Middleton et al. 2017), a comparison of IR occurrences in 11 vertebrate species has shown that in 50-75 % of multi-exonic genes are affected vertebrates (Braunschweig et al. 2014). IR is also widespread in fungi, insects, viruses and represents the most frequent form of AS in plants

(Kim, Magen, and Ast 2007; McGuire et al. 2008). In rice, for example, IR occurs in 47% of all AS events (Zhang et al. 2010).
In Saccharomyces cerevisiae, orchestrated IR occurs during the transition from vegetative growth to sporulation as 13 meiosis-specific introns are incompletely spliced during exponential growth in rich media (Juneau et al. 2007). Post-transcriptional regulation of the transition from mitosis to meiosis via IR is essential for yeast in order to maintain active growth. IR is also widespread during parasite differentiation, which was shown in analyses of the intron-rich genomes of apicomplexan parasites. Moreover, IR prevents translation of stage specific isoforms of glycolytic enzymes in T. gondii (Lunghi et al. 2016).
Thus, it has been known for a while that IR is widespread in plants, fungi and unicellular eukaryotes (Ner-Gaon et al. 2004; Marquez et al. 2012; Sebe-Pedros et al. 2013). The omnipresence in vertebrate and mammalian species became only apparent when next-generation sequencing technologies became available (Schmitz et al. 2017; Braunschweig et al. 2014).

Alternative fates of intron-retaining transcripts

It was shown recently that ~80% of coding genes can be affected by IR in human (Middleton et al. 2017), however the fate of intron-retaining transcripts is not always the same. Nuclear detained intron-retaining transcripts are either target of nuclear degradation pathways or comprise a pool of "sentinel" RNAs that are ready to be processed upon environmental stimuli facilitating rapid protein translation (Wong et al. 2015). Some intron-retaining transcripts might even become blueprints for new protein isoforms (Gontijo et al. 2011). Moreover, introns can carry signals that facilitate the specific subcellular localization of the intron-retaining transcript (Buckley et al. 2011).
An overview about IR, fates of intron-retaining transcripts, and other forms of alternative splicing is provided in Figure 1.

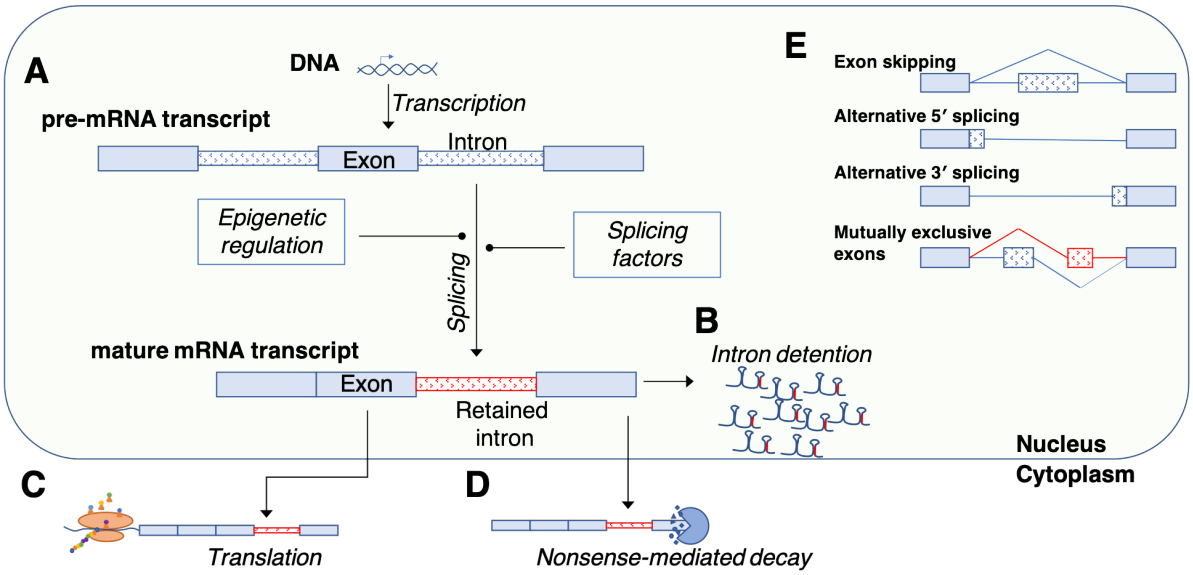

Figure 1 Biogenesis and fate of intron-retaining transcripts. ***(A)*** *IR is an alternative splicing event that leads to the inclusion of an intron in the mature mRNA transcript. Gene splicing occurs co-transcriptionally and trans-regulators, such as DNA methylation, histone marks and the availability of splicing factors can modulate IR events.* ***(B)*** *The accumulation of intron-retaining transcripts in the nucleus is referred to as intron detention.* ***(C)*** *In rare cases, intron-retaining transcripts are translated and produce new protein isoforms.* ***(D)*** *The majority of intron-retaining transcripts is degraded in the cytoplasm by nonsense-mediated decay. This leads to the reduction in target gene expression, which is why IR is considered a mechanism of post-transcriptional gene regulation.* ***(E)*** *Other forms of alternative splicing include exon skipping, alternative 3′ or 5′ splice site selection, and the mutual exclusive expression of exon pairs.*

IR is tissue-specific and aberrant in disease

Data suggest a tissue-specific regulation of IR leading to varying frequencies of IR observed between cell types. Methods to predict alternative splicing in a cellular or disease context have been developed, but have primarily focused on splice site mutations and their impact on splicing (Leung et al. 2014) (Xiong et al. 2015; Jaganathan et al. 2019; Baeza-Centurion et al. 2019).

Aberrant IR was found in multiple human diseases including diverse cancers (Perfetti et al. 2014; Lacroix et al. 2012; Dvinge and Bradley 2015). Often, somatic mutations are the cause for aberrant IR, resulting in mis-splicing and as a consequence in partial or complete IR. In cancer, IR-inducing somatic mutations often affect tumour suppressor genes (Jung et al. 2015).

In summary, an increasing number of studies have identified IR as a fundamental physiological process of gene regulation important in normal biology and disease. While advances in next-generation sequencing technologies have revealed the extent to which alternative splicing (including IR) enhances transcriptomic and proteomic complexity (Pan et al. 2008), consensus workflows or best practises for IR detection and quantification are currently lacking. In this chapter, after introducing experimental techniques used for alternative splicing and IR research, we provide an overview about currently available tools and statistical approaches for differential IR analyses and challenges associated with IR detection and quantification.

Experimental approaches for the investigation of intron retention

When we consider experimental techniques used in IR research, we have to differentiate between methods for IR identification and quantification, as well as methods to study the regulation and consequences of IR.

For the transcriptome-wide identification and quantification of IR, RNA sequencing is widely used, mostly as part of other whole transcriptome analyses, such as gene expression and alternative splicing. However, for the accurate identification of IR events optimized sample and library preparation, as well as sequencing protocols, are essential (Vanichkina et al. 2017). For an unbiased identification of IR events RNA samples have to be cleared from nascent RNA and DNA contamination, e.g. by DNAse treatment and poly-A enrichment protocols. For compartmental localization of IR, cellular fractionation protocols can be applied prior to RNA sequencing.

Using data from bulk short-read RNA sequencing experiments, we and others were able to unravel specific sequential and structural characteristics associated with retained introns and their host genes (Edwards et al. 2016; Wong et al. 2013; Schmitz et al. 2017). For short-read protocols stranded paired-end sequencing is the preferred method and a high sequencing depth is crucial (Vanichkina et al. 2017). The dependency of novel splice junction discovery on sequencing depth and thus the reliable detection of IR events is illustrated in Figure 2.

Long-read sequencing protocols, such as PacBio's Single- Molecule, Real-Time (SMRT) Sequencing or Oxford Nanopore sequencing, can be used to study whole transcript isoforms (Rhoads and Au 2015; Byrne et al. 2017; Wang et al. 2016). With the aim of sequencing full-length transcripts, these techniques provide the opportunity to identify single-molecule patterns of IR, such as mutually exclusive IR events, interdependent IR events, or IR switches.

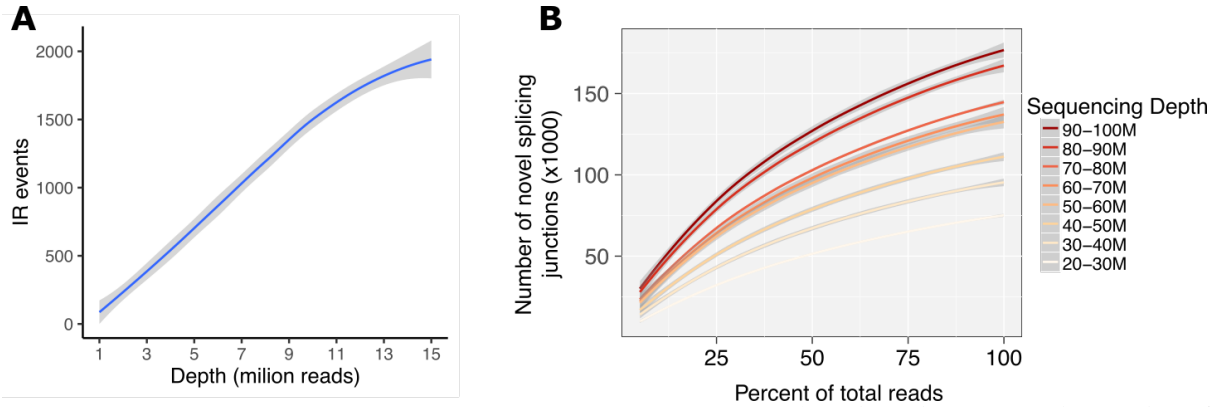

Figure 2 The dependence of alternative splicing discovery on sequencing depth. ***(A)*** *Subsampling RNA-seq data into bins of increasing depth illustrates how an increasing number of IR events is detected with greater depth.* ***(B)*** *Sufficient sequencing depth is crucial to perform alternative splicing analysis in general, where saturated RNA-seq data rediscovers most annotated splice junctions. Junction saturation of RNA-seq data can be determined using the junction_saturation.py module of the RSeQC RNA-seq quality control package (rseqc.sourceforge.net/). Sequencing depth is indicated in million reads.*

IR is a low-frequency transcription event and apart from a high sequencing depth (>80 million reads), adequate read coverage is essential. For that reason, single-cell RNA sequencing does not yet fulfil basic requirements and is thus at present not suitable for IR detection and quantification. Due to various constraints (such as budget or RNA concentration), conventional mRNA sequencing is sometimes as well not adequate to quantify IR with sufficient precision. Moreover, sequenced reads are typically quenched by transcripts from highly-expressed genes. A medium-throughput solution, for the accurate quantification of IR, would be RNA Capture sequencing (CaptureSeq) for a selected panel of IR events. CaptureSeq uses a custom panel of oligonucleotide probes designed to bind complementary sequences specific to transcripts of interest (Mercer et al. 2014). While not suitable for *de novo* identification of IR events, this technique enables a strong increase in sequencing depth of the targeted transcripts. Despite this advantage, CaptureSeq has not been used for this purpose to date. Instead, qRT-PCR is the most widely used method for IR validation and quantification.
In some situations, it is desired to also determine the cellular location of intron-retaining transcripts, which can be achieved by applying subcellular fractionation prior to RNA sequencing (Wong et al. 2013) or by using microscopy-based approaches, such as single-molecule RNA FISH (fluorescence in situ hybridization).

Bioinformatic approaches for IR identification and quantification

Custom computational workflows are essential for IR detection and quantification (Vanichkina et al. 2017). Like with any RNA sequencing data analysis, it is vital that established quality control software, such as fastqc (www.bioinformatics.babraham.ac.uk/projects/fastqc/), multiqc (multiqc.info)(Ewels et al. 2016), Piccard (broadinstitute.github.io/picard/), or RSeQC (rseqc.sourceforge.net), are used before any other data analysis is performed.
Given the quality control step confirmed positive attributes of the raw and mapped sequencing reads, the next data pre-processing steps are transcript identification and quantification. Only then further analysis steps can be applied, which in this case is the alternative splicing analysis or direct IR analysis. Best practices for pre-processing and analysis of RNA sequencing data have recently been summarized by Conesa and co-authors (Conesa et al. 2016).
The analysis of IR events in RNA sequencing data differs from the analysis of other alternative splicing events, such as exon skipping or alternative splice site selection. However, available alternative splicing analysis tools report IR as part of their analysis reports. An overview of some alternative splicing software is provided in Table 1.

Table 1 Overview of algorithms for the analysis of alternative splicing events.

Tool/Resource	Purpose/Method	Website	PMID
MISO (Mixture of Isoforms) (Differential)	gene isoform expression analysis; determines intronic percent spliced in (PSI) levels	genes.mit.edu/burgelab/miso	21057496
rMATS (Multivariate Analysis of Transcript Splicing)	Differential alternative splicing analysis	rnaseq-mats.sourceforge.net	25480548
spliceR	AS identification/quantification	bioconductor.org/packages/spliceR	24655717
Psichomics	Alternative splicing quantification and analysis	bioconductor.org/packages/psichomics	30277515
Whippet	Fast AS detection and quantification algorithm	github.com/timbitz/Whippet.jl	30220560
SUPPA2	Fast differential splicing analysis	github.com/comprna/SUPPA	29571299
MAJIQ	Detection and quantify of local splicing variations from RNA-Seq data	majiq.biociphers.org	29236961
VAST-TOOLS (Vertebrate Alternative Splicing and Transcription Tools)	Toolset for profiling and comparing alternative splicing events in RNA-Seq data	github.com/vastgroup/vast-tools	28855263

IR analysis software

IR is a special form of alternative splicing and its identification requires specific considerations. Only a handful of tools that are dedicated to IR identification and quantification incorporate these considerations into their algorithms (Middleton et al. 2017; Pimentel, Conboy, and Pachter 2015) (Oghabian, Greco, and Frilander 2018).

While most alternative splicing analysis tools follow a splice-junction or coverage-based approach for isoform identification and exon inclusion quantification, IR analysis requires a combination of both (Vanichkina et al. 2017). To date, three software tools have been developed specifically for IR detection and quantification (IRFinder, kma, and IntEREst; see Table 2 for details). These software tools have not been systematically benchmarked yet. However, in the following, we provide a brief overview and compare some of their key features.

Models of gene structure

While exonic sequences are well annotated for widely used model organisms the definition of introns remains fuzzy. The R package (Intron– Exon Retention Estimator, a.k.a. IntEREst) comes with a function for preparing a reference genome with defined gene structures (Oghabian, Greco, and Frilander 2018). The user can select to collapse all gene isoforms to avoid that intronic regions in one isoform belong to an exon of an alternative isoform. IRFinder too includes tools for preparing a custom reference genome (Middleton et al. 2017). All introns are derived from a given annotation file in GTF or GFF format and are defined as the regions between neighbouring exons in any transcript. To avoid false-positive predictions of IR events, IRFinder excludes regions within the intron that are covered by a non-intron feature (e.g.

miRNAs or snoRNAs). Moreover, the IRFinder output includes warnings indicative of overlapping isoforms or overlapping anti-sense genes and leaves it to the user to decide whether an IR event is real or not. kma determines intronic coordinates from a given genome reference (FASTA) and an annotation file, and, similar to IntEREst, excludes exonic regions from other isoforms (Oghabian, Greco, and Frilander 2018). However, kma adds a small region of the neighbouring exons to the intron coordinates to include reads spanning the intron-exon junctions for intron expression quantification (Pimentel, Conboy, and Pachter 2015).
Other studies used derivatives of the intron models proposed by the three established tools. For example, Ni et al. consolidated transcript isoforms from RefSeq annotations and considered only shared intronic and exonic regions in their gene models to determine IR levels. Genes that contain non-coding transcripts or overlap with another genes or antisense transcripts were removed from the analysis.

Metrics for IR quantification

Most IR detection algorithms report the fraction of intron retaining transcripts in all transcripts of the same gene or gene isoform. Hence, the emphasis is not on the intron abundance. IntEREst, however, uses Fragments Per Kilobase of transcript per Million mapped reads (FPKM) for intron expression quantification, which is normalized for intron length and the total number of introns in a gene (Oghabian, Greco, and Frilander 2018). Moreover, IntEREst determines the relative intron inclusion level, also known as percentage spliced-in metric (PSI or Ψ)(Katz et al. 2010), to quantify IR. Ψ is determined based on the number of reads mapped to introns divided by the number of reads spanning the intron (or mapping exons flanking the intron) (Oghabian, Greco, and Frilander 2018).

IRFinder uses a similar metric, which is called the IR-ratio:

$$IR_{ratio} = \frac{intronic\ abundance}{intronic\ abundance + exonic\ abundance} \quad (1)$$

The IR-ratio considers the abundance of the retained intron and its flanking exons. The exonic abundance refers to the number of read fragments spliced across the exon-exon junction. The intronic abundance is the median number of reads that map to an intron. As indicated before, IRFinder excludes overlapping features as well as the highest and lowest 30% of values from the intronic abundance. Moreover, exonic and intronic abundance are filtered for feature length (Middleton et al. 2017). Therefore, although Ψ and IR-ratio are similar measures determined by reads mapping to introns, across splice sites, and to the flanking exons, software-specific filtering criteria are applied leading to slightly varying IR quantification measures.

kma measures intronic abundance using either the transcripts per million (TPM) or FPKM metrics. For that, kma can be used with established transcript quantification tools such as Bowtie (Langmead and Salzberg 2012) or eXpress (Roberts and Pachter 2013). However, kma also provides Ψ as a readout, which is the ratio between intron expression and expression of the overlapping transcripts plus the intron expression (Pimentel, Conboy, and Pachter 2015).

Other names for very similar approaches have been used, such as the Percent Intron Retention metric (PIR; (Braunschweig et al. 2014) or the Intron Retention Index (IRI) proposed by (Ni et al. 2016), where IRI is the ratio of the read density of intronic regions and that of exonic regions shared by all transcript isoforms of the same gene. Ni et al. also determined the intron retention

percentage (IRP) as the fraction of all reads that map to a junction (i.e. across-junction + spliced) (Ni et al. 2016).

Table 2 Overview of IR detection/quantification algorithms.

Tool/Resource	Purpose/Method	Website	Reference
IRFinder	Detecting intron retention from RNA-Seq experiments	github.com/williamritchie/IRFinder	(Middleton et al. 2017)
Keep Me Around (kma)	R package for IR Detection	github.com/pachterlab/kma	(Pimentel, Conboy, and Pachter 2015)
IntEREst	IR quantification	github.com/gacatag/IntEREst	(Oghabian, Greco, and Frilander 2018)

In summary, the key question in most IR studies, as with most alternative splicing analyses, is about the proportion of transcripts that are affected by IR.

Challenges in the identification and quantification of IR events

A few confounders, i.e. transcriptional "noise" introduced by DNA contamination or unprocessed pre-mRNA transcripts, have to be considered in the analysis of IR events. IRFinder detects DNA contamination by computing the ratio of reads mapped to intergenic regions to the number of reads that mapped to coding regions (Middleton et al. 2017). In case, the ratio is above 10%, IRFinder emits a warning informing the user that the sample may not be suitable for IR detection.

It is important to enrich RNA libraries for polyadenylated RNA (mature mRNA) to minimize pre-mRNA contamination. Pre-mRNA contamination would inflate IR-ratios and by counting the number of reads that map to a list of non-polyadenylated genes (small nucleolar RNAs and histone genes) IRFinder can identify samples that were not poly-A enriched prior to RNA sequencing. Again, in this case, the user is informed through a warning message.

Another obstacle in IR quantification is low coverage or highly variable coverage in either the intronic or exonic regions or both. A reason for variable coverage could be repetitive sequences such as Long and Short Interspersed Nuclear Elements (LINEs and SINEs), DNA transposons, tandem and low complexity repeat sequences. kma removes introns with highly variable coverage using coverage filters (Pimentel, Conboy, and Pachter 2015), while IntEREst allows users to exclude repeat regions from the analysis (Oghabian, Greco, and Frilander 2018). However, it the user's responsibility to provide a table of repeat coordinates, which can, for example, be retrieved from the Dfam database of repetitive DNA families (dfam.org). IRFinder determines regions of poor unique mappability, which include repetitive sequences, and excludes these from the IR quantification (Middleton et al. 2017).

Statistical approaches for differential IR analysis

In many scientific scenarios, it is desirable to assess changes to IR pattern between two or more conditions. For example, we and others have determined differences in IR in hematopoietic cell differentiation (Edwards et al. 2016; Ni et al. 2016; Wong et al. 2013). Important insights were also gained by comparing IR pattern in tumours versus adjacent normal tissues (Dvinge and Bradley 2015).

For the analysis of differential IR multiple statistical approaches have been proposed. For example, IRFinder is equipped with the Audic and Claverie test (Audic and Claverie 1997), which is suitable for scenarios in which only one or two replicates per sample are available. This was very often the case when RNA sequencing was expensive and labs could not afford

to sequence multiple replicates. In its current version, IRFinder provides scripts that prepare the IRFinder output for the use with the R Bioconductor package DESeq2 (Love, Huber, and Anders 2014). DESeq2, normally used for differential gene expression analysis in RNA sequencing data, fits read counts to a negative binomial generalized linear model and employs Wald statistics or the likelihood ratio test to determine differential gene expression or in this case differential IR. IntEREst too uses functions from established digital gene expression analysis tools. Differential IR can be determined using either edgeR (Robinson, McCarthy, and Smyth 2010), DEXSeq (Anders, Reyes, and Huber 2012), or DESeq2 (Love, Huber, and Anders 2014).

The general assumption of most alternative splicing analysis tools is that splicing events, including IR, follow a binomial distribution, while the variability among replicates is considered to be normally distributed as well. However, some tools assume non-normally distributed intron inclusion levels and therefore use non-parametric tests, such as the Wilcoxon rank-sum, Kruskal–Wallis rank-sum, or Fligner–Killeen tests, to determine differences in mean intron inclusion levels between two condition.

Often thousands of introns are tested for differential retention. Thus, multiple testing correction is required to reduce the chance of false-positives (or Type 1 errors). Popular methods for multiple testing correction are the Benjamini-Hochberg, Holm–Bonferroni, and False Discovery Rate methods, however, none of the IR quantification tools provides multiple testing correction. Hence, the user has to make sure that multiple testing correction is applied.
An overview of statistical tests provided by different IR quantification tools is provided in Table 3.

Table 3 For the analysis of differential IR multiple statistical approaches have been proposed.

Software	Statistical test	Description
IRFinder	Audic and Claverie test	suitable for scenarios in which only one or two replicates per sample are available
IRFinder + DESeq2	Wald statistics or likelihood ratio test	fits read counts to a negative binomial generalized linear model and employs Wald statistics or the likelihood ratio test to determine differential gene expression or in this case differential IR
IntEREst	various	differential IR can be determined using either edgeR, DEXSeq, or DESeq2
rMATS	likelihood-ratio test	uses the binomial distribution for modelling the estimation uncertainty in individual replicates and the normal distribution for modelling the variability among replicates based on inclusion read counts, skipping read counts, and intron inclusion levels
psichomics	Wilcoxon rank-sum, Kruskal–Wallis rank-sum, Fligner–Killeen tests	assume non-normally distributed intron inclusion levels and therefore use non-parametric tests to determine differential IR

Experimental validation of IR events

Acceptable candidates for qRT-PCR validation should have at least raw read counts IR_{count} > 20 and consistent read coverage throughout the intron, while the flanking exons should also be well expressed ($exon_{count}$ >200). The next step is to generate cDNA from RNA extracted from the selected cell line/tissue. This step is crucial for IR validation as any DNA contamination would interfere with the detection of the mRNA-containing-intron as DNA can be used as a template for amplification. Therefore, effective DNAse treatment is essential to eliminate any

DNA contamination from the RNA extraction step. For cDNA synthesis, oligo(dT) is used for selectively reverse transcribe mature RNA transcripts containing retained introns. Finally, for qRT-PCR validation, two specific sets of primers are designed to validate IR events. One set of primers targets the flanking exons to determine the exonic expression of the intron-retaining gene. Ideally, one of the primers should anneal across the exon-exon boundary to make sure that the spliced variant is detected (without the intron). The second set of primers aims to detect the intronic expression of the retained intron. Similar to the first set of primers, one primer should anneal across the exon-intron boundary. Finally, to calculate the abundance of intronic expression over the flanking exons expression (or % of IR), the expression of the intron and exon are normalised first to the housekeeper gene ($2^{-\Delta Ct}$). Then the IR-ratio is computed (eq. 1).

Phylogenetic IR analyses

With the help of phylogenetic IR analyses, one can determine the evolutionary and functional conservation of IR events across multiple taxa and in different cell systems. Several studies have demonstrated the relevance of IR conservation, e.g. in the innate immunity (Braunschweig et al. 2014; Boutz, Bhutkar, and Sharp 2015; Wong et al. 2013). In a phylogenetic analysis of alternative splicing in 7 organs from 11 vertebrate species, Barbosa-Morais et al. found that transcriptomic complexity increased in species evolutionarily closer to primates (Barbosa-Morais et al. 2012). In this context, we have shown that the fraction of intron-retaining transcripts strongly anti-correlates with the number of protein-coding genes in vertebrate genomes, suggesting that IR compensates for the lack of transcriptomic complexity in species with fewer protein-coding genes (Schmitz et al. 2017). Moreover, we have shown that not just the characteristics of retained introns, such as their short length, high GC content, weak splice sites, etc. are strongly conserved, but also the characteristics of intron-retaining genes, such as their larger number of introns, longer 3′ untranslated regions, and bi-directional promoters (Schmitz et al. 2017).

However, there are a few obstacles to deal with when performing a phylogenetic IR analysis. These include, for example, the likely event that sequencing depths vary between samples from different species. Moreover, the quality and depth of genome annotations vary between model organisms and therefore conservation of IR in gene orthologs is difficult to assess.
Bias in the detection of IR event frequencies introduced by differences in annotation qualities can be avoided by generating *de novo* exon-intron structures from the same number of random reads for each sample (Barbosa-Morais et al. 2012). Stringent filtering based on coverage, depth, and read distribution can further reduce the risk of false intron retention calls due to mis-annotation, or insufficient precision due to lack of coverage (Braunschweig et al. 2014; Barbazuk, Fu, and McGinnis 2008).
Another factor that needs to be considered is that the number of IR events detected depends not just on the sequencing depth (Figure 2) but is also dependent on the number of transcripts per gene (Chen et al. 2014). Therefore, a method for transcript number normalization on a gene-by-gene basis is required in comparative analyses across taxa.
Intron sequences are poorly conserved (unlike exons). Low-conservation of most intron sequences has previously led to the conclusion that IR is merely transcriptional noise (e.g. due to errors in splicing) and has no functional implications (Mendell et al. 2004). It can, therefore, be difficult to determine orthologous introns for phylogenetic analyses of IR. However, the intron positions in the gene structure are often shared between species of the same lineage (Rogozin et al. 2003). Therefore orthologous introns could be considered as those occurring in the same position in orthologous genes or could be determined based on the orthology relationship or conservation of their flanking exons (Zhang and Edwards 2012).

Differences in the size and number of introns in different species require 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.
There are currently no tools available that could be employed for the analysis of lineage-specific IR events as well as their downstream effects and due to the relatively sparse data available for some model organisms, extrapolations are required to assess the evolutionary conservation of IR.

Functional consequences of IR can be gauged using functional enrichment analysis for the genes affected by IR, with resources such as GSEA (Subramanian et al. 2005), DAVID (Sherman et al. 2007), or PANTHER (Mi, Muruganujan, and Thomas 2013). All of these and others use similar statistical approaches and can be used with predefined genomic background data or customised backgrounds, e.g. based on expressed genes. Nevertheless, since the annotation qualities vary, conclusions from a cross-species comparison of cellular processes or pathways affected by IR have to be made with caution. Generally, functional enrichment analyses of alternative splicing events are strongly confounded by detectability, to which expression is the biggest contributor (Timmons, Szkop, and Gallagher 2015).

Analysis of IR regulation

The exact mechanisms that lead to IR events are not yet fully understood. However, several *cis*- and *trans*-regulatory elements that have an impact on IR are known (Monteuuis et al. 2019). Moreover, somatic mutations near splice sites are responsible for increased IR occurrences in multiple human cancers, often negatively affecting the expression of tumour suppressor genes (Jung et al. 2015). Differential expression of splicing factors and components of the nonsense-mediated decay pathway can also explain some of the aberrant IR patterns observed in human cancers (Dvinge and Bradley 2015).

Experimental approaches to find regulators of IR

For the discovery of regulators of IR multiple experimental and computational approaches are available. For example, advanced next-generation sequencing technologies provide opportunities to study intrinsic and extrinsic regulators of IR. Whole-genome or whole exome sequencing paired with RNA sequencing experiments can be used to identify genomic variants causing or inhibiting IR events (Jung et al. 2015; Maselli et al. 2014). Using whole-genome bisulfite sequencing, we found that reduced DNA methylation around splice sites and within the intron body can be favourable for IR (Wong 2017). Analysis of ChIP-seq data has shown that certain histone marks are enriched near splice sites of retained introns, suggesting an epigenetic mechanism of IR regulation (Braunschweig et al. 2014). Hence, to find genomic and epigenomic regulators of IR the same approaches as for gene regulation in general can be employed.
Moreover, based on transcriptomics data of splicing factor knockdown experiments (encodeproject.org) we identified *trans*-regulators causing a drastic increase in IR (Middleton et al. 2017). Another transcriptomic approach for the identification of *trans*-regulators of IR is RNA crosslinking immunoprecipitation sequencing used for example for RNA binding protein footprinting. Widely used derivatives of this technique are HITS-CLIP, PAR-CLIP, and iCLIP (Lagier-Tourenne et al. 2012; Bergeron et al. 2015). A growing resource of such data that can be mined to find *trans*-acting or epigenetic regulators of IR is the Encyclopedia of DNA Elements – ENCODE project (encodeproject.org).

Bioinformatics analysis of IR regulators

For the computational identification of regulators of IR custom workflows have to be implemented. For the analysis of intrinsic features of retained introns and their host transcripts one can take advantage of the many tools and code libraries that are available for the analysis of sequence composition, motif discovery, and structural characterisation of RNA molecules (e.g. bedtools - bedtools.readthedocs.io; BioPython - biopython.org, BioPerl - bioperl.org, Bioconductor - bioconductor.org), ViennaRNA Package - tbi.univie.ac.at/RNA). Our own analysis of intrinsic features of IR regulation revealed conserved characteristics, such as the shorter length, higher GC content, weaker splice sites of retained compared to non-retained introns (Schmitz et al. 2017). In this context, the maximum entropy model of short sequence motifs proposed by Yeo and Burge can be used to estimate the strengths of donor and acceptor sites (Yeo and Burge 2004). Intron-retaining genes were found to have longer 3′ UTR sequences, are enriched in bi-directional promoters, and have on average more introns than non-intron-retaining genes.

The analysis of epigenomic regulators of IR can be performed analogously to the analysis of epigenomic gene expression regulation. Methods for the analysis of epigenomics data including DNA methylation (e.g. WGBS)(Bock 2012), histone modifications (e.g. ChIP-seq)(Bailey et al. 2013), and chromatin structure (3C-based technologies, MNase-seq, DNase-seq, FAIRE-seq, ATAC-seq)(Chang et al. 2018) data have been critically reviewed before.

In studies that investigate the potential role of DNA methylation as a regulator of IR, methylation of CpG sites around 5′ and 3′ splice sites and the middle of an intron are assessed (Amit et al. 2012; Gelfman et al. 2013; Wong 2017; Gascard et al. 2015). The methylation signal (as percentage of methylated cytosines) is usually aggregated into non-overlapping bins or sliding windows and either parametric or non-parametric testing is performed to assess differential methylation between retained and non-retained introns. Whilst analysing DNA methylation as percentage is widely adopted and incorporated into a vast number of computational pipelines (Hansen, Langmead, and Irizarry 2012; Akalin et al. 2012; Dolzhenko and Smith 2014), normalised methylation fraction does not account for the potential inconsistency in sequencing depth of the different regions of the genome (Lea et al. 2017). To mitigate the potential coverage bias, differential methylation analysis can be performed on the raw count of methylated and unmethylated cytosines (that are calculated per sliding window) using an appropriate statistical model whose assumptions would satisfy the attributes of the data. Desirable characteristics of such a model would include the ability to deal with the correlated measures (as counts are aggregated into sliding windows) and to handle the unbalanced observations (which is especially relevant as the size of IR samples is usually smaller compared to samples of non-retained introns) and missing data (absence of a CpG site at the genomic loci of interest). One of the models that meets these parameters is the Binomial Generalised Linear Mixed Model (GLMM), where the raw counts of methylated and unmethylated cytosines for retained and non-retained introns are modelled through the logit link function and the hypothesis test is performed using a Wald test. The binomial GLMM procedure can be applied on Bismark output files, which contain the counts of methylated and unmethylated cytosines. In the absence of raw counts (majority of publicly available WGBS experiment datasets, including ENCODE, provide information on the methylation ratios and the read depth only), the binomial GLMM procedure can be carried out using the methylation ratio as a response variable and the read depth as the observational weight.

Most peak calling tools for ChIP-seq data report the location of the mapped reads in BED format, a tab-delimited text file format to represent genomic coordinates. The next step in the

analysis of histone marks as IR regulators is to overlay these coordinates onto the genome and identify corresponding genetic features. Discovery and annotation of the sequenced genome remain a major ongoing challenge in the post-human genome era. Fortunately, the annotated coordinates of genetic features such as introns and exons for a variety of species can be readily downloaded from resources such as the Ensembl consortium (ensembl.org/info/data/ftp). Specialised tools such as the Linux command line software *bedtools* and the R Bioconductor package *GenomicRanges* provide a range of utilities to efficiently intersect, merge and sort genomic intervals, which aid in constructing a 2x2 contingency table to chart the frequencies of ChIP-seq peaks against binary IR event outcome. Statistical methods for enrichment analysis including Fisher's exact test (for small sample size), Chi-square test or hypergeometric test (for large sample size) are commonly used to identify chromatin marks significantly associated with IR. Sampling procedures such as bootstrapping or subsampling procedures such as 'm out of n' bootstrapping should also be employed to increase statistical robustness of hypothesis testing.

For a holistic approach integrative "omics" analysis pipelines involving the above-mentioned methods should be applied. Methods for multi-omics data integration and associated challenges have been discussed in recent reviews (Gomez-Cabrero et al. 2014; Qin et al. 2016; Huang, Chaudhary, and Garmire 2017).

Modelling IR-mediated gene regulation

In the previous section, we have discussed methods used for the identification of IR regulators. Given that regulators of IR become known one could predict the occurrence of IR events. Indeed, multiple machine learning-based approaches have been developed to predict exon usage, however, IR prediction tools have not been developed to date. Barash et al. used a Bayesian neural network to decipher the "splicing code", which consists of hundreds of RNA sequence and structural features which can predict tissue-specific changes in exon usage (Barash et al. 2010). Later, the same group managed to improve prediction accuracy by applying a deep neural network (Leung et al. 2014). Other machine- and deep learning methods were developed to predict cryptic splicing as a result of somatic mutations (Xiong et al. 2015; Jaganathan et al. 2019; Baeza-Centurion et al. 2019).

IR enhances gene regulatory complexity through an increased sophistication in gene expression fine-tuning (Figure 3a/b) and also induces complexity on a molecular network level (i.e. gene regulatory networks, metabolic networks, signalling networks) by introducing dose-dependent nonlinear dynamics (Figure 3c/d). Premature termination codons within introns mediate nonsense-mediated decay of intron-retaining transcripts. Therefore, orthotopic IR may serve to regulate overexpressing genes towards levels that are desired for a particular phenotype (Figure 3a) or causes target repression towards ineffective levels (Figure 3b). Other downstream effects of IR in regulatory cascades other regulatory motifs are conceivable (Figure 3c/d).
To study the dynamics of IR, it's regulation and downstream consequences, a systems biology approach can be employed using either stochastic or deterministic modelling formalisms. Systems biology has been successfully implemented before, to study microRNA-mediated gene regulation, which is another form of post-transcriptional gene regulation (Schmitz et al. 2014; Lai et al. 2013; Lai et al. 2018).

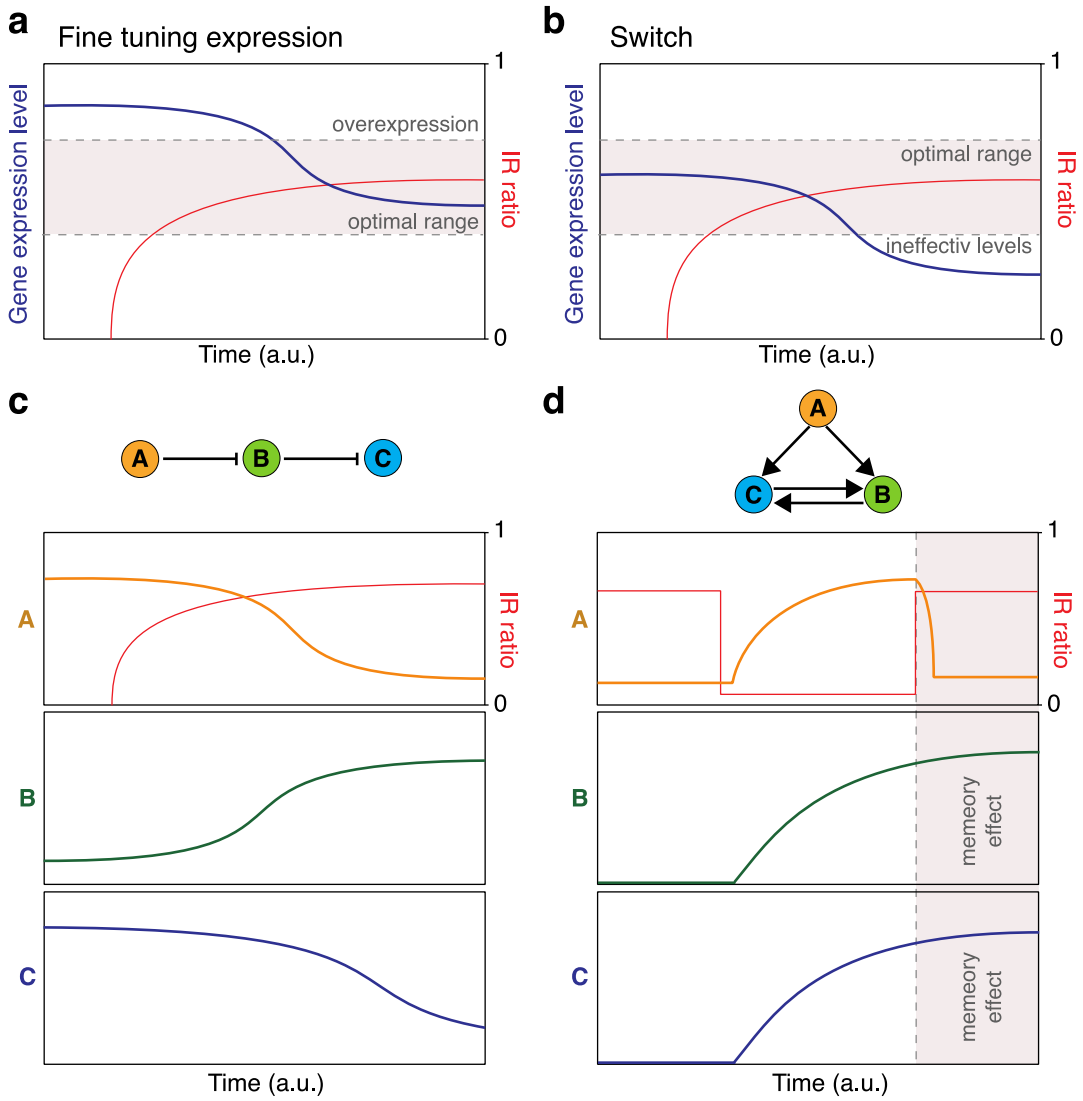

Figure 3 IR-mediated fine-tuning of gene expression and network dynamics. *The illustrations depict the possible consequences of IR-mediated gene regulation that add to the gene regulatory complexity of a cell. (a) Orthotopic IR may serve to regulate overexpressing genes towards levels that are desired for a particular phenotype. (b) In this scenario, orthotopic IR causes target repression towards ineffective levels (the target is switched off). (c) The effect of IR in a cascade of sequential repression. (d) In a gene regulating a double positive feedback loop, IR may induce a memory effect causing the loop to lock irreversibly into a steady-state (expression of B and C is activated). a.u. = arbitrary unit. This figure has been adopted from the Supplementary Materials of the article (Schmitz et al. 2017) which is published under the Creative Commons Attribution license (CC-BY).*

Conclusion

Introns have gained more attention recently and are now recognised as part of the complex gene regulatory network. IR, previously considered transcriptional noise, is in fact introducing additional transcriptomic complexity and variability in gene expression. IR is important in various stages of development, in cell differentiation, and diseases such as cancer. The sophistication of computational methods for IR identification and quantification is increasing constantly but there are still some challenges to overcome (Vanichkina et al. 2017). A major focus in the coming years of IR-related research lies on the integration of various experimental and computational approaches to facilitate comprehension of the complex regulation of IR and its intricate interplay with other forms of gene regulation. Systems biology will play an

important role when we aim to gain a mechanistic understanding of the regulation of IR and its consequences.

References

Akalin, A., M. Kormaksson, S. Li, F. E. Garrett-Bakelman, M. E. Figueroa, A. Melnick, and C. E. Mason. 2012. 'methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles', *Genome Biol*, 13: R87.

Amit, M., M. Donyo, D. Hollander, A. Goren, E. Kim, S. Gelfman, G. Lev-Maor, D. Burstein, S. Schwartz, B. Postolsky, T. Pupko, and G. Ast. 2012. 'Differential GC content between exons and introns establishes distinct strategies of splice-site recognition', *Cell Rep*, 1: 543-56.

Anders, S., A. Reyes, and W. Huber. 2012. 'Detecting differential usage of exons from RNA-seq data', *Genome Res*, 22: 2008-17.

Audic, S., and J. M. Claverie. 1997. 'The significance of digital gene expression profiles', *Genome Res*, 7: 986-95.

Baeza-Centurion, P., B. Minana, J. M. Schmiedel, J. Valcarcel, and B. Lehner. 2019. 'Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing', *Cell*, 176: 549-63 e23.

Bailey, T., P. Krajewski, I. Ladunga, C. Lefebvre, Q. Li, T. Liu, P. Madrigal, C. Taslim, and J. Zhang. 2013. 'Practical guidelines for the comprehensive analysis of ChIP-seq data', *PLoS Comput Biol*, 9: e1003326.

Barash, Y., J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe, and B. J. Frey. 2010. 'Deciphering the splicing code', *Nature*, 465: 53-9.

Barbazuk, W. B., Y. Fu, and K. M. McGinnis. 2008. 'Genome-wide analyses of alternative splicing in plants: opportunities and challenges', *Genome Res*, 18: 1381-92.

Barbosa-Morais, N. L., 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, and B. J. Blencowe. 2012. 'The evolutionary landscape of alternative splicing in vertebrate species', *Science*, 338: 1587-93.

Bergeron, D., G. Pal, Y. B. Beaulieu, B. Chabot, and F. Bachand. 2015. 'Regulated Intron Retention and Nuclear Pre-mRNA Decay Contribute to PABPN1 Autoregulation', *Mol Cell Biol*, 35: 2503-17.

Bock, C. 2012. 'Analysing and interpreting DNA methylation data', *Nat Rev Genet*, 13: 705-19.

Boutz, P. L., A. Bhutkar, and P. A. Sharp. 2015. 'Detained introns are a novel, widespread class of post-transcriptionally spliced introns', *Genes Dev*, 29: 63-80.

Braunschweig, U., N. L. Barbosa-Morais, Q. Pan, E. N. Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia, and B. J. Blencowe. 2014. 'Widespread intron retention in mammals functionally tunes transcriptomes', *Genome Res*, 24: 1774-86.

Buckley, P. T., M. T. Lee, J. Y. Sul, K. Y. Miyashiro, T. J. Bell, S. A. Fisher, J. Kim, and J. Eberwine. 2011. 'Cytoplasmic intron sequence-retaining transcripts can be dendritically targeted via ID element retrotransposons', *Neuron*, 69: 877-84.

Byrne, A., A. E. Beaudin, H. E. Olsen, M. Jain, C. Cole, T. Palmer, R. M. DuBois, E. C. Forsberg, M. Akeson, and C. Vollmers. 2017. 'Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells', *Nat Commun*, 8: 16027.

Chang, P., M. Gohain, M. R. Yen, and P. Y. Chen. 2018. 'Computational Methods for Assessing Chromatin Hierarchy', *Comput Struct Biotechnol J*, 16: 43-53.

Chen, L., S. J. Bush, J. M. Tovar-Corona, A. Castillo-Morales, and A. O. Urrutia. 2014. 'Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity', *Mol Biol Evol*, 31: 1402-13.

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szczesniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. 2016. 'A survey of best practices for RNA-seq data analysis', *Genome Biol*, 17: 13.

Dolzhenko, E., and A. D. Smith. 2014. 'Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments', *BMC Bioinformatics*, 15: 215.

Dvinge, H., and R. K. Bradley. 2015. 'Widespread intron retention diversifies most cancer transcriptomes', *Genome Med*, 7: 45.

Edwards, C. R., W. Ritchie, J. J. Wong, U. Schmitz, R. Middleton, X. An, N. Mohandas, J. E. Rasko, and G. A. Blobel. 2016. 'A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages', *Blood*.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. 'MultiQC: summarize analysis results for multiple tools and samples in a single report', *Bioinformatics*, 32: 3047-48.

Gascard, P., 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. Jones, M. A. Marra, T. D. Tlsty, J. F. Costello, and M. Hirst. 2015. 'Epigenetic and transcriptional determinants of the human breast', *Nat Commun*, 6: 6351.

Gelfman, S., N. Cohen, A. Yearim, and G. Ast. 2013. 'DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure', *Genome Res*, 23: 789-99.

Gomez-Cabrero, D., I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegner. 2014. 'Data integration in the era of omics: current and future challenges', *BMC Syst Biol*, 8 Suppl 2: I1.

Gontijo, A. M., V. Miguela, M. F. Whiting, R. C. Woodruff, and M. Dominguez. 2011. 'Intron retention in the Drosophila melanogaster Rieske Iron Sulphur Protein gene generated a new protein', *Nat Commun*, 2: 323.

Hansen, K. D., B. Langmead, and R. A. Irizarry. 2012. 'BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions', *Genome Biol*, 13: R83.

Hirose, T., and J. A. Steitz. 2001. '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: 12914-9.

Huang, Sijia, Kumardeep Chaudhary, and Lana X Garmire. 2017. 'More is better: recent progress in multi-omics data integration methods', *Frontiers in genetics*, 8: 84.

Jaganathan, K., S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, and K. K. Farh. 2019. 'Predicting Splicing from Primary Sequence with Deep Learning', *Cell*, 176: 535-48 e24.

Juneau, K., C. Palm, M. Miranda, and R. W. Davis. 2007. 'High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing', *Proc Natl Acad Sci U S A*, 104: 1522-7.

Jung, H., D. Lee, J. Lee, D. Park, Y. J. Kim, W. Y. Park, D. Hong, P. J. Park, and E. Lee. 2015. 'Intron retention is a widespread mechanism of tumor-suppressor inactivation', *Nat Genet*, 47: 1242-8.

Katz, Y., E. T. Wang, E. M. Airoldi, and C. B. Burge. 2010. 'Analysis and design of RNA sequencing experiments for identifying isoform regulation', *Nat Methods*, 7: 1009-15.

Kim, E., A. Magen, and G. Ast. 2007. 'Different levels of alternative splicing among eukaryotes', *Nucleic Acids Res*, 35: 125-31.

Kim, Y. K., and V. N. Kim. 2007. 'Processing of intronic microRNAs', *EMBO J*, 26: 775-83.

Lacroix, M., L. Lacaze-Buzy, L. Furio, E. Tron, M. Valari, G. Van der Wier, C. Bodemer, A. Bygum, A. C. Bursztejn, G. Gaitanis, M. Paradisi, A. Stratigos, L. Weibel, C. Deraison, and A. Hovnanian. 2012. 'Clinical expression and new SPINK5 splicing defects in Netherton syndrome: unmasking a frequent founder synonymous mutation and unconventional intronic mutations', *J Invest Dermatol*, 132: 575-82.

Lagier-Tourenne, C., M. Polymenidou, K. R. Hutt, A. Q. Vu, M. Baughn, S. C. Huelga, K. M. Clutario, S. C. Ling, T. Y. Liang, C. Mazur, E. Wancewicz, A. S. Kim, A. Watt, S. Freier, G. G. Hicks, J. P. Donohue, L. Shiue, C. F. Bennett, J. Ravits, D. W. Cleveland, and G. W. Yeo. 2012. 'Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs', *Nat Neurosci*, 15: 1488-97.

Lai, X., A. Bhattacharya, U. Schmitz, M. Kunz, J. Vera, and O. Wolkenhauer. 2013. 'A systems' biology approach to study microRNA-mediated gene regulatory networks', *Biomed Res Int*, 2013: 703849.

Lai, X., S. K. Gupta, U. Schmitz, S. Marquardt, S. Knoll, A. Spitschak, O. Wolkenhauer, B. M. Putzer, and J. Vera. 2018. 'MiR-205-5p and miR-342-3p cooperate in the repression of the E2F1 transcription factor in the context of anticancer chemotherapy resistance', *Theranostics*, 8: 1106-20.

Langmead, B., and S. L. Salzberg. 2012. 'Fast gapped-read alignment with Bowtie 2', *Nat Methods*, 9: 357-9.

Lareau, L. F., M. Inada, R. E. Green, J. C. Wengrod, and S. E. Brenner. 2007. 'Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements', *Nature*, 446: 926-9.

Lea, A. J., T. P. Vilgalys, P. A. P. Durst, and J. Tung. 2017. 'Maximizing ecological and evolutionary insight in bisulfite sequencing data sets', *Nat Ecol Evol*, 1: 1074-83.

Leung, M. K., H. Y. Xiong, L. J. Lee, and B. J. Frey. 2014. 'Deep learning of the tissue-regulated splicing code', *Bioinformatics*, 30: i121-9.

Love, M. I., W. Huber, and S. Anders. 2014. 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biol*, 15: 550.

Lunghi, M., F. Spano, A. Magini, C. Emiliani, V. B. Carruthers, and M. Di Cristina. 2016. 'Alternative splicing mechanisms orchestrating post-transcriptional gene expression: intron retention and the intron-rich genome of apicomplexan parasites', *Curr Genet*, 62: 31-8.

Marquez, Y., J. W. Brown, C. Simpson, A. Barta, and M. Kalyna. 2012. 'Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis', *Genome Res*, 22: 1184-95.

Maselli, R. A., J. Arredondo, J. Nguyen, M. Lara, F. Ng, M. Ngo, J. M. Pham, Q. Yi, J. M. Stajich, K. McDonald, M. A. Hauser, and R. L. Wollmann. 2014. 'Exome sequencing detection of two untranslated GFPT1 mutations in a family with limb-girdle myasthenia', *Clin Genet*, 85: 166-71.

Mauger, O., F. Lemoine, and P. Scheiffele. 2016. 'Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity', *Neuron*, 92: 1266-78.

McGuire, A. M., M. D. Pearson, D. E. Neafsey, and J. E. Galagan. 2008. 'Cross-kingdom patterns of alternative splicing and splice recognition', *Genome Biol*, 9: R50.

Mendell, J. T., N. A. Sharifi, J. L. Meyers, F. Martinez-Murillo, and H. C. Dietz. 2004. 'Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise', *Nat Genet*, 36: 1073-8.

Mercer, T. R., M. B. Clark, J. Crawford, M. E. Brunck, D. J. Gerhardt, R. J. Taft, L. K. Nielsen, M. E. Dinger, and J. S. Mattick. 2014. 'Targeted sequencing for gene discovery and quantification using RNA CaptureSeq', *Nat Protoc*, 9: 989-1009.

Merkin, J., C. Russell, P. Chen, and C. B. Burge. 2012. 'Evolutionary dynamics of gene and isoform regulation in Mammalian tissues', *Science*, 338: 1593-9.

Mi, H., A. Muruganujan, and P. D. Thomas. 2013. 'PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees', *Nucleic Acids Res*, 41: D377-86.

Middleton, Robert, Dadi Gao, Aubin Thomas, Babita Singh, Amy Au, Justin J-L Wong, Alexandra Bomane, Bertrand Cosson, Eduardo Eyras, John E. J. Rasko, and William Ritchie. 2017. 'IRFinder: assessing the impact of intron retention on mammalian gene expression', *Genome Biology*, 18: 51.

Monteuuis, G., J. J. L. Wong, C. G. Bailey, U. Schmitz, and J. E. J. Rasko. 2019. 'The changing paradigm of intron retention: regulation, ramifications and recipes', *Nucleic Acids Res*.

Naro, C., A. Jolly, S. Di Persio, P. Bielli, N. Setterblad, A. J. Alberdi, E. Vicini, R. Geremia, P. De la Grange, and C. Sette. 2017. 'An Orchestrated Intron Retention Program in Meiosis Controls Timely Usage of Transcripts during Germ Cell Differentiation', *Dev Cell*, 41: 82-93 e4.

Ner-Gaon, H., R. Halachmi, S. Savaldi-Goldstein, E. Rubin, R. Ophir, and R. Fluhr. 2004. 'Intron retention is a major phenomenon in alternative splicing in Arabidopsis', *Plant J*, 39: 877-85.

Ni, T., 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, and J. Zhu. 2016. 'Global intron retention mediated gene regulation during CD4+ T cell activation', *Nucleic Acids Res*.

Nilsen, T. W., and B. R. Graveley. 2010. 'Expansion of the eukaryotic proteome by alternative splicing', *Nature*, 463: 457-63.

Oghabian, A., D. Greco, and M. J. Frilander. 2018. 'IntEREst: intron-exon retention estimator', *BMC Bioinformatics*, 19: 130.

Pan, Q., O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. 2008. 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nat Genet*, 40: 1413-5.

Perfetti, A., S. Greco, P. Fasanaro, E. Bugiardini, R. Cardani, J. M. Garcia-Manteiga, M. Riba, D. Cittaro, E. Stupka, G. Meola, and F. Martelli. 2014. 'Genome wide identification of aberrant alternative splicing events in myotonic dystrophy type 2', *PLoS One*, 9: e93983.

Pimentel, H., M. Parra, S. L. Gee, N. Mohandas, L. Pachter, and J. G. Conboy. 2016. 'A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis', *Nucleic Acids Res*, 44: 838-51.

Pimentel, Harold, John G Conboy, and Lior Pachter. 2015. 'Keep me around: intron retention detection and analysis', *arXiv preprint arXiv:1510.00696*.

Qin, Jing, Bin Yan, Yaohua Hu, Panwen Wang, and Junwen Wang. 2016. 'Applications of integrative OMICs approaches to gene regulation studies', *Quantitative Biology*, 4: 283-301.

Rhoads, A., and K. F. Au. 2015. 'PacBio Sequencing and Its Applications', *Genomics Proteomics Bioinformatics*.

Roberts, A., and L. Pachter. 2013. 'Streaming fragment assignment for real-time analysis of sequencing experiments', *Nat Methods*, 10: 71-3.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26: 139-40.

Rogozin, I. B., Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, and E. V. Koonin. 2003. 'Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution', *Curr Biol*, 13: 1512-7.

Schmitz, U., X. Lai, F. Winter, O. Wolkenhauer, J. Vera, and S. K. Gupta. 2014. 'Cooperative gene regulation by microRNA pairs and their identification using a computational workflow', *Nucleic Acids Res*, 42: 7539-52.

Schmitz, U., N. Pinello, F. Jia, S. Alasmari, W. Ritchie, M. C. Keightley, S. Shini, G. J. Lieschke, J. J. Wong, and J. E. J. Rasko. 2017. 'Intron retention enhances gene regulatory complexity in vertebrates', *Genome Biol*, 18: 216.

Sebe-Pedros, A., M. Irimia, J. Del Campo, H. Parra-Acero, C. Russ, C. Nusbaum, B. J. Blencowe, and I. Ruiz-Trillo. 2013. 'Regulated aggregative multicellularity in a close unicellular relative of metazoa', *Elife*, 2: e01287.

Sherman, B. T., W. Huang da, Q. Tan, Y. Guo, S. Bour, D. Liu, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. 2007. 'DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis', *BMC Bioinformatics*, 8: 426.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. 2005. 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles', *Proc Natl Acad Sci U S A*, 102: 15545-50.

Timmons, J. A., K. J. Szkop, and I. J. Gallagher. 2015. 'Multiple sources of bias confound functional enrichment analysis of global -omics data', *Genome Biol*, 16: 186.

Vanichkina, D. P., U. Schmitz, J. J. Wong, and J. E. J. Rasko. 2017. 'Challenges in defining the role of intron retention in normal biology and disease', *Semin Cell Dev Biol*.

Wang, B., E. Tseng, M. Regulski, T. A. Clark, T. Hon, Y. Jiao, Z. Lu, A. Olson, J. C. Stein, and D. Ware. 2016. 'Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing', *Nat Commun*, 7: 11708.

Wong, A. C. H., J. E. J. Rasko, and J. J. Wong. 2018. 'We skip to work: alternative splicing in normal and malignant myelopoiesis', *Leukemia*, 32: 1081-93.

Wong, J. J., A. Y. Au, W. Ritchie, and J. E. Rasko. 2015. 'Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology', *Bioessays*, 38: 41-49.

Wong, J. J., W. Ritchie, O. A. Ebner, M. Selbach, J. W. Wong, Y. Huang, D. Gao, N. Pinello, M. Gonzalez, K. Baidya, A. Thoeng, T. L. Khoo, C. G. Bailey, J. Holst, and J. E. Rasko. 2013. 'Orchestrated intron retention regulates normal granulocyte differentiation', *Cell*, 154: 583-95.

Wong, J.J.-L.; Gao, D.; 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. 2017. 'Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment', *Nature Communications*, 8.

Xiong, H. Y., B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe, and B. J. Frey. 2015. 'RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease', *Science*, 347: 1254806.

Yeo, G., and C. B. Burge. 2004. 'Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals', *J Comput Biol*, 11: 377-94.

Zhang, G., G. Guo, X. Hu, Y. Zhang, Q. Li, R. Li, R. Zhuang, Z. Lu, Z. He, X. Fang, L. Chen, W. Tian, Y. Tao, K. Kristiansen, X. Zhang, S. Li, H. Yang, J. Wang, and J. Wang. 2010.

'Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome', *Genome Res*, 20: 646-54.

Zhang, Q., and S. V. Edwards. 2012. 'The evolution of intron size in amniotes: a role for powered flight?', *Genome Biol Evol*, 4: 1033-43.