Long-read transcriptomic profiling across human tissues and scales: from bulk to single-cell resolution

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

The University of East Anglia

Sofia Kudasheva

September 2025

Thesis Abstract

The advent of long-read sequencing technologies offers an unprecedented opportunity to systematically characterise transcript diversity, yet analytical frameworks for accurate isoform discovery and quantification remain underdeveloped.

This thesis leverages Oxford Nanopore long-read RNA sequencing to develop and benchmark computational approaches for isoform-level analysis, applying them to diverse biological contexts ranging from post-mortem human brain tissue to induced pluripotent stem cell (iPSC)-derived neuronal and cardiac models.

In Chapter 2, I employed targeted long-read CaptureSeq across multiple human brain regions and conditions to generate a comprehensive isoform catalogue. By benchmarking competing annotation pipelines and integrating orthogonal evidence, I identified novel isoforms of risk genes, some of which were differentially enriched between brain regions and in the brains of donors with psychiatric disease. The pipeline developed in this chapter was then tailored and applied to the datasets in chapters 3 and 4.

In Chapter 3, I adapted these workflows for single-cell long-read sequencing, enabling the construction of transcriptome references for iPSC-derived neurons, astrocytes, and microglia from a shared genotype. This work revealed cell–type–specific isoform regulation and splicing heterogeneity, and highlighted both the opportunities and limitations of single-cell long-read methods for resolving transcript diversity.

In Chapter 4, I investigated the role of the small nucleolar RNA *SNORD116* in cardiomyocyte differentiation using long-read RNA-seq of knockout and control iPSC-derived cardiomyocytes. This analysis uncovered differentiation stage-specific changes in splicing and alternative polyadenylation, pointing to putative targets of *SNORD116* regulation during cardiac development.

Together, these studies show that long-read sequencing is a powerful approach for uncovering isoform diversity and its functional relevance in the human brain and in stem cell–derived models of neuronal and cardiac differentiation. This work improves both the interpretability of ONT data and the robustness of downstream analyses. It also begins to establish methodological foundations for integrating long-read transcriptomics with orthogonal functional assays, moving towards a systematic characterisation of the molecular and phenotypic impact of specific RNA isoforms in health and disease.

Acknowledgements

I would like to begin by expressing my sincere gratitude to my supervisors, Dr Wilfried Haerty and Dr Iain Macaulay, for their guidance throughout my PhD. Their expertise and thoughtful feedback have shaped this thesis and my development as a researcher. A particular thank you to Wilfried for being an exceptionally present and supportive PI, always generous with his very limited time. To the members of the Haerty group, past and present: thank you for creating such a kind and collaborative environment. Your friendship, humour, and support made the PhD journey far more enjoyable. I am especially grateful to Dr David Wright for his tireless willingness to read drafts and provide meticulous feedback, and for always listening with patience when things felt overwhelming. I am indebted to collaborators, including Dr Nicola Hall, Dr Terri Lee-Holmes and Fran Cervilla Martinez, whose expertise and data were central to this project. I have learned so much from all of you. Thank you to Kamil, who repeatedly rescued me from the despair of failed software installations.

I would like to thank the Biotechnology and Biological Sciences Research Council (BBSRC) for funding this project, my training and conference attendance.

My deepest thanks go to my friends — Jess, Kate, Mia, Lu, Becky, and Insect George. Our work lunches at 12:30 were a daily highlight and a much-needed group therapy session. You reminded me that a PhD is not only about research, but also about the community of slightly nerdy peers you build throughout it who make the journey possible.

To my forever friends, Marusia and Margo: thank you for being my anchors, the people I look to when I need grounding, perspective, or simply the comfort of unwavering friendship. I thank my brother, Max, for his help in speeding up the otherwise tedious process of referencing. I am grateful to Jack, who has believed in me, cheered me on, and reminded me that life exists outside the PhD. Finally, I am forever indebted to my parents, who instilled in me a love of learning and have always supported my education.

Contents

List of Figures

List of Tables

List of Abbreviations

APA alternative polyadenylation.

AS alternative splicing.

ASD Autism Spectrum Disorder.

BD Bipolar Disorder.

Capture-Seq targeted RNA capture sequencing.

CB Cell Barcode.

CM cardiomyocyte.

DE differentially expressed.

DLPFC dorsolateral prefrontal cortex.

DTE differential transcript expression.

DTU differential transcript usage.

FSM Full-Splice Match.

GWAS genome-wide association study.

iPSC induced pluripotent stem cell.

IR intron retention.

ISM Incomplete-Splice Match.

lncRNA long non-coding RNA.

MDD major depressive disorder.

mRNA messenger RNA.

MS mass spectrometry.

NAS nonsense-associated altered splicing.

NIC Novel In Catalog.

NMD nonsense-mediated decay.

NNC Novel Not in Catalog.

ONT Oxford Nanopore Technologies.

OPC oligodendrocyte progenitor cell.

ORF open reading frame.

PacBio Pacific Biosciences.

PAS polyadenylation site.

PCR Polymerase Chain Reaction.

PTC premature termination codon.

PWS Prader-Willi syndrome.

QTL quantitative trait loci.

RBP RNA-binding protein.

RNA-seq RNA sequencing.

SCZ Schizophrenia.

snoRNA small nucleolar RNA.

TSO Template Switching Oligo.

TSS Transcription Start Site.

TTS Transcription Termination Site.

UMI Unique Molecular Identifier.

UTR untranslated region.

VGCC voltage-gated Ca2+ channel.

Chapter 1

General Introduction

Alternative splicing (AS) is a fundamental mechanism of post-transcriptional gene regulation that expands the coding capacity of the human genome. Through the selective use of alternative splice sites, a single gene can give rise to multiple tissue- and developmental stage-specific transcript isoforms. It is now estimated that over 95% of multi-exon human genes undergo alternative splicing, with an average of more than three distinct transcripts produced per locus (Pan et al., 2008). AS can have significant effects on coding potential and resulting protein products, such as altering their localisation, post-translational modifications, protein-protein interactions, and binding affinities (Baralle and Giudice, 2017).

1.1 Splicing Mechanism

A defining feature of eukaryotic protein-coding genes is that, unlike prokaryotic genes, they generally contain longer complex non-coding regions, such as the untranslated region (UTR) and introns interspersed between coding exons. For a mature messenger RNA (mRNA) to form, introns have to be removed from the newly transcribed precursor messenger RNA (pre-mRNA), via a process known as splicing. RNA splicing is mostly co-transcriptional and is coupled to other pre-mRNA processing events such as 5' end capping, 3' end cleavage and polyadenylation (Herzel et al., 2017). Introns were once considered "junk DNA" (Wong et al., 2000). However, subsequent work has shown that they can give rise to functional non-coding RNAs and "serve as repositories of *cis* elements", serving as (Chorev and Carmel, 2012; Shaul, 2017).

In humans, the vast majority (>99.5%) of introns are excised by the major spliceosome, a large ribonucleoprotein (RNP) complex that catalyses intron removal and exon ligation (Jurica and Roybal, 2013). It comprises five small nuclear RNPs (U1, U2, U4, U5, and U6), and a large number of auxiliary proteins. Conserved cis-acting elements known as the core splicing signals: short 5' donor and 3' acceptor splice sites (SS), branch point sequence and polypyrimidine tract, guide the interactions between spliceosomal components and the nascent pre-mRNA (Singh, 2002). The GU and AG dinucleotides, which are contained within the splice sites, define the 5' and 3' intron boundaries, respectively.

Spliceosome assembly proceeds through a series of well-defined steps. In vertebrate cells, where exons are short and introns are long, the splicing machinery initially assembles across an exon rather

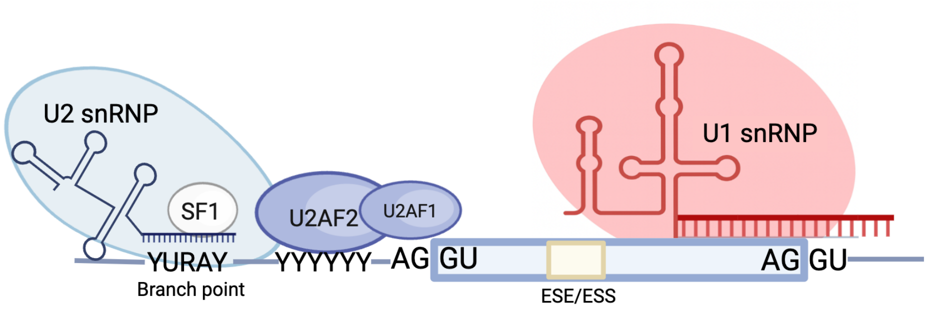


Figure 1.1: **Splice site recognition.** Interactions between the branch point and 5' splice site (5'ss)) on pre-mRNA and small nuclear RNA (snRNA) components of small nuclear ribonucleoprotein (snRNP) complexes (U2 and U1, respectively) and U2 auxiliary factors U2AF2, which binds to the polypyrimidine tract, and U2AF1, which binds to the 3' splice site (3'ss) AG. Created with Biorender.com, adapted from Rogalska et al. (2022)

than directly across the intron. The process begins with formation of the E (early) complex, in which U1 snRNP binds to the 5'SS GU sequence, splicing factor 1 (SF1) binds to the branch point sequence, and U2 auxiliary factor 2 (U2AF2) binds to the polypyrimidine tract and U2AF1 to the 3'SS AG dinucleotide (Figure 1.1, Wang et al. (2015); Chen and Manley (2009)). This is followed by the formation of the A complex (pre-spliceosome), in which U2 snRNP replaces SF1 at the branch point in an ATP-dependent step, forming stable base pairing with the branch point adenosine. The branch point then becomes covalently linked to the 5' end of the intron, generating a branched intron, or lariat. Finally, during the last catalytic step, the free 3' hydroxyl of the 5' exon attacks the 3' splice site, ligating the two exons and releasing the intron lariat, which is subsequently de-branched and degraded.

1.1.1 Non-canonical splicing

Most introns (91% in human and 92% in mouse) follow the canonical GU donor and AG acceptor dinucleotide rule (GT–AG in DNA, Irimia and Roy (2008)). However, a small proportion of splice junctions deviates from this consensus. For example, approximately 1% of annotated introns contain non-canonical GC–AG splice site pairs, in which the T→C substitution at the 5' donor site creates a mismatch with U1 snRNA, making these sites 'weaker' than their GT–AG counterparts (Kralovicova et al. (2011)).

A distinct subset of non-canonical introns belongs to the U12-dependent (minor-class) family, which is defined by extended and highly conserved consensus sequences at both the 5' splice site and the branch point (Sibley et al. (2016)). The splicing mechanism for these introns is nearly identical to that of the major class, but the minor spliceosome is assembled around four distinct snRNAs — U11, U12, U4atac, and U6atac (Olthof et al. (2019); Montañés-Agudo et al. (2021)). Most U12-type introns conform to an AT–AC or GT–AG dinucleotide combination, although rarer variants such as GT–CT also occur.

Comparative genomics indicates that the U12-type spliceosome co-evolved with its target introns, with multiple independent loss events across eukaryotic lineages but strict conservation in metazoans and land plants. In these groups, minor intron positions are more highly conserved than those of major introns, suggesting strong selective pressure to retain precise regulation of minor intron-containing genes despite the metabolic cost of maintaining two splicing machineries. Those genes are enriched in genes encoding voltage-gated Na+ channel (VGSC) and voltage-gated Ca2+ channel (VGCC) $\alpha1$ subunit gene families, and inhibition of minor intron splicing via U6atac knockdown in rat cardiomyocytes reduces Scn5a and Cacna1c protein expression (Montañés-Agudo et al., 2021). In humans, partial loss of function due to mutations in *RNU4ATAC*, encoding the U4atac snRNA, underlies developmental disorders such as Roifman syndrome and Lowry–Wood syndrome, in which the severity of growth defects and brain hypoplasia correlates with the extent of minor spliceosome impairment (Baumgartner et al., 2019; Farach et al., 2018; Merico et al., 2015). These findings demonstrate that the activity level of the minor spliceosome has direct consequences for organ and tissue development, likely mediated through the regulation of progenitor cell populations. Although most current isoform detection methods are optimised for canonical splice junctions, as they account for 91% of all splicing events, rare non-canonical events remain important to investigate, and tailored approaches will be needed to fully assess their roles in transcript stability and novel protein isoform production.

1.2 Alternative splicing

During constitutive splicing, introns are removed, and exons are ligated in the order in which they appear in a gene. Alternative splicing (AS), on the other hand, is a highly regulated process where different pairs of splice sites are selected by the splicing machinery to produce multiple distinct RNA and protein isoforms from a single gene (Mathur et al., 2019). AS can also lead to changes in protein subcellular localisation, protein post-translational modifications, or protein binding affinities to ligands (Zeng and Hamada, 2020). There are multiple mechanisms for the formation of alternative splice variants from the same genomic locus, including cassette exon skipping; alternative 3' SS; alternative 5' SS; and intron retention (IR)(1.2B). Approximately 10% of cassette exons – exons that are either included or skipped from the transcript – exist in pairs or groups in which only one of the two exons or groups is included in the mRNA. These mutually exclusive exons (MXEs) are highly homologous and are believed to have originated from exon duplications (Pohl et al., 2013). Although MXEs maintain the size and general structure of the protein, they are usually not functionally redundant, and their inclusion is tightly regulated. This is supported by instances where mutations in MXEs have been shown to cause diseases such as a missense mutation in the CACNA1C gene resulting in Timothy Syndrome (Splawski et al., 2004).

Transcriptomic and proteomic studies indicate that under 5% of human protein-coding genes consist of a single exon and 95% of the remaining multi-exon genes undergo AS, with the brain exhibiting the most diverse repertoire of splice variants (Pan et al., 2008; Wang et al., 2008). Long non-coding RNAs (lncRNAs) also undergo alternative splicing and are enriched for the 'weaker' GC-AG junctions, which are more prone to AS (Abou Alezz et al., 2020). Although weaker splicing signals and shorter exon length are reported to be predictors of whether an exon would be alternatively spliced (Zheng et al., 2005), splice site selection is controlled by many separate components, including non-spliceosomal splicing regulatory factors (Matera and Wang, 2014), RNA secondary structures (McManus and Graveley, 2011), RNA polymerase elongation speed (Fong et al., 2015), and epigenetic regulation (Luco et al., 2010). Trans-acting splicing factors bind to enhancer or silencer motifs close to splice sites to promote or prevent the usage of a particular splice site, and fall into two main conserved families, serine-arginine rich (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs) (Blencowe, 2000). However, the function of these proteins may vary considerably, depending on binding position and co-activity with other splicing regulators (1.2A). SR proteins generally bind to exonic splicing enhancers (ESEs) through their RNA-recognition motifs (RRMs) and recruit splicing machinery to the new RNA transcript, whilst the hnRNPs often act as antagonists to SR-protein-regulated

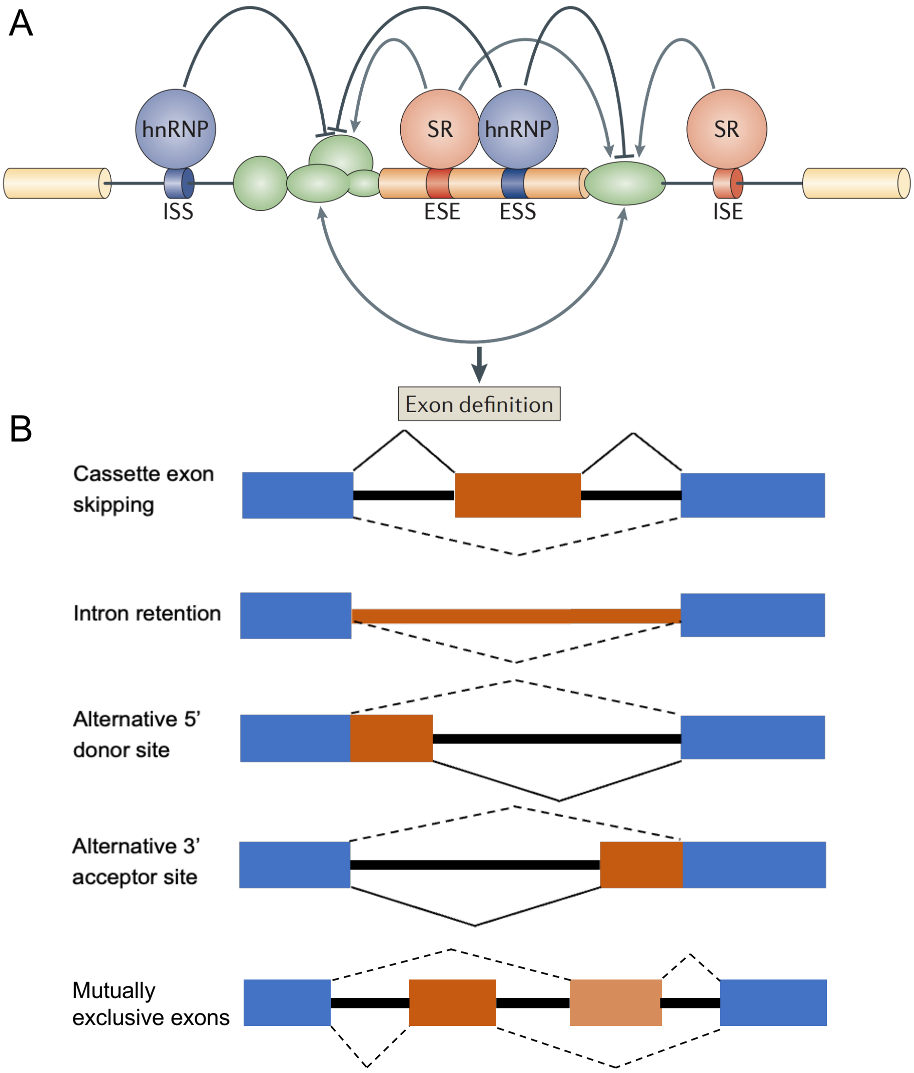


Figure 1.2: **Regulation and types of alternative splicing.** (A) An overview of the regulation of exon definition during splicing. Ser/Arg-rich (SR) proteins bind to exonic splicing enhancers (ESEs) and intronic splicing enhancers (ISEs) to stimulate the binding of the spliceosome components (shown in green) to the splice site. hnRNPs bind exonic and intronic splicing silencers (ESSs and ISSs) and prevent binding of the snRNPs. Taken from (Kornblihtt et al., 2013); (B) Five main types of AS events. Blue boxes, flanking exons; orange, alternatively spliced regions; solid lines, splice junctions supporting the inclusion isoform; dotted lines, splice junctions supporting the exclusion isoforms.

alternative splicing events, by binding to exon splicing silencers (ESSs) and inhibiting the inclusion of exons (Wang et al., 2015; Dvinge, 2018). hnRNPs can inhibit splicing by directly competing with SR activators for overlapping binding sites (Nazim et al., 2017). Alternatively, binding of hnRNP I, known as polypyrimidine-tract-binding protein (PTB), or hnRNP A1, to several sites flanking a silenced exon can cause the pre-mRNA to loop out, making it unavailable for splicing (Lamichhane Rajan et al., 2010). Data from RNASeq and cross-linking and immunoprecipitation (CLiP) experiments have been combined to create RNA splicing maps, which identify the splice factor binding motifs and their location in target genes (D. Ray et al. 2013; Van Nostrand et al. 2020). These studies revealed that most splicing factors recognise short and degenerate sequence motifs which occur frequently in pre-mRNAs. For example, neuron-specific NOVA (neuro-oncological ventral antigen) family regulators bind to clusters of YCAY (Y = C/T) (Ule et al., 2006) and SRSF1 preferentially binds to GGAGA consensus sequences within exonic regions (Pandit et al., 2013). Many RBPs utilise their modular structures to bind RNA with multiple RNA-recognition motifs (RRM) and thus improve their binding specificity. Although primarily known to be regulators of splice-site selection, SR protein family members have also been implicated in all crucial steps of mRNA processing. Aberrant expression and mutations in binding sites of these splicing factors have been associated with cancer and neurological disorders (Naro and Sette, 2013; Zhang et al., 2021). For example, the prototypical SR splice factor SRSF1 (previously known as SF2/ASF), plays roles in nonsense-mediated decay (NMD), mRNA export, translation and miRNA processing (Das and Krainer, 2014) and is a known proto-oncogene frequently overexpressed in multiple cancers (Anczuków et al., 2015). SR and hnRNP gene expression is therefore tightly controlled by autoregulatory feedback loops, where splicing factors bind their own pre-mRNAs and promote unproductive splicing events, such as IR or inclusion of alternative exons containing premature termination codon (PTC). PTC-containing transcripts are exported to the cytoplasm and targeted to the NMD pathway (Lareau et al., 2007; Ni et al., 2007).

1.2.1 Alternative splicing and nonsense-mediated decay

NMD was originally identified as an mRNA surveillance mechanism responsible for degradation of aberrant transcripts, arising from mutations or splicing errors, thus preventing their translation into C-terminal truncated (and potentially toxic) proteins (Maquat, 1995). Two principal models have been proposed for how PTCs are discriminated from a normal termination codon in mammalian cells. The exon–junction complex (EJC) model suggests that a stop codon located ∼50 nucleotides upstream of an exon–exon junction is marked as premature due to the presence of an EJC downstream. In contrast, the EJC-independent model posits that unusually long 3' untranslated regions (UTRs) downstream of a stop codon act as a signal to trigger NMD (Bühler et al., 2006; Hogg and Goff, 2010). In both cases, UPF proteins act as central effectors by initiating RNA deadenylation, decapping, and subsequent degradation.

More recent studies have revealed that NMD is not limited to aberrant transcripts but also serves as a regulated pathway that modulates the abundance of physiological isoforms during development and homeostasis. For example, many splicing factors, such as SR proteins and hnRNPs, autoregulate their expression by producing isoforms with 'poison exons' that contain PTCs and are degraded by NMD (Goetz and Wilkinson, 2017). In this way, AS coupled to NMD (AS–NMD) acts not only to diversify protein isoforms but also as a key post-transcriptional mechanism controlling gene expression levels (Fair et al., 2024).

A particularly important form of AS–NMD involves intron retention (IR). Once regarded as transcriptional noise in mammals, IR is now recognised as a widespread regulatory strategy for tuning transcript levels during differentiation (Yap et al., 2012; Pimentel et al., 2016). IR often introduces PTCs, targeting transcripts for NMD, however depending on the location of the IR event within the transcript and whether it disrupts the ORF, it can also lead to transcript stabilisation or nuclear retention and degradation via the exosome complex (Jacob and Smith, 2017; Yap et al., 2012). In mice, dietary restriction led to an increase in PTC-introducing IR in genes enriched not only in mRNA splicing, but also metabolism and innate immunity pathways (Tabrez et al., 2017). In general, proliferative precursor cells exhibit lower levels of IR, whereas differentiating cells display increased IR, consistent with its role in down-regulating non-physiological isoforms. An exception is terminal erythropoiesis, which features a dynamic IR programme where some IR clusters show a progressive increase before the cells reach a terminal erythrocytic stage that has lower IR compared to the precursor cells (Pimentel et al., 2016; Edwards et al., 2016). Similar to cassette exons, retained introns are shorter than average introns and associated with weaker splice sites and silencer elements. It is therefore expected that IR events are co-regulated by activity of various proteins, including RNA polymerase II, epigenetic regulators and specific RNA-binding protein (RBP)s such as PTBP1, hnRNPLL (Cho et al., 2014; Wong et al., 2017). Some alternatively spliced introns, referred to as exitrons, are present entirely within regions annotated as exons. Exitrons are rarely spliced under normal conditions and usually lack PTCs, but when removed, they can alter protein structure or induce frameshifts that generate premature stop codons (Marquez et al., 2015). Together, these examples highlight how AS and NMD are functionally coupled, with outcomes ranging from transcript degradation to protein diversification, depending on the context of the splicing event.

1.2.2 "Noisy splicing" or regulatory isoforms

Accurate identification and quantification of novel isoforms might assist the design of targeted therapies or identify AS events that can act as disease biomarkers. Therefore, efforts to discover novel isoforms continue. High-throughput RNA-seq routinely identifies thousands of novel alternatively spliced transcripts, but their functional relevance is debated. The "noisy splicing" model argues that many of these isoforms represent errors of the splicing machinery rather than bona fide functional transcripts (Pickrell et al., 2010). Supporting this view, large-scale proteomics and ribosome profiling studies suggest that only a subset (approximately 40–50%) of annotated alternative isoforms are detectably translated (Saudemont et al., 2017; Tress et al., 2017; Reixachs-Solé et al., 2020). Furthermore, both RNA-seq and mass spectrometry indicate that most genes exhibit a single dominant isoform expressed at significantly higher levels than others, raising the possibility that low-abundance isoforms often reflect biological or experimental noise (Gonzàlez-Porta et al., 2013; Ezkurdia et al., 2015).

This perspective, however, is contested. Critics emphasise the limited sensitivity and coverage of MS-based methods, particularly in complex or heterogeneous tissues, and argue that the absence of proteomic evidence does not equate to non-functionality (Blencowe, 2017). Analyses of the human interactome indicate that most isoform pairs share fewer than half of their protein–protein interaction partners (Yang et al., 2016; Jaffe et al., 2018). Moreover, as discussed above, even isoforms not translated into stable proteins may play important regulatory roles, for example, by modulating mRNA stability, competing for RNA-binding proteins, or regulating the abundance of coding isoforms through NMD. Recent work also highlights the breadth of AS–NMD beyond splicing factors, extending to genes involved in chromatin regulation and neuronal differentiation (Karousis et al., 2021; Zhuravskaya et al., 2024). In addition, nonsense-associated altered splicing (NAS) provides a complementary mechanism whereby recognition of a PTC by NMD factors can influence splice site choice and remove a PTC-containing exon or alter the reading frame, thereby "rescuing" transcript function and creating a feedback loop that shifts isoform production towards more functional variants (Wang et al., 2002; Abrahams et al., 2021).

Taken together, these findings suggest that transcript function cannot be defined solely by protein-coding potential or relative abundance. Non-coding or unproductive isoforms may exert substantial influence through post-transcriptional regulatory mechanisms, including modulation of mRNA stability, translation, and competition for RNA-binding proteins. Restricting analyses to the major protein-coding isoforms overlooks a potentially vast and functionally important layer of gene regulation.

1.2.3 Alternative splicing in the developing nervous system

The biological significance of alternative splicing has become more evident since the discovery of tissue-specific events and splicing factors that regulate them. Brain tissues display especially diverse developmentally dynamic and cell-type specific AS. Changes in the concentration of neuron-specific Nova, PTBP1/2 and RBFOX splicing factors control programs of isoform switches in genes essential for multiple aspects of neurodevelopment, including neurogenesis, synaptogenesis, cellular migration and axon guidance (Makeyev et al., 2007; Ule et al., 2005). Recently developed spatially resolved single cell methods have revealed brain region- and cell type-specific isoform and splice factor regulation (Joglekar et al., 2021; Feng et al., 2021).

Nova proteins, Nova-1 and Nova-2, were the first mammalian neuron-specific splicing regulators identified (Buckanovich et al., 1993; Yang et al., 1998). In humans, *NOVA1* is enriched in the spinal cord and cerebellum, particularly at postnatal stages, whereas *NOVA2* is more highly expressed in the cortex (Meldolesi, 2020). Both proteins recognise the YCAY motif (where Y is a pyrimidine, typically U or C), whose conservation is strongly predictive of brain-specific splicing patterns. Depending on binding position, Nova proteins can either promote or repress exon inclusion: binding within exons favours skipping, whereas binding to intronic YCAY clusters downstream of a regulated exon promotes its inclusion (Zhang et al., 2010). Through these mechanisms, Nova proteins regulate transcripts important for synaptogenesis, including neurotransmitter receptors (Park et al., 2011a), and also autoregulate their own expression by modulating exon usage (Dredge et al., 2005). The functional importance of Nova has been underscored by experimental replacement of human *NOVA1* with a Neanderthal variant, which carries a single-nucleotide substitution and reduces synaptic protein expression in cortical organoids (Trujillo et al., 2021).

RBFOX and PTBP proteins, although not strictly neuronal, are crucial regulators of the transition from neural progenitors to mature neurons. The dynamic interplay between these RBPs underpins precise temporal control of splicing switches during differentiation. RBFOX proteins recognise a long, well-defined (U)GCAUG motif, which they bind within intronic enhancer elements downstream from alternatively spliced exons to promote their inclusion (Jin et al., 2003). In contrast, PTBP1 binding upstream has been shown to repress inclusion of certain neural-regulated microexons (Quesnel-Vallières et al., 2015; Li et al., 2015). PTBP1 is largely absent from mature neurons but highly expressed in neural stem cells, where it suppresses inclusion of poison exon 10 in its neuronal paralog PTBP2, introducing a premature stop codon and leading to NMD (Boutz et al., 2007; Spellman et al., 2007). As neuronal differentiation progresses, PTBP1 is downregulated by miR-124, relieving repression of *PTBP2* and enabling its accumulation (Makeyev et al., 2007; Yeom et al., 2018). Recent work has extended this model by showing that AS-NMD acts more broadly to shape neuronal transcriptomes. In mouse pluripotent stem cell–derived neurons, AS-NMD was found to silence non-neuronal gene programmes rather than merely fine-tuning transcript abundance (Zhuravskaya et al., 2024). These events were strongly enriched for PTBP1-regulated cassette exons. In progenitors, PTBP1 represses the inclusion of poison exons, stabilising expression of their host genes. As PTBP1 levels decline, these exons are incorporated, triggering NMD and driving downregulation of transcripts such as *Fmnl3*, *Iqgap1*, and *Ripk1*.

This regulatory handover is reinforced by RBFOX2, which increases in abundance in differentiated neurons. RBFOX2 promotes inclusion of neuron-specific exons and stabilises the mature neuronal transcriptome, while also cross-regulating PTBP2 via AS-NMD (Jangi et al., 2014). The timing of its activity is critical: premature Rbfox2 expression in the embryonic mouse neocortex disrupts PTBP-regulated programmes, causing early inclusion of PTBP2-repressed exons in pathways such as Reelin signalling and leading to impaired neuronal migration and delayed differentiation (Weißbach et al., 2024).

Neuronal splice factors orchestrate extensive transcriptomic remodelling during development, enabling the unparalleled isoform diversity characteristic of the central nervous system. This regulation is exemplified by the neurexin family, a paradigm for AS–driven molecular diversity in synaptogenesis. The three neurexin genes (*NRXN1–3*) encode presynaptic cell adhesion molecules that interact with an array of postsynaptic partners to shape synaptic specificity. Alternative splicing at six canonical splice sites (SS1–SS6) generates extraordinary combinatorial diversity, with more than 3,000 predicted protein isoforms (Nguyen et al. (2016); Treutlein et al. (2014)). Among these, exon 20 of *NRXN1* (SS4) is the most extensively studied due to its central role in modulating postsynaptic binding affinities. This exon encodes a peptide overlapping the sixth LNS (for Laminin, Neurexin and Sex hormone-binding globulin) domain. Inclusion or exclusion of SS4 alters the surface chemistry of the LNS domain, thereby shifting binding specificity from neuroligins (SS4-) to cerebellin–GluD complexes (SS4+).

1.2.4 Microexons

One class of cassette exons highly enriched in neurons are the extremely short (3-27 nt) microexons. Microexons represent a notable exception to the 50–250 nt size constraints that normally define efficient exon definition (Figure 1.1, De Conti et al. (2013)). Unlike AS events that add or remove entire functional protein domains, microexons generally preserve the open-reading frame and are thought of as modulators that "fine-tune" protein activity and protein–protein interaction and are enriched in intrinsically disordered regions. Although neural microexons are regulated by the standard neural splicing programme involving RBFOX and PTBP1 proteins, they are generally too short to contain exonic splicing enhancers and thus for their inclusion to be regulated solely by conventional spliceosomal interaction mechanisms (Li et al. 2015). Their inclusion is enabled by nSR100 (also known as SRRM4) acting as the master regulator for most mammalian neural microexons (Quesnel-Vallières et al., 2015; Head et al., 2021). Unlike other SR proteins that bind to exonic enhancers, SRRM4 binds to intronic enhancers located in the polypyrimidine tract upstream of the targeted microexons, thereby bypassing the requirements of exon definition. Despite their involvement in protein–protein interactions (Ellis et al. (2012)), the detection of protein variation associated with differential microexon inclusion using proteomics is currently challenging. The functional significance for most microexons is unknown, but in cases where roles have been described, microexons appear essential in various aspects of mammalian neuronal development, such as neurite outgrowth, axon guidance, and neurogenesis (Irimia et al., 2014; Torres-Méndez et al., 2021; Lee et al., 2021).

1.3 Splicing and neuropsychiatric disorders

Given the essential roles of alternative splicing in the developing brain outlined above, it is unsurprising that disruptions to splicing patterns are implicated in a range of neurodevelopmental and psychiatric disorders, including autism spectrum disorder (ASD), bipolar disorder (BD), and schizophrenia (SCZ) (Zhang et al. (2021); Gandal et al. (2018)). Mis-splicing and increased skipping of neuronal microexons are often involved in ASD and SCZ (Gonatopoulos-Pournatzis and Blencowe (2020); Gandal et al. (2018)). In ASD, aberrant microexon skipping is likely due to a decrease in SR100 expression induced by increased neuronal activity. The *CACNA1C* gene codes for the pore-forming VGCC α subunit 1C (Clark et al. 2020). All ten mammalian *CACNA* genes are thought to have evolved from multiple gene duplications and are important risk genes for multiple neuropsychiatric disorders. Variants in *CACNA1C* are associated with SCZ, BPD, and Timothy syndrome. Aberrant splicing in neuropsychiatric disorders is further discussed in section 2.1.2.

1.4 Targeting RNA splicing as a therapeutic strategy

Timothy syndrome is just one example of many rare disorders caused by mutations that affect splicing. Despite major advances in rare disease diagnostics, a significant proportion of patients remain without a molecular diagnosis, and many identified variants are classified as variants of unknown significance (VUSs). RNA sequencing has proven invaluable in clarifying the pathogenicity of such variants, particularly when they affect splicing(Jaramillo Oquendo et al. (2024)). Existing short-read splice junction detection tools vary in sensitivity, particularly for variants that cause low-abundance or leaky events, where only some of the resulting isoforms are spliced incorrectly. Long-read RNA sequencing has the potential to overcome many of these challenges by characterising events that are often difficult to detect with short reads, such as intron retention, exon extension, and multi-exon skipping (Wu et al. (2023a)). Accurate baseline isoform annotation is critical in this context, as it provides a reference against which novel or patient-specific events can be interpreted.

Beyond diagnostics, modulation of aberrant splicing is emerging as a promising therapeutic strategy in neurological disorders. One of the most notable successes in this area is the development of antisense oligonucleotides (ASOs) that correct disease-causing splicing patterns. In spinal muscular atrophy (SMA), loss of function of the *SMN1* gene results in reduced levels of the SMN protein, leading to degeneration of motor neurons. The paralogous gene *SMN2* predominantly produces a truncated, non-functional protein due to exclusion of exon 7. The ASO therapy nusinersen (marketed as Spinraza) binds to a splicing silencer in *SMN2*, promoting exon 7 inclusion and restoring production of full-length SMN protein (Chiriboga (2017); Hua et al. (2010)). Newer therapeutic modalities are being developed, including selective small molecules, such as risdiplam (marketed as Evrysdi), the first orally administered splicing modifier drug approved for treatment of SMA (Ratni et al. (2018); Dhillon (2020)). Risdiplam stabilises the interaction between U1 snRNA and the weak non-canonical 5'ss of SMN2 exon 7.

These advances underscore the potential of splicing modulation as a therapeutic strategy in neurological diseases. Identifying appropriate therapeutic targets requires precise knowledge of the splicing landscape, including disease-specific isoforms and their functional relevance. In this context, accurate isoform annotation, particularly in tissues with extensive splicing complexity such as the brain, is essential for uncovering new therapeutic opportunities.

1.5 Currently available approaches for the annotation and quantification of AS events

In order to study splicing variance between health and disease states, we need to polish the current methods of annotation and isoform identification.

1.5.1 Limitations of short-read RNA sequencing

The increasing accessibility of high-throughput sequencing has greatly accelerated the discovery of novel AS events. However, recent long-read transcriptome studies continue to show that even our best annotations from short-read RNA-seq miss large numbers of relevant splice variants, particularly low-abundance and tissue-specific isoforms.

Whilst short-read sequencing remains a cost-effective, accurate, and high-coverage method for transcript quantification, it involves fragmentation of cDNA into sequences that are typically 100–150 bases long. Meanwhile, average human transcript lengths range between 1000-2500 bases, and many are much longer (Lopes et al. (2021)). Additionally, short-read sequence fragments have to be computationally assembled by inference into potential transcripts. Even when all constituent splice junctions and exons are identified and described, complex isoform structures remain difficult to resolve.

Short reads are particularly limited in defining alternative 5' and 3' UTRs, which can affect translation efficiency, mRNA stability and localisation (Hughes 2006). Historically, this was partially caused by coverage heterogeneity across transcript bodies (Steijger et al. (2013)). Earlier Illumina RNA-seq library preparation protocols often showed uneven coverage across transcripts, with pronounced 3' or 5' biases depending on the method used. Improvements in modern library preparation kits have substantially reduced these effects, producing more uniform coverage. However, subtle coverage variability that arises from Polymerase Chain Reaction (PCR) amplification of cDNA still occurs. This leads to different exons being sequenced at different coverage levels (Stark et al. (2019)). Beyond UTR resolution, short-read data also under-represent the structural complexity of alternative splicing events. As highlighted by Nanni et al. (2024), alternative exon cassettes often occur in combination with alternative donor or acceptor sites rather than as isolated events. This trend is consistent across species examined, including *Z. mays, C. elegans, D. melanogaster, D. simulans,* and *H. sapiens*, underscoring the need for methods capable of capturing coordinated splicing patterns, rather than individual exon inclusion events. Isoforms from the same gene typically have a high degree of sequence overlap, and so it can be difficult to determine which isoform the short read originated from. This limitation of short-read quantification has long been recognised, but only recently formally described as the phenomenon of "transcript flipping" (Wissel et al. (2025)). When isoforms are highly similar, short-read quasi-mapping methods (e.g. Salmon (Patro et al., 2017), Kallisto (Bray et al., 2016)) can assign inconsistent expression across replicates, with some runs reporting near-zero abundance and others mid-range values for the same transcript.

1.5.2 Long-read RNA-seq

Recent advances in long-read sequencing have addressed many of the challenges associated with short-reads; Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies (ONT) sequencing can generate reads >10 Kb and therefore sequence full-length transcripts. Long reads eliminate the need for transcript inference from short fragments, allowing accurate resolution of exon connectivity and isoform composition.

However, long-read platforms have their own challenges. PacBio, while offering high per-base accuracy in HiFi mode, has lower throughput, which limits the detection of lowly expressed isoforms. ONT offers higher sequencing depth but historically suffered from higher per-base error rates (especially indels), although recent chemistry and basecalling advances (R10.4.1, Kit 14, Dorado) have greatly improved accuracy. For both platforms, low-abundance isoforms remain difficult to capture without enrichment due to the dominance of highly expressed genes and isoforms in the RNA pools. These dominant isoforms are often orders of magnitude more highly expressed than their minor counterparts (Amarasinghe et al., 2020). Target enrichment approaches, such as PCR amplification of specific gene isoforms (amplicon sequencing) or hybrid capture of target cDNAs (Capture-Seq), address this challenge by selectively enriching for genes of interest and thereby reducing the sequencing depth—and associated costs—needed to achieve sufficient coverage (Mercer et al., 2011, 2014). These capture-based protocols are particularly effective when the targets are well defined. In cases where the targets are unknown, depletion strategies such as Jumpcode can be employed instead, which achieve a similar outcome by removing highly abundant yet less informative transcripts (Pandey et al. (2022)).

In addition to cDNA sequencing, ONT direct RNA sequencing enables the detection of RNA modifications, providing insights into post-transcriptional regulation. While no official multiplexing protocols are available for direct RNA sequencing, recent academic protocols also allow multiplexing of direct RNA libraries through the incorporation of DNA barcodes ligated to RNA adaptors. These barcodes are not basecalled, but can be detected as distinct electrical signals by specialised demultiplexing tools (Smith et al., 2020; van der Toorn et al., 2025). Additionally, integration of long-read sequencing and proteomic data can facilitate improved characterisation of human protein isoform diversity (Miller et al., 2022).

Complex splicing and isoform diversity in long genes

Interestingly, pathways associated with neuronal development and neurodevelopmental multigenic disorders are often enriched for genes with longer transcripts and many splice variants (Sahakyan and Balasubramanian, 2016; Lopes et al., 2021). While the average human protein-coding gene contains around nine exons and encodes 3–4 isoforms, some, such as cell adhesion molecules (CAMs) and VGCC, produce hundreds of multi-kilobase isoforms with distinct functional properties (Ray et al., 2020; Lipscombe et al., 2013). The Drosophila Down Syndrome Cell Adhesion Molecule gene (Dscam) has 95 cassette exons arranged into clusters, with one of them involving 48 mutually exclusive exons (Figure 1.3, Graveley 2005). This leads to 38,016 theoretical splicing variants and 18,496 isoforms confirmed using long read technology(Sun et al. 2013). The psychiatric risk gene *CACNA1C*, mentioned earlier in this introduction, is another example of such a large and complex gene, with at least 50 annotated exons and 31 predicted transcripts (Clark et al., 2020a). While long-read sequencing enables the recovery of full-length sequences for such genes, accurately quantifying their transcripts and their proportional contribution to overall gene expression remains challenging for both long- and short-read sequencing approaches.

1.5.3 Computational methods for splicing and isoform analysis

Alignment to reference genome

Long-read sequencing facilitates the direct identification of the complete set of transcripts and their expression for these complex genes that cannot be reliably reconstructed from short-read data. However, accurate downstream analysis and quantification remain challenging, especially for ONT data.

As ONT-based methods are most relevant to the work described in this thesis, this chapter will focus on briefly describing the state of the art bioinformatics approaches utilised for these data. The long-read analysis workflow begins with basecalling (e.g. Dorado for ONT, CCS for PacBio)

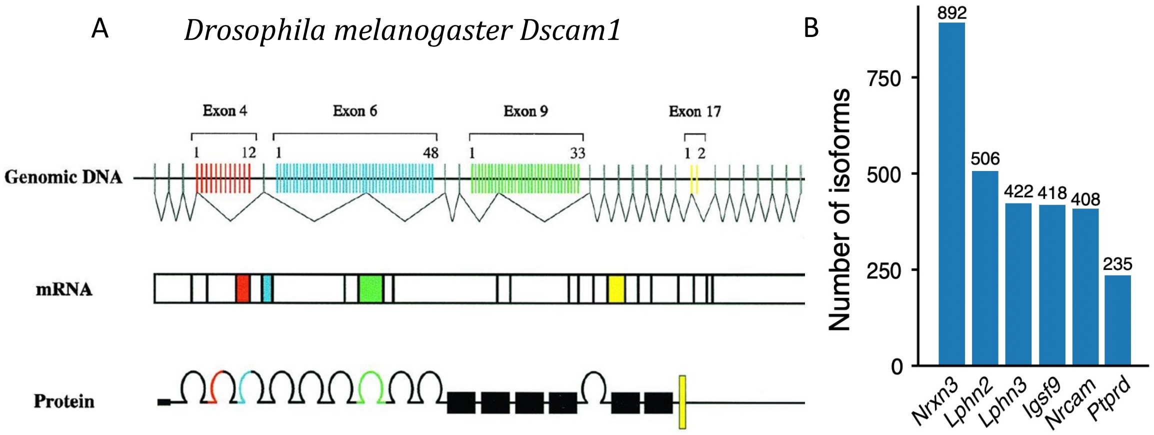


Figure 1.3: **Splicing complexity of long neuronal genes**(A) Drosophila melanogaster Dscam1 (Down syndrome cell adhesion molecule) gene, mRNA, and protein structure. Taken from Wojtowicz et al. (2004), (B) Number of isoforms identified for human neural cell adhesion molecule genes. Taken from Ray et al. (2020)

to convert raw signals into nucleotide sequences. At the time of analysis for Chapter 2, Guppy was the most up-to-date ONT basecalling tool. Guppy converts raw signal data stored in FAST5 format into demultiplexed FASTQ reads. Dorado, which has since superseded Guppy, operates on the newer POD5 format and can output either FASTQ or BAM files. This is followed by read filtering and spliced alignment to the reference genome. Aligners optimised for long reads (e.g. minimap2, deSALT) are required, as short-read aligners perform poorly with the higher indel rates of ONT and the splicing complexity of full-length cDNAs (Li, 2018; Liu et al., 2019). This is because bias correction of short-read approaches requires high read coverage and is affected by small range misalignments, inserts and deletions, especially prevalent in ONT data (Amarasinghe et al., 2020).

Minimap2, the most widely used long-read aligner, uses a seed-and-extend approach based on minimisers, which are small representative k-mers selected to efficiently index the reference genome and reduce the search space. During alignment, seeds are chained to identify candidate mapping regions and penalties for mismatches or the opening of insertions, including introns, are used to handle spliced alignments. To avoid excessive false intron calls, minimap2 will only introduce an intron when the alignment containing the intron is substantially better than an alignment without it. This conservative approach, while reducing spurious splice detection, means that very short exons (<20bp) often cannot compensate for the penalty of introducing an additional intron (Figure1.4). This is especially the case for ONT reads, where the bonus for aligning a short exon is lowered by sequencing errors. Minimap2 also prioritises canonical splice junctions, which improves alignment accuracy for 98% of human transcripts but can limit detection of non-canonical junctions, which are prevalent in microexons and non-coding RNAs.

Transcript detection

Once reads are aligned, the next step is to assign mapped reads to the transcripts from which they most likely originated. This can be approached using two main strategies: reference-guided or reference-free transcriptome annotation. Each approach has distinct advantages and limitations, depending on the experimental design, the quality of the reference annotation for the organism, and the goals of the analysis.

Reference-guided approaches, which were applied in this thesis, identify isoforms from genome

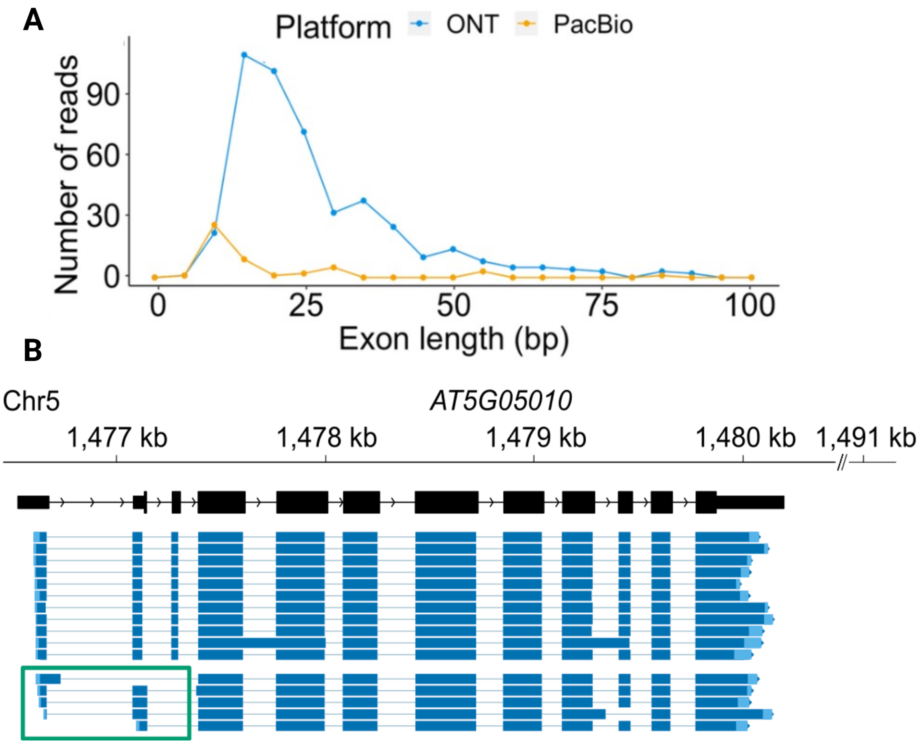


Figure 1.4: **Platform-specific biases in ONT read alignment.** (A) Comparison of skipped exon lengths detected in PacBio versus ONT reads (adapted from Mikheenko et al. (2022)). (B) Example of ONT alignment artefacts where fragments of small exons are misaligned to the ends of flanking exons (adapted from Parker et al. (2021)).

alignments by comparing them to existing gene models. These methods benefit from the accuracy and completeness of well-curated annotations, making them the most widely used class of isoform detection tools in human and mouse studies. Common examples include ToFU, FLAIR, IsoQuant, Bambu, TAMA, TALON, and StringTie (Tang et al., 2020; Kuo et al., 2020; Wyman et al., 2020; Prjibelski et al., 2023; Chen et al., 2022). Within this category, the most common subgroup comprises cluster-based methods, which group reads by similarity (e.g. sequence, splice junction pattern, or exon–intron structure) and collapse them into consensus transcripts. Quantification is then performed by counting the reads in each cluster. Examples include TALON, FLAIR, and Cupcake ToFU (Wyman et al., 2020; Tang et al., 2020). These methods typically incorporate extensive sequencing error correction prior to collapsing.

Bambu (Chen et al., 2022) groups aligned reads into read classes (RC) based on shared exon–intron structures and evaluates the likelihood that each RC represents a valid transcript using a supervised machine learning model to generate a Transcript Probability Score (TPS). Novel transcript detection is controlled by a single parameter, the Novel Discovery Rate, which is calculated as FDR (False Discovery Rate) plus VDR (Valid Discovery Rate) - i.e. proportion of true novel transcripts among predictions. In well-annotated genomes such as human, the NDR can be interpreted as an upper bound on the FDR, as the proportion of false novel calls exceeds that of true novel transcripts (FDR > VDR). Using the SG-NEx dataset as a benchmark (removing chromosome 1 annotations as ground truth), the authors recommend NDR values up to 0.4, as the FDR increases steeply beyond this threshold due to sequencing and alignment artefacts.

Rule-based methods, such as TAMA, filter low-confidence splice junctions according to criteria such as junction ranking or the number of mismatches in the surrounding alignment (Kuo et al., 2020).

This approach has the advantage of making the rationale for filtering specific exons or splice junctions transparent and reproducible. Graph-based methods model the transcriptome as a splice graph, with nodes representing exons or exonic segments and edges representing splice junctions. Reads are traversed through the graph to reconstruct and quantify isoforms. Examples include IsoQuant and StringTie2 (Prjibelski et al., 2023; Kovaka et al., 2019).

Although the false positive rate of novel isoform detection tools remains incompletely characterised, studies using GTEx cell lines have reported substantial overlap in novel isoforms detected across samples (Glinos et al., 2021), and many of these isoforms have been validated by RT–PCR (Tardaguila et al., 2018; Pardo-Palacios et al., 2024b).

Benchmarking by Pardo-Palacios et al. (2024b) identified Bambu and IsoQuant as the only two tools capable of reconstructing very long transcripts (>10,000 nt). However, Bambu showed lower experimental support for transcript ends (CAGE/QuantSeq), as well as reduced long-read coverage (LRC = 60%), i.e. basis in data it used to generate the models, compared with IsoQuant (LRC = 75%).

Reference-free tools, such as isONform and RATTLE, are particularly useful for species or tissue types lacking well-curated reference annotations (de la Rubia et al., 2022; Petri and Sahlin, 2023). FLAIR can also be run without a reference annotation. These methods cluster and assemble reads based solely on sequence similarity, bypassing the need for genome alignment. They often place strong emphasis on read error correction, which improves the accuracy of transcript consensus sequences but can lead to over-collapsing of isoforms at low sequencing depths.

Transcript quantification

Several of the tools described above combine isoform reconstruction from genome-aligned reads with quantification in a single run (e.g. IsoQuant, StringTie, TALON). These typically operate under the assumption that each full-length read corresponds to one transcript molecule, allowing quantification to proceed as a direct counting process.

At the commencement of this PhD, the ONT-recommended workflow involved unspliced alignment to the transcriptome, followed by quantification using Salmon in alignment-based mode. Similarly, NanoCount applied an expectation–maximisation (EM) algorithm after transcriptome alignment (Gleeson et al., 2022). However, these workflows were computationally intensive when novel isoform discovery was required, as they involved two alignment steps.

A much more lightweight alternative is pseudoalignment followed by EM. However, up until recently, pseudoalignment tools such as kallisto and Salmon were only available for short reads and were not optimised for the error profiles of long-read data. The recently released lr-kallisto (Loving et al., 2024a) extends kallisto's pseudoalignment to long reads by adapting the handling of transcript compatibility classes (TCCs) and tuning parameters for long-read error rates. Key modifications include increasing the k-mer size to 63 and, in cases where the intersection of TCCs is empty due to sequencing errors or variants, selecting the most frequent TCC among mapping k-mers, with priority given to uniquely mapping k-mers.

1.5.4 Considerations and limitations

Regardless of method, isoform detection and quantification can be influenced by technical biases and biological complexity. Both GC content and biases arising from PCR amplification can distort the estimated gene and transcript expression levels. Thus, synthetic RNA spike-ins have been developed to act as internal references for expression measurements. They have an entirely artificial sequence with no homology to natural reference genomes, and instead align to an artificial *in silico* chromosome. Additionally, some of these programs can utilise corresponding short-read splice junction information to inform splice junction correction and improve isoform quality. The resulting high-confidence isoform annotations can be used for quantification.

1.6 Thesis Outline

Novel splicing events and transcripts are continually being discovered, with some human genes now containing hundreds of annotated transcripts. However, most of the annotated isoform models, particularly for low-abundance and tissue-specific isoforms, remain incomplete as they arise from short-read sequencing studies. RNA isoform expression profiles of different cell types and their effect on cellular differentiation also remain poorly understood. Newly emerging transcriptomics methods, such as long-read and single-cell sequencing, can allow us to resolve full isoforms and describe the cell-to-cell splicing heterogeneity. The aim of my PhD is to leverage the advantages of these novel methods to develop and improve approaches to annotate functionally relevant isoforms in complex human tissues.

To achieve this, in Chapter2, I benchmarked multiple long-read transcriptomic tools and developed a reproducible long-read RNA-seq analysis workflow for rare isoform detection and differential transcript usage analysis from bulk ONT data. This workflow was evaluated using targeted CaptureSeq data from post-mortem human brain samples. Experimentally validated isoforms from the dataset were used to refine the workflow and optimise parameters for accurate isoform quantification.

In Chapter 3, I adapted the bulk long-read workflow developed in Chapter 2 to single-cell RNA-seq data. This optimised single-cell pipeline was applied to iPSC-derived neural and glial cell cultures to annotate brain cell types and identify both cell–type–specific isoforms and splicing patterns during differentiation. I also discuss the technical and computational limitations of long-read single-cell data analysis, highlighting the need for improved workflows and statistical frameworks.

In Chapter 4, I applied the bulk workflow developed in Chapter 2 to long-read RNA-seq data from iPSC-derived cardiomyocytes to investigate the role of SNORD116 in alternative splicing and 3' end processing during cardiomyocyte differentiation. I further extended the pipeline to include additional analytical steps for 3' UTR and poly(A) tail length analysis, and the results were integrated with matched proteomics data to explore the relationship between transcript isoforms and protein expression.

Finally, I draw together the results of these data chapters in my discussion chapter 5, where I highlight that together, they demonstrate the advantages of long-read sequencing for isoform-level analysis across diverse experimental systems and provide the community with refined workflows to support the discovery of functionally important isoforms in health and disease.

Chapter 2

Comprehensive transcriptome annotation across human brain regions and conditions using long-read CaptureSeq data

This chapter was carried out in collaboration with the Tunbridge and Harrison group (University of Oxford). All cDNA library preparation and Oxford Nanopore sequencing were done by Dr Nicola Hall. All bioinformatic analyses presented here were performed by the author.

The supplementary tables and custom scripts produced for this chapter are freely available at `https://github.com/skudashev/Capture-analysis`.

2.1 Introduction

2.1.1 Splicing variation during brain development

Splicing variation is particularly prominent in the developing human brain, where it is regulated by neural-enriched splicing regulators (Makeyev et al. (2007); Raj and Blencowe (2015); Yeo et al. (2004)). As such, between 22% to 38% of genes expressed in the prefrontal and cerebellar cortices exhibit development-related differential splicing, with over 70% of the observed splicing variation occurring during early postnatal development (Mazin et al. (2013)). BrainSpan (https://www.brainspan.org/) and BrainSeq Phase I (https://eqtl.brainseq.org/phase1/) short-read RNA-seq datasets revealed that the largest numbers of differentially expressed transcripts (DTE) in cortical regions were detected across the prenatal–to-postnatal transition, with 23.7% of the developmentally regulated genes showing opposing isoform-specific patterns of expression (Jaffe et al., 2018; Chau et al., 2021).

In the second phase of the BrainSeq Consortium, RiboZero RNA-seq was used to explore expression differences between the dorsolateral prefrontal cortex (DLPFC) and hippocampus across 551 individuals, reporting that 55% of expressed genes displayed either differentially used exons or splice junctions between these brain regions across development (Rasetti et al. (2014); Collado-Torres et al. (2019)). These studies collectively demonstrate that splicing changes are pervasive during the prenatal–postnatal transition and throughout normal brain development; however, a comprehensive characterisation of the full isoform repertoire in the fetal human brain is still lacking.

2.1.2 Splicing disruption in neurodevelopmental and psychiatric disorders

Independent of gene expression changes, dysregulation of alternative splicing has been strongly implicated in the risk of schizophrenia and neurodevelopmental disorders (NDDs, Gandal et al. (2018); Parikshak et al. (2016); Zhang et al. (2021)). Many of the associated expression quantitative trait loci (eQTLs) are isoform-specific, frequently involving previously unannotated exons or splice junctions. Walker et al. (2019) analysed RNA-seq and genotypes from 201 prenatal cortical samples, identifying gene-level eQTLs (7,962) alongside splicing quantitative trait loci (QTL) (4,635 in 2,132 genes (sGenes)). There was partial overlap between eQTL and sQTL genes, however 50% of sGenes were unique, so splicing regulation was largely independent of overall gene expression regulation. Similarly, Qi et al. (2022) showed that in the adult brain (PsychENCODE dataset), about 61% of cis-sQTL signals were distinct from eQTLs.

Recent studies integrating genome-wide association study (GWAS) data with isoform-level expression profiles have reinforced this point (Bhattacharya et al. (2023); Wen et al. (2023)). Analyses that consider full-length isoform usage, thereby capturing multiple, co-occurring splicing events within the same transcript, consistently outperform approaches restricted to individual local splicing events in pinpointing risk loci for neuropsychiatric disorders.

Autism Spectrum Disorder

Out of all NDDs, alternative splicing disruption is most well-described in Autism Spectrum Disorder (ASD) (Engal et al. (2024)), with numerous pathogenic variants either altering splice sites directly or affecting splicing regulators such as RBFOX1 and PTBP1 (Gauthier et al. (2009); Smith and Sadee (2011)). Although the functional consequences of many of these splicing events remain incompletely understood, converging evidence points towards perturbations in synaptic function, particularly affecting the excitation–inhibition (E/I) balance. One pathway consistently affected is the splicing of synaptic adhesion molecules such as NRXN1 and NLGN4X. These genes undergo extensive isoform switching in early-stage excitatory neurons, with splice variants conferring distinct synaptic binding preferences and response properties (Patowary et al. (2024); Cao et al. (2017)).

Disruption of E/I balance in ASD is also exemplified by Timothy syndrome, a rare condition with overlapping ASD phenotypes. Mutations in the *CACNA1C* gene lead to prolonged opening of the CaV1.2 calcium channel and reduced synaptic inhibition. *CACNA1C* contains mutually exclusive exons 8 and 8A, which encode different versions of a region within the calcium channel's pore-forming domain (Tang et al. (2011)). ASD- and Timothy syndrome–associated mutations have been shown to bias splicing towards exon 8A inclusion (Chen et al. (2024b); Panagiotakos et al. (2019)), altering channel kinetics and excitability.

As discussed in Chapter 1, RBFOX and PTBP proteins act antagonistically to regulate the inclusion of neuronal microexons in the human brain (Li et al. (2015); Quesnel-Vallières et al. (2015)). RBFOX1 promotes inclusion, whereas PTBP1 represses it. During brain development, PTBP1 expression declines as RBFOX proteins are upregulated, enabling neuron-specific microexon inclusion. Disruption of this developmental transition has been described in ASD Chau et al. (2021).

The functional consequences of microexon dysregulation in ASD are exemplified by CPEB4, a translational regulator whose microexon 4 is frequently skipped in idiopathic ASD. This microexon encodes a peptide segment that regulates histidine-rich domain interactions and prevents aggregation. Its inclusion allows CPEB4 to reversibly control gene expression in response to neuronal stimulation. Skipping of microexon 4 disrupts this regulation and affects downstream target mRNAs, many of which are themselves ASD risk genes (Garcia-Cabau et al. (2025)).

Single-cell and single-nucleus transcriptomic studies have provided additional resolution to these findings. Large-scale datasets reveal that splicing dysregulation in ASD is cell-type–specific, with intratelencephalic neurons and inhibitory interneurons showing the most pronounced changes and enrichment for ASD risk genes (Gandal et al. (2022); Velmeshev et al. (2020, 2023)). Inhibitory interneurons (*INT_5_SST*) in ASD brains show downregulation of genes related to synaptic signalling, and mRNA splicing, including *RBFOX1*, *SLM1*, and *SLM2* (Wamsley et al. (2024)).

Schizophrenia

Schizophrenia (SCZ) is a severe psychiatric disorder characterised by disruptions in thought, perception, and behaviour, typically emerging in late adolescence or early adulthood. Despite the later onset of clinical symptoms, converging genetic and epidemiological evidence supports a neurodevelopmental origin of the disorder. The disorder is highly heritable and polygenic, with many risk variants shared with other NDDs (Harrison, 1997; Harrison and Weinberger, 2005; Rees et al., 2021). In contrast to ASD, where risk loci are often enriched for genes with many isoforms, schizophrenia risk loci show less enrichment overall. However, certain high-impact loci stand out (Patowary et al., 2024). The most frequent single-gene mutation associated with SCZ is *NRXN1* (Lowther et al., 2017; Sebastian et al., 2023). The gene is transcribed into two major isoforms, longer α and shorter β, from independent promoters (Jenkins et al., 2016). Beyond these canonical forms, *NRXN1* contains at least five alternative splice sites and undergoes extensive alternative splicing, generating hundreds of transcript variants (Treutlein et al., 2014). Only 65 of these isoforms are annotated in GENCODE v47.

Common SCZ-associated variants show significant enrichment for splicing QTLs in both the developing and adult human brain (Takata et al., 2017). At least 12 schizophrenia-associated genes, including *DRD2*, *GRM3*, and *DISC1*, undergo aberrant splicing, producing distinct transcript isoforms with altered or disrupted function. *GRM3* encodes a metabotropic glutamate receptor involved in modulating synaptic transmission. Expression of GRM3Δ4 splice variant is increased in the DLPFC of individuals carrying a schizophrenia-associated SNP (Sartorius et al., 2008). Exon 4 is skipped in this variant, resulting in a truncated C-terminal domain that is predicted to alter receptor signalling (García-Bea et al., 2017).

Multiple SCZ-associated SNPs in *ERBB4*, which encodes a NRG1 receptor tyrosine kinase implicated in interneuron development, are isoform-specific. Distinct isoforms of *ERBB4* differ in their juxtamembrane and cytoplasmic domains, with exon 16 and exon 26 included isoforms being increased in the DLPFC of SCZ patients (Law et al. (2007); Veikkolainen et al. (2011)). MIAT (also known as Gomafu), a long non-coding RNA (lncRNA) implicated in numerous cardiovascular and neurologic disorders (Zeinelabdeen et al. (2024)), interacts with splicing factors such as QKI, SF1, and CELF3. In SCZ, MIAT expression is dysregulated in a brain region–specific manner: it is reduced in the superior temporal gyrus, frontal cortex, and temporal cortex (Barry et al. (2014); Teng et al. (2023)), but increased in the DPLFC and nucleus accumbens (Teng et al. (2023)). Knockdown of MIAT has been shown to promote the expression of SCZ-associated isoforms of ERBB4 and DISC1 (Zakutansky and Feng (2022); Ip et al. (2016); Barry et al. (2014)).

Bipolar Disorder

Bipolar Disorder (BD) shares many features with SCZ, including adolescent onset, high heritability, and substantial overlap in genetic risk loci with other major psychiatric disorders (Harrison et al., 2018; O'Donovan and Owen, 2016). The symptomatic similarities, particularly during the prodromal (pre-psychotic) phase, characterised in both by cognitive difficulties, social withdrawal, anxiety and depressive symptoms, can make differential diagnosis challenging (Kafali et al., 2019; Yang et al., 2024).

GWASs have consistently identified *CACNA1C* and *ANK3* as key BD risk genes. A splice-site variant (rs41283526) within an alternatively spliced exon of *ANK3* has been shown to lead to exon skipping. Interestingly, this variant was found to be strongly protective against BD and SCZ (Hughes et al. (2016); Holmgren et al. (2022)). Using BrainSpan data, the authors found that this exon is only included in a minor isoform primarily expressed in early adolescence, a period coinciding with myelin maturation.

Another gene implicated in BD, *NRG3* (neuregulin 3), has nine annotated isoforms grouped into three classes based on sequence homology, each with distinct developmental and disease-associated expression patterns (Paterson et al., 2017). In the DLPFC, Class I peaks during neonatal and infant stages, Classes II and IV during fetal and neonatal periods, while Class III expression remains stable. Different classes were found to be upregulated depending on disorder: classes I and II in BD and classes I and III in major depressive disorder patients (Paterson et al., 2017).

More recently, Fahey and Lopez (2024) used pathway-based polygenic score analysis to show that the minor splicing pathway is enriched for genetic variation shared between BD and chronotype. The minor (U12-type) spliceosome processes a small subset of evolutionarily conserved introns, many located in genes essential for neuronal function. Several of its components, including SF3B1, SF3B2, and PRPF6, harbour BD-associated regulatory variation. Because spliceosomal activity is partly regulated by circadian clock genes, disruption of minor splicing may contribute to the circadian rhythm disturbances characteristic of BD.

Major Depressive Disorder

Major depressive disorder (MDD) is the most prevalent of the major psychiatric disorders and is characterised by recurrent episodes of low mood and anhedonia. In addition, ASD and SCZ are frequently comorbid with, or can predispose individuals to, depression (Samsom and Wong (2015)). Like other common psychiatric disorders, MDD has a major heritable component, with twin study estimates of around 30–50%, but environmental influences such as stress play a comparatively larger role in disease onset and progression (Tsuang et al. (2004); Cui et al. (2024)).

Large-scale GWASs have identified over 100 risk loci for MDD. These include genes involved in dopaminergic neurotransmission (*DRD2*), glutamatergic (*GRIK5, GRM5*) and calcium channel signalling (*CACNA1E, CACNA2D1*) (Wray et al. (2018); Howard et al. (2019)). However, the specific mechanisms by which these loci increase risk are less well understood.

In the largest postmortem human brain transcriptomic study of MDD to date, Goes et al. (2025) performed high-depth short-read RNA sequencing of tissue from the subgenual anterior cingulate cortex (sACC) and the amygdala, regions central to mood regulation. Using this approach, they revealed extensive transcript-level differences, particularly in the amygdala. Splicing analyses with Leafcutter (Li et al. (2018)) at the level of intron inclusion identified differentially spliced clusters enriched for pathways related to synaptic function and calcium signalling, including genes such as *CACNA1A*. QTL mapping of recently identified MDD risk loci revealed that splicing QTLs were more numerous and sometimes independent of gene-based signals. Many GWAS loci were linked to transcript-level or splicing variation, particularly in synaptic genes such as *NRXN1*, *NLGN1*, *RBFOX1*, and *CACNA1C*. These results suggest that splicing variation is a key mediator of MDD genetic risk. However, as in most large-scale human postmortem datasets, these analyses relied on short-read RNA sequencing.

2.1.3 Brain-specific isoforms as drug targets

Importantly, emerging evidence indicates that targeting brain-specific isoforms may represent a promising therapeutic avenue. One such candidate class are the voltage-gated calcium channels (VGCCs). Observational studies indicate that brain-penetrant calcium channel blockers (CCBs) are associated with modest reductions in the incidence and recurrence of psychiatric diagnoses compared to non-brain-penetrant CCBs (Hayes et al. (2019); Colbourne et al. (2021)).

However, their clinical use in psychiatry is limited by off-target cardiovascular effects such as hypotension and arrhythmia. These effects arise because the heart and other muscles are also excitable tissues that express high levels of L-type Ca^{2+} channels, particularly Cav1.2 encoded by *CACNA1C*, and share many downstream signalling pathways with neurons; systemic CCBs therefore inhibit calcium influx in cardiomyocytes and vascular smooth muscle. Long-read sequencing studies have revealed that many previously unannotated *CACNA1C* transcripts are highly expressed in the brain relative to the heart and aorta, with some exceeding the abundance of annotated forms. Several of these brain-enriched isoforms are predicted to encode channels with altered functional or pharmacological properties (Clark et al. (2020a); Hall et al. (2021b); Harrison et al. (2022)). Such isoforms could offer a novel therapeutic avenue for psychiatric disorders by modulating calcium signalling in the brain while sparing the heart and other excitable tissues.

2.1.4 Characterising the brain transcriptome

Identifying brain-enriched isoforms and accurately characterising their developmental and disease-specific expression requires comprehensive and reliable transcript annotations, as these have a major impact on downstream differential expression analyses. However, current human gene annotations and our knowledge of developmentally regulated neural isoform expression are largely based on short-read RNA sequencing (RNA-seq). While short-read RNA-seq is cost-effective and provides accurate, high-coverage data for detecting local splicing events, it is poorly suited to resolving full-length isoforms or identifying transcript start and termination sites due to cDNA fragmentation. This limitation is particularly problematic for genes such as *CACNA1C*, a major psychiatric risk locus, which have unusually complex isoform structures that are challenging to reconstruct from fragmented short-read data (Sahakyan and Balasubramanian, 2016; Patowary et al., 2024).

Long-read sequencing platforms such as PacBio SMRT and ONT overcome the need for transcript reconstruction by directly sequencing full-length molecules. These approaches have revealed that existing annotations of the human brain transcriptome are incomplete, especially for low-abundance and region-specific isoforms (Clark et al., 2020b; Wright et al., 2021). However, long-read sequencing has its own limitations: PacBio offers high read accuracy but relatively shallow sequencing depth, while early ONT chemistries provided higher depth but were hampered by higher indel error rates that affected splice junction and open reading frame (ORF) annotation (Amarasinghe et al., 2020; Stark et al., 2019). Carefully selected support filtering steps are therefore required to extract reliable full-length isoforms from ONT data generated using older chemistries (<v14/ R10.4). Early studies, therefore, relied on hybrid strategies that combined long-read and short-read data to ensure reliable transcript quantification (Tilgner et al., 2014; Au et al., 2013). First, SMRT-seq was used to construct a comprehensive transcript set, followed by the alignment of sample-matched high-coverage Illumina sequencing reads to the long-read transcriptome.

Alternatively, some researchers explored targeted RNA capture sequencing (Capture-Seq) an approach initially developed for short-read, involving biotinylated probe-based cDNA hybridisation capture (Mercer et al., 2014; Clark et al., 2015). Capture-Seq was consequently successfully coupled with long-read sequencing (Lagarde et al., 2017; Dainis et al., 2019; Deveson et al., 2018; Sheynkman et al., 2020; Schwenk et al., 2023).

In summary, transcriptome-wide long-read sequencing lacks the depth required to comprehensively capture the full isoform repertoire of the human brain, particularly for low-abundance transcripts, and deeper sequencing is prohibitively costly. Current annotations underestimate the isoform diversity of genes strongly implicated in psychiatric disorders. To address this gap, with our collaborators in the Tunbridge laboratory (University of Oxford), we employed ONT long-read Capture-Seq to profile transcripts from 1,469 genes of psychiatric relevance. The probe design included both protein-coding genes (e.g. calcium channels and RNA-binding proteins) and lncRNAs overlapping GWAS risk loci. This dataset comprised 52 brain tissue samples from three brain regions of 12 patients with a major psychiatric disorder (SCZ, BD, or MDD) and 8 control individuals without a known history of psychiatric disease (Figure 2.1).

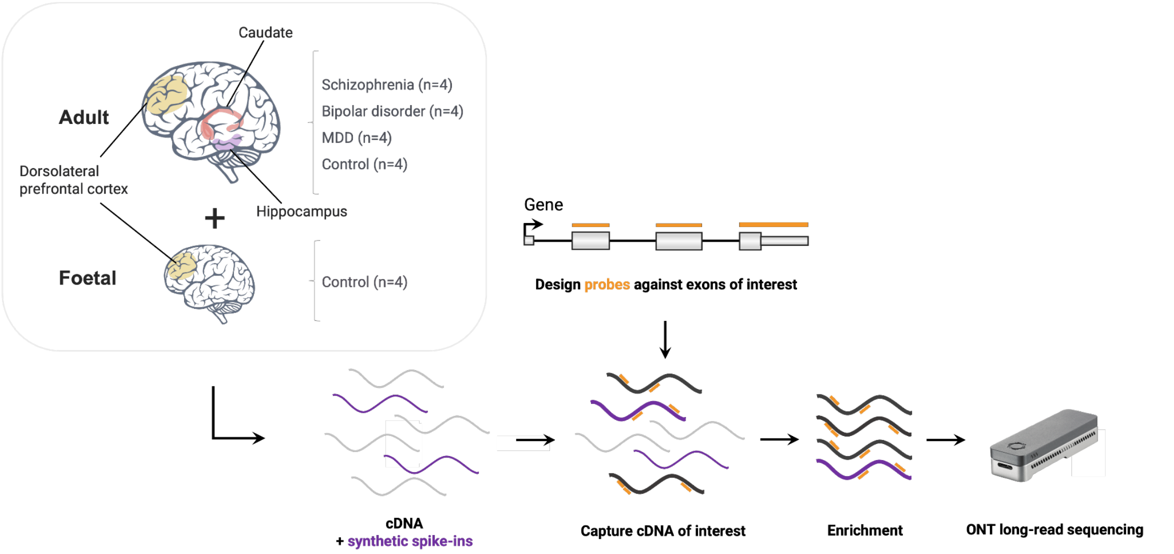


Figure 2.1: **Long-read Capture-Seq experimental design and sample overview.** A comprehensive probe library targeting GWAS-identified psychiatric risk lncRNA and protein-coding genes was developed by the Tunbridge group. These oligonucleotide probes were used to enrich cDNA libraries for target transcripts.

Within this chapter, I developed and benchmarked an analysis workflow for comprehensive isoform discovery and quantification using these data. Multiple tools were evaluated across all stages of the pipeline, from quality control to quantification, and the workflow was iteratively refined based on manual curation. Particular attention was given to the identification and filtering of potential technical artefacts, such as truncated isoforms resulting from mispriming and RNA degradation. The resulting transcript models were further assessed against Lieber Institute short-read RNA-seq data and publicly available reference Transcription Start Site (TSS) and Transcription Termination Site (TTS) databases.

This approach enabled the annotation of 56,249 novel isoforms across 1,055 of the 1,444 targeted genes detected in the samples. Synthetic spike-in controls confirmed that capture enrichment did not compromise quantification accuracy, allowing extension to all transcripts. Across brain regions and developmental stages, we identified 90 significant isoform switches between adult brain regions and 214 between fetal and adult samples, 47 of which involved novel transcripts. These results demonstrate that isoform diversity at psychiatric risk loci is substantially greater than represented in current annotations.

2.2 Methods

Tool	**Version**	**Reference**
edgeR	4.4.2	Chen et al. (2025b)
IsoformSwitchAnalyzer	2.6.0	Vitting-Seerup and Sandelin (2019)
Minimap2	2.24	Li (2018)
Restrander	1.0.0	Schuster et al. (2023)
Salmon	1.8.0	Patro et al. (2017)
Samtools	1.15.1	Danecek et al. (2021)
SeqKit	2.5.1	Shen et al. (2016)
SQANTI3	5.0	Pardo-Palacios et al. (2024a)
kallisto	0.50.0	Loving et al. (2024b)
bustools	0.42.0	Melsted et al. (2021)
2passtools	0.3	Parker et al. (2021)
PFAM	–	Mistry et al. (2021)
DeepLoc2	2	Ødum et al. (2024)
DeepTMHMM	1.0.21	Hallgren et al. (2022)
Trim Galore	0.6.6	Krueger (2018)
STAR	2.7.10	Dobin et al. (2013)
Guppy	4.4.0	https://github.com/nanoporetech
TAMA	–	Kuo et al. (2020)
cDNA Cupcake	28.0	Tang et al. (2020)
FLAIR	1.5.0 / 1.6.2	Tang et al. (2020)
Bambu	2.0.0	Chen et al. (2022)
IsoQuant	3.6.1	Prjibelski et al. (2023)

Table 2.1: Tools, versions and associated references used in Chapter 2.

2.2.1 Library preparation and sequencing

All Capture-Seq library preparation and sequencing was performed by Dr Nicola Hall (Tunbridge group, University of Oxford). RNA was extracted from post-mortem human brain tissue provided by the Lieber Institute for Brain Development (LIBD, Baltimore, MD), comprising 16 adult samples from the dorsolateral prefrontal cortex (DLPFC), hippocampus, and caudate (n=4 each for control, major depression, bipolar disorder, and schizophrenia), as well as 4 foetal DLPFC samples. Synthetic spike-ins ('sequins') (Hardwick et al. (2016), Garvan Institute, Sydney, Australia) v2 Mix A or B were added at a final concentration of 1:10,000, followed by poly(A) selection and cDNA synthesis using the Nanopore reverse transcription and strand-switching protocol. A custom barcoding strategy was implemented, enabling multiplexing of up to seven samples per capture reaction. Hybridisation-based capture was performed using SeqCap EZ probes and a modified version of the protocol from Mercer et al. (2014), with post-capture PCR using M13-tagged primers. Capture probes were designed against constitutively spliced exons to ensure recovery of isoforms across targeted loci. Target enrichment was confirmed by qPCR (minimum 10-fold), and sequencing was carried out using Oxford Nanopore Technologies (SQK-LSK109) on FLO-MIN106 flow cells (R9.4.1 pores).

For short-read validation, we used Illumina RNA-seq data generated by the LIBD and provided through personal communication. Libraries were sequenced as 100 bp paired-end reads with a minimum depth of 100 million reads per sample. The dataset comprised control fetal as well as control and SCZ adult DLPFC samples. Detailed sample descriptions and data access are provided in the PsychENCODE publications (PsychENCODE Consortium et al. (2015); Wang et al. (2018)).

Some code snippets were drafted or refined with the assistance of ChatGPT (OpenAI, 2025), and subsequently validated by the author.

2.2.2 Oxford Nanopore read processing

Basecalling and demultiplexing of barcoded reads were performed using Guppy v4.4.0. Reads were filtered for length $\geq$ 200bp and quality (Q) $\geq$7 using SeqKit. Full-length cDNA reads were identified and oriented with Restrander (Schuster et al. (2023)). Sequencing metrics were generated using pycoQC. The trimmed and reoriented reads were aligned to the human genome (GRCh38, UCSC), modified to include the synthetic sequin chromosome (chr IS), using minimap2 (v2.24) with parameters `"-ax splice --cs=long -k 13"`. NanoPlot was used to calculate alignment statistics. A high confidence set of splice junctions was generated with 2passtools filter, and the resulting set was provided for second pass alignment with minimap2 `"-ax splice -k 14 --secondary=no -ub --junc-bonus 11 -G 500000 --junc-bed"`. Samtools view (v1.15) was used to filter out supplementary alignments (0x900).

2.2.3 Isoform detection

To construct a high-confidence isoform set from ONT long-read data, we benchmarked five widely used isoform detection tools: TAMA, cDNA Cupcake, FLAIR, IsoQuant, and Bambu (Kuo et al. (2020); Tang et al. (2020); Prjibelski et al. (2023); Chen et al. (2022)). Tool performance was assessed both across all targeted genes and by evaluating recovery of a previously validated novel alternative first exon of *CACNA1C* exon 1d (Dr Nicola Hall, personal communication). As a first exon, its 5' splice junction also serves as a transcript start site. Although ONT reads show high variability at

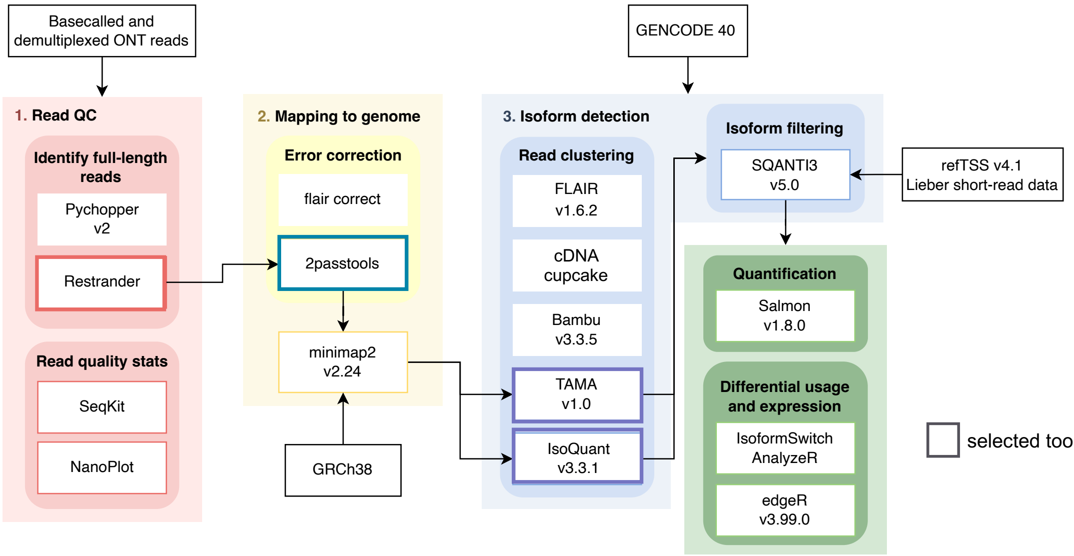


Figure 2.2: **Schematic diagram of computational workflow.** Tools selected for final analysis highlighted by thick outlines.

both 5' and 3' ends, it is especially difficult to experimentally generate ONT reads with reliable 5' ends as the process of 5' capping is more challenging than poly(A) selection and there is evidence that ONT reads are particularly susceptible to 5' truncation as a result of reverse transcription artifacts (Calvo-Roitberg et al. (2023), 2.6B).

TAMA TAMA was selected for its highly customisable and transparent rule-based algorithm (1.5.3). Both default and optimised parameters were evaluated (`-x no_cap -c 90 -i 80 -a 200 -m 20 -z 200 -icm ident_cov -rm low_mem -log log_off -sjt 20 -lde 5`). Due to the high variability in 5' and 3' ends in ONT data, particularly in cDNA libraries lacking cap-selection, the 5'/3' end thresholds (-a and -z) were increased. Reads with more than five errors within 20 bp of splice junctions were excluded. The suggested method of speeding up the process is to first run `tama_collapse.py` with the "capped" argument, followed by `tama_merge.py` with "no_cap" selected. This reduces the complexity during the collapse step and then resolves $5'$ degraded models during merging. However, as the ground truth in this case was a $5'$ terminal exon, I chose to run `tama_collapse.py` in the "no_cap" mode to ensure I captured the complexity.

To further reduce computational burden, the aligned BAM was initially split by chromosome using `tama_mapped_sam_splitter.py`, followed by `tama_collapse.py` on each subset and merging with tama_merge.py. As high coverage at targeted loci frequently led to substantial slowdowns, a custom script (`split_bam_by_windows.py`) was later developed to enable splitting into non-overlapping genomic bins, further improving scalability.

cDNA Cupcake To enable parallelisation, similarly to TAMA, the aligned sorted BAM file is split into regions and processed concurrently. Transcript collapsing was performed using `collapse_isoforms_by_sam.py` with default parameters and with the same coverage and identity parameters as TAMA `"-c 0.9 -i 0.8"`.

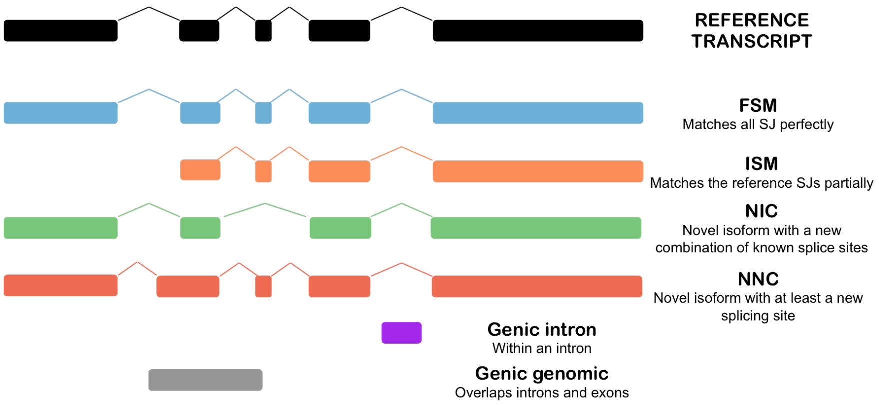

REFERENCE TRANSCRIPT
FSM
Matches all SJ perfectly
ISM
Matches the reference SJs partially
NIC
Novel isoform with a new combination of known splice sites
NNC
Novel isoform with at least a new splicing site
Genic intron
Within an intron
Genic genomic
Overlaps introns and exons

Figure 2.3: SQANTI isoform classification

FLAIR (v1.5.0 and v1.6.2) `flair collapse` (v1.5.0) was run with default parameters and FLAIR (v1.6.2) with `"--no_gtf_end_adjustment --isoformtss --filter comprehensive --support 2 --end_window 200"`.

Bambu Bambu was evaluated at two novelty detection rate thresholds (NDR=0.4 and NDR=0.7). While Bambu's machine learning-based classification of novel isoforms is advantageous, v2.0 outputs all reference transcripts by default, including unexpressed isoforms. This, combined with limited transparency in filtering decisions, made it less suitable for our aim of understanding potential capture-specific isoform detection biases.

cDNA Cupcake, FLAIR and Bambu were not utilised downstream (see Section 2.3.2).

After running each isoform detection tool, SQANTI3 (v5.0) was used to classify identified isoforms into structural categories by comparing them to the GENCODE (release 40) and sequin reference annotation (Tardaguila et al. (2018); Pardo-Palacios et al. (2024a)). The isoforms are classified into standard SQANTI structural categories. These include: Full-Splice Match (FSM): Isoforms in which the complete splice junction chain (SJC) exactly matches a reference transcript; Incomplete-Splice Match (ISM): Isoforms where the SJC is a subset of an annotated transcript; Novel In Catalog (NIC): Novel isoforms of known genes that contain new combinations of annotated junctions; Novel Not in Catalog (NNC): Novel isoforms with at least one novel donor and/or acceptor splice site (2.3. Additional isoform classes are antisense, fusion, genic intron (lies entirely within an intron of a gene), genic genomic (spans both exons and introns) or intergenic.

Initially, TAMA was selected as the primary tool due to its flexibility and strong performance in capturing rare validated isoforms. However, certain longer annotated CACN full-length isoforms, which had high coverage according to bedtools coverage analysis and had supporting reads, were missing from the TAMA detected set. With the release of IsoQuant (Prjibelski et al. (2023)), we re-evaluated the pipeline. IsoQuant improved reconstruction of long transcripts (>10 kb), aligning with recent benchmarking Pardo-Palacios et al. (2024b). It was therefore incorporated into the final workflow as a complementary method.

IsoQuant `isoquant.py` was run with parameters optimised for pre-corrected ONT reads: `--model_construction_strategy assembly --splice_correction_strategy conservative_ont --data_type nanopore --matching_strategy loose`. IsoQuant output was merged with TAMA isoforms using `tama_merge.py` (no wobble). We selected this two-tool strategy to balance sensitivity

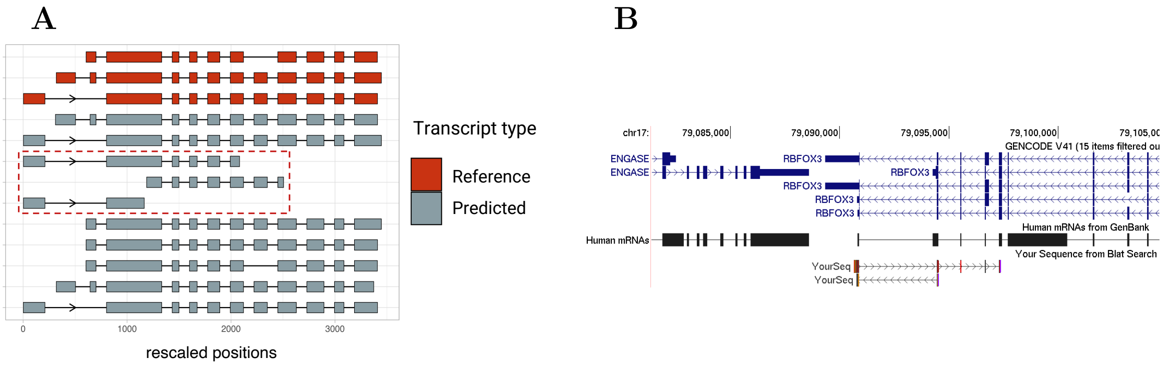


Figure 2.4: **Artefacts in long-read cDNA sequencing data.** A) Truncated transcripts for sequin R2_53. B) UCSC browser screenshot of an example of a foldback artefact read.

and precision.

2.2.4 Artefact detection and filtering

An unexpectedly high proportion of truncated isoforms was detected across multiple tools. To determine whether these represented genuine biological transcripts or artefacts, we assessed predicted sequin isoforms. As synthetic molecules of known structure, sequins provided an unambiguous reference: truncated isoform calls in sequins indicated technical artefacts rather than biological variation (Figure 2.4A). Inspection of alignments revealed a recurrent "foldback" artefact, where a read contained a primary alignment in one strand orientation and a supplementary alignment in the opposite orientation at the same locus. This pattern was consistent with RT-mediated self-priming or hairpin formation, generating chimeric molecules.

To remove these artefacts, a custom Python script (filter_RT.py, `https://github.com/skudashev/Capture-analysis`) was implemented to flag reads meeting any of the following criteria:

1. Primary alignment contains large deletions (> 20 bp) indicative of RT switching artefacts introduced due to secondary structures (Cocquet et al. (2006); Houseley and Tollervey (2010)).
2. Primary alignment covers $< 30\%$ of total read length, suggesting a truncated fragment of the original molecule.
3. Supplementary alignment maps to the opposite strand of the same locus with overlapping genomic coordinates, consistent with foldback artefacts.

2.2.5 Short-read RNA-seq data processing

Short-read RNA-Seq raw reads from independent Lieber Institute brain samples were trimmed with TrimGalore (v0.6.6, Krueger (2018)) with default parameters. Trimmed reads were aligned to the GRCh38 reference genome using STAR (v2.7.10, Dobin et al. (2013)) two-pass mapping using default parameters. The resulting BAM files and `SJ.out.tab` files containing high-confidence splice-junctions and numbers of reads spanning them were used as input for sqanti3_qc.py to filter novel isoform models.

2.2.6 Transcriptome annotation and validation

The resulting merged set of unique isoforms was characterised and classified using SQANTI3 (Pardo-Palacios et al. (2024a)) with respect to GENCODE (release 40) gene annotation, STAR splice-junction and TSS short-read support from independent LIBD samples. The transcriptome was filtered based on coordinate overlap with the 1,469 genes that were included in the study design, rather than based on SQANTI3 predicted gene ID. Filtering criteria were tailored to SQANTI3 structural categories:

Full-splice match (FSM): Retained if the proportion of genomic adenosines downstream of the transcript 3' end was $\leq$ 79%.

Incomplete-Splice Match (ISM), Novel In Catalog (NIC) and Novel Not in Catalog (NNC): Retained if

- downstream adenosine content was $\leq$ 69%,
- no RT-switch junctions were detected,
- the ratio of short-read coverage at the TSS was $\geq$ 1.1 or a refTSS CAGE peak within ±1 kb of the TSS was present
- either a PolyASite atlas (Herrmann et al. 2020) polyA peak or a polyA motif within ±50 bp of the TTS
- the transcript contained at least two exons.

Other categories (fusion, antisense, genic, intergenic, genic_intron): Retained if

- downstream adenosine content was $\leq$ 69%,
- no RT-switch junctions were detected,
- the ratio of short-read coverage at the TSS was $\geq$ 1.1 or a refTSS CAGE peak within ±1 kb of the TSS was present
- either a PolyASite atlas (Herrmann et al. 2020) polyA peak or a polyA motif within ±50 bp of the TTS
- the transcript contained at least two exons.
- all junctions have at least 2 short reads covering it.

2.2.7 Gene and isoform expression quantification

To evaluate transcript-level expression, we applied two quantification strategies: an alignment-based approach using **Salmon** for initial spike-in enrichment testing, and a pseudoalignment-based approach using **lr_kallisto**. Both were used for isoform quantification and differential transcript usage (DTU) analysis; however, the final analysis pipeline was based on lr-kallisto, which provided greater computational efficiency and lower memory requirements. While the dataset analysed in this study comprised 52 samples, the workflow was designed to be scalable and applicable to larger cohorts. A recent benchmarking study using PacBio Kinnex data and SIRV spike-ins further demonstrated that lr-kallisto achieved one of the lowest false discovery rates for differential transcript expression (DTE) detection, with accuracy comparable to Illumina short-read data and to IsoQuant (Wissel et al. (2025), Figure S2).

Alignment-based quantification with Salmon All demultiplexed FASTQ reads—including reads that did not pass the stringent quality-control thresholds applied prior to transcriptome construction—were aligned to the custom transcriptome FASTA using minimap2 with parameters optimised for ONT data in unspliced mode (`-ax map-ont -N 50`). A high value of `-N` was chosen to retain multiple secondary alignments, which are common in long-read datasets due to incomplete or partially degraded transcripts. Aligned reads were quantified using Salmon in alignment-based mode with ONT-specific parameters (`--ont`), 30 bootstrap iterations, and both length correction and the error model disabled (`--noLengthCorrection --noErrorModel`). Thirty bootstrap replicates were generated to estimate technical variance, providing a balance between accuracy and computational efficiency, as increasing to 100 replicates substantially increased runtime without improving precision. Length correction was disabled, given that ONT reads typically span full-length transcripts, and the short read tailored error model was disabled to better accommodate the elevated per-base error rates of ONT sequencing.

Pseudoalignment-based quantification with lr-kallisto Index was generated from the long-read generated transcriptome using k=63. Pseudoalignment of all demultiplexed FASTQ reads was then performed using `kallisto bus --long --threshold 0.8`, followed by sorting and counting with `bustools sort` and `bustools count`. Transcript-level abundance estimates were derived from transcript compatibility counts using `kallisto quant-tcc --long --platform ONT`.

2.2.8 Analysis of on-target enrichment

To assess the lower limit of quantification, we examined the relationship between measured expression values and known input concentrations of synthetic spike-in controls (sequins). The on-target rate was defined as the ratio of the number of distinct reads mapping to targeted genomic regions (excluding sequin RNA spike-ins) to the total number of mapped reads. The number of reads overlapping targeted regions was calculated directly from the minimap2 genomic alignment BAM file with `bedtools intersect`. The correlation (R^2) between normalised transcript counts, the number of probes designed per transcript, and input spike-in concentration were calculated and visualised using linear regression and scatter plot visualisation in R.

2.2.9 Differential transcript usage and expression analysis

Transcript-level counts and abundances generated by Salmon were imported into an R 4.3.0 environment using tximport, which scales them to library size (scaledTPM). Per-sample distributions of isoform novelty were calculated by first taking genes with overall >5 counts in each sample, and subsequently calculating the percentage of reads mapping to novel isoforms for each gene and sample. DIU analysis was performed using IsoformSwitchAnalyzeR (v2.6.0), which utilises the DEXSeq negative binomial. For each pair-wise comparison, counts were filtered to remove single isoform genes and genes with average expression in both conditions under 0.05 TPM. Isoform switches were defined as significant if the absolute change in isoform fraction was over 10% and the q-value was under 0.05. ORF nucleotide sequences were extracted and coding potential predicted with CPC2 (Kang et al. 2017). The ORFs of significant DIU isoforms with coding probability ≥0.45 were translated into amino acid sequences and then scanned for protein domains using pfamscan.py (v1.6) (Mistry et al. (2021)) and topology using DeepTMHMM (Hallgren et al. (2022)).

To assess the impact of potential technical and biological confounders on isoform usage, surrogate variable analysis (SVA) as implemented as part of `importRdata()` in IsoformSwitchAnalyzeR.

Specifically, we tested whether including covariates such as sex, sequencing barcode, or sequencing run reduced the number of inferred latent variables (surrogate variables, SVs). The inclusion of barcode as a covariate did not yield results, likely due to insufficient replication across barcode-group combinations. When sex was included, the number of inferred SVs remained unchanged, indicating limited explanatory power. In contrast, inclusion of Run as a covariate reduced the number of inferred SVs from 4 to 1, suggesting that this factor accounted for a substantial portion of the unwanted technical variation.

2.3 Results

2.3.1 Capture-Seq reads are strongly enriched for the targeted genes and spike-ins

To initially evaluate the specificity and effectiveness of the capture enrichment process, I first performed quantification using the Salmon in alignment mode against the sequin reference annotation. A total of 160 synthetic spike-ins (sequins) were added to cDNA at known concentrations spanning a 3,733,025-fold range, serving as quantification controls. Sequins included in the RNA mixtures but not targeted by capture probes acted as negative controls. Although probed spike-ins showed substantially higher read counts, a considerable fraction of reads also aligned to negative-control sequins, indicating off-target capture. This suggests that the enrichment procedure is "leaky", potentially due to sequence similarity between gene bodies. Despite this off-target signal, targeted sequins were markedly enriched, and their measured expression correlated strongly with input concentrations ($R^2 = 0.82$), compared with lower correlation for untargeted sequins ($R^2 = 0.56$; Figure 2.5a).

For endogenous genes, comparison of transcript abundances with GTEx v8 short-read RNA-seq data from frontal cortex and hippocampus revealed strong capture efficiency, with mean fold-enrichment values of 53.0 in cortex and 69.8 in hippocampus for protein-coding genes targeted by capture probes, relative to non-targeted genes (Figure 2.5b). When stratified by functional class, enrichment of lncRNA genes was not significant (Figure 2.5c), likely reflecting their low endogenous expression and suboptimal annotation, which complicates probe design. Finally, probe count per gene showed only a weak relationship with observed transcript levels ($R^2 = 0.14$ in cortex, 0.07 in hippocampus; Figure 2.5d), indicating that the number of probes per gene was not a major determinant of capture efficiency and therefore low lncRNA counts were more likely due to low endogenous expression.

After establishing the capture performance with Salmon, I subsequently revised my pipeline to use lr-kallisto quantification. Repeating the sequin-based analysis with this updated workflow yielded highly consistent results (Figure S3), but with improved correlation for untargeted sequins as well. In addition, I assessed the coefficient of variation across sequins as a function of input concentration, which gave comparable results for both Salmon and lr-kallisto (Figure S2). This analysis showed that below an input concentration of 1 amol/µL, CV was consistently over 0.5 and therefore the measured TPMs are potentially too noisy/unreliable to interpret quantitatively below that.

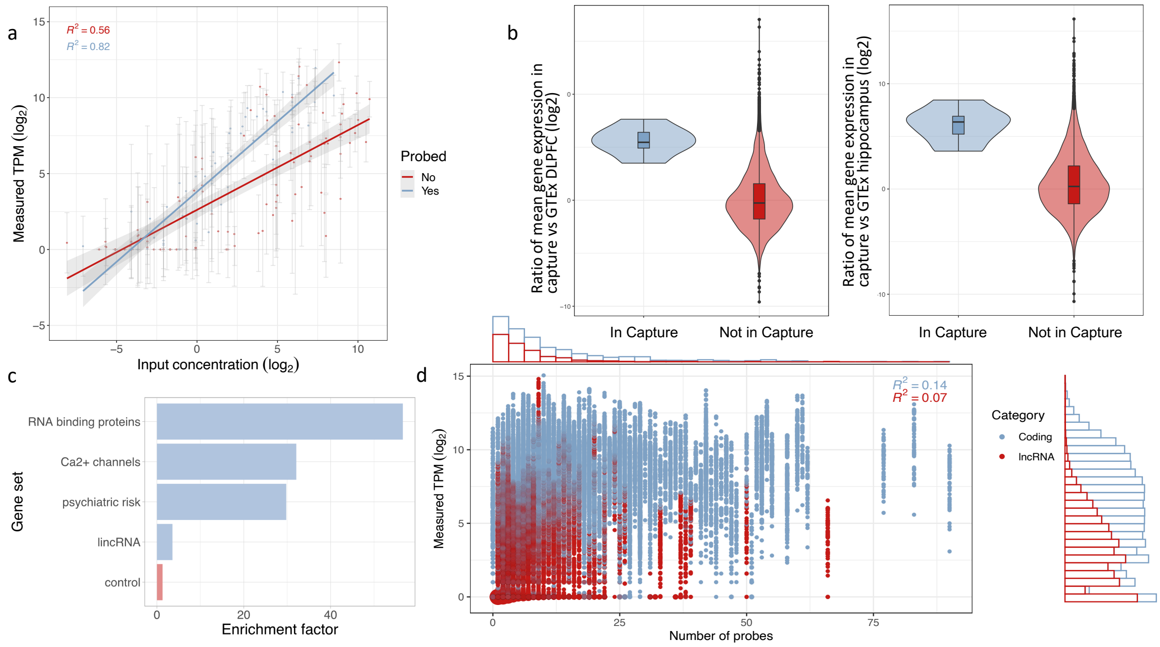
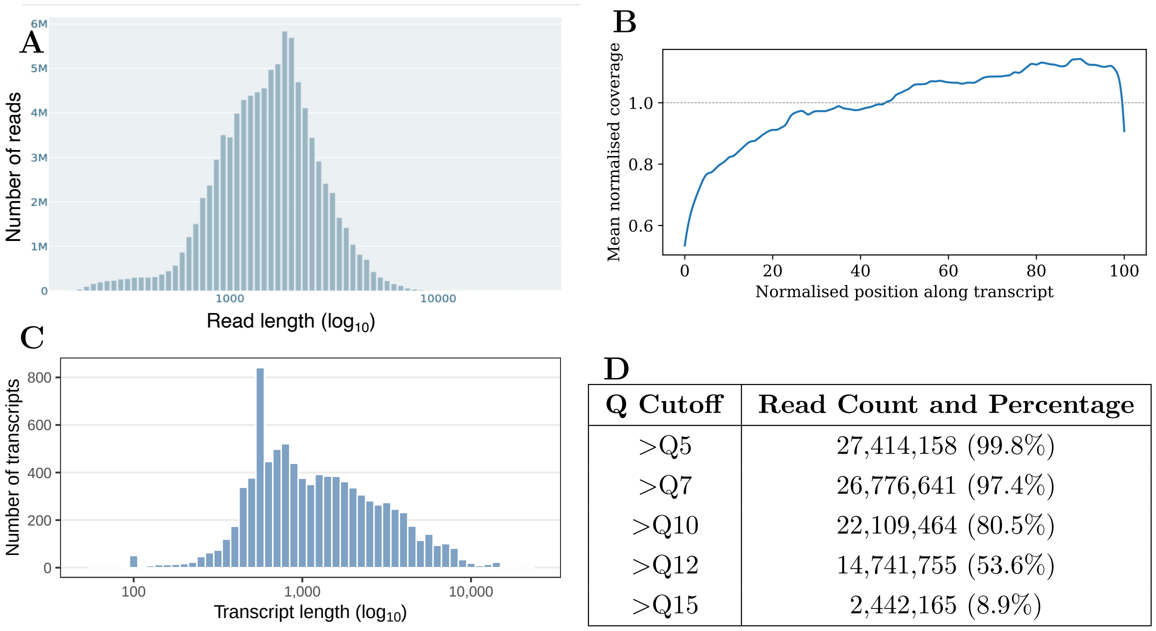


Figure 2.5: **Assessment of capture enrichment based on `Salmon` quantification.** (a) Correlation between estimated transcript abundance and input concentration of synthetic spike-in controls (MixB). Of the 160 spike-ins, 61 were targeted by capture probes (blue) and 99 were untargeted (red). (b) Fold-change in mean expression of protein-coding genes ($\log_2$(TPM)) between capture data and GTEx v8 short-read data from frontal cortex and hippocampus. Genes included in the capture panel are shown in blue, while untargeted genes are shown in red. The control set comprises untargeted genes with elevated expression in brain tissue according to GTEx. (c) Gene class enrichment analysis comparing targeted and untargeted genes. The control group represents untargeted genes normally enriched in brain tissue. (d) Relationship between the number of probes designed per gene and the measured expression level ($\log_2$(TPM)), showing no evidence of correlation.



Q Cutoff	Read Count and Percentage
>Q5	27,414,158 (99.8%)
>Q7	26,776,641 (97.4%)
>Q10	22,109,464 (80.5%)
>Q12	14,741,755 (53.6%)
>Q15	2,442,165 (8.9%)

Figure 2.6: **Read quality and transcript coverage metrics** A) Read length distribution. B) The average relative coverage is shown at each relative position along the transcripts' length. C) Length distribution of annotated isoforms of targeted genes. D) Number and percentage of demultiplexed reads above quality cut-offs

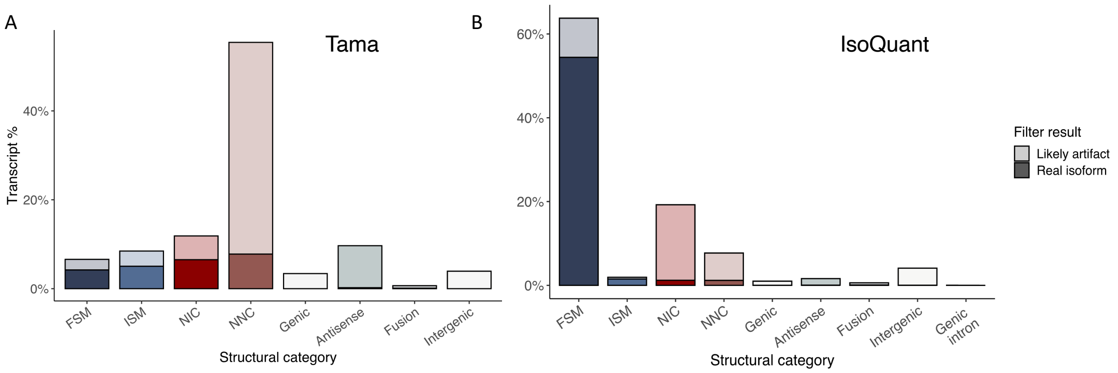


Figure 2.7: **SQANTI3 classification of assembled transcripts from A) TAMA and B) IsoQuant.**

2.3.2 Selection of isoform detection methods

I next generated a custom reference annotation of the human brain, merging data across all samples. Between 16 – 52.4% of quality pass reads were lost during demultiplexing as Guppy could not assign a barcode to them, therefore all basecalled reads were used for custom transcriptome construction. 33,276,536 (96.2%) of reads passed the read quality (Q) ≥ 7 cut-off. A shallow peak at 250bp suggested an abundance of truncated and incomplete partially sequenced reads, resulting from RNA degradation and nanopore current blockage.

To minimise the impact of artefactual truncated reads on transcript assembly, I applied quality control measures at the read-processing stage, rather than relying solely on post hoc filtering. While tools such as SQANTI3 can remove artefactual isoform models after transcript assembly, this approach allows erroneous reads to contribute to false models. Instead, I sought to remove artefacts before transcriptome construction. Full-length reads were first identified and reoriented using Restrander Schuster et al. (2023), which improves transcript identification accuracy, particularly in loci where protein-coding and lncRNA transcripts are transcribed antisense to each other. Restrander was also used to detect and exclude template-switching artefacts, including Template Switching Oligo (TSO)–TSO artefacts and RTP–RTP artefacts, the latter arising from second-strand cDNA priming at internal poly-T tracts. Subsequently, read alignments were further filtered for signatures of reverse transcription artefacts using a custom script.

After processing, 31,293,184 reads (90.5%) were successfully reoriented, with an average length of 1,560 bp. These reads were mapped to the human genome using minimap2 to generate a high-confidence set of splice junctions, followed by remapping, which produced 30,129,485 primary alignments. Of these, 13,485,222 (44.8%) uniquely mapped to the targeted regions.

Recent benchmarking studies (Su et al. (2023); Dong et al. (2022); Pardo-Palacios et al. (2024b); Wang et al. (2023)) have provided valuable insights into the selection of transcript reconstruction tools and reveal that these tools can be broadly categorised into two main groups. Some tools heavily rely on the reference annotation and orthogonal support to correct their transcript models. Others prioritise sensitivity and rely solely on sequencing read support and base quality to correct predictions. This sensitivity can be crucial for capturing transcript diversity present within a sample. However, it's worth noting that these tools may be more susceptible to false positives caused by sequencing errors. The focus of this paper is a set of genes involved in neural development and multigenic disorders. As mentioned above, this set is enriched for long genes with complex structures (Sahakyan and Balasubramanian (2016)). And despite being fundamental for brain function, these genes often have low to medium abundance in the brain. For example, *CACNA1C* generates hundreds of transcripts expressed across a wide range of concentrations, with some reaching over 13,000 nucleotides (Clark et al. (2020a)).

Although even the longest sequins did not exceed lengths of 5-7kb, they provided a reference set of expected isoforms against which artefacts, such as truncations or misassemblies, could be evaluated independently of biological variability.

Each assembled transcriptome was compared to GENCODE reference transcripts (v40) and classified for novelty using SQANTI3. GENCODE/Ensembl annotation was chosen over RefSeq due to its broader coverage of novel splice variants, including intron retention events. Consistent with the sensitivity–precision trade-off, TAMA identified a larger proportion of novel not in catalog (NNC) isoforms, though many lacked orthogonal support or contained flagged features. IsoQuant, in contrast, predominantly recovered reference isoforms (Figure 2.7).

Stringent read artefact filtering by cDNA Cupcake and FLAIR (v1.5.0) run with default parameters led to consistent exclusion of exon 1d, misclassified as a shifted terminal exon due to ONT alignment variability. Cupcake was designed for PacBio data and operates under the assumption that transcript ends are reliable, and therefore fails to collapse reads that only partially cover full-length transcripts and are subsets of previously annotated junctions. Given the high coverage of targeted genes intrinsic to Capture-Seq, and anticipating the generation of newer, higher-throughput datasets, we re-evaluated the performance of cDNA Cupcake and FLAIR (v1.6.2). This decision was motivated by the observation that TAMA exhibited substantial slowdowns or failed to complete transcript reconstruction at loci with extremely high read coverage. FLAIR (v1.6.2) improved handling of truncated isoforms. In this version, isoforms that are subsets of another isoform are filtered out if they are less abundant than the most expressed superset of isoforms for that particular gene. Although not selected for this pipeline, this updated algorithm can be a good alternative to TAMA in the future.

TAMA recovered exon 1d when run in `no_cap` mode, which tolerates greater 5' variability. While TAMA was selected for its strong recovery of rare validated isoforms, its complete reliance on reads for isoform construction and therefore its inability to reconstruct long annotated transcripts (>10 kb) motivated integration with IsoQuant, which can use incomplete reads for reconstruction as long as there is at least one full-length supporting read.

Merging outputs from TAMA and IsoQuant using `tama_merge.py` (without wobble) yielded a high-confidence isoform set, balancing sensitivity and precision, and capturing both rare low-abundance and long, complex isoforms.

2.3.3 Validation and comparison with existing annotations

The collapsed set of isoforms was filtered based on long-read support across samples and support from independent data sources. Novel transcripts of all protein-coding genes were validated by checking for presence of refTSS CAGE peaks within ±500bp of each TSS and polyA peaks from PolyASite atlas within ±50bp of each TTS.

To maximise novel isoform detection, particularly given that most panel genes were lncRNAs (1,080), I systematically tested multiple `minimap2` alignment parameters. The final choice of the `-un` option, which disables canonical splice site matching, yielded 39,447 novel isoforms (10.7% of the total) that passed quality control and were retained for downstream analyses. The filtered transcriptome comprised: GENCODE-annotated isoforms (5.4%), incomplete splice matches (ISM; 1.2%), novel in catalog (NIC; 0.8%), and novel not in catalog (NNC; 92.5%). This approach led to improved sample separation in principal component analysis (PCA). On average, 98.89% of reads mapped to the transcriptome across the 49 samples. For quantification, I used the then-recommended approach of double mapping with `minimap2` followed by `Salmon` in alignment-based mode. Using our updated transcriptome annotation increased the proportion of inter-sample variance explained by the first two principal components (14% and 8%, respectively) and produced clearer clustering by tissue, with the most marked improvement in fetal samples (Figure S4b). This improved separation is partially expected, as quantification was based on capture-enriched reads that predominantly originate from our targeted genes, and the updated annotation includes 39,447 additional isoforms for these targets compared with GENCODE, many of which are enriched in fetal tissue, a less accessible sample type.

Of the novel transcripts observed, 13,646 exhibited protein coding potential according to SQANTI, with at least one complete putative ORF. However, a subset of transcripts aligning to annotated lncRNA genes was predicted to be protein-coding by the GeneMarkS-T. Therefore, the coding potential of isoforms that were identified as alternatively spliced in the downstream analysis was recalculated using CodAn Nachtigall et al. (2021). Interestingly, the GENCODE protein-coding transcript translation sequences FASTA contained isoforms of lncRNA genes.

Manual inspection of isoform structures for protein-coding genes of interest and coding sequence prediction for downstream structural modelling with AlphaFold revealed that, in some cases, misaligned junctions, in cases where the end of one exon and the start of the next were sufficiently similar, interrupted open reading frames. To address this, the entire analysis was re-run using `minimap2` with the `-ub` option, which matches canonical junctions and improves alignment.

This updated approach detected transcripts for 1,055 target genes. Among the genes for which no confident isoforms were identified after enrichment, 372 were annotated as lncRNAs and 17 as protein-coding genes. In several protein-coding cases, transcript models were present in the TAMA output but were excluded during quality control. For example, five predicted isoforms were detected for *GATA3*, but all were filtered due to insufficient 3' end support, no poly(A) site or poly(A) motif within 50bp. All mono-exonic transcripts (n = 17,019) were excluded from downstream analyses. The largest fractions of these belonged to the antisense (31.0%) and genic genomic (24.6%) categories, with manual inspection indicating that the latter largely represented unspliced transcripts. Within the 4,823 mono-exonic transcripts classified as FSM, ISM, or NIC, 4,184 (86.7%) overlapped the annotated 3' terminal exon of annotated transcripts of their corresponding genes, and 1,019 (21.1%) were entirely contained within the terminal exon. These most likely represent the result of RNA degradation, which starts from the 5'end. A total of 848 of these short transcripts also overlapped or lay within 50 bp of CAGE peaks; however, these signals might be attributable to 3' UTR re-capping events rather than genuine transcription start sites (Haberman et al. (2023)). Although mono-exonic transcripts are often excluded due to challenges in distinguishing genuine isoforms from artefacts, it is important to note that 4.9% of human protein-coding genes are mono-exonic. Across the capture panel, 98 of the 1,444 detected genes were represented exclusively by mono-exonic transcripts, all of which corresponded to lncRNAs (S1). Thus, while a more nuanced filtering strategy may be appropriate in other contexts, here we applied strict exclusion.

After SQANTI filtering, a total of 111,443 novel isoforms (89.4% of all detected) were retained for downstream analyses (Figure 2.8A). The final combined transcriptome comprised the following isoform categories: GENCODE-annotated isoforms (10,736, 8.6%), incomplete splice matches (ISM; 2,325, 1.9%), novel in catalog (NIC; 14,337, 11.5%), and novel not in catalog (NNC; 96,719, 77.9%).

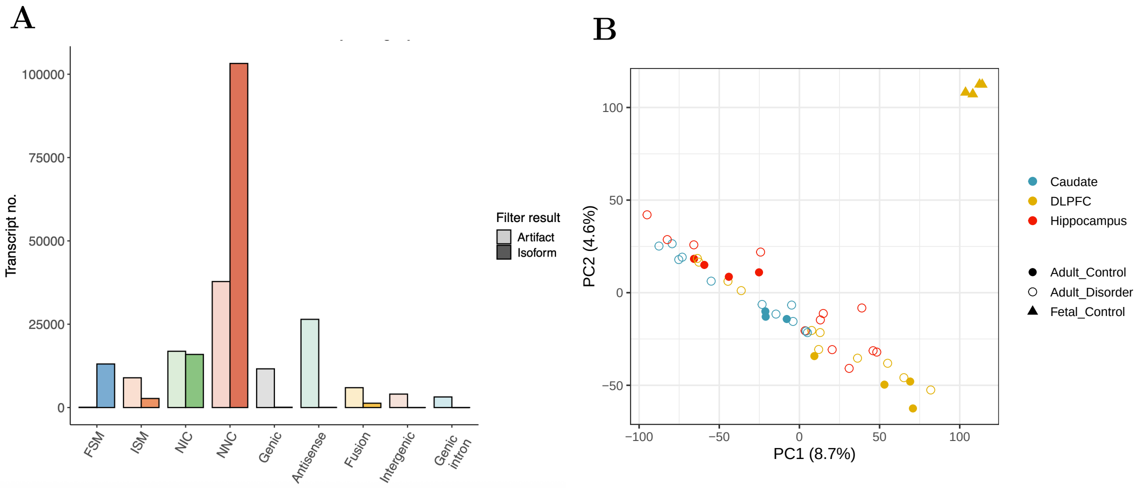


Figure 2.8: **Characterisation of isoforms and transcript-level variation in the combined transcriptome** A) Total isoforms and artefacts by SQANTI category. B) Transcript-level principal component analysis.

Only 1,940 FSM transcripts matched the start and end coordinates of their associated reference transcript. The high prevalence of NNC isoforms is consistent with previous observations that cDNA ONT data contain more NNC junctions, whereas PacBio cDNA data yield more reference-matching splice donors and acceptors Keil et al. (2024). In this study, we did not pursue analysis of the 1,232 putative fusion isoforms, many of which spanned adjacent genes and were likely artefacts arising from transcript overlap or chimeric cDNA molecules; instead, only isoforms corresponding to target genes were retained.

Given this large number of novel isoforms detected, I sought to characterise their patterns of expression and usage across samples. By the time of reanalysis, the pipeline was updated to be more efficient and used `kallisto` for long reads, which is equivalent to `Salmon` in quasi-mapping mode. The resulting PCA, although different as both mapping and quantification methods were different, again showed clear separation of fetal samples, and control samples separated well by brain region, whereas disorder samples exhibited greater overlap (Figure 2.8 A). In this analysis, the variance explained by the first two principal components was lower (Figure 2.8 B, 8.7% and 4.6%), which may partly reflect additional biological variation in the dataset, such as donor sex and a great variation in adult donor ages (21–76 years). Although of note that all brains were selected for the absence of overt ageing pathology.

After applying an expression filter requiring non-zero counts in at least three samples, the isoform set was reduced to **61,117** transcripts, of which **56,249** were classified as novel.

2.3.4 Isoform switches in control brains

To investigate region-specific splicing patterns in healthy human brain tissue, I performed DTU analysis on control samples. Surrogate variable analysis identified two hidden factors, which were accounted for by including the sequencing run number in the model matrix. This adjustment could only be applied to the fetal–adult comparison, as the inter-regional comparisons did not have sufficient biological replicates per group to accommodate the additional covariate.

Prior to DTU testing, lowly expressed and single-isoform genes were removed using pre-filtering criteria of a minimum gene expression of 3 TPM in all tested samples and no minimum isoform expression threshold. This filtering excluded 57,779 transcripts (90.68%), leaving 5,940 isoforms for the fetal–adult comparison. For inter-regional comparisons, 50,746 transcripts (88.07%) were removed, resulting in 6,871 isoforms for analysis. Fewer transcripts were excluded in the latter case as these comparisons involved only adult samples, where gene expression levels are more uniform.

We identified 90 significant isoform switches across three adult brain region comparisons; $p < 0.05$). However, the majority (214) of switching events were found between fetal and adult splicing in the DLPFC, in agreement with previous findings that the majority of transcriptomic changes in the brain occur in the pre to postnatal transition. Of the 268 isoforms involved in these developmental switches, 120 (44.8%) were novel (Figure 2.9A). Given that fetal transcriptomes are typically less well annotated, and that PCA with the updated annotation showed clearer separation of fetal samples, we hypothesised that novel isoforms would contribute disproportionately to fetal expression. The median proportion of counts from novel isoforms was slightly higher in fetal DLPFC (57.5%) than in adult (52.5%), but the distribution was largely similar (Figure 2.9B).

Although their enrichment in the samples was not significantly increased with probe capture, a proportion of significant splicing switches included lncRNA genes. For example, a NNC transcript of *LINC01135* (JUN-DT, JUNI) was differentially used between adult Caudate and DLPFC.

Table 2.2: Differential isoform switching in control brains

Comparison	nrIsoforms	nrSwitches	nrGenes
Caudate vs DLPFC	44	45	37
Caudate vs Hippocampus	5	5	4
DLPFC vs Hippocampus	28	40	24
Adult vs Fetal (with Run)	90	111	77
Adult vs Fetal	268	214	190

Across genes, isoform usage generally followed the well-described "major isoform" pattern (Tung et al., 2022), whereby a single isoform accounts for most of a gene's expression and the remaining isoforms contribute progressively less, often in an exponential decay pattern (Figure 2.9C). Importantly, the identity of the dominant isoform was not fixed but varied across tissues and developmental stages. We identified 67 fetal–adult isoform switches and 25 inter-regional switches that specifically involved a change in the dominant isoform, either through a switch with a different transcript or through a major shift in isoform fraction.

Examining selected genes with significant switches (Figure 2.9D) illustrates these distinct patterns. For *CACNA1G*, the major isoform identity changes but the overall distribution remains similar between fetal and adult. In contrast, both *TMEM161B* and *CACNB3* showed markedly higher overall gene expression in fetal brain compared with adult, with the majority of this fetal–adult difference driven by the dominant fetal isoform. For *TMEM161B*, the major isoform contributed nearly 80% of total expression in fetal samples but was substantially reduced in adults, resulting in a more evenly distributed isoform profile. A similar pattern was observed for *CACNB3*, where the same isoform remained dominant across development but decreased from 67% of total expression in fetal brain to 10% in adults.

A comparison of switching results obtained with different mapping and quantification strategies revealed limited overlap. Using minimap2 without the GT-AG matching preset and Salmon for quantification, versus minimap2 with the preset and kallisto for quantification, only 5 out of 63 inter-regional comparisons overlapped with the original results. Similarly, in the fetal–adult comparison, only 47 out of 190 switching genes were consistent across analyses. Although isoform detection was performed with both IsoQuant and TAMA, the choice of quantification framework had a strong impact on the final set of significant switches. This observation aligns with broader benchmarking efforts, which consistently demonstrate that software choice has a substantial influence on isoform detection and quantification. For example, results from the LRGASP consortium showed very limited overlap in the set of transcripts identified by any two pipelines (Pardo-Palacios et al., 2024b), underlining the challenges of achieving consensus in isoform-level analyses.

2.3.5 Capture-Seq identifies candidate isoform switches in psychiatric brains.

Building on previous findings that established a correlation between splicing dysregulation and multiple psychiatric disorders, including ASD, BD, and SCZ (Gandal et al. (2018)), we took the full advantage of the dataset to assess the specific isoform population in BD, SCZ and MDD brains. To account for regional heterogeneity in transcript expression, the isoform abundance data were stratified by brain region, and pairwise comparisons were performed between each disorder and matched controls. While principal component analysis (PCA) revealed no marked separation at the gene expression level between diagnostic groups (Figure S5a), DTU analysis identified 14 significant isoform switching events in 12 genes in SCZ, 17 in BD (13 genes), and 18 in 17 genes in MDD (Figure 2.10). These events were not uniformly distributed across brain regions. For example, no significant DTUs were observed in the hippocampus for SCZ, whereas the majority of DTUs in BD were hippocampus-specific. However, given the modest sample size and associated limitations in statistical power, these differences in the number and distribution of significant events should be interpreted cautiously. It is plausible that some of the observed variation reflects residual technical confounders, such as sequencing run and barcode batch effects and limited within-condition replication, rather than underlying biological differences.

Of particular note, in the DLPFC, we identified two novel isoforms of lncRNA genes that were differentially used between healthy and psychiatric disorder brains. A NNC isoform G63625.20.nnc of *LINC01004* was specifically enriched in MDD samples, while the isoform G66627.40.nic of *LINC01299* was absent in SCZ samples. These findings highlight the potential for long-read sequencing to uncover condition-associated isoforms of poorly characterised lncRNA loci, which remain under-annotated in current transcriptome references. Among other significant events detected, in the hippocampus novel isoform `G29432.125.nnc` carries a PTC, which could explain the reduced overall expression of *CACNG4* observed in BD samples, potentially through NMD (Figure 2.11). *CACNG4* encodes a γ-subunit of the VGCC complex, a class of proteins with established relevance in neuropsychiatric disease through their roles in synaptic transmission and neuronal excitability.

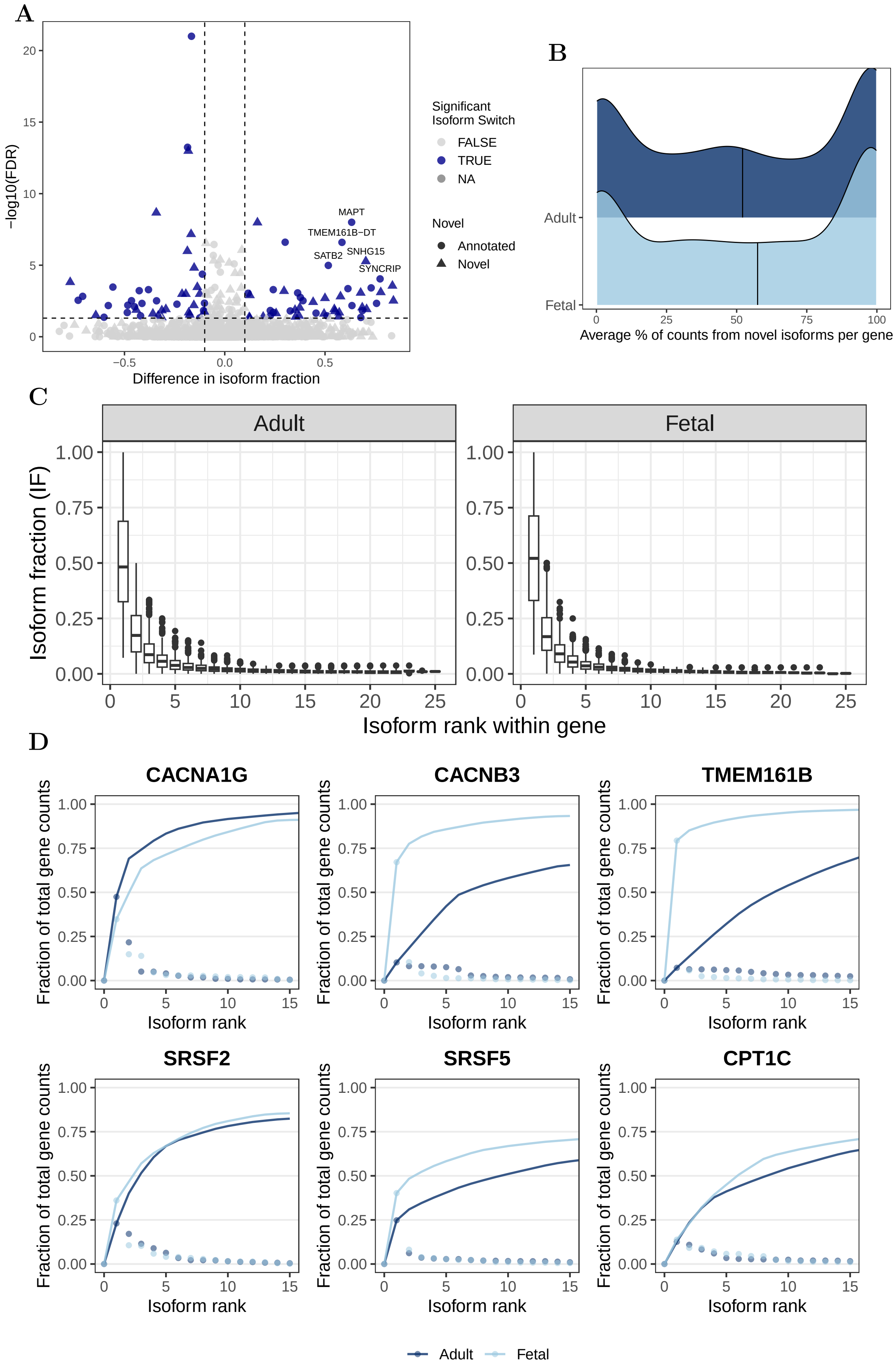


Figure 2.9: **Isoform usage and novelty in DLPFC across development.** A) Differential isoform usage (DTU) between fetal and adult DLPFC. B) Average percentage of counts per gene arising from novel isoforms. Vertical lines indicate group medians. C) Isoform-fraction (IF) rank profiles in DLPFC. Isoforms were ranked within genes by IF; box plots show the IF distribution for each rank, illustrating that most genes have a dominant isoform and a rapidly decaying contribution from lower-ranked isoforms. D) Cumulative distribution of isoform fraction of total gene reads from highest (left) to lowest (right).

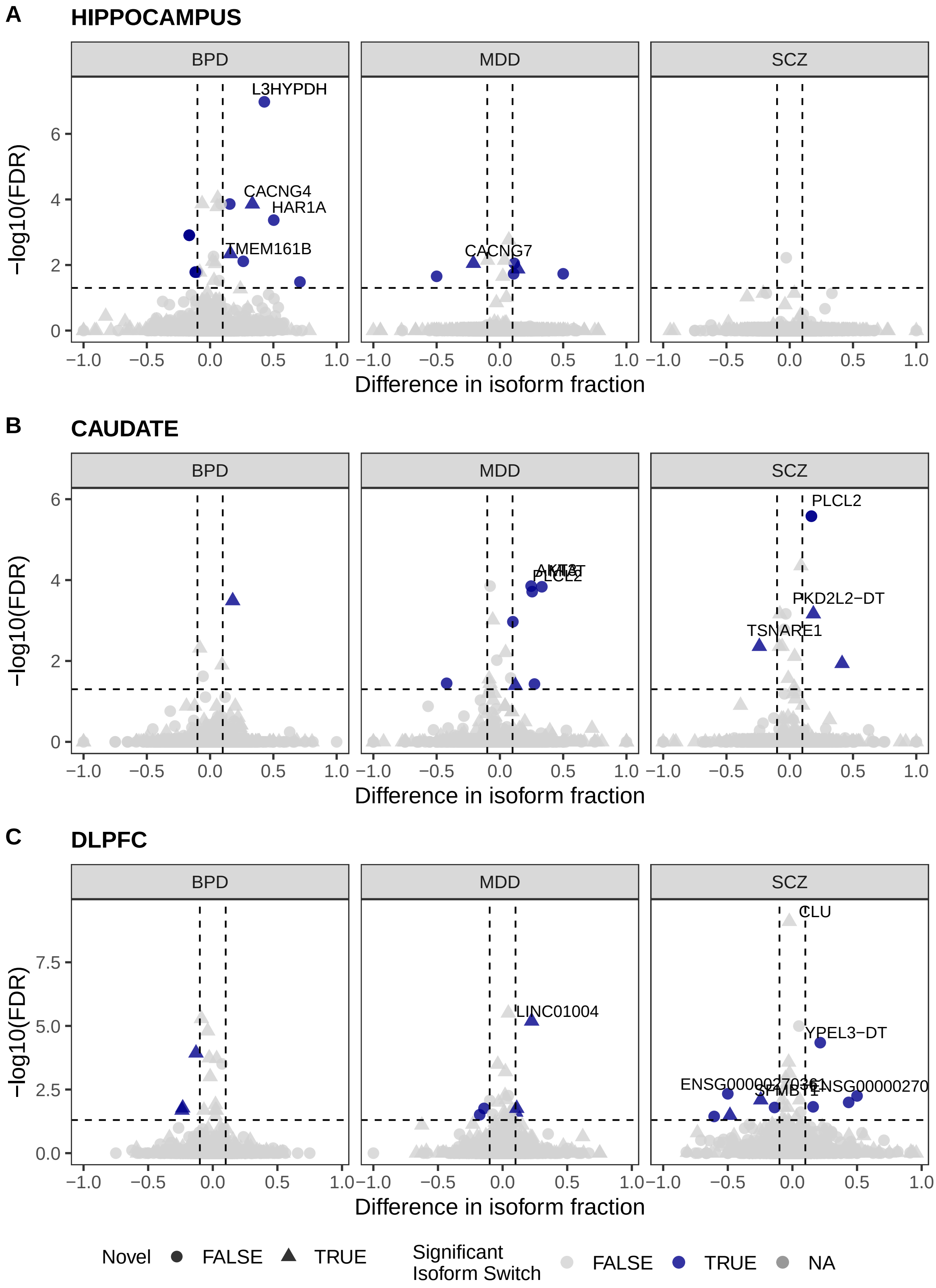


Figure 2.10: **Isoform switching events across brain regions in psychiatric conditions.** Differential transcript usage was assessed between each disorder and control within three brain regions: (A) Hippocampus, (B) Caudate, and (C) Dorsolateral Prefrontal Cortex (DLPFC). Conditions tested include bipolar disorder (BPD), major depressive disorder (MDD), and schizophrenia (SCZ). Each point represents an isoform, with significant isoform switches (FDR < 0.05, $|dIF| > 0.1$) highlighted in dark blue. Novel isoforms (NIC, NNC, or ISM) are indicated by triangle shapes. Lowest FDR isoforms are annotated with their corresponding gene symbol.

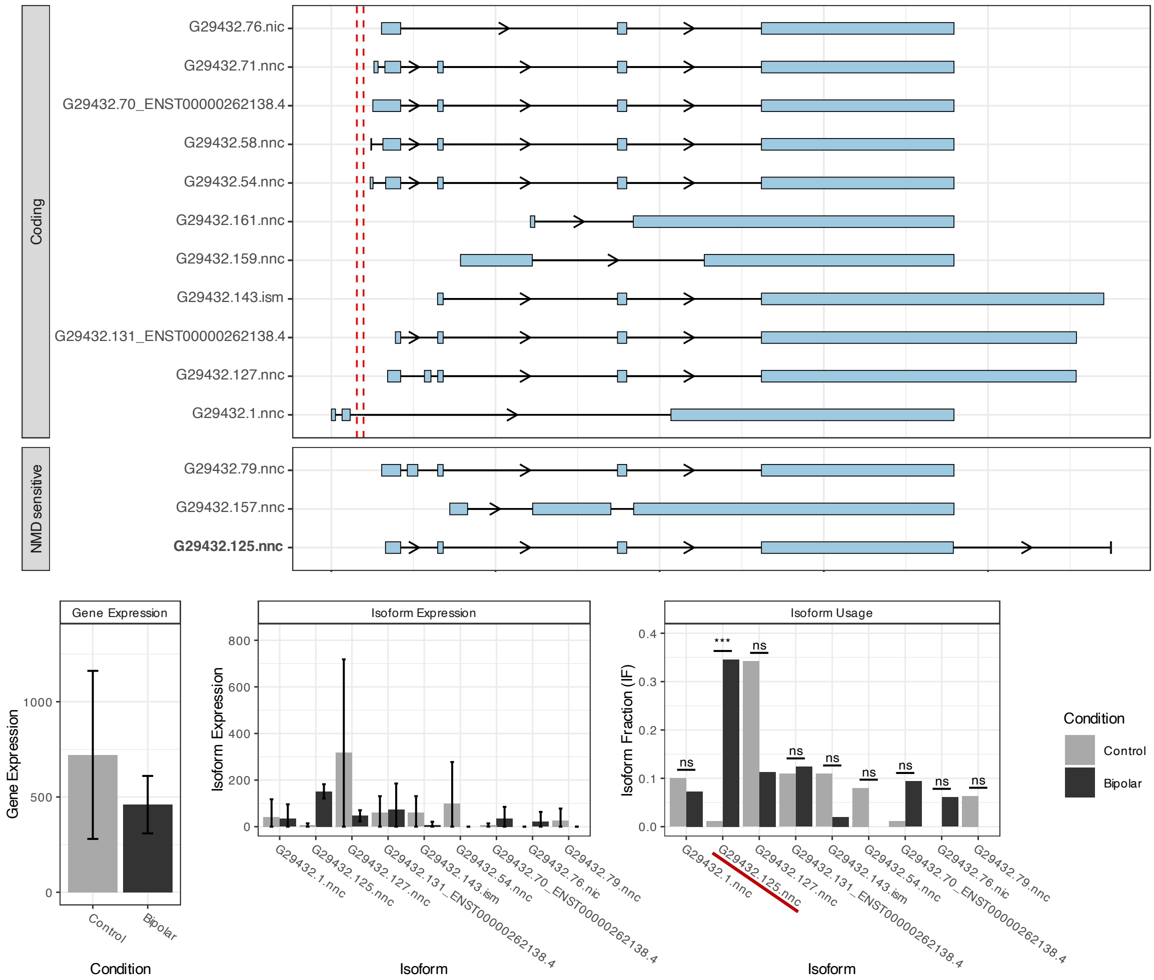


Figure 2.11: **CACNG4 isoform switching in the hippocampus of BD donors.** Exons are shown as boxes. Variant positions are indicated by red dashed lines. The isoform with significant differential usage is highlighted in bold

2.4 Discussion

In this study, I developed a comprehensive computational pipeline to support high-confidence isoform annotation from targeted long-read Capture-Seq data. The primary objective of the pipeline was to enable accurate detection of previously unannotated transcripts, while addressing the known limitations and technical artefacts inherent to Oxford Nanopore data. Capture-based enrichment results in high, localised coverage across a predefined set of genes of interest. This offers increased power to resolve complex splicing patterns of lowly expressed genes, but also introduces computational challenges, particularly in handling high read depth and artefactual transcript models. To mitigate these issues, the pipeline incorporated several layers of filtering, correction, and validation. For example, primary alignments arising from potential template-switching artefacts were removed based on their CIGAR string structure. Novel transcript models were excluded unless they could be validated by orthogonal datasets.

The resulting transcriptome contains 56,249 novel isoforms across 1,055 targeted genes. This represents a substantial expansion of the known transcriptomic landscape in the human brain and demonstrates the utility of long-read Capture-Seq in resolving transcript diversity.

2.4.1 Balancing sensitivity and specificity in isoform detection

In this study, I focused on using the high coverage Capture-Seq data produced by our collaborators to detect low-abundance isoforms and therefore prioritised sensitivity over precision. This choice initially motivated the use of a more permissive mapping strategy. However, manual inspection of the *CACN* gene family revealed spurious splice junction assignments, prompting a switch to a mapping approach with stronger bias towards canonical junctions. While the revised pipeline remains relatively sensitive, this experience illustrates why researchers might alternatively favour more stringent filtering strategies, particularly those incorporating short-read support.

A representative example of the trade-offs involved is provided by *CACNG4*. Among the novel isoforms detected, **G29432.1.nnc** illustrates both the potential value and the limitations of our approach. Although this isoform was not classified as differentially used, it could have been of interest because it recontextualises two intergenic variants previously associated with bipolar disorder (rs17645023 and rs9905054), situating them within intronic regions of a transcribed locus (Figure 2.11). The model, however, was supported by a single spliced read and was retained solely due to the presence of a CAGE peak 17 bp downstream of the predicted TSS, which did not overlap any annotated gene or transcript start site. Even if genuine, this transcript spans seven genes and is more likely to represent a readthrough, as its TSS lies 35,091 bp upstream of any annotated *CACNG4* start site.

Such readthrough transcripts, arising from consecutive genes on the same strand rather than interchromosomal fusion events, were widespread in our dataset. In total, we observed 2,799 isoforms from 416 genes with transcription start or end sites located more than 5 kb away from the nearest annotated boundary and overlapping multiple genes, which are features strongly suggestive of readthrough transcription. While some of these may reflect genuine biological phenomena, consistent with recent reports that readthroughs are particularly prevalent in the healthy human brain (Caldas et al., 2024), others are plausibly artefacts introduced during targeted capture. Dedicated tools have recently been developed to systematically identify and characterise such events (Chen et al., 2024a), but we did not pursue this analysis further in the present work.

These observations underscore both the value and limitations of highly sensitive approaches: they may preserve potentially novel and disease-relevant isoforms, but at the risk of retaining artefactual models. To accommodate different analytical needs, I generated both SJ–filtered and unfiltered novel transcript annotations. In this study, I used the transcriptome where short-read data was used to only filter novel isoforms that were not ISM, NIC, or NNC. This choice was informed by cases such as a 35-exon CACNA1C exon 1d isoform, which was excluded solely because junction 34 lacked short-read support, despite retention of shorter isoforms with 6–8 exons. Together, these findings emphasise the importance of manual curation and context-specific filtering strategies in long-read isoform studies.

2.4.2 Downstream utility of isoform annotations

Although the pipeline presented here focuses on transcript detection and annotation, the downstream utility of these isoforms extends to multiple layers of functional analysis. For example, improved isoform annotations can enhance peptide–isoform matching in proteomics, increasing the sensitivity of mass spectrometry-based detection of functionally relevant protein products. Conversely, proteomic evidence can in future be leveraged to prioritise novel isoforms for experimental validation and functional studies (Abood et al., 2024; Mehlferber et al., 2022). In ongoing collaborative work with Dr Wilsenach and Dr Ahnert, AlphaFold is being applied to switching isoforms identified in this chapter to assess how alternative transcript sequences translate into alterations in protein structure, with potential consequences for stability or function (Figure S6). In parallel, tools such as TRIFID, which integrate evolutionary conservation and protein feature annotations to predict isoform functionality (Pozo et al., 2021a), offer complementary strategies for prioritising candidate isoforms.

2.4.3 Scalability and future directions

It is important to note that due to the heterogeneity and polygenicity of neuropsychiatric conditions, large sample numbers are typically required to identify statistically significant disease-associated transcriptomic changes (Tesfaye et al. (2024)). In this study, I focused on developing a robust annotation framework rather than performing large-scale differential expression analysis. Nonetheless, the methods described here are scalable and can be applied to larger datasets as they become available.

As long-read sequencing technologies continue to improve in terms of accuracy and throughput, the stringency of current transcriptome filtering steps may be relaxed. The pipeline described in this work provides a flexible and modular foundation for future studies of isoform-level regulation. All code and documentation are made publicly available through a dedicated GitHub repository (https://github.com/skudashev/Capture-analysis), allowing the workflow to be reproduced and adopted by the wider community.

Chapter 3

Splicing heterogeneity across hiPSC-derived brain cell types

This chapter was done in collaboration with Dr Andrew Bassett and Dr Sarah Cooper, Cellular Operations, Wellcome Sanger Institute, who generated the human induced pluripotent stem cell (iPSC)-derived cell lines and performed the 10x single-cell library preparation and Oxford Nanopore sequencing. The results of my analysis were benchmarked against the output of Pacific Biosciences sequencing, which was analysed by Francisco Cervilla-Martinez as part of the collaborative project. All bioinformatic analyses presented here were performed by the author.

The supplementary tables and custom scripts produced for this chapter are freely available at `https://github.com/skudashev/scLR-isoform-analysis`.

3.1 Introduction

Long-read sequencing technologies enable the resolution of full-length isoforms, while single-cell approaches offer insights into cell-type-specific expression and cellular heterogeneity. However, new computational methods are needed to address the unique challenges posed by these emerging technologies. The overarching aim of my thesis is to benchmark and adapt existing analysis methods to extract reliable biological conclusions from long-read RNA sequencing data.

In Chapter 2, I introduced a pipeline for isoform annotation and quantification using targeted Oxford Nanopore Capture-Seq on bulk RNA from post-mortem human brain regions. The approach enabled sensitive detection of lowly expressed isoforms and differential isoform usage across regions and conditions. However, bulk measurements average transcript abundance over millions of cells. As a result, they can obscure cell-type-specific isoform regulation, particularly in rare populations that contribute little to the mean signal. For many questions about heterogeneous tissues such as the brain, this masking effect limits biological interpretability.

Historically, a combination of morphology, electrophysiology and laminar position was utilised to delineate the major brain cell classes: excitatory and inhibitory neurons, vascular cells and glia that include oligodendrocytes, oligodendrocyte progenitor cell (OPC)s, microglia, astrocytes and ependymal cells. Modern single-cell transcriptomics now resolves much finer granularity and helps determine functionally relevant cell states and subpopulations based on their expression profiles (Darmanis et al., 2015). For example, the cerebral cortex has a layered and regionally specialised architecture with marked differences in cellular composition across laminae and areas. Recent large-scale atlases report extensive neuronal and macroglial diversity, with up to roughly 250 transcriptionally defined cell subtypes described in the primate brain (Chen et al., 2023), and fine-grained maps of human DLPFC highlighting spatially patterned gene expression across layers (Maynard et al., 2021). Against this backdrop of cellular complexity, it is likely that isoform programmes vary not only between brain regions but also within regions between major brain cell subtypes and cortical layers. Bulk assays cannot resolve such structure, which makes them underpowered for questions that require attribution of isoform usage to specific cell populations.

This limitation is particularly relevant for neurodevelopmental and psychiatric disorders where genetic risk and molecular pathology are enriched in defined cell types. For example, schizophrenia associations show over-representation among genes highly expressed in excitatory glutamatergic neurons of the cortex and hippocampus (including pyramidal CA1 and CA3 cells), in cortical inhibitory interneurons, and medium spiny-like neurons from the amygdala (Trubetskoy et al., 2022; Duncan et al., 2025).

In Chapter 3, I therefore utilise single-cell sequencing of models for brain cell types to characterise isoform and gene expression at cellular resolution.

3.1.1 Single-cell RNA sequencing

A variety of single-cell RNA sequencing (scRNA-seq) methods are available, with the most widely recognised being Smart-seq2 (Picelli et al., 2014) and 10x Genomics Chromium (Zheng et al., 2017).

Smart-seq2 is a plate-based approach that uses fluorescence-activated cell sorting (FACS) to isolate individual cells into wells of microtiter plates, followed by full-length cDNA synthesis and amplification. This approach provides better coverage across transcripts and high sequencing depth per cell, enabling detection of lowly expressed genes and isoforms, but is limited by higher cost and lower throughput compared to droplet-based methods.

10x Genomics Chromium is a droplet-based platform that encapsulates individual cells in oil droplets containing a single barcoded gel bead Klein et al. (2015). Each bead is preloaded with barcoded oligonucleotides, consisting of a Cell Barcode (CB) and a Unique Molecular Identifier (UMI). These are incorporated into cDNA during reverse transcription, allowing each read to be traced back to its cell and molecule of origin. This early multiplexing enables high throughput and lower cost per cell. However, the trade-off is lower sequencing depth per cell and the fact that standard short-read 10x libraries capture only the 3' or 5' end of transcripts. Therefore, any inference into alternative splicing using short-read scRNA-seq has been limited to alternative polyadenylation and 3'end splice variants (Fansler et al. (2024); Kang et al. (2023); Gao et al. (2021); Ake et al. (2025)).

3.1.2 Long-read scRNA-seq

While Smart-seq2 is designed to capture full-length mRNA and therefore offers improved isoform coverage compared to 10x short-read data, it still relies on short-read sequencing. The fragmentation inherent to short-read platforms makes isoform reconstruction indirect and computationally challeng-

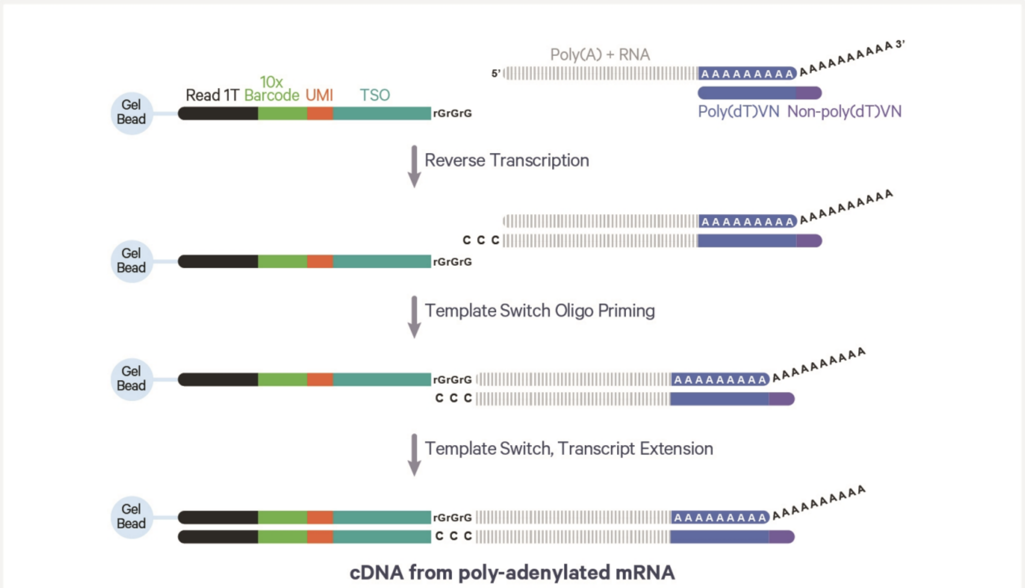


Figure 3.1: **Overview of RNA capture and cDNA synthesis workflow using the Chromium Next GEM Single Cell 5' Kit v2.** Diagram adapted from Oxford Nanopore Technologies nanoporetech.com

ing, limiting the resolution of cell-to-cell isoform heterogeneity and alternative splicing dynamics. To address this challenge, the 10x protocol has been adapted to be used with long-read sequencing. For ONT, the original 3' protocol was adapted from Lebrigand et al., which used primers with Ns at the 5' ends (Lebrigand et al., 2020). In this approach, single-cell cDNA generated by the 10x Chromium is amplified with primers that contain short stretches of degenerate bases ("N" spacers) at the 5' ends. These heterogeneity spacers reduce sequence uniformity, thereby limiting unintended complementarity to the TSO and reducing reverse-strand bias. In the 3' workflow, a biotin pull-down step is included to deplete TSO–TSO artifacts that arise when the TSO is incorporated at both ends of a cDNA molecule. In the 5' protocol this step is unnecessary, as the TSO is already embedded within the gel-bead primer together with the sequencing primer, CB, and UMI (Figure 3.1). Amplified cDNA is then adapted for ligation with ONT sequencing adapters. For PacBio, 10x barcoded cDNA is converted to SMRTbell libraries and sequenced with circular consensus (HiFi) to retain the cell barcode and UMI per read. To boost throughput, PacBio's Kinnex (previously MAS-Seq) concatenates multiple (usually 16 for single-cell applications) cDNA inserts into a single molecule before SMRTbell ligation, so each HiFi read contains multiple transcript units (Al'Khafaji et al., 2024; Biosciences, 2023).

Many more studies have since implemented 10x library preparation in combination with both ONT and PacBio sequencing to profile isoforms, alternative UTR usage, gene fusions and structural variants in single cells (Jin et al. (2022); Dai et al. (2023); Joglekar et al. (2023); Mincarelli et al. (2023)).

3.1.3 Technical challenges in long-read scRNA-seq and methods to address them

Experimental and technical artefacts inherent to cDNA sequencing, such as errors introduced during reverse transcription and PCR, have amplified consequences in single-cell applications, where input material is limited.

Lower accuracy

The lower accuracy of long-read platforms, particularly ONT, not only affects the reliable detection of splice junctions but also poses significant challenges for accurate identification and correction of CBs and UMIs, which are essential as the CB ensures reads are correctly assigned to their cell and the UMIs allow for accurate quantification. Experimental methods such as synthesis of barcodes and UMIs using homodimers, sequencing the same read multiple times to concatemeric consensus (R2C2) and using matched Illumina datasets have been employed to correct these sequencing and basecalling errors (Philpott et al. (2021); Volden et al. (2018); Volden and Vollmers (2022); Tian et al. (2021). Although effective, these approaches increase costs and prolong the data generation process. With the recent improvements in error profiles of nanopore reads, computational methods that can be used to recover CBs and UMIs without the reliance on additional short-read sequencing runs, such as BLAZE and scNanoGPS have been developed (You et al. (2023); Shiau et al. (2023).

Amplification bias

As outlined in Chapter 1 (1.5.2), amplification is an additional source of technical uncertainty. Single-cell long-read sequencing is especially vulnerable because of its low throughput and the ultra-low amounts of input RNA. For example, Gupta et al. (2018) reported that since roughly 1 µg of cDNA was required per Minion flow cell, extra PCR cycles would be necessary when starting from single cells. Such limited input leads to uneven, stochastic sampling of templates during reverse transcription and PCR, introducing amplification bias. Unique molecular identifiers (UMIs) can partly mitigate these effects by enabling removal of PCR duplicates from transcript counts and by informing models of amplification bias (Islam et al., 2014).

Capture inefficiency and zeros

Whilst useful for amplification bias, UMIs cannot resolve underlying sampling biases nor can they be used to detect the origin of “dropout” whereby the abundance of a gene can be recorded as zero due to true lack of biological expression or to experimental factors such as low cell sequencing coverage or inefficient mRNA capture Grün et al. (2014). Unfortunately, this is an even bigger issue for isoform-level quantification since individual transcripts are inherently expressed at a level less than or equal to the genes that they originate from. Although droplet scRNA-seq data appears robust to technical “dropout” or “zero inflation”, it is still rich in stochastic sampling zeros, which should be taken into consideration during analysis, as this can provide useful biological information Svensson (2020); Jiang et al. (2022). The sparsity in single-cell datasets also leads to an almost linear relationship between sequencing depth and library complexity, affecting not only the inferred gene expression levels, but also the variation in genes detected. Methods developed for bulk RNA-seq, such as limma and DESeq, that were discussed in the previous chapter, can be used to normalise raw read counts by applying the same transformation to each count to adjust for differences in sequencing depth between cells within the sample and make the data more normally distributed Cole et al. (2019). Initially, Scanpy and Seurat employed this global-scaling strategy, where gene or transcript counts for each cell are divided by the total counts, multiplied by the fixed scale factor (10,000 by default), and then the result is log-transformed. To avoid undefined values when attempting to log zero, a pseudo-count (1 by default) is added to all counts. Log transformation reduces the disproportionate influence of high-abundance genes on differential expression analysis. However, it is less effective in cells with low total UMI counts.

As global transformations assume that most genes are not differentially expressed between cells and each cell contains the same number of RNA molecules, they result in an imbalance in the effect of the correction on genes with different abundances, where lowly expressed genes tend to be over-corrected Bacher et al. (2017). To address this, *sctransform v2* in Seurat models UMI counts for each gene using a negative binomial generalised linear model (GLM), with sequencing depth as a covariate. Gene-specific parameters (intercept, slope, and dispersion), are initially estimated and then regularised by smoothing against neighbouring genes with similar average expression. This reduces sensitivity to sampling noise and prevents overfitting, particularly for lowly expressed genes Hafemeister and Satija (2019); Choudhary and Satija (2022). The resulting parameters are used to transform the UMI counts into Pearson residuals, which represent the difference between each observed count and estimated mean expression, and serve as normalised expression values. These residuals reflect the difference between observed and expected counts, scaled by the gene's estimated variance, thereby accounting for both mean expression and overdispersion.

Batch effects

Another major consideration in single-cell studies is batch effects caused by samples being processed separately, making it difficult, and potentially even impossible, to distinguish true biological differences from experimental artifacts. While batch effects are also an issue in bulk RNA-seq data, they are much more problematic in single-cell data due to the heightened sensitivity to technical variance. In single-cell RNA-seq, each cell's data is treated individually, leading to greater variability and the potential for significant biases. These are especially difficult when the batches represent different conditions and contain different cell types, confounding biological variation with unwanted technical variability. An example of technical variability can be differences in detection rates between sequencing runs, which have been found to correlate with estimated distances between cells and therefore lead to false cluster discoveries (Finak et al. (2015); Hicks et al. (2018)). While there are some experimental strategies that can minimise batch effects, such as multiplexing sample libraries across flow cells, there are often unavoidable biases that can only be resolved computationally. In recent years, several *in-silico* methods have been developed to remove batch effects. These methods can be divided into two categories based on the input data: (1) dimensionality-reduced data, and (2) the original gene expression matrix. Some of the most popular ones include the Mutual Nearest Neighbours (MNN) algorithm, a PCA-based clustering method implemented in the package HARMONY, and a method that combines canonical correlation analysis (CCA) and MNN implemented in Seurat (Butler et al. (2018); Stuart et al. (2019); Korsunsky et al. (2019a). Seurat and MNN work by learning a shared population structure, whereby MNN pairs cells from different batches that belong in each other's set of nearest neighbours and assumes that they belong to the same cell type. It then uses the difference between the cells in these pairs to estimate the size of the batch effect and outputs an expression matrix with the same dimensions as the raw input counts (Haghverdi et al. (2018)). Harmony instead identifies cluster-specific correction vectors in a low-dimensional embedding by iteratively aligning shared structure across batches. More recently, deep learning models have been developed. scVI uses a variational autoencoder to model gene expression while conditioning on batch labels Svensson et al. (2020). DESC goes a step further by using a stacked autoencoder so as to perform simultaneous batch effect correction and clustering Li et al. (2020).

Stochasticity

Another source of heterogeneity in transcript expression between isogenic cells is intrinsic biological noise, which, while potentially complicating downstream analyses, also presents an opportunity for mechanistic insight. Single-molecule fluorescence microscopy studies revealed that transcription and translation in individual cells are highly stochastic and occur in sporadic "bursts", potentially due to each cell having low numbers of regulatory proteins expressed at any one time Elowitz et al. (2002); Zhang et al. (2024). The bursting kinetics of a particular gene are regulated by the local chromatin environment and cis-regulatory elements Dar et al. (2012); Ochiai et al. (2020). At the same time, most highly expressed protein-coding genes have a dominant isoform expressed in any given tissue Ezkurdia et al. (2015). Long-read scRNA-seq enables us to start investigating whether certain genes deviate from this general pattern by displaying greater variability in the relative usage of isoforms across individual cells. Such analyses could reveal genes for which isoform choice is more susceptible to stochastic regulation or responsive to subtle shifts in the regulatory environment. scRNA-seq protocols are also known to introduce high numbers of intronic internally primed reads when compared to bulk sequencing (Haque et al. (2017)).

3.1.4 Biological context: iPSC-derived models of the developing human brain

Microglia, astrocytes, and neurons each have distinct transcriptional programmes and perform specialised functions in the central nervous system. Single-cell transcriptomic atlases across multiple developmental stages have been generated with the aim of describing dynamics of these programmes (Eze et al. (2021); Chen et al. (2024c)). More recently, single-cell long-read technologies have allowed research to shift focus to begin cataloguing the isoform diversity of human brain cell types Patowary et al. (2024); Shimada et al. (2024).

Access to fresh human brain tissue is highly restricted, and patient-derived post-mortem samples are typically obtained at late stages of disease when extensive neuronal loss has already occurred. For example, in Parkinson's disease (PD), over 60% of midbrain dopaminergic neurons are lost by the time of diagnosis, severely limiting the availability of viable patient-derived neurons for molecular analysis Novak et al. (2022). Consequently, most transcriptomic studies rely on frozen post-mortem brain material.

However, dissociating frozen brain tissue into intact single cells is technically challenging. As a result, many studies have turned to single-nucleus RNA sequencing (snRNA-seq) Lake et al. (2016); Nagy et al. (2020), which captures nuclear transcripts enriched in pre-mRNAs and partially spliced intermediates. While snRNA-seq has been transformative for large-scale cell type–specific gene expression profiling, it has critical limitations for isoform-level analysis, since cytoplasmic, fully processed transcripts are under-represented. To overcome these issues, the Tilgner lab (Weill Cornell Medicine) developed a series of methods for snRNA-seq that use Linear/Asymmetric PCR (LAP) to selectively amplify polyA-containing, barcoded molecules, followed by exome capture (CAP) to remove purely intronic or artefactual cDNA and enrich for spliced isoforms Hardwick et al. (2022); Joglekar et al. (2024). Nonetheless, even with such enrichment strategies, snRNA-seq remains inherently limited in its ability to recover cytoplasmic transcripts, which include fully processed isoforms involved in essential cellular functions such as translation regulation, mRNA localisation, and stress granule formation. In addition, post-mortem samples often suffer from sample/RNA degradation induced by death and frozen storage.

Research into the developing brain is further constrained by limited access to early-stage human samples. Consequently, much of our understanding of brain development derives from rodent models. Although these models provide valuable information on general neurodevelopmental principles, human brain formation differs in key aspects, including progenitor diversity, neurogenic period length, and splicing regulation Ásgrímsdóttir and Arenas (2020). Species-specific differences in gene regulation and splicing programmes limit the ability of rodent models to faithfully recapitulate human-specific molecular phenotypes, especially in neurodevelopmental disorders.

iPSCs provide a renewable, genetically matched source of human brain cell types, including neurons, microglia, and astrocytes. Differentiation from the same iPSC line ensures an isogenic background, so observed differences in gene or isoform expression reflect regulatory rather than genetic variation. This makes iPSC models a powerful system for uncovering how post-transcriptional mechanisms such as alternative splicing and polyadenylation contribute to cell identity Novak et al. (2022); Bello et al. (2023).

However, iPSC-derived models have well-recognised limitations. Two-dimensional monolayer neuron cultures lack the cytoarchitectural complexity of the *in vivo* brain, including distinct cortical layers and long-range connectivity. Cell maturity is another constraint, as many iPSC-derived neurons and glia remain in an immature state that may not fully reflect adult splicing programmes. Protocol variability can further confound interpretation: even within the same donor line, differences in culturing or differentiation conditions across laboratories can markedly alter the cellular transcriptome Reed et al. (2021). The Bassett group has improved reproducibility for differentiating iPSCs into microglia and neurons, yet their studies also show that chromatin context and the transcriptional baseline influence CRISPR activation efficacy in iPSCs Washer et al. (2022); Wu et al. (2023b).

3.1.5 Aims of this chapter

Given these constraints, it is essential to characterise the transcriptomic profiles of our induced cell types. A comprehensive, cell–type–resolved reference for each differentiated population provides the necessary baseline for interpreting changes induced by CRISPR-based perturbations. The aims of this chapter are twofold.

Biological objective. I aim to construct transcriptome references for hiPSC-derived neurons, microglia, and astrocytes generated from a shared genotype. We hypothesise that, even under isogenic conditions, these cell types will exhibit significant differences in isoform expression and proportional usage, reflecting cell–type–specific post-transcriptional regulation. By profiling these populations in a controlled isogenic culture environment, I can disentangle regulatory contributions from underlying genetic variation, and contrast these results with the post-mortem dataset from Chapter 2, where cellular diversity and donor heterogeneity are confounding factors.

Technical objective. In parallel, I seek to benchmark single-cell long-read analysis workflows and evaluate limitations of competing sequencing platforms. We hypothesise that differences in chemistry and error profiles between PacBio and ONT platforms will yield systematic variation in isoform detection and quantification.

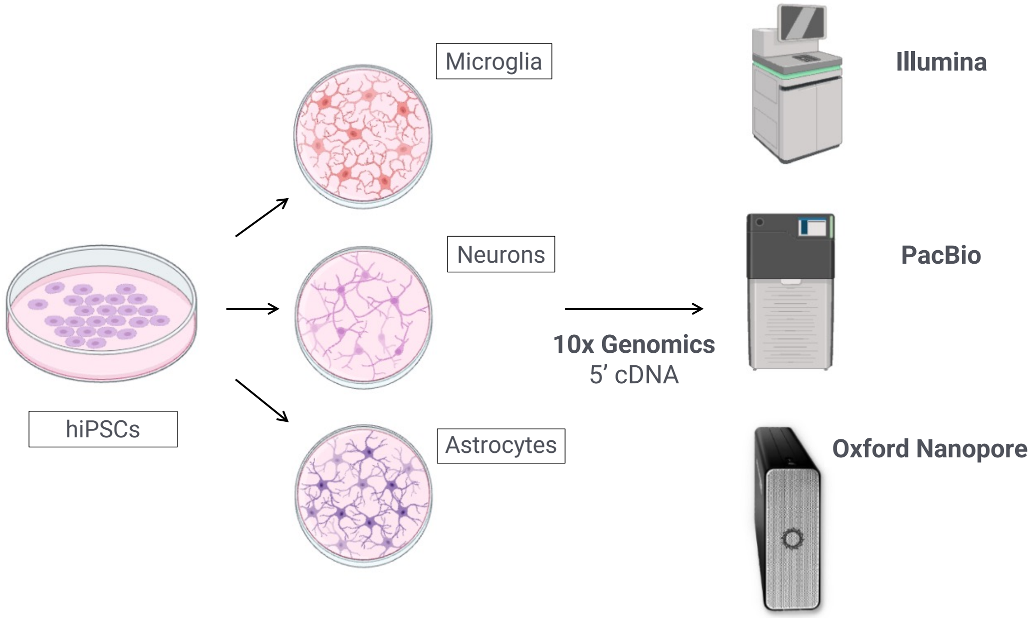


Figure 3.2: **Study design and sample overview.** hiPSC-derived cell populations were profiled using single-cell ONT, PacBio and Illumina sequencing. Created with BioRender.com.

To address these aims, I adapted the bulk Capture-Seq pipeline for single-cell ONT sequencing and profiled barcoded neuronal and glial populations generated from human iPSCs (hiPSCs). In parallel, I collaborated with a study that employs PacBio sequencing to systematically evaluate state-of-the-art single-cell long-read tools and evaluate transcriptomes resulting from competing methods.

3.2 Methods

Human iPSCs (KOLF2.1S, commercially available) were differentiated into neurons, astrocytes and microglia and sequenced by the Cellular and Gene Editing Research group (Wellcome Sanger Institute; Sarah Cooper and Andrew Bassett). iNeurons and iAstrocytes were generated by transcription-factor–driven (NFIA–SOX9 for astrocytes and NGN2 for neurons) protocols using doxycycline-inducible cassettes integrated at the CLYBL safe-harbour locus, followed by blasticidin selection and mApple-positive sorting. iAstrocytes differentiation protocol is adapted from Cvetkovic et al. (2022). Microglia were produced via embryoid body (EB) differentiation as described previously (Washer et al. (2022); van Wilgenburg et al. (2013)), using Aggrewell-based EB formation in OxE8 with BMP4, VEGF and SCF, with Y-27632 to prevent cell death. Primitive microglia precursors appearing in the supernatant after approximately 3–5 weeks were collected and matured for 14 days with IL-34, TGF-β1 and M-CSF, with GM-CSF as indicated and CD200/CX3CL1 added from day 10.

Single-cell libraries were prepared using the 10x Genomics 5' cDNA workflow. Sequencing was performed on Illumina NovaSeq 6000 and long-read platforms (Oxford Nanopore PromethION and PacBio Revio). For each differentiated cell type (iNeurons, iAstrocytes, and iMicroglia), libraries were sequenced in parallel on three flow cells allocated per cell type on both ONT and PacBio platforms.

Some code snippets were drafted or refined with the assistance of ChatGPT-4 and ChatGPT-5 (OpenAI, 2025), and subsequently validated by the author.

Tool	Version	Reference
wf-single-cell	0.2.8	https://github.com/epi2me-labs/wf-single-cell
2passtools	0.3	Parker et al. (2021)
minimap2	2.24	Li (2018)
IsoQuant	3.6.1	Prjibelski et al. (2023)
SQANTI3	5.2.2	Pardo-Palacios et al. (2024a)
UMI-tools	1.1.4	Smith et al. (2017)
Isosceles	0.2.0	Kabza et al. (2024)
STAR	2.7.10	Dobin et al. (2013)
Seurat	5.0.0	Stuart et al. (2019); Hao et al. (2024)
leiden	0.4.3	Kelly (2023)
presto	1.0.0	Korsunsky et al. (2019b)
uwot	0.2.2	Melville (2022)
scDblFinder	1.20.0	Germain et al. (2021)
CellTypist	1.7.1	Domínguez Conde et al. (2022)
refTSS	4.1	Abugessaisa et al. (2019)
PolyASite	2.0	Herrmann et al. (2020)

Table 3.1: Tools, versions, and associated references used in Chapter 3.

3.2.1 Transcript annotation

To generate initial read- and barcode-level quality metrics, the merged raw FASTQ file was first processed using the EPI2ME wf-single-cell Nextflow pipeline(v0.2.8) with the parameters: `--kit_name 5prime --kit_version v1 --expected_cells 9000 --gene_assigns_minqv 20 --barcode_min_quality 15 --stringtie_opts="-c 1.5 -m 200"`. During this step, uncorrected cell barcodes were filtered based on a minimum base quality score of 15. Barcodes that did not initially match the whitelist were then corrected by assigning them to the closest-matching whitelist barcode, provided the Levenshtein edit distance was ≤ 2. One of the generated output files is a bam file containing genomic primary alignments with the following tags: `CB` (corrected cell barcode sequence), `CR` (uncorrected cell barcode sequence), `UB` (corrected UMI sequence), and `UR` (uncorrected UMI sequence).

This tagged BAM file was used to generate a ”bulk” transcriptome using the pipeline developed in Chapter 2 (see section 2.2). This included splice junction correction using `2passtools`, genome alignment with `minimap2`, artefact filtering, transcript model construction using `IsoQuant`, and transcript classification and filtering with `SQANTI3`. `TAMA` was excluded due to performance limitations when processing high-depth data with multimapping reads.

IsoQuant first assigns reads to reference isoforms by constructing splice junction and exon profiles and matching them to known annotations. The reads that align ambiguously or with inconsistencies are used to detect potential novel isoforms. As splice junction correction had already been performed with 2passtools, the least stringent read correction and model construction options were selected to increase transcript detection sensitivity: `--data_type nanopore --splice_correction_strategy assembly --model_construction_strategy sensitive_ont --matching_strategy loose --polya_requirement auto --report_canonical all`. Poly(A) tail requirement was set to auto. Poly(A) tails were detected in 232,302,639 (96.4%) alignments and were therefore required for novel multiexon and all monoexon transcripts to be reported.

Transcript models generated with IsoQuant were then classified by SQANTI3 v5.2.2 with reference to GENCODE v46. The Reference Transcription Starting Sites v4.1 (refTSS, Abugessaisa et al. (2019)) and PolyASite v2.0 (Herrmann et al., 2020) databases were used to validate transcript start (TSS) and termination sites (TTS), respectively. The refTSS dataset was constructed by integrating publicly available Cap Analysis of Gene Expression (CAGE) datasets and removing low-confidence peaks based on read support and distance from TATA-box motifs and known promoter regions. Annotated isoforms (full-splice matches) were filtered only for internal priming ($\geq$ 70% adenines downstream of TTS). Novel isoforms were stringently filtered:

- <70% genomic As downstream of the 3' end
- No RT-switch flagged junctions
- 5' end supported by refTSS proximity (within a 1 kb window) or short-read pileup ratio > 1.1
- 3' end within 50 bp of polyA site or canonical polyA motif (from SQANTI3 resource)

Short-read support Short-read data from Illumina NovaSeq paired-end libraries (n=3) were aligned to the GRCh38 genome with STAR using the following parameters: `--soloType CB_UMI_Simple --soloBarcodeMate 1 --clip5pNbases 39 0 --soloCBstart 1 --soloCBlen 16 --soloUMIstart 17 --soloUMIlen 10 --soloStrand Reverse --outSAMattributes CB UB --soloCellFilter EmptyDrops_CR`. The resulting BAM files and SJ.out.tab files with high-confidence junctions and the number of reads mapping across these junctions were supplied to SQANTI3 for short-read validation of splice junctions and 5' ends.

3.2.2 Transcript quantification

UMI deduplication of the tagged bam file was performed using UMI-tools (v1.1.4) with the following settings for the group: `--per-cell --per-gene --extract-umi-method=tag --umi-tag=UB --cell-tag=CB --gene-tag=GN`. Within each group, the read with the highest mapping quality and the fewest secondary alignments was retained. Deduplicated BAM files were then used as input for Isosceles Kabza et al. (2024). Shortly, the BAMs were parsed to extract splicing structures using `bam_to_read_structures`, and transcript models were generated with `prepare_transcripts`, using a minimum splice read count of 1. Transcript compatibility counts (TCCs) were computed per cell using `bam_to_tcc` in de_novo_loose mode, with spliced transcripts extended by 100 bp and a minimum read count = 1. Final transcript-level quantification was performed using the EM algorithm via `tcc_to_transcript`, with 250 maximum iterations and a convergence threshold of 0.01. The resulting SummarisedExperiment was converted to a SingleCellExperiment object and used for downstream single-cell expression and isoform usage analysis.

3.2.3 Single-cell analysis

To identify and exclude potential doublets, the transcript-level SummarizedExperiment object was converted to a SingleCellExperiment object. Doublet detection was performed using the scDblFinder R package (v1.20.0) with default parameters. Artificial doublets were generated, and a k-nearest neighbour classifier was trained to identify true doublets. Cells classified as ”doublet” were excluded from further analysis; the resulting object contained 8638 cells.

The filtered transcript-level SingleCellExperiment object was converted to a Seurat object for normalisation and variance stabilisation using SCTransform on originalexp. PCA was computed on the SCT assay (RunPCA, npcs = 50), and the default 50 selected PCs were used to construct a k-nearest neighbour graph with Annoy (FindNeighbors, Euclidean, n.trees = 50). Clusters were identified using Leiden (FindClusters, algorithm = 4, resolution = 0.7). UMAP embeddings were generated with uwot on the chosen PCs (RunUMAP). FindNeighbours and RunUMAP were parallelised via the future package, and a fixed seed (42) ensured reproducibility.

3.2.4 Cell type and mitochondrial content annotation

Transcript-level clustering results were transferred to the gene-level SummarizedExperiment object by matching cell barcodes. A new column `tx_cluster` was appended to the gene-level metadata, representing each cell's transcript-derived cluster identity. Gene identifiers were converted from Ensembl gene IDs to gene symbols using a transcript-to-gene mapping file. Columns corresponding to novel genes or genes with unresolved names were excluded. The resulting table was exported as a CSV file for classification with `CellTypist` (Domínguez Conde et al. (2022)). We tested several CellTypist models: Developing_Human_Hippocampus, Developing_Human_Brain, Adult_Human_PrefrontalCortex. All models were run with the `--majority-voting` option. Within each major cell class, I retained subclusters with ≥100 cells, aggregated transcript-level counts per subcluster using Seurat's `AggregateExpression`, computed Spearman rank correlations between subcluster profiles and visualised clustered correlation heatmaps with `pheatmap`.

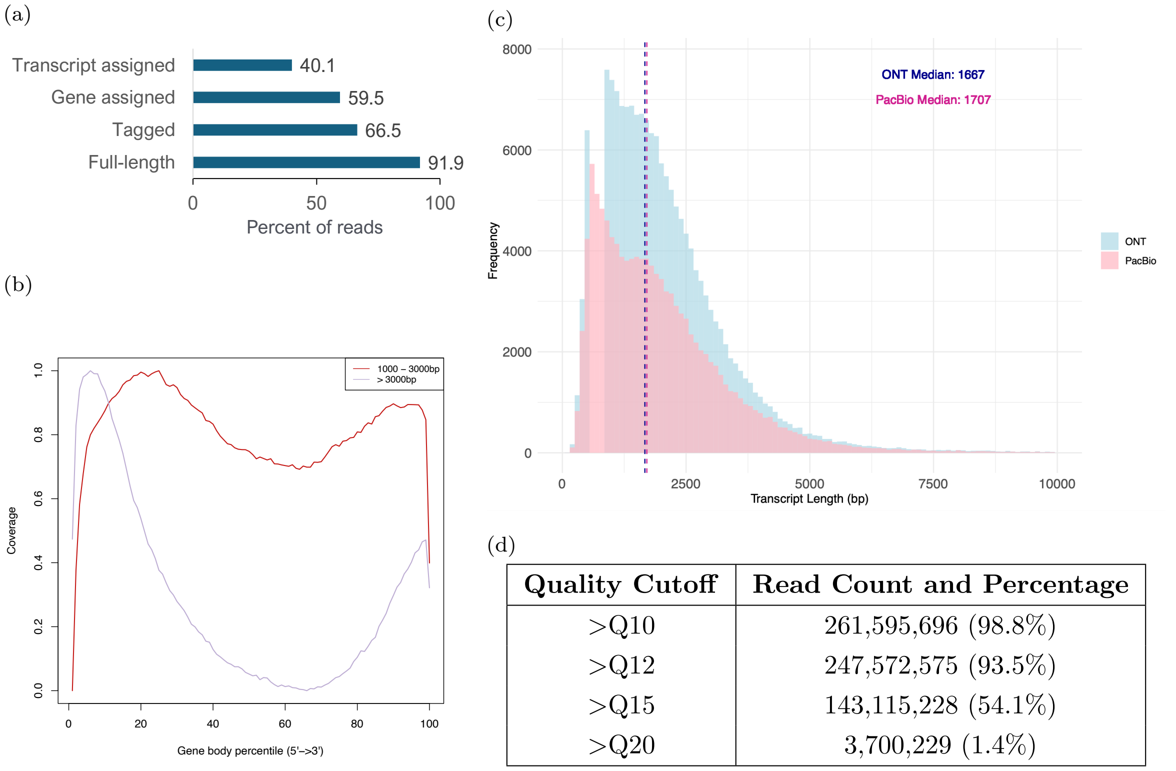
Within this gene-level Seurat object, I quantified the per-cell proportion of mitochondrial transcripts using Seurat's `PercentageFeatureSet` with features defined as genes whose symbols begin with `MT-`, and used these values to compare mitochondrial content across transcript-derived clusters.

To assess enrichment of previously defined microglial states, we used published marker gene sets from Mancuso et al. (2024). Gene symbols were mapped to Ensembl identifiers using a transcript–gene reference generated during isoform quantification, retaining only markers present in the single-cell expression matrices. For each programme, genes overlapping with “general” microglial markers (e.g. CSF1R, C3, CD74) were removed to avoid redundancy across signatures. Module scores were computed per cell with Seurat's `AddModuleScore`. The resulting enrichment scores were added to the cell metadata, visualised on UMAP embeddings, and summarised by cluster as z-scored means. Expression of selected individual markers from Mancuso et al. (2024) were additionally displayed with `DotPlot`.

3.2.5 Detection of differential isoform usage between cell clusters

To detect lineage-enriched isoforms, `FindAllMarkers` with a two-tailed Wilcoxon rank-sum statistical test (min.pct = 0.25, logfc.threshold = 0.25, adjusted p-value < 0.05) was used after changing cell identities to lineage.

To identify differential isoform usage between subpopulations within major lineages, I used a modified FindIsoforms() script (adapted from Ake et al. (2025)). Isoform counts were aggregated per cluster with `AggregateExpression`, then filtered as follows: retain genes with ≥2 expressed isoforms after filtering; keep isoforms that reach ≥10% of the gene's total counts in at least one condition (threshold_abund = 0.1). It then applies a chi-squared test to identify genes with significant changes



Quality Cutoff	Read Count and Percentage
>Q10	261,595,696 (98.8%)
>Q12	247,572,575 (93.5%)
>Q15	143,115,228 (54.1%)
>Q20	3,700,229 (1.4%)

Figure 3.3: **Read retention, quality, and coverage metrics from wf-single-cell analysis** A) Read retention by stage of wf-single-cell analysis. B) Gene body coverage for transcripts of varying lengths. C) Read length distribution. D) Number and percentage of reads above quality cut-offs.

in isoform proportions. P values were Benjamini–Hochberg adjusted.

3.3 Results

3.3.1 Pipeline performance

I began by assessing the performance of ONT's own recommended scRNA-seq pipeline (`epi2me-labs/wf-single-cell`), which performs all steps of analysis from cell barcode and UMI detection to generation of Seurat compatible counts matrices. While this approach provides a streamlined workflow for long-read single-cell RNA-seq analysis, I observed substantial read loss at multiple steps of the pipeline, potentially impacting downstream analyses. Namely, only 66.5% of reads were successfully assigned corrected UMI tags. This loss likely arises, in part, from the exclusion of reads that could not be confidently assigned to a gene or genomic interval, a filtering step intended to reduce UMI collision risk and limit the correction search space (Figure 3.3a. An additional 26.4% of reads were lost during the transcript assignment step, most likely due to the pipeline's reliance on existing transcript annotations and its inability to support novel isoform detection.

3.3.2 High-confidence pooled transcriptome

To address these limitations and enable the discovery of unannotated isoforms, I adapted the long-read isoform annotation pipeline described in Chapter 2. Isoform discovery was performed on pooled "bulk"

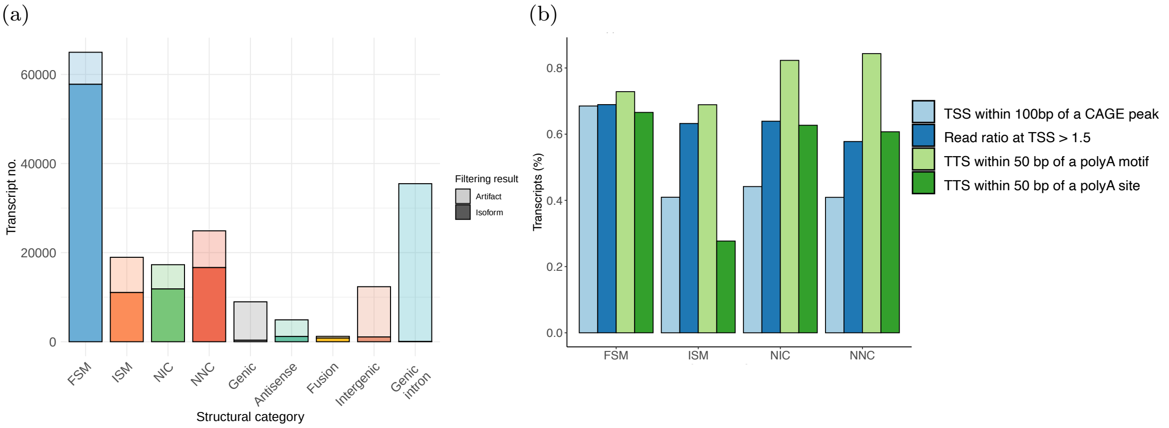


Figure 3.4: **Overview of bulk isoform catalog generated with IsoQuant.** A) Summary of predicted artefact and supported transcripts in different SQANTI3 categories. B) Orthogonal support for transcript start (TSS) and termination (TTS) sites.

reads, without incorporating cellular barcodes or UMIs, to maximise sensitivity for low-abundance isoforms expressed in only a small subset of cells. Compared to the R9 chemistry data used in Chapter 2 (Figure 2.6), the R10.4 data showed a substantial improvement in read quality 3.3d, with a median read quality score of 15 and 54.1% of reads exceeding this threshold—versus only 8.9% in the earlier dataset. This improvement allowed us to apply more relaxed read correction parameters in IsoQuant, while still using `2passtools` to correct splice junctions prior to transcript assembly.

To assess transcript novelty relative to GENCODE reference, I classified IsoQuant-assembled transcripts using SQANTI3 into standard structural categories. Out of 189,031 total isoforms, a substantial fraction fell into the `genic intron` (n = 35,480; 18.8%) and `intergenic` (n = 12,354; 6.5%) categories, representing transcripts entirely within introns of annotated genes or located outside known gene boundaries, respectively. These categories are typically enriched for artefactual transcripts arising from library preparation artefacts, and have been widely reported in studies using 10x Genomics assays (Ding et al. (2020); La Manno et al. (2018)). Although these artefacts have been primarily characterised in short-read datasets (see 10x technical note), many of the underlying mechanisms, such as internal poly-A priming and poly(dT)-mediated strand invasion, are equally relevant to long-read data.

The 10x 5' protocol does mitigate some of the artefacts observed in the 3' protocols. In particular, internal polyA priming events are markedly reduced compared to 3' assays, with only 20% of antisense intronic reads in 5' libraries exhibiting a downstream poly-A tract, versus 65–80% in 3' datasets. The equivalent proposed mechanism of TSO mispriming at homologous sites is almost nonexistent. Nonetheless, the 5' protocol is still susceptible to artefacts arising from first-strand cDNA poly-A priming. This can occur when RNA degradation exposes internal poly-A stretches on the cDNA, allowing poly(dT) primers or residual TSO to initiate spurious reverse transcription. Most of the isoforms of `genic intron`, `genic` and `intergenic` classes were filtered out based on lack of orthogonal support (Figure 3.4a). After filtering, the remaining transcriptome was largely composed of the four main structural categories with good TSS and TTS support: `Full-Splice Match` (FSM, n = 57,788), `Incomplete-Splice Match` (ISM, n = 11,567), `Novel In Catalog` (NIC, n = 12,700), and `Novel Not in Catalog` (NNC, n = 16,780) (Figure 3.4b), together comprising a sample-specific high-confidence but cell-type-agnostic reference transcriptome.

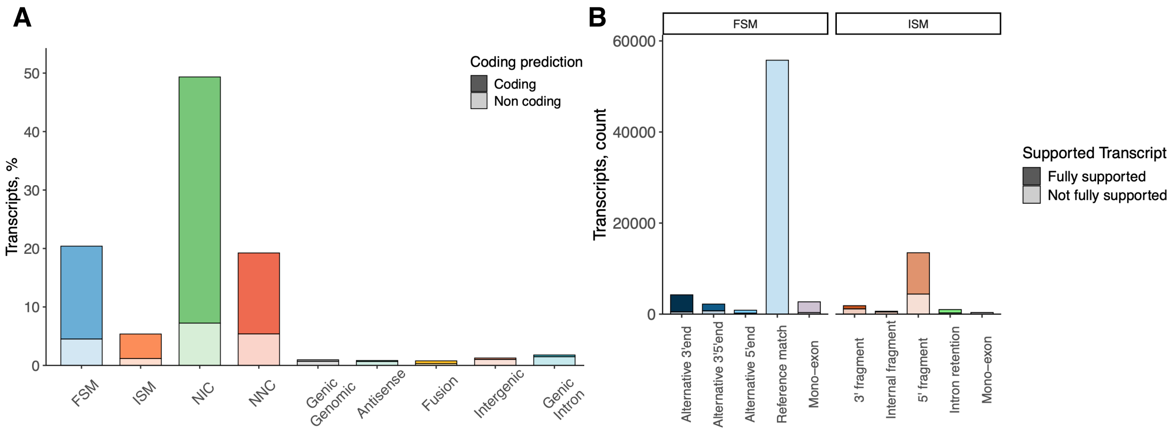


Figure 3.5: **Overview of single-cell isoform catalog generated with Isosceles with IsoQuant transcriptome as reference annotation** A) SQANTI3 structural categories B) Transcript end variations for full and incomplete splice match categories.

3.3.3 Quantification at single-cell level

To enable single-cell resolved isoform quantification, I next evaluated tools capable of incorporating cellular barcode information. Based on benchmarking results from the Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP; Pardo-Palacios et al. (2024b)), which identified IsoQuant as a top-performing tool for cDNA-ONT quantification, I initially evaluated its performance using the `--read_group tag:CB` option, which groups reads by cell barcode. Although IsoQuant provides cell-aware quantification, it adopts a conservative quantification strategy that excludes reads mapping ambiguously or inconsistently to annotated isoforms and only reports transcripts with at least one uniquely assigned read. This approach is suboptimal for our dataset, as we found that gene body coverage declined with increasing transcript length and therefore longer isoforms had fewer full-length supporting reads (Figure 3.3b). This might be the result of biases introduced during 10x library preparation, which can result in truncated reads and over-representation of transcripts around 1kb (Figure 3.3c).

To recover reads with ambiguous assignment, I quantified transcripts with Isosceles (v0.2.0), which uses an expectation–maximisation algorithm, and adopted the Isosceles + IsoQuant workflow for transcript discovery. I selected this configuration based on benchmarking results that reported this strategy achieved the highest overall performance across annotated and novel transcripts (TPR 84.5% and F1-score 89.7%), outperforming Isosceles alone Kabza et al. (2024). The resulting single-cell isoform catalog contained 322,484 transcripts.

SQANTI3 classified 65,781 transcripts as FSM (20.4%), 17,321 as ISM (5.4%), 159,123 as NIC (49.3%) and 62,040 as NNC (19.2%), with remaining transcripts mostly split between Genic Intron and Intergenic (Figure 3.5A). Within the FSM set, 7,319 transcripts showed alternative TSS (5'end) and/or TTS (3'end) relative to the reference. Among ISMs, 13,478 were 5' fragments, consistent with the 5' library preparation (Figure 3.5B) and indicative of bias towards 3' truncation.

3.3.4 Population heterogeneity within iPSC-derived lineages

Doublet detection with scDblFinder v1.20.0 identified 263 (3.0%) and 373 (4.2%) putative doublets from the transcript- and gene-level counts respectively, with 198 overlapping. These values are somewhat higher than the 1–2% often reported for standard 10x Chromium experiments and might be due to sample multiplexing. To preserve isoform-specific signal, I did not perform multimodal (isoform–gene) clustering with Weighted Nearest Neighbours; all downstream clustering and annotation used the transcript-level object. After doublet filtering, I detected 318,931 expressed transcripts (≥1 count in at least one cell), of which 290,475 were expressed in at least five cells. In total, 8,638 cells were clustered and the UMAP revealed four expected major groups (Figure 3.6).

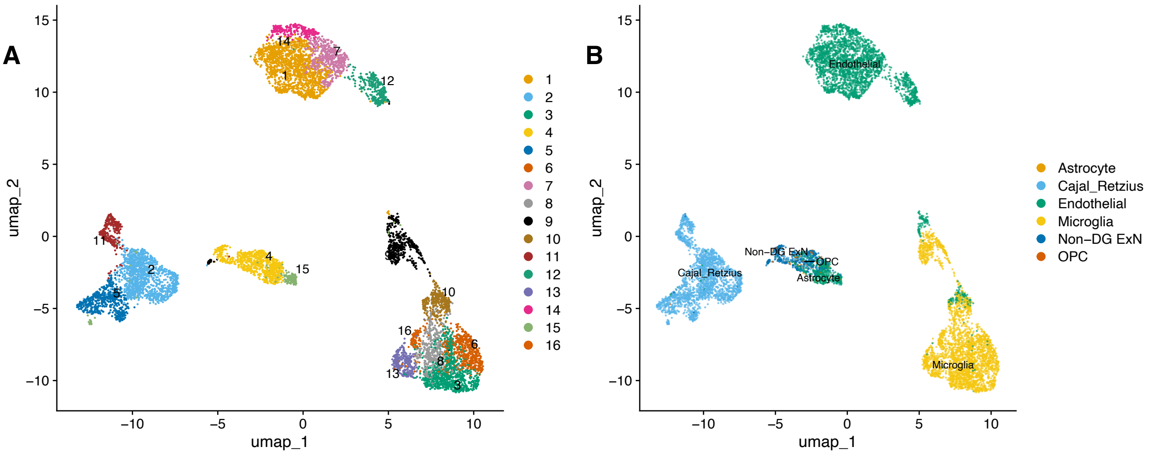


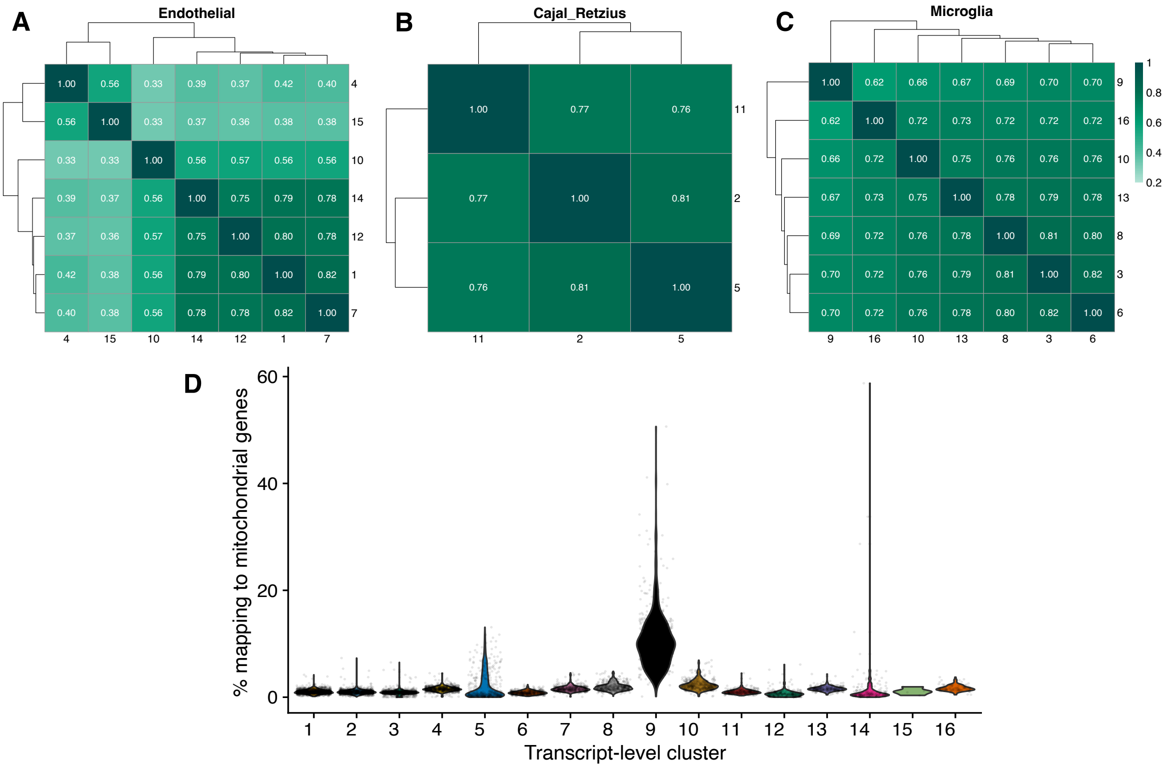
Figure 3.6: **UMAP plots of major cell populations.** A) Clusters identified using Leiden clustering on isoform-level expression. B) Cell type annotations based on gene-level expression assigned using CellTypist, overlaid on the same UMAP projection.

To identify cluster markers and assign cell types, transcript-level clusters were transferred to the gene-level object. Given that the dataset comprised iPSC-derived cultured cells rather than whole tissue, I evaluated several CellTypist v1.7.1 models to check that the expected populations were recovered and to compare their relative abundance and heterogeneity. The *Developing Human Hippocampus* model showed the best agreement with UMAP structure, returning three major groups: Microglia, Endothelial/iPSC-like, and Cajal–Retzius (CR), plus a smaller fourth cluster (1,038 cells) with heterogeneous labels, including endothelial (867), excitatory neuronal (Non-DG ExN, 389), OPC (45) and Astrocyte (10) annotations.

I next identified one-vs-rest gene markers for each major group using `FindAllMarkers()` and cross-referenced them to a collaborator-curated panel (Supplementary Table 3.1). Although the overlap was modest (n = 26 genes), concordance was high for CR and Microglia (16 of 26, 62% matched class to group). CR markers were strongly neuronal, including *DCX*, *DLG4*, *MAP2*, *MAPT*, *TUBB3*, *SYP*, *SYT1*, *STMN2*, *SOX11* and *SOX4* (for example, *DCX* avg_log2FC = 4.49). Microglia showed canonical signatures (*CX3CR1*, *TLR2*, *RUNX1*, *CD83*, *CCL2*; for example, *CX3CR1* avg_log2FC = 6.68). Two curated astrocyte markers *GJA1* and *SOX9* were identified as unique endothelial group markers.

Transcript-level Spearman correlation across subclusters of CellTypist-identified groups further clarified relationships between populations. Subclusters corresponding to CR, microglial and endothelial populations showed high inter-cluster correlation overall (Figure 3.7A-C). Within microglia, cluster 9 stood out with lower correlations to other microglial subclusters (Spearman $\rho \approx 0.62$-0.70). This cluster comprised 11.4% of microglial cells (334 of 2,918) and likely represents stressed or dying cells as it exhibited high mitochondrial content (median percent.mt = 10.1%; Figure 3.7D).

The small fourth cluster was initially interpreted as astrocyte-like. This interpretation resulted from two observations: cells annotated as endothelial within its constituent subclusters (4 and 15) showed relatively low similarity to the large endothelial/iPSC-like clusters (Spearman's $\rho \approx 0.33$-0.42), and a minority in this group carried an astrocyte label. Re-evaluation with gene-level FeaturePlots for transcription factors used for differentiation protocols (NFIA, SOX9, NGN2) resolved the ambiguity. *SOX9* (and *NFIA*) showed strong, diffuse expression across the large endothelial/iPSC-like cluster on the transcript-derived UMAP (Fig.3.8), consistent with undifferentiated astrocyte-lineage cells embedded within that larger cluster. In contrast, the small cluster was enriched for canonical glutamatergic neuronal markers, including *SLC17A6* (VGLUT2) and *GRIN1* (NMDAR1), supporting reassignment as an excitatory neuronal population rather than astrocytes. After re-annotation, the final grouping comprised 3,118 neurons, 3,044 microglia and 2,476 astrocytes. Clusters 5 and 14 also exhibited higher mitochondrial transcript abundance, which is consistent with excitatory neurons increasing mitochondrial gene expression and ATP production in response to activity-dependent calcium influx and signalling (Kwon et al., 2016).



	4	15	10	14	12	1	7
4	1.00	0.56	0.33	0.39	0.37	0.42	0.40
15	0.56	1.00	0.33	0.37	0.36	0.38	0.38
10	0.33	0.33	1.00	0.56	0.57	0.56	0.56
14	0.39	0.37	0.56	1.00	0.75	0.79	0.78
12	0.37	0.36	0.57	0.75	1.00	0.80	0.78
1	0.42	0.38	0.56	0.79	0.80	1.00	0.82
7	0.40	0.38	0.56	0.78	0.78	0.82	1.00

	11	2	5
11	1.00	0.77	0.76
2	0.77	1.00	0.81
5	0.76	0.81	1.00

	9	16	10	13	8	3	6
9	1.00	0.62	0.66	0.67	0.69	0.70	0.70
16	0.62	1.00	0.72	0.73	0.72	0.72	0.72
10	0.66	0.72	1.00	0.75	0.76	0.76	0.76
13	0.67	0.73	0.75	1.00	0.78	0.79	0.78
8	0.69	0.72	0.76	0.78	1.00	0.81	0.80
3	0.70	0.72	0.76	0.79	0.81	1.00	0.82
6	0.70	0.72	0.76	0.78	0.80	0.82	1.00

Figure 3.7: **Correlation and QC metrics.** Spearman correlation between subcluster transcript expression within major groups (A) Endothelial, (B) Cajal–Retzius, (C) Microglia.(D) Mitochondrial proportion.

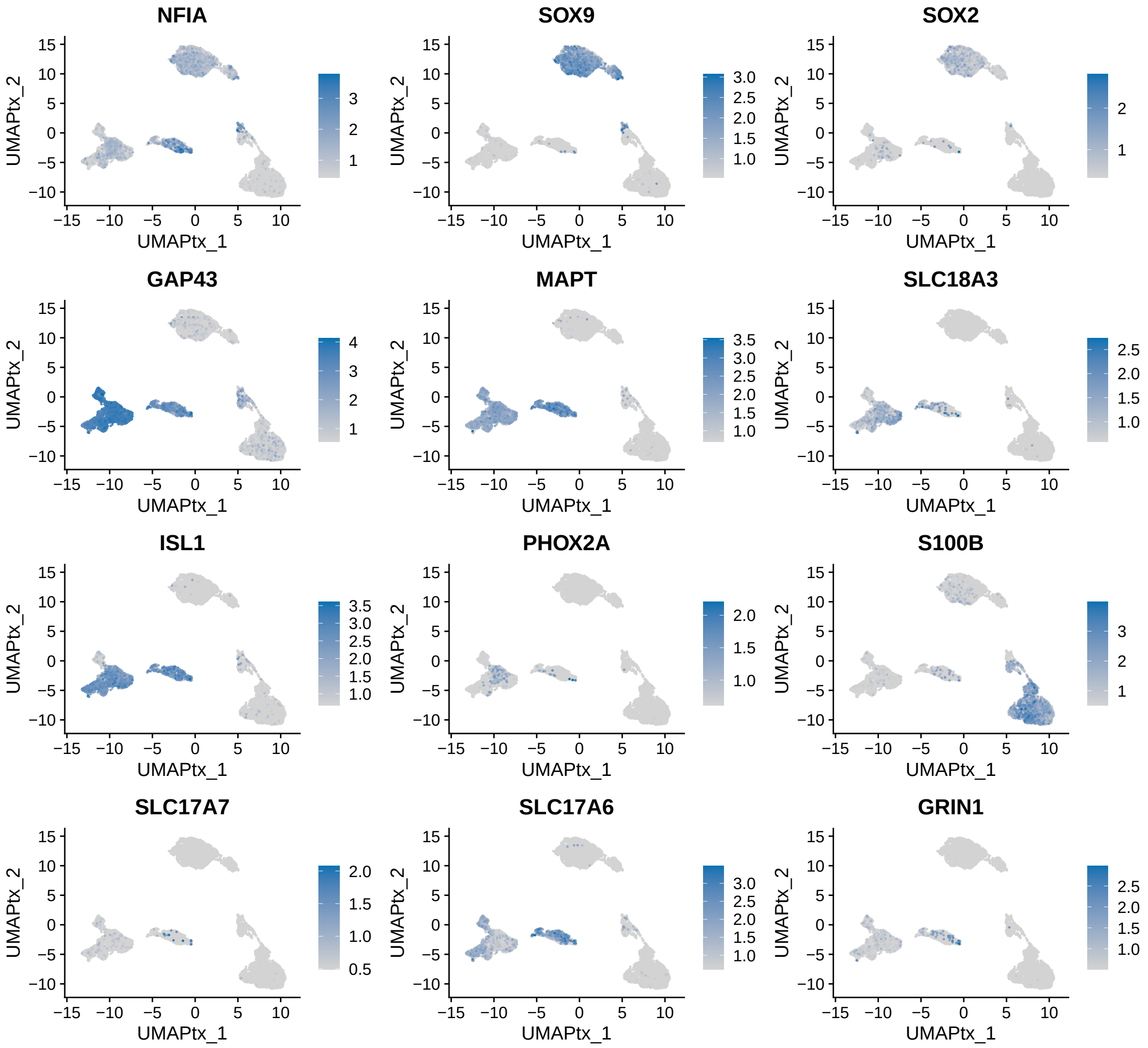


Figure 3.8: **UMAP projection of cells coloured by expression of selected marker genes.** Transcript-level UMAP coordinates from the Seurat transcript-based object were transferred to the gene-level data. Each panel shows the normalised expression of a marker gene (grey = low expression, blue = high expression) across all cells. Genes shown include progenitor and glial markers (*NFIA*, *SOX9*, *SOX2*, *S100B*), neuronal differentiation and axonal growth markers (*GAP43*, *MAPT*), and excitatory neuronal markers (*SLC17A7*, *SLC17A6*, *GRIN1*).

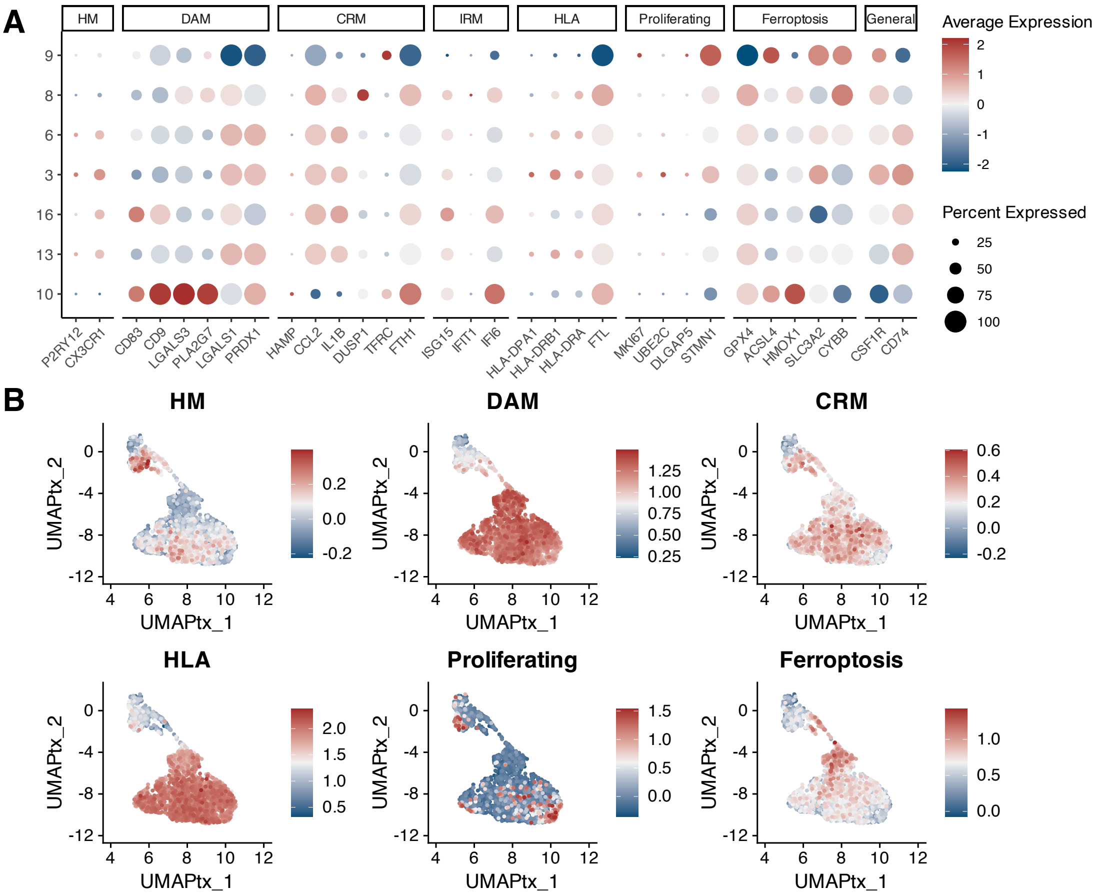


Figure 3.9: **Marker gene expression and module scores identify microglial subtypes.** (A) Expression of representative marker genes across microglial clusters. (B) UMAP visualisation of aggregated module scores for curated microglial subtype gene sets (HM, DAM, CRM, IRM, HLA, Proliferating). Colours indicate relative enrichment.

To characterise microglial heterogeneity, I applied marker-programme scoring using the reference sets from Mancuso et al. (2024), prioritising genes highlighted by Perez-Alcantara et al. (2025). In this framework, microglial states are organised into five major programmes: homeostatic (HM), cytokine response (CRM), interferon response (IRM), disease-associated response (DAM) and human leukocyte antigen-presenting response (HLA). Clusters 3 and 6 accounted for the largest fractions of microglia (884 and 551 cells) and exhibited broadly similar signatures, with high expression of *LGALS1*, *PRDX1*, *CSF1R* and *CD74*, and subset of cluster 3 expressing high levels of *HLA-DRB1* and *HLA-DRA* displaying an activated but more HLA-like state (Figure 3.9).

Subcluster 10 showed strong enrichment for DAM, with elevated *CD9*, *LGALS3* and *PLA2G7*, consistent with an activated phenotype. Notably, a subset of the putative dying cluster 9 scored as enriched for the HM programme according to the Mancuso gene set, yet it expressed comparatively low levels of canonical HM markers *P2RY12* and *CX3CR1*. The original HM set contained nine mitochondrial genes; removing these did not abolish the HM enrichment in that subset of cluster 9 cells. Despite this, converging evidence pointed to a dying state, so to confirm that clusters 9–10 represent trajectories towards cell death, curated gene sets for programmed cell-death pathways (apoptosis, necroptosis, ferroptosis, pyroptosis) were added to analysis (Matsudaira and Prinz, 2022). Among these, the ferroptosis programme was most strongly enriched in cluster 10, suggesting a pre-death, inflammation-responsive state. This interpretation is supported by upregulation of *HMOX1* in cluster 10, a gene that accelerates ferroptosis in microglia, while cluster 9 showed higher *CYBB* and *SLC3A2*

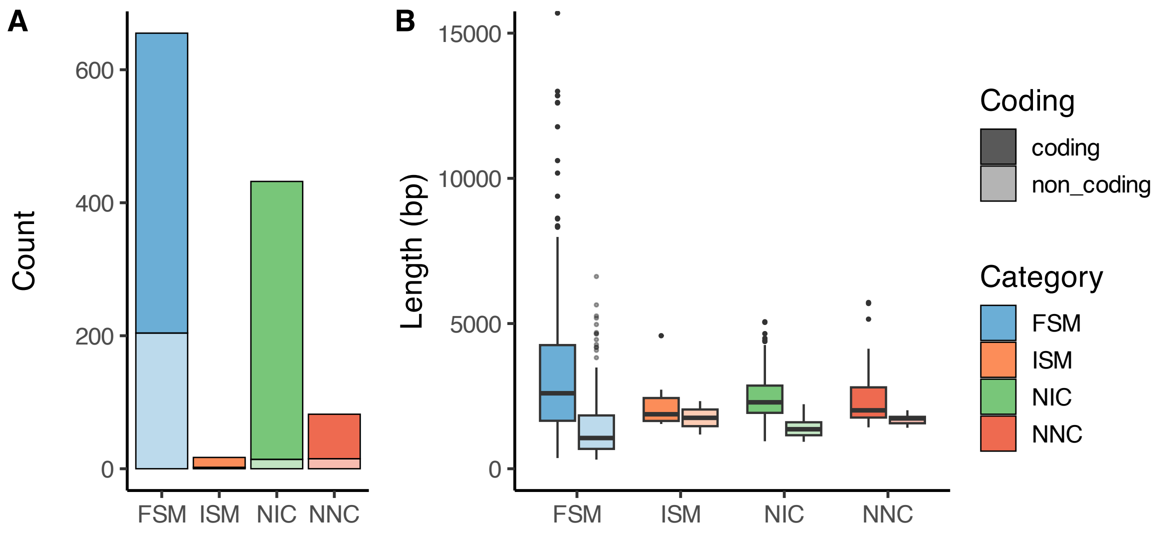


Figure 3.10: **Structural categories and length distribution of isoforms detected in both Capture-Seq and single-cell long-read data.** (A) SQANTI3 structural categories. (B) Transcript length (bp) by category.

expression, consistent with oxidative stress signalling (Fernández-Mendívil et al., 2021).

3.3.5 Overlap with CaptureSeq bulk transcriptome

To assess concordance between isoforms detected in single-cell long-read data (iPSC-derived neurons, microglia and astrocytes) and bulk CaptureSeq from post-mortem human brain (Chapter 2), I matched isoforms by exon junction chains, permitting a 500 nt difference at transcript starts and ends. Out of the 1,201 overlapping transcripts for 369 genes, the majority were classified as full-splice matches (655) and 951 were coding (Figure 3.10), consistent with expectations that a core set of well-annotated, highly expressed isoforms would be reproducible across both experimental contexts. Consistent with this, many shared isoforms were among the most prevalent per gene in the single-cell data: 432 ranked within the top three by mean expression in at least one lineage, and 266 were top-three in all three lineages (Figure S2). The overall overlap was small (1,201 out of 290,475 isoforms; 0.41%), which is expected given differences in library preparation and sample type. Bulk CaptureSeq was performed on heterogeneous post-mortem brain tissue spanning multiple regions and developmental stages and focused on a defined set of genes, whereas the single-cell data derive from relatively homogeneous, *in vitro* cell cultures with sequencing performed transcriptome-wide.

3.3.6 Differential isoform usage

To investigate lineage-specific splicing patterns, transcript-level clustering was collapsed into three major groups corresponding to microglia, neurons, and astrocytes. Within this reduced annotation, we applied `FindAllMarkers` at the transcript level to identify both positive and negative markers across lineages. Candidate markers were filtered to retain only those genes for which distinct isoforms emerged as positive markers in different lineages, thereby highlighting cases where isoform identity, rather than gene-level expression, determined lineage specificity. This analysis identified 795 genes with evidence of lineage-variable isoform regulation, represented by 2,303 distinct isoforms (Supplementary Table 3.2).

Results for ten representative genes from the Capture panel are shown in Figure 3.11. For example, two NIC isoforms of *HNRNPA2B1* displayed opposing lineage associations: one was enriched in neurons (ISOT-ba46-3b95-f807-e2ff:s26190300:e26200750:NC:FL; avg_log$_2$FC = 2.21), while another was preferentially expressed in astrocytes (ISOT-6b6b-537d-7379-eaeb:s26190100:e26200750:NC:FL; avg_log$_2$FC = 1.55).

I next examined isoform dynamics within subclusters. Using pseudo-bulked counts aggregated with `AggregateExpression()`, we identified isoform switching genes using chi-square statistical test (FDR adjusted p-value < 0.05), requiring that each gene retained at least two expressed isoforms and that any isoform contributed $\geq$10% of counts in at least one cluster. This analysis identified DTU in 1,095 genes in microglia, 260 in neurons, and 1,617 in astrocytes (Supplementary Tables 3.3-3.5). Restricting to microglia and excluding the stressed cluster 9 still yielded 1,006 DTU genes, confirming that the signal was not driven solely by the stress-associated population.

Several examples highlight these patterns. In microglia, *CD83* underwent a stress-associated switch: cells in cluster 9 favoured two non-coding, intron-retaining isoforms (transcript22871.chr6.nic and transcript22870.chr6.nic), while predicted coding isoforms (transcript22724.chr6.nnic and transcript22699.chr6.nnic) were depleted (Fig. 3.12A). Increased intron retention is a recognised hallmark of stressed or senescent cells (Yao et al., 2020; Hadar et al., 2022). In neurons, DTU was driven primarily by differences between excitatory glutamatergic clusters (4 and 15) and the remaining neuronal subtypes. For instance, an alternative *FNBP1L* transcript (ENST00000370253.6) was enriched in excitatory cluster 15 relative to the canonical isoform (ENST00000271234.13), which encodes the primary protein product (Fig. 3.12B) (Steyn et al., 2024).

I next asked whether isoforms that showed DTU between single-cell subclusters, and were therefore more likely to represent differences in maturation, rather than lineage, also displayed DTU between adult and fetal DLPFC. As outlined in the introduction (Section 3.1), the DLPFC is a six-layered, cell-type-rich region, so bulk measurements aggregate signals from diverse neuronal and non-neuronal populations. Given this complexity, and the tendency of iPSC-derived cells to resemble earlier developmental states, we hypothesised that isoform usage *in vitro* would align more closely with fetal patterns.

The results were mixed rather than uniformly fetal-like. For *SYNCRIP*, the annotated isoform ENST00000616122.5 was higher in fetal than adult DLPFC in CaptureSeq and varied across microglial subclusters in the single-cell data (Fig.3.13A). By contrast, ENST00000355238.11 was the most abundant *SYNCRIP* isoform in adult DLPFC. Its full-splice match in the single-cell data, `ISOT-0b2d-632d-9002-f2ae:s85612550:e85642950:EC:FL`, carried alternative 5' and 3' ends with a much shorter terminal exon at the 3' end (4588 vs 349 bp). Although I initially suspected that the shorter form might reflect truncation due to the 5'-end library preparation, its 3' end falls within a supported poly(A) site and includes a canonical polyadenylation motif.

As mentioned previously, in neurons isoform shifts were largely driven by excitatory clusters. For the splice factor *ZRANB2*, the shared annotated isoform ENST00000370920.8 was enriched in adult DLPFC in comparison to fetal and constituted the predominant transcript in glutamatergic cluster 15 (Figure 3.13 B). For *HNRNPA2B1*, ENST00000354667.8 (alternative 3' end) was adult-enriched in bulk and prominent in clusters 11, 2 and 5, whereas clusters 15 and 4 favoured a novel NIC isoform (Fig. 3.13C).

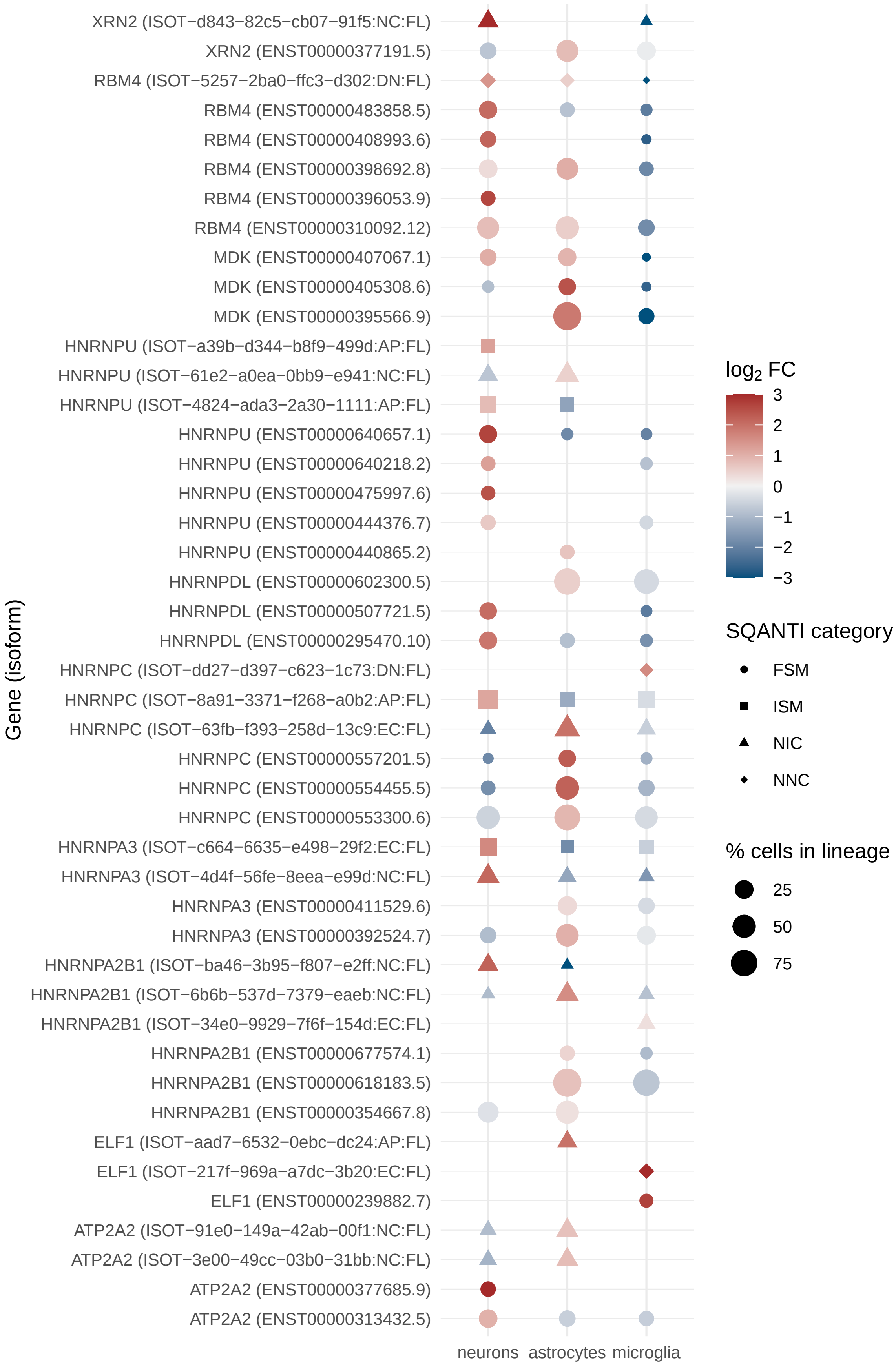


Figure 3.11: **Isoform-level markers (both positive and negative) across lineages of genes on the Capture panel with different isoforms significantly enriched in at least two lineages.**

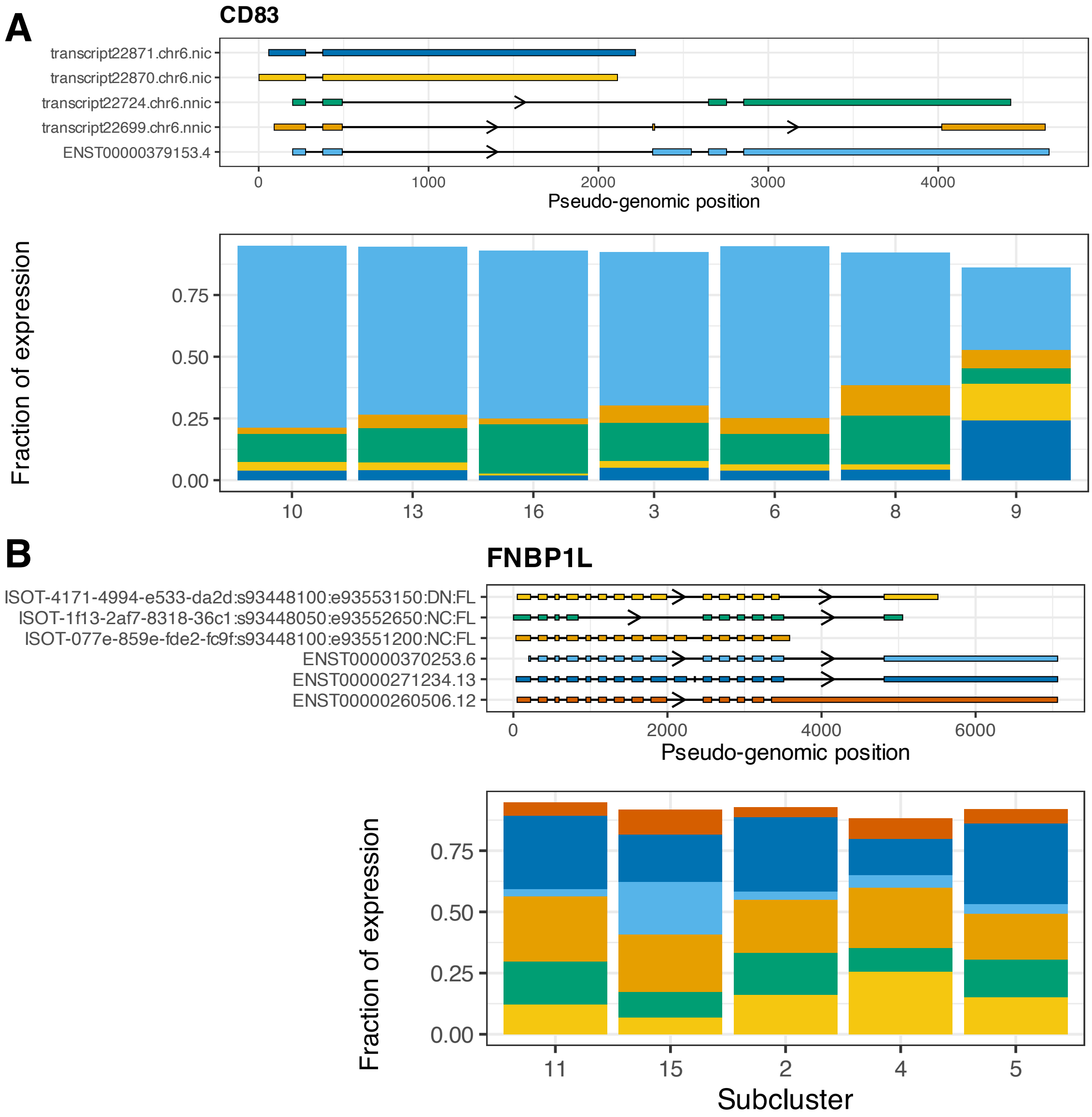


Figure 3.12: **Isoform usage exemplars in microglia and neurons.** (A) *CD83* isoform usage across microglial subclusters; (B) *FNBP1L* isoform usage across neuronal subclusters. For each gene, the top panel shows transcript structures and the bottom panel shows the fractional contribution of each isoform within subclusters.

3.4 Discussion

In this chapter, I adapted the pipeline developed in Chapter 2 and applied it to a single-cell long-read dataset to characterise transcriptomic differences among hiPSC-derived neurons, astrocytes, and microglia, as well as their subpopulations. I tested the hypothesis that sensitive, tailored analysis of single-cell long-read data can reveal significant isoform shifts without relying on short-read support. By prioritising comprehensive isoform detection and retaining information often discarded by default workflows, I captured a more nuanced view of isoform dynamics across isogenic populations.

This approach demonstrated that the intrinsic variability within isogenic iPSC-derived cultures—typically considered a limitation—can instead be leveraged to study post-transcriptional regulation in lineage specification, maturation, and stress responses. Cross-referencing single-cell isoform programmes with bulk CaptureSeq from post-mortem brain showed that cultured cells can reproduce isoform-level variation relevant to brain development and disease, while also placing novel isoforms identified in bulk tissue into a cellular context.

3.4.1 Population heterogeneity in iPSC-derived models

Despite advances in differentiation methods, iPSC-derived cultures remain heterogeneous, both between lines and within cultures, reflecting differences in genetic background, epigenetic state, maturity, and local signalling environments. In our data, NFIA–SOX9 induced iAstrocytes were the least concordant with mature astrocyte signatures, whereas iPSC-derived neurons and microglia mapped more closely to *in vivo* reference states. Within neurons, variation was driven by differences between excitatory clusters 4 and 15 and the remaining clusters. In microglia, clusters 9 and 10 stood out and were further resolved into DAM-enriched and stressed subsets through programme scoring. While protocols for iPSC-derived neurons and microglia are relatively well established and produce cell types that closely match *in vivo* counterparts, astrocyte differentiation remains variable and poorly standardised. For example, a head-to-head comparison of a long, serum-free protocol versus a shorter, serum-containing method yielded astrocyte populations with distinct transcriptomic and phenotypic profiles (Mulica et al., 2023).

Within-culture maturation heterogeneity, on the other hand, reflects the reality of biological systems and can be leveraged to study differentiation trajectories and lineage decisions and to ask how specific isoforms and splicing programmes emerge along these axes. Mapping such *in vitro* states back to *in vivo* brain heterogeneity helps anchor interpretation and represents an important future direction.

3.4.2 Linking iPSC models to *in vivo* brain signatures

To contextualise *in vitro* states, I used gene-level profiles from short-read gene-level reference atlases. Going forward, this can be strengthened by leveraging single-cell and single-nucleus long-read datasets from post-mortem brain, which provide isoform-resolved cell state profiles (Joglekar et al., 2024; Shimada et al., 2024), and emerging spatial long-read methods that measure isoform usage directly in intact tissue. Short-read spatial studies were already showing region- and layer-specific astrocyte programmes across cortex and striatum (Bayraktar et al., 2020; Hasel et al., 2021; Hodge et al., 2019; O'Dea and Hasel, 2025), while long-read studies now demonstrate within-cell type isoform variation across cortical layers and regions of mouse and human brain (Michielsen et al., 2025; Foord et al., 2025).

In this study, I did not attempt deconvolution of bulk post-mortem brain data into cell-type proportions (Wang et al., 2019). The datasets are not directly comparable in their design, sampling, and targets, so deconvolution would have risked over-interpretation. Instead, I compared differential transcript usage signals across bulk and single-cell data wherever there was a well-defined overlap. There were 1,201 overlapping transcripts between the targeted bulk and the transcriptome-wide single-cell dataset, which constitutes only 2% of isoforms detected with CaptureSeq for the targeted genes. This low proportion is expected: Chapter 2 profiled a targeted gene panel enriched for lowly expressed lncRNAs, whereas the single-cell experiment surveyed the whole transcriptome and therefore captured a larger share of abundant protein-coding isoforms. Despite this difference in scope, the overlap was informative. Notably, many shared transcripts were novel relative to the reference annotation, including 432 NIC and 82 NNC isoforms. This shows that single-cell long reads can recover previously uncharacterised transcript structures that also appear in targeted bulk profiling of human tissue.

Genes with concordant isoforms across single-cell and targeted bulk sequencing of post-mortem brain

We focused on three a priori genes of interest: the RNA-binding protein *SYNCRIP* and two splice regulators previously implicated in schizophrenia, *ZRANB2* and *HNRNPA2B1*. *SYNCRIP* (also known as hnRNP Q; Section 1.2) encodes three canonical protein isoforms: Q3 (longest, with full RGG-rich tail), Q2 (containing a deletion within RRM2), and Q1 (shorter C-terminus with an alternative terminal coding exon, typically containing a single nuclear localisation signal Mourelatos et al. (2001)). In our single-cell data, ENST00000355238.11 corresponds to hnRNP Q1 (UniProt O60506-3), while ENST00000369622.8 corresponds to hnRNP Q3 (UniProt O60506-1). The transcript ISOT-0b2d-632d-9002-f2ae:s85612550:e85642950:EC:FL retains the complete splice junction chain and open reading frame of ENST00000355238.11, consistent with a protein-identical isoform carrying an alternative 3' UTR. ENST00000355238.11 was found to be the top isoform in adult DLPFC in Capture-Seq data (Figure 3.13A).

Although most *SYNCRIP* isoforms detected here are already annotated, single-cell long-read sequencing provides the ability to quantify them directly at transcript level within defined cell populations. This allows us to separate isoform differences involving coding changes (Q1 versus Q3) from UTR-only variants, and to detect regulated alternative 3' ends in specific cellular contexts. Such fine-grained resolution illustrates the potential of single-cell long-read approaches to bridge the gap between bulk sequencing and *in vivo* transcriptomic diversity.

In neurons, we detected subpopulation variable isoforms in the splice regulators ZRANB2 and HNRNPA2B1 (Figure 3.13B-C), both included in our capture panel as they had been implicated in schizophrenia. These isoform differences were driven by glutamatergic subclusters, which is notable given recent evidence that schizophrenia risk genes show high expression specificity in layer 4 glutamatergic neurons (Tume et al., 2024). We can infer that lineage-specific changes in splicing machinery could plausibly propagate broader changes in exon and isoform expression across disease-relevant genes. Cohen et al. (2016) showed that a schizophrenia-risk variant in DRD2 (rs1076560) that alters ZRANB2 binding correlates with shifts in D2 short-to-long isoform ratios and is also linked to altered activity and functional connectivity in the DLPFC during working-memory tasks.

For *HNRNPA2B1*, the direction of change in expression in SCZ studies is contradictory across models. (Puvogel et al., 2022) used short-read 10x snRNA-seq to profile endothelial nuclei from schizophrenia post-mortem tissue and found reduced expression of *HNRNPA2B1.* In schizophrenia-derived brain organoids, on the other hand, hnRNP proteins, including A2/B1 were upregulated (Nascimento et al., 2022). These findings need not conflict once isoforms are considered. Gene-level measures conflate productive and non-productive transcripts. Long-read RNA-seq can separate ORF–intact, protein-productive HNRNPA2B1 transcripts from intron-retaining or NMD–sensitive transcripts, especially since nuclei are enriched for non-productive isoforms. In CaptureSeq data, two of the four most highly expressed isoforms in DLPFC, ENST00000677574.1 and ENST00000490912.6, are likely targeted for NMD.

It is important to acknowledge that iPSC-derived cells are an artificial system and do not fully recapitulate the complexity of the adult human brain. Nevertheless, they remain an indispensable model for studying relevant human disease molecular phenotypes and responses under controlled conditions (Leng et al., 2022). They enable mechanistic investigations within a defined genetic background and are increasingly used to test disease-relevant perturbations at high throughput. Co-culture strategies offer a promising route to enhance physiological relevance. In a recent study, Chen et al. (2025a) incorporated iPSC-derived microglia into brain organoids composed of cells of ectodermal origin (neurons and astrocytes), generating a long-term microglia-containing organoid model. In this system, microglia enhanced calcium signalling and other neuronal functions. These findings show that multi-lineage co-cultures offer a promising route to not only more accurately model cell interactions in the human brain, but also promote maturation and activity of constituent lines (Guttikonda et al., 2021; Park et al., 2018; Yang et al., 2023).

3.4.3 Future directions for single-cell and spatial long-read transcriptomics

Single-cell long-read sequencing remains a rapidly evolving field without gold-standard protocols or analysis pipelines. While short-read single-cell methods have become more widely adopted, many analytical aspects are still debated, such as normalisation and statistical testing of differential transcript expression remain debated. Long-read methods introduce additional challenges, such as reduced coverage, sequencing and mapping errors. Discrepancies between software often yield differences in isoform detection and quantification, and systematic benchmarking is still limited. Nevertheless, the field is progressing quickly, with improvements in throughput, error profiles, and computational workflows expected to increase reproducibility.

Spatial long-read approaches represent a particularly exciting avenue. Methods such as Spl-IsoQuant and its successor Spl-IsoQuant2 (Michielsen et al., 2025; Foord et al., 2025) that now combine spatially barcoded cDNA with long-read sequencing demonstrate that isoform regulation is not only cell-type specific but also spatially organised at micro-anatomical scales. Both studies report within–cell-type spatial programmes of splicing and polyadenylation that align with cortical layers, subregions, and even follow gradients that do not respect anatomical boundaries. Despite these advances, technical limitations remain. In single-cell long-read sequencing, low coverage leads to moderate isoform differences in low-abundance transcripts often being missed, and in spatial studies, long-read sparsity prevents robust deconvolution of spatial spots using long reads alone; Foord et al. had to rely on Illumina data for cell-type and layer assignment. Michielsen et al. found that many of their spatially variable isoforms were only detected as significant in their targeted panel. Therefore, a practical next step would be to increase sensitivity with targeted approaches through gene panel capture (CaptureSeq) curated gene sets (e.g., disease loci, synaptic genes) and these can be layered onto whole-exome protocols to balance scope and depth.

Cross-platform comparisons: ONT and PacBio

Comparative analyses of long-read platforms (ONT and PacBio) applied to cell barcoded cDNA remain limited, and there is still no clear consensus on best practices. I have contributed to addressing this gap both in Ahlert Scoones et al. (2025) and in the present study, albeit with different emphases. In Ahlert Scoones et al. (2025), the evaluation centred on company-recommended workflows under their default settings, reflecting typical user practice. Here, by contrast, we applied highly sensitive isoform detection to a deeply sequenced dataset generated on the most recent platform versions, providing a best-case scenario for isoform discovery. Using this approach, we found that 35,317 of 39,394 junction chains (89.7%) detected by both ONT and PacBio corresponded to previously annotated isoforms. A systematic comparison is still needed to assess platform-specific biases in isoform detection and quantification.

Within the broader aims of this PhD—to use long-read sequencing to annotate functionally relevant isoforms in complex human tissues—this chapter demonstrates that moving to single-cell resolution provides a promising route to evaluate the biological significance of novel isoforms. To progress beyond cataloguing, however, these findings will need to be integrated with orthogonal approaches such as protein structure prediction, ribosome and proteomic profiling, and targeted CRISPRi/a and recently developed CRISPR–Cas13d mediated perturbation experiments (Konermann et al., 2018; Schertzer et al., 2023).

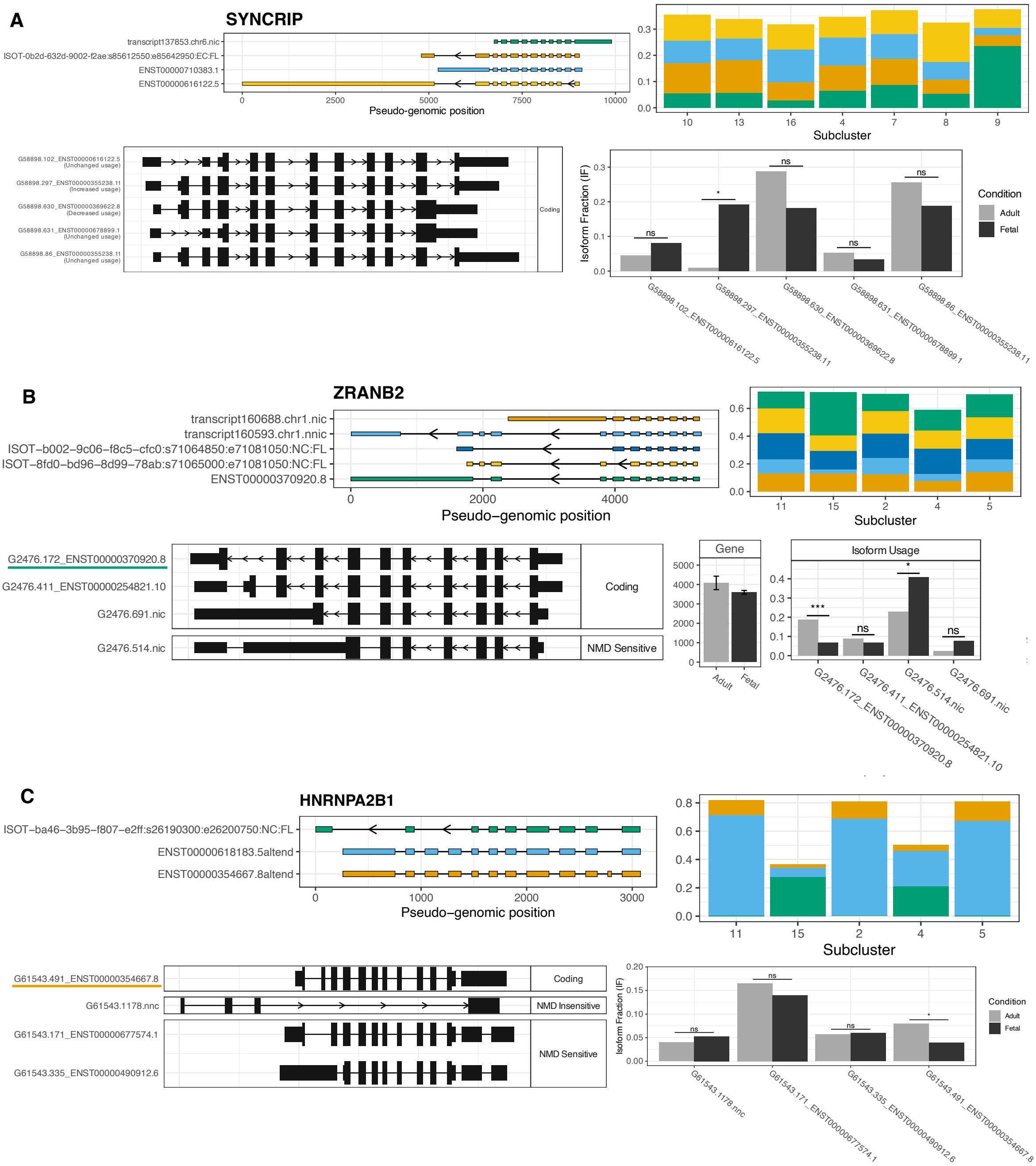


Figure 3.13: **Cross-dataset isoform usage case studies.** Top panels show isoform usage in iPSC-derived single-cell long-read data; bottom panels show targeted bulk CaptureSeq from DLPFC (adult and fetal). In each panel, the left subpanel depicts transcript (exon–intron) structures, and the right subpanel shows relative isoform usage (fractional contribution within the gene) across subclusters (top) and adult vs fetal DLPFC (bottom). (A) *SYNCRIP* usage across microglial clusters. (B) *ZRANB2* and (C) *HNRNPA2B1* usage across neuronal clusters.

Chapter 4

Role of SNORD116 in RNA processing during hiPSC-derived cardiomyocyte differentiation

This chapter was carried out in collaboration with several contributors. All cell culture and cardiomyocyte differentiation were performed by Dr Terri Holmes, University of East Anglia. RNA extraction, library preparation, and Oxford Nanopore sequencing were carried out by Vanda Knitlhoffer, Earlham Institute. LC–MS/MS proteomic data generation and database searching against both UniProt and the custom transcriptome-derived protein database were carried out by Prof. Mandy Peffers' group, University of Liverpool. All downstream data integration and bioinformatic analyses were performed by the author.

4.1 Background

4.1.1 pre-mRNA processing

Investigating the precise changes in gene expression and their regulation during cellular differentiation is crucial for our understanding of normal human development and developmental disorders. Central to this is the processing of nascent RNA. This mechanism includes 5' capping, splicing, and $3'$ end cleavage followed by polyadenylation. Although transcription initiation marks the onset of gene expression, splicing and polyadenylation are required to generate mature RNA that can be exported from the nucleus and shielded from degradation. As cells progress through developmental states, alternative splicing (AS) and alternative polyadenylation (APA) fine-tune gene expression and the mRNA isoform repertoire, with APA generating isoforms that differ in their 3' ends by cleavage at alternative polyadenylation sites (Ji et al., 2009; Agarwal et al., 2021; Kiltschewskij et al., 2023).

4.1.2 Nascent RNA processing is coupled to transcription

Both of these processes are tightly coupled—both mechanistically and kinetically—to RNA Polymerase II (Pol II) transcription and termination (Hirose and Manley, 1998; Bentley, 2014). During transcription, the C-terminal domain (CTD) of Pol II serves as a dynamic scaffold, recruiting a range of RNA processing complexes to the nascent transcript (Bentley, 2005). Specific CTD phosphorylation patterns, collectively referred to as the "CTD code", help recruit different RNA processing factors and influence splicing and cleavage site choice (Bentley, 2014). These include the multi-subunit protein complexes that make up the $3'$ end processing complex, including the cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF) and cleavage factor I (CFIm) (Shi et al., 2009). CPSF recognises the A(A/U)AAA polyadenylation signal. Once Pol II transcribes the poly(A) signal, the CPSF complex binds the polyadenylation site (PAS) and initiates cleavage. This not only triggers the synthesis of the poly(A) tail by poly(A) polymerase (PAP) but also creates an entry point for a $5'$ -to-$3'$ RNA exonuclease Xrn2, which displaces Pol II from the template and ultimately leads to transcription termination (Bentley, 2005; Glover-Cutter et al., 2008; Corden and Patturajan, 1997; West et al., 2004; Connelly and Manley, 1988). A direct consequence of this coupling is that the rate of Pol II elongation can influence poly(A) site selection, where a slower rate tends to favour usage of upstream (proximal) poly(A) sites, while faster elongation facilitates transcription to more distal sites Geisberg et al. (2022); Fong et al. (2015)). This is partly due to the temporal window available for the recruitment of the cleavage machinery.

4.1.3 Coordination of alternative polyadenylation and splicing

A growing body of evidence suggests that this effect may also be concurrently influenced by splicing activity. Specifically, inhibition of splicing components such as U1, U4, and U6 snRNPs using antisense morpholinos (AMOs) has been shown to induce widespread premature cleavage and polyadenylation (PCPA), particularly within the first intron Feng et al. (2025); Yang et al. (2025). This has led to two non-mutually exclusive mechanisms through which splicing may regulate cryptic polyA site use: (i) steric hindrance of cleavage and polyadenylation factors by spliceosomal proteins, and (ii) modulation of RNA Pol II elongation. The former suggests that snRNPs physically bind near cryptic PASs, blocking access to the cleavage machinery. The latter is supported by U1 has been shown to act as a transcription elongation factor, especially for AT-rich intronic sequences.

In addition to suppression of cryptic PASs, splicing and APA are often co-regulated by shared trans-acting RBPs Proteins such as HNRNPC, PTBP1, and ELAVL1 act on both splicing and polyadenylation by binding to overlapping sequence motifs Bak et al. (2024); Ji et al. (2013); Dai et al. (2012). Recent long-read sequencing studies in Drosophila and humans suggest that poly(A) site selection often correlates with splicing of upstream cassette exons in a cell-type-specific manner, especially in neuronal tissues (Hardwick et al., 2022; Herzel et al., 2018; Edwalds-Gilbert et al., 1997). Resolving these coordinated events at the level of individual transcripts is challenging with short-read sequencing, which fragments RNA and requires inference of isoform structures. Long-read sequencing overcomes this limitation by capturing full-length RNA molecules, thereby enabling the unambiguous identification of splicing and polyadenylation events that occur on the same transcript.

4.1.4 Determinants of poly(A) site selection

The core PAS hexamer motif (AAUAAA or its variants) is located 21nt upstream of the cleavage site and is flanked by the auxiliary downstream and upstream U/GU-rich sequences (Shi and Manley (2015); Lianoglou et al. (2013); Legendre and Gautheret (2003)). Together, these motifs and their positions relative to the cleavage site determine the strength of a poly(A) site (Stroup and Ji (2023)). Poly(A)-binding protein nuclear 1 (PABPN1), which recognises A-rich sequences and canonically acts by binding the poly(A) tail and stabilising PAP until the tail is synthesised, can also directly bind to the A(A/U)AAA hexamer and mask it from cleavage factors (Jenal et al. (2012)). In most genes, the distal poly(A) site tends to have a stronger sequence context that prevents PABPN1 from outcompeting the cleavage and polyadenylation complex (Tang et al. (2022)). Depending on the location of the poly(A) sites in the gene, APA can change the transcript's coding sequence or the 3′ UTR (Di Giammartino et al., 2011). As 3'UTRs contain many binding sites for RBPs and miRNAs, APA can affect mRNA stability, localisation, nuclear export, or protein translation efficiency due to altering the presence of these binding sites on the mRNA (Tian and Manley (2017)). These interactions between mRNA 3'UTRs and miRNAs can be further regulated by certain lncRNAs, which can act as competing endogenous RNAs (ceRNAs) and prevent miRNAs from binding to their target sites. For instance, in diabetic cardiomyopathy, the lncRNA MIAT promotes cardiomyocyte apoptosis by sponging miR-22-3p, which normally represses pro-apoptotic DAPK2. In contrast, HOTAIR alleviates oxidative stress in myocytes by sponging miR-34a and thereby upregulating SIRT1 Zhou et al. (2017); Gao et al. (2019).

4.1.5 Roles of APA in development and differentiation

APA plays a critical role in regulating gene expression during cell differentiation and development. Numerous studies have reported global 3′ UTR and poly(A) tail lengthening across tissues and cell types during these processes (Hilgers et al., 2011; Ji et al., 2009; Miura et al., 2013). However, recent advances in 3′ -tagged single-cell RNA-seq technologies have revealed important exceptions, highlighting stage- and cell-type-specific patterns of APA (Shulman and Elkon, 2019; Agarwal et al., 2021; Mitschka and Mayr, 2022). For example, analysis of 2 million nuclei from the mouse organogenesis cell atlas dataset confirmed the overall trend of 3′ UTR lengthening but revealed pronounced 3′ UTR shortening in haematopoietic and spermatogenic lineages, cell types characterised by rapid turnover and proliferation (Cao et al., 2019; Agarwal et al., 2021; Sandberg et al., 2008).

Generally, 3′ UTR lengthening is associated with decreased mRNA stability due to the gain of additional miRNA binding sites and AU-rich elements (AREs), leading to reduced gene expression (Zhang et al., 2023; Khajuria et al., 2023). AREs are cis-regulatory elements that are bound by ARE-binding proteins. These proteins, including AUF1 (ARE-Binding Protein 1/ HNRNPD) and HuR/ELAVL1, often compete or cooperate with other RBPs and miRNAs to fine-tune mRNA decay and translation. ELAVL1 typically protects RNA from degradation, whereas AUF1 promotes RNA decay by recruiting components of the RNA degradation machinery. However, certain AUF1 isoforms can enhance the stability and translation of specific target mRNAs (Yoon et al. (2014); White et al. (2013); Wang et al. (2001)).

Cardiac development

Among other tissues, 3'UTR length changes and APA have been implicated in normal cardiac differentiation and disease (Yang et al. (2022)). Yet, the regulatory mechanisms and functional consequences of APA during cardiomyocyte (CM) differentiation remain incompletely understood. Recent bulk RNA-seq studies have provided a more nuanced view of APA dynamics in differentiating CMs. Yang et al. used fuzzy c-means clustering to identify distinct, pathway-specific APA patterns in differentiating human induced pluripotent stem cell-derived CMs (hiPSC-CMs) (Yang et al. (2022)). One cluster (cluster 5, Figure 4.1A) exhibited a sharp transient $3'$ UTR shortening in the first 24 hours of cell differentiation, which the authors suggest is caused by upregulation of poly(A) machinery gene expression. In contrast, a separate study using human embryonic stem cell-derived CMs (hESC-CMs) found that the most 3'UTR lengthening occurred during mesodermal commitment (days 1–4), followed by a progressive shift back to short UTRs at later stages (days 9–15). Many transcripts involved in these UTR changes did not show altered expression levels but did exhibit differences in polysome association, linking APA during CM differentiation to translational control (Hansel-Frose et al. (2024)). APA regulation in the heart is also coupled to transcription and is influenced by key cardiac transcription factors, such as Nkx2–5. Knockdown of Nkx2–5 in mouse embryos resulted in heart development genes expressing mRNAs with longer $3'$UTRs, likely via increased recruitment of the exonuclease Xrn2, which promotes proximal PAS usage (Nimura et al., 2016). In addition, multiple RBPs contribute to APA regulation during cardiac differentiation. For instance, CELF1 (CUGBP1), a multifunctional RBP associated with splicing and translation, is implicated in conditions such as dilated cardiomyopathy and myotonic dystrophy type 1 (Degener et al., 2022; Chang et al., 2017). *Celf1* is downregulated during postnatal heart development, and its knockout in developing mouse hearts led to changes in AS and upregulation of cell cycle genes with Celf1 binding sites in their 3'UTRs (Giudice et al., 2016; Ladd, 2016; Kalsotra et al., 2008). In vitro studies have shown that CELF1 binds GU-rich elements in $3'$ UTRs and promotes poly(A) tail shortening by recruiting the poly(A) deadenylase PARN, suggesting that it might destabilise cell cycle gene transcripts through PARN recruitment (Moraes et al., 2007). RBFOX2 and PTBP1 are RBPs that antagonistically regulate AS in various tissues, including neurons and muscle. Using long-read and poly(A)click-seq (PAC-Seq), Cao et al demonstrated that RBFOX2 and PTBP1 modulate the splicing of terminal exons of *Trpm1*, a gene essential for muscle contraction, during rat heart development (Cao et al., 2021b,a). Notably, genome-wide $3'$ UTR shortening has also been observed in mouse and rat models of cardiac hypertrophy, representing a reversal of the lengthening trends seen during normal differentiation and suggesting a reversion to a more stem-like, proliferative state (Soetanto et al., 2016; Park et al., 2011b).

4.1.6 Small nucleolar RNAs (snoRNAs)

While proteins and cis-regulatory elements within transcripts are well-established regulators of $3'$ end processing and APA, emerging evidence highlights the involvement of non-coding RNAs (ncRNAs), including small nucleolar RNA (snoRNA), in modulating the activity of the $3'$ processing machinery. One type of ncRNAs recently implicated in $3'$ end processing are small nucleolar RNAs (snoRNAs) (Huang et al. 2017). snoRNAs are enriched in Cajal bodies or nucleoli and are classically known for their role in the post-transcriptional modification of other RNAs, such as tRNA, ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs). They are generally classified as either C/D box snoRNAs (SNORDs) or H/ACA box snoRNAs (SNORAs) based on their secondary structure and conserved sequence motifs. Although there are other subclasses of snoRNAs, such as small Cajal body-specific RNAs (scaRNAs), which primarily localise to nuclear Cajal bodies (Baldini et al., 2021). scaRNAs share sequence and structural features with either C/D- or H/ACA-box snoRNAs or both, which explains their described ability to guide both pseudouridylation and 2'-O-methylation of spliceosomal snRNA (Meier, 2017; Beneventi et al., 2021). Bittel group demonstrated that 12 scaRNAs that target U2 and U6 spliceosomal snRNAs are reduced in ventricular cardiomyocytes of infants with tetralogy of Fallot, a congenital heart disease, and that knockdown of these scaRNAs in zebrafish disrupts mRNA splicing and heart development, at least in part due to loss of scaRNA1-directed pseudouridylation (Patil et al., 2015; Nagasawa et al., 2018, 2020).

Most snoRNAs in humans are encoded within intronic regions of protein-coding host genes and generated when the intron is excised and degraded during splicing. The 5' and 3' exonuclease is blocked from degrading the whole intron by snoRNAs forming stable ribonucleoprotein (RNP) complexes with core RBPs. These snoRNP canonically recruit enzymes, which they guide to specific target RNAs for modification. Specifically, SNORDs preferentially associate with NOP56, NOP58, and SNU13 (NHP2L) and 2'-O-methyltransferase fibrillarin FBL (Figure 4.1B), while SNORAs associate with NOP10, NHP2, GAR1 and the pseudouridine synthase DKC1 (Hayano et al., 2003; Kiss, 2001; Song et al., 2024).

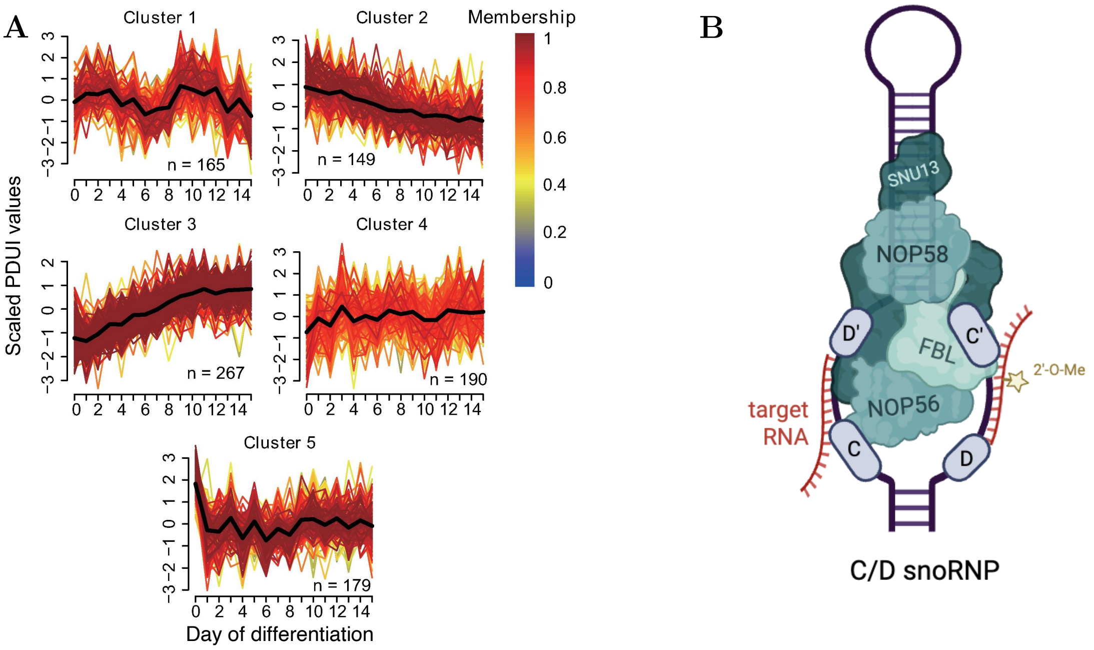


Figure 4.1: **Alternative polyadenylation dynamics and snoRNP structure.** (A) Dynamics of alternative polyadenylation during iPSC-derived cardiomyocyte differentiation (adapted from Yang et al. (2022)). (B) Schematic representation of the C/D box snoRNA–protein complex. Created with Biorender.com.

4.1.7 Emerging functions of snoRNAs in RNA processing

Emerging evidence suggests that some SNORDs can also regulate RNA processing independently of chemical modification through direct binding to target transcripts and proteins. For example, *SNORD27* was shown to regulate splicing of transcription factor *E2F* through direct RNA–RNA interaction without Fibrillarin catalysing methylation of its RNA. *SNORD27* knockdown in HeLa cells increased the inclusion of weak alternative exons in *MAP4K3*, *ZBTB37*, *FER*, and *ABCA8* pre-mRNAs (Falaleeva et al., 2016). *SNORD88C*, formerly known as *HBII-180C*, was shown to be processed into shorter miRNA-like fragments, some of which contain a region highly complementary to pre-mRNA sequences, termed the M-box. Expression of an antisense construct targeting the M-box of *SNORD88C* altered the splicing pattern of *FGFR3*, suggesting that *SNORD88C* regulates splicing through direct base-pairing with intronic sequences (Scott et al., 2012; Ono et al., 2011). Analysis of large human RNA–RNA interaction datasets identified that *SNORD2*, encoded from an intron of *EIF4A2*, regulates AS of its host transcript by forming an RNA duplex with a conserved intronic region near an alternative exon, masking the branch point, which leads to decreased inclusion of that exon (Bergeron et al., 2023). Using a pull-down assay, Huang et al. (2017) identified nine box C/D snoRNAs associated with the mRNA 3′ processing complex, independent of the core snoRNP proteins. Further investigation revealed that the U/A-rich *SNORD50A* inhibits mRNA 3′ end processing by competing with the PAS for binding to Fip1, a component of the CPSF complex (Huang et al. 2017; J. Shi et al. 2018).

To improve the identification of noncanonical snoRNA interactions, Deschamps-Francoeur et al. developed snoGlobe, a predictive model based on a gradient boosting classifier (Deschamps-Francoeur et al., 2022). Trained on previously identified box C/D snoRNA–target interactions from high-throughput crosslinking and sequencing datasets, snoGlobe searches for potential interactions in the whole snoRNA sequence. Transcriptome-wide search revealed that interaction sites were enriched in exons and exon–intron junctions, particularly within 5′ and 3′ UTRs, relative to their general distribution across the transcriptome. To experimentally validate its predictions, *SNORD126* was knocked down in HepG2 cells, resulting in alternative splicing events in several of its predicted target genes, overlapping the predicted binding sites. Notably, three of these genes, *CPT1B*, *MR1*, and *DDX11*, were also differentially expressed, suggesting that *SNORD126* may influence RNA stability through its effects on splicing regulation.

4.1.8 SNORD116

Many snoRNAs, however, have no described function or experimentally confirmed targets and are therefore classed as ‘orphan’. One of the most notable examples is *SNORD116*, a C/D box snoRNA cluster located within the imprinted Prader-Willi syndrome (PWS) critical region on chromosome 15q11–13. The cluster consists of 30 paralogous copies embedded within the long non-coding host gene *SNHG14*, which spans $\sim$ 600 kb. Loss of paternal expression of genes within 15q11.2–q13 accounts for the majority of PWS cases. While most arise from large 5-6Mb deletions, smaller microdeletions restricted to the *SNORD116* cluster alone are sufficient to recapitulate the core features of PWS, including neonatal hypotonia, hyperphagia, obesity, and cognitive impairment (Tan et al., 2020; Bieth et al., 2015).

In contrast, deletion of the neighbouring SNORD115 cluster does not cause PWS (Runte et al., 2005). Although early studies in mice suggested that *Snord115* regulates alternative splicing of the serotonin receptor 2C (*Htr2c*) pre-mRNA (Kishore et al., 2010; Cavaillé et al., 2000), subsequent knockout models reported no detectable effect on *Htr2c* splicing *in-vivo* (Hebras et al., 2020). Several other snoRNAs are generated from the introns of *SNHG14*, including *SNORD115*, *SNORD107*, *SNORD64*, *SNORD108*, *SNORD109A*, and *SNORD109B*. Beyond mature snoRNAs, the *SNHG14* locus also gives rise to hybrid lncRNAs such as SPA-lncRNAs and sno-lncRNAs, which are formed from introns flanked by snoRNAs. In sno-lncRNA transcripts, snoRNAs act as stabilisers instead of the 5′ cap and poly(A) tail (Yin et al., 2012). These PWS region lncRNAs sequester splicing regulators such as TDP43, RBFOX2, and hnRNPM, thereby modulating alternative splicing (Wu et al., 2016).

Mechanism of action

The precise molecular function of *SNORD116* remains unclear, but recent studies point to diverse, potentially context-dependent roles (Holmes et al., 2025). Using snoKARR-seq, Liu (2025) identified 28 RNA targets of *Snord116* in mouse cortex, including an mRNA (*Iqcg*), lncRNAs (*Meg3, Malat1, Gm44066*), tRNAs, and snRNAs. Notably, RNA interactions of most snoRNAs identified in the study varied across cell lines, highlighting the importance of cellular context in defining *SNORD116* targets. *SNORD116* has been linked to RNA stability and alternative splicing. In mouse neuronal cells, *Snord116* overexpression altered the decay rate of *Nhlh2* mRNA, dependent on the length and type of 3′UTR used in the *Nhlh2* construct (Kocher et al., 2021), while knockdown in a human HeLa S3 cell line increased exon 3 inclusion in *Nlgn3* mRNA (Baldini et al., 2022). Other evidence suggests that *SNORD116* may act primarily at the level of RNA stability or translational regulation. A recent study in LUHMES cells, a human dopaminergic neuronal progenitor line, reported that *SNORD116* KO affected mRNA stability and protein synthesis, with some genes showing altered protein levels despite unchanged mRNA (Helwak et al., 2024). No strong evidence was found that *SNORD116* regulates alternative splicing; changes to transcript structure were mostly due to alternative TSS and TTS sites, not exon inclusion. Comparative sequence analyses have shown that these paralogs can be grouped according to sequence similarity and evolutionary conservation, raising the possibility of subgroup-specific functions (Good and Kocher, 2017; Baldini et al., 2022). Figure S1 illustrates representative predicted secondary structures of three SNORD116 paralogs from each of the three subgroups, as annotated in RNAcentral (Sweeney et al., 2021). The structural models reveal subgroup-specific differences in stem–loop configurations and predicted guide regions. However, the extent to which individual groups contribute to post-transcriptional regulation remains poorly understood.

While much of the research on *SNORD116* has focused on neural tissues, emerging clinical evidence points to a high incidence of congenital heart defects in individuals with PWS, suggesting that *SNORD116* may also play a role in cardiac development. Supporting this, *SNORD116* expression has been found to be elevated in a hiPSC model of cardiomyopathy. Despite these associations, no studies to date have directly investigated *SNORD116*'s role in RNA regulation within the developing heart.

4.1.9 Aims of this chapter

We set out to determine whether *SNORD116* regulates pre-mRNA processing during human cardiac differentiation, and if so, through which mechanisms and on which transcript targets. To address this, Oxford Nanopore long-read sequencing of human induced pluripotent stem cell-derived cardiomyocytes (hiPSC-CMs) was generated at three differentiation stages. My contribution focused on the bioinformatic analysis of these data, which enabled the simultaneous quantification of transcript features that are difficult to resolve with short-read approaches, including alternative splicing, cleavage and polyadenylation, and poly(A) tail length, within a single experimental framework. I further prioritised candidate *SNORD116* targets using binding-site prediction and by integrating isoform-level changes with evidence for altered translation from high-throughput proteomics (provided by collaborators).

4.2 Methods

SNORD116$^{-/-}$ human induced pluripotent stem cells (hiPSCs) were developed and kindly provided by the Talkowski Lab (Broad Institute, MIT & Harvard Universities). All the following cell culture and differentiation were carried out by Dr Terri Holmes and Dr James Smith (UEA Medical School). Briefly, human cardiomyocytes were differentiated from hiPSCs using a small-molecule-based monolayer protocol adapted from (Dark et al., 2023). Differentiation was initiated in B8 medium, followed by staged treatment with CHIR99021, BMP4, Activin A, and Wnt inhibitors to guide mesodermal and cardiac lineage specification. At Day 20, the differentiation medium was replaced with glucose-depleted, fatty acid-enriched media adapted to promote metabolic maturation of the cardiomyocytes.RNA was extracted at days 2, 6 and 30 of differentiation.

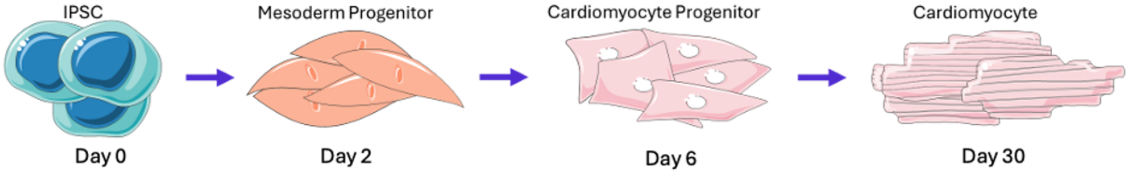


Figure 4.2: **Schematic representation of different stages of cardiomyocyte differentiation from iPSCs.** Figure generated by Terri Holmes.

Library preparation and sequencing were carried out by Vanda Knitlhoffer and are detailed here in short. Barcoded cDNA libraries were prepared from RNA with the Oxford Nanopore PCR Barcoding kit (SQK-PCB114.24), pooled and sequenced on two PromethION R10 flow cells. Basecalling was performed through MinKNOW v24.06.16 in real time with Dorado v7.4.14 high-accuracy model `dna_r10.4.1_e8.2_400bps_hac@v4.3.0` with minimum mean quality score of the bases in the read = 9 (https://github.com/nanoporetech/dorado).

Some code snippets were drafted or refined with the assistance of ChatGPT-4 and 5 (OpenAI, 2025), and subsequently validated by the author.

4.2.1 Custom transcriptome

Full-length reads were identified and reoriented with Restrander (Schuster et al., 2023). Minimap2 (Li, 2018) was used to align the reoriented reads to the human reference genome (GRCh38) twice. First with: `-ax splice --cs=long -G 500k -ub --secondary=yes`. This was followed by filtering high-confidence junctions using 2passtools (Parker et al., 2021): `score and filter --exprs 'count > 2 and jad > 3 and (decision_tree_1_pred or decision_tree_2_pred)'`. These were then passed to the second alignment step using:`-ax splice -k 14 -ub -G 500k --junc-bed $filt_juncs --junc-bonus 15 --secondary=yes`. Unmapped reads and supplementary alignments, primary alignments with large deletions (> 20bp) in the CIGAR string or supplementary alignments that reverse overlap the primary alignment were removed. Filtered alignments were clustered into unique isoforms with IsoQuant v3.6.1 (Prjibelski et al., 2023) using: `--data_type nanopore --splice_correction_strategy conservative_ont --model_construction_strategy sensitive_ont --polya_requirement auto --report_canonical all` Transcript models generated with IsoQuant were then classified by SQANTI3 v5.2.2 with reference to GENCODE v46.The Reference Transcription Starting Sites v4.1 (refTSS, (Abugessaisa et al., 2019)) and PolyASite v2.0 ((Herrmann et al., 2020))databases were used to validate transcript start and end sites, respectively.

Tool	**Version**	**Reference**
Bedtools	2.30.0	Quinlan and Hall (2010)
clusterProfiler	4.14.6	Wu et al. (2021)
edgeR	4.4.2	Chen et al. (2025b)
IsoformSwitchAnalyzer	2.6.0	Vitting-Seerup and Sandelin (2019)
IsoQuant	3.6.1	Prjibelski et al. (2023)
Minimap2	2.24	Li (2018)
Nanoplen	0.9.0	Dar et al. (2024)
PicardTools	2.23.9	https://broadinstitute.github.io/picard/
Restrander	1.0.0	Schuster et al. (2023)
Samtools	1.15.1	Danecek et al. (2021)
SeqKit	2.5.1	Shen et al. (2016)
SQANTI3	5.1	Pardo-Palacios et al. (2024a)
kallisto	0.50.0	Loving et al. (2024b)
bustools	0.42.0	Melsted et al. (2021)
2passtools	0.3	Parker et al. (2021)
DaPars2_LR	2.1	Feng et al. (2018)
LAPA	1.0.0	Çelik and Mortazavi (2022)
SnoGloBe	1.1	Deschamps-Francoeur et al. (2022)
CPC2	–	Kang et al. (2017)
PFAM	–	Mistry et al. (2021)
DeepLoc2	2	Ødum et al. (2024)
DeepTMHMM	1.0.21	Hallgren et al. (2022)

Table 4.1: Tools, versions and associated references used in Chapter 4.

4.2.2 Gene and isoform abundance estimation

The custom transcriptome fasta was used to build a kallisto index with kmer length of 63 (Loving et al., 2024b). lr-kallisto is more computationally efficient than IsoQuant and Bambu. Pseudoalignment of reads to the index was performed with kallisto bus –long, followed by bustools sort and bustools count. Finally, the estimated counts were generated from transcript-compatibility-counts with `kallisto quant-tcc --long --platform ONT`.

4.2.3 Differential expression

edgeR catchKallisto() function was used to import kallisto count estimates into R v 4.4.3 Baldoni et al. (2024). catchKallisto uses bootstrap resamples to estimate read-to-transcript ambiguity dispersion for each transcript. The DGEList object was created using scaled counts, generated by dividing Kallisto counts by this overdispersion parameter. Lowly expressed genes and transcripts were then filtered out from the count matrices, so that only genes with at least 10 counts in all samples of one group and transcripts with at least 3 counts in one group were kept. Data dispersions, including quasi dispersions from kallisto were estimated and model fitted with edgeR glmQLFit() using a model 0 + group, where the group represented a combination of genotype and day. `glmTreat() with lfc_threshold=`$\log_2$`(1.5)` was used to determine differentially expressed genes and transcripts. `edgeR::plotMDS(method="logFC")` function was used to explore the amount of variance contributed by different variables. As differentiation time was the main contributor to variance, instead of including it as a model term, we instead proceeded with pairwise comparisons and used GENCODE IDs to determine genes and transcripts that were consistently differentially expressed.

4.2.4 Gene ontology analysis

Differentially expressed genes and $\log_2$fold changes identified with edgeR were used for gene ontology (GO) enrichment analysis with enrichGO() function clusterProfiler(v 4.10.1), with options for pAdjustMethod = ‘BH’ and qvalueCutoff = 0.05. org.Hs.eg.db (v3.18.0) was used to convert gene identifiers. The universe used was the list of all genes expressed in each DGEList object. GO results were simplified using the simplify() function from clusterProfiler with cutoff = 0.7. Dotplots were generated using enrichplot (v1.22.0) dotplot function, ordered by GeneRatio.

4.2.5 Differential isoform usage

To detect isoform switching events, we applied the DEXSeq model via `IsoformSwitchAnalyzer` (Anders et al., 2012; Vitting-Seerup and Sandelin, 2019). Similarly to DE analysis, lowly expressed genes and isoforms were filtered to minimise noise in downstream analyses. Specifically, genes were kept only if they had at least 5 counts in both comparison groups, and all single-isoform genes were removed, removing 51,196 transcripts (65.22%), leaving 27,667 isoforms for testing. Changes in isoform composition were identified by comparing isoform fractions (IF, relative isoform expression within a gene). Significant switching events were defined as those with a ΔIF > 0.1 and FDR < 0.05. Coding potential, functional domains, subcellular localisation and topology of transmembrane proteins were predicted using `CPC2` (Kang et al., 2017), `PFAM` (Mistry et al., 2021), `DeepLoc2` (Ødum et al., 2024), and `DeepTMHMM` (Hallgren et al., 2022), respectively.

4.2.6 Alternative polyadenylation

Alternative polyadenylation analysis was carried out using two tools, `DaPars2_LR` and `LAPA` (Dondi et al., 2023; Çelik and Mortazavi, 2022). The former applies to the category of tools initially developed for short-read sequencing that identify potential PAS sites by detecting drops in read coverage at the $3'$ ends of transcripts. Both tools require a transcript model annotation to identify the coordinates of $3'$UTRs, so the custom annotation's format was adapted to fit the tools' requirements. `LAPA` requires the UTRs to be annotated as either `five_prime_utr` or `three_prime_utr`. `DaPars2`, due to working with coverage values, does not deal well with genes with overlapping $3'$UTRs. `DaPars2` therefore has a script `DaPars2_LR_Filter_Anno.py` that detects and filters out genes with overlapping $3'$UTRs. The script removes entire genes instead of resolving overlaps at the isoform level.

The degree of difference in APA usage is quantified as a change in Percentage of Distal poly(A) site Usage Index (ΔPDUI), which can then be used to infer $3'$UTR lengthening (positive) or shortening (negative). Differences in ΔPDUI were considered statistically significant where adjusted P-value $<$ 0.05, and ΔPDUI > 0.1.

With the `-ub` argument enabled, `minimap2` will attempt to infer the strand of the transcript the read originated from by checking for the canonical splice junction sequence (GT–AG) at the boundaries of long gaps (introns). Instead of flipping the alignment, `minimap2` preserves the original sequence direction in the CIGAR and alignment and instead annotates transcript strand information in the `ts` tag, where `ts:A:+` indicates that the read strand matches the transcript strand, and `ts:A:-` indicates the read strand is opposite to the transcript strand. The strand flag (`0x10`), on the other hand, reflects the alignment strand relative to the reference genome. ONT cDNA protocol is unstranded. While `restrander` partially mitigates this, we followed the suggestion of the `LAPA` authors that the stranding information in the alignment file be pre-processed so that the `ts` tag and the flag agree.

Lower-quality alignments with `MAPQ < 10` and all secondary alignments were filtered out. Only reads with poly(A) tails over 10 bp long detected in soft-clipped regions were used for poly(A) site detection, and clusters were filtered out if they were present in under three samples. Barcode and primer trimming before tail detection is not required as `LAPA` uses dynamic scoring to output the longest high-scoring segment of A/T matching bases. Poly(A)-site usage is quantified as the percentage of read-end counts per site relative to total reads for the gene, and statistical tests such as Fisher's exact test assess differences between conditions.

4.2.7 Poly(A) tail length

Recent versions of Dorado (>v0.8.0) have the initial support for estimating poly(A) tail lengths. Dorado struggles to call long homopolymers accurately, so tail lengths are estimated directly from the raw current signal. For each read, the barcode and estimated poly(A) tail lengths were obtained using `pysam` from the `BC` and `pt:i` tags, respectively, generated by Dorado's in-built `--estimate-poly-a` parameter. The reads were then annotated with transcript IDs using IsoQuant's `transcript_model_reads.tsv` assignment file. These were then filtered to only include transcripts that have at least three reads with tails over 10 nt in each of the samples in the per-day comparisons. To calculate differences in poly(A) tail length between conditions, the resulting transcript-to-tail length table was used as input for `Nanoplen`'s Wilcoxon test (`p < 0.05; |logFC| > 0.1`) (Dar et al., 2024).

4.2.8 Binding site prediction

SnoGloBe (Deschamps-Francoeur et al., 2022) was used to predict potential snoRNA interaction sites in all the genes expressed in the samples, i.e. those present in the DGEList object after expression level filtering. To reduce the search space, targets were only searched for in exons of the expressed genes. The searching was parallelised by splitting the genome fasta by chromosome. As suggested by the authors, we filtered the predicted sites to only include those having at least 3 consecutive windows (equivalent to 15 bp) with probability greater than or equal to 0.98 (-t 0.98 -m -w 3). bedtools intersect -f 0.5 -F 0.5 -e was used to get overlaps between predicted interaction sites, with a minimum 50% overlap required for either of the intervals.

4.2.9 Metagene analyses

A custom script was used to divide regions of transcripts (5′UTR, CDS and 3′UTR) into bins with number of bins proportional to the median length of the region. Sites with more than one potential target isoform were filtered to retain only the top-most highly expressed WT isoform for density plotting.

4.2.10 Transcriptome-proteome analysis

Proteomic profiling was performed by the Peffers laboratory using a label-free LC–MS/MS workflow. Protein extracts from the same cell populations used for RNA-seq were digested with trypsin, separated by reverse-phase liquid chromatography, and analysed by tandem mass spectrometry. The resulting raw spectra were imported into Progenesis QI (Nonlinear Dynamics) for preprocessing. Features were aligned, followed by peak detection and quantification of peptide ion intensities. Peptide identification was carried out by database searching against two references: the UniProt human proteome, and the custom database generated from predicted coding sequences of the long-read transcriptome. Differential protein abundance was then assessed within Progenesis QI using normalised ion intensities of non-conflicting peptides.

To enable integrated transcriptome–proteome comparisons, proteomic quantifications were linked to RNA-seq expression values. Gene- and transcript-level counts from `kallisto` were pre-processed in edgeR, removing lowly expressed genes and computing normalised counts-per-million (CPM). For each sample, corresponding label-free quantification (LFQ) protein intensities were averaged across replicates within WT and KO conditions. Protein-to-RNA (P/R) ratios were then calculated for each gene and transcript, and differential regulation was assessed as the $\log_2$ fold-change in KO relative to WT. Thresholds for significant dysregulation were set at $|\log_2(\Delta\text{ratio})| \geq 2$ for gene-level comparisons and $|\log_2(\Delta\text{ratio})| \geq 1.2$ for transcript-level comparisons. For peptides mapping to multiple isoforms in the custom database, the most highly expressed RNA isoform was selected for downstream comparison.

4.3 Results

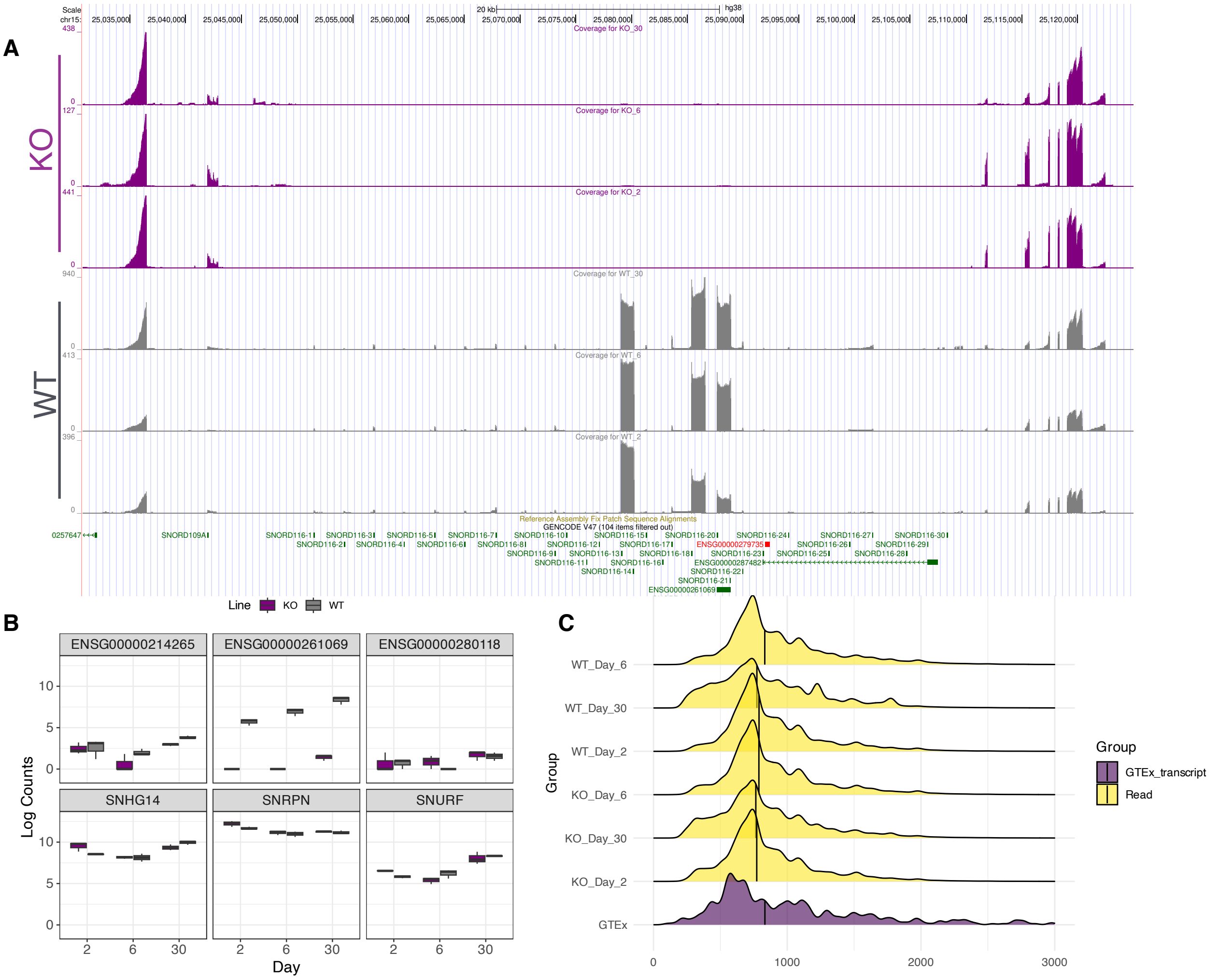


Figure 4.3: **Transcriptomic landscape of the chr15q11–q13 region in iPSC-derived cells.** A) UCSC Genome Browser image displaying read coverage tracks from each day and genotype. Purple tracks show coverage on the + strand from the knockout; grey tracks show coverage from the WT samples. GENCODEv47 gene annotations are shown at the bottom; noncoding genes are shown in green, and To Be Experimentally confirmed (TEC) biotype genes are shown in red. B) Box and whisker plots of log_2 count for a subset of genes in the chromosome 15q11-q13 region. C) GTEx transcript length distribution of the 50 most highly expressed genes in the left ventricle in purple.

To assess the effects of the deletion of SNORD116 on gene and transcript expression, we performed Nanopore long-read RNA sequencing at three time points: days 2, 6, and 30. Due to snoRNAs being monoexonic, their transcripts were removed from downstream differential expression analysis, but coverage analysis confirmed the accurate deletion of the SNORD116 cluster. The overall gene expression of SNHG14 was only significantly reduced at Day 30, however, a different set of its transcripts was expressed. Expression of other protein-coding genes in the region, SNURF-SNRPN was retained. Expression of ENSG00000261069, a sno-lncRNA capped with SNORD116-20 and SNORD116-2, was reduced.

4.3.1 Construction of cardiomyocyte transcriptome

On average, 13 million reads per sample passed the Q-score > 8 filter. According to Lopes et al. 2021 human transcript length distribution peaks at 2,065 bp. To ensure the read length distribution captures the biologically relevant transcripts we compared it to the transcript length distribution of the 50 most highly expressed genes in the human left ventricle, as reported by GTEx v10 (Figure 4.3C, GTEx Consortium (2020)). Although a bias towards lengths under 1000bp is present, the distributions are comparable. The median read length across our samples was 752 bp, with an N50 of 952 bp, compared to a median transcript length of 876 bp in the GTEx left ventricle dataset. Full-length reads containing both 5' and 3' primers were extracted and pooled across all samples. They were then aligned and collapsed into a unified set of transcripts using IsoQuant. Short-read RNA-seq data generated for Johnson et al. (2024) were used for support-based filtering. Following quality control and filtering with SQANTI3, a total of 81,149 isoforms were identified. The 3' and 5' ends of a significant proportion of novel isoforms are supported by refTSS and PolyASite atlas.

4.3.2 Differential gene and transcript expression

Lowly expressed genes and transcripts were filtered out from the count matrices, resulting in 13,922 genes and 45,044 isoforms used for differential expression. Differential expression was evaluated pairwise between wildtype and KO at each time point. Initially, a model taking both time and genotype into account was applied, however, time or differentiation stage was the more significant driver of variance between samples (Principal Component (PC) 1 captures 53% of variation based on the largest log-fold changes, Figure 4.4A), with most of PC2 driven by Day 6 (early spontaneous contracting CMs). edgeR glmQLFTest() was first tested to determine differentially expressed (DE) genes and transcripts; followed by a filtering by log-fold change. However, it was later replaced with glmTreat(), which still conducts a QL F-test but it incorporates the fold-change threshold into the statistical model, shifting the null hypothesis from: $H_0 : \log_2 \mathrm{FC} = 0$ to $H_0 : |\log_2 \mathrm{FC}| \leq \text{threshold}$ (Chen et al. 2025). Across all time points, between 19.2% and 31.0% of DE transcripts were not annotated in GENCODE, comprising novel isoforms classified as novel in catalog (NIC) and novel not in catalog (NNIC, Figure 4.4B). Proportion of novel isoforms decreased with differentiation stage, which is expected as mature tissues are better annotated than differentiated cells. These findings underscore the limitations of relying solely on existing transcript annotations and highlight the importance of custom isoform annotation enabled by long-read sequencing.

The number of genes and transcripts differentially expressed between KO and wildtype cells was highest at Day 6 early cardiomyocyte stage (914 genes; Figure 4.4, Supplementary tables S1-S6). We observed greater overlap in DE genes and transcripts between days 2 and 6 than between days 6 and 30 (Figure 4.4C–D). Overall, only 27 genes were consistently dysregulated across all stages, indicating that most effects of SNORD116 loss are stage-specific. DE genes (Supplementary Sheets 1–3) were classified into up- and downregulated sets for GO term enrichment analysis. At Day 2, genes upregulated in *SNORD116*$^{-/-}$ samples were significantly enriched for metal ion response pathways, including "*response to cadmium ion*" and "*transition metal ion homeostasis*" (adjusted $p < 0.05$). At Day 30, upregulated DE genes were enriched for terms such as "*negative regulation of peptidase activity*".

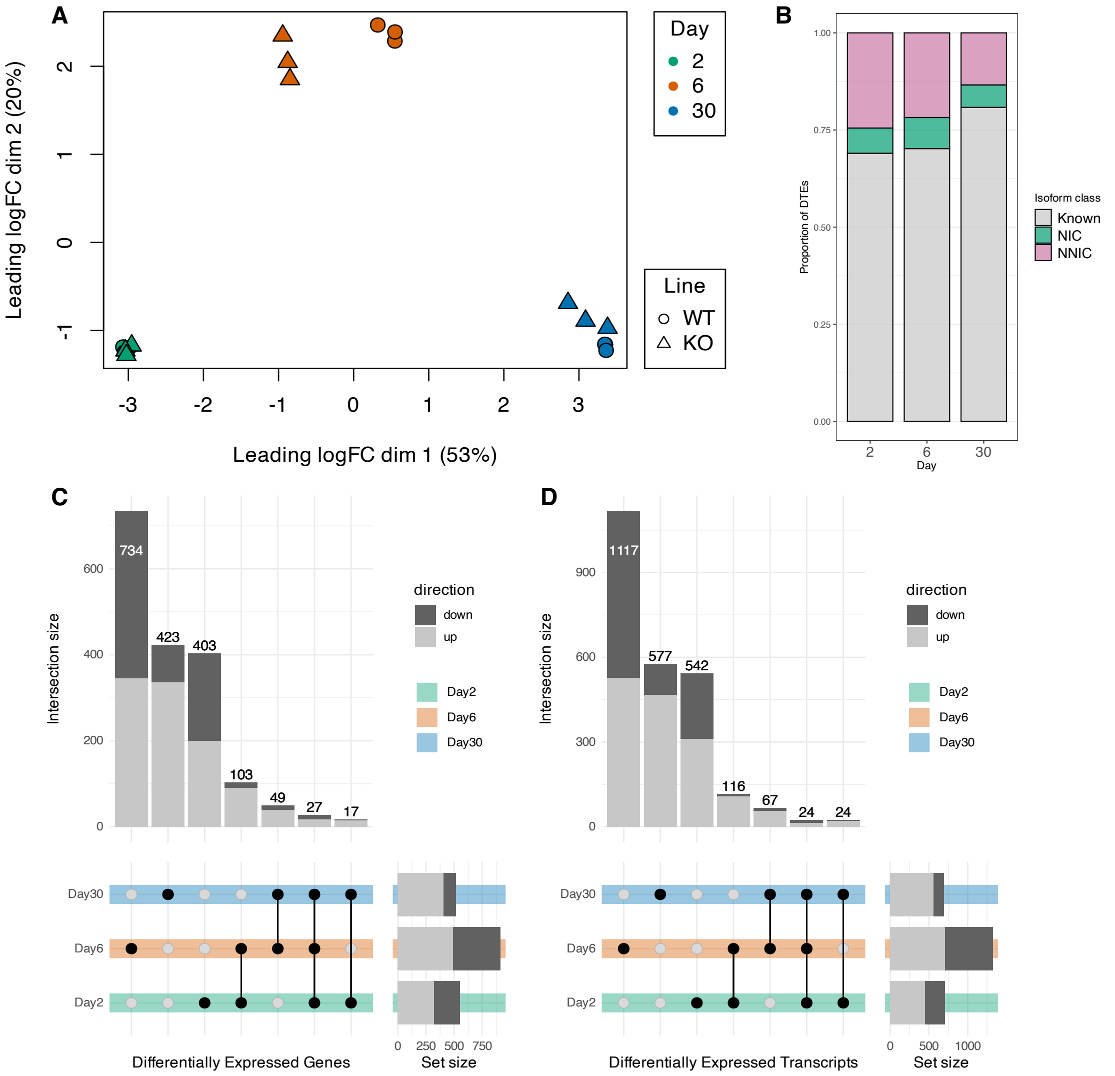


Figure 4.4: **Differential expression in *SNORD116* knockout versus wildtype cardiomyocytes.** A) A multidimensional scaling (MDS) plot generated after expression-level filtering and normalisation, where distances between points (samples) represent the leading $\log_2$-fold changes of transcripts. Wild-type (WT, circles) and *SNORD116* knockout (KO, triangles). Dimension 2 is driven by Day 6. B) Proportions of novel isoforms differentially expressed for each day (NIC – novel in catalog; NNIC – novel not in catalog). C-D) UpSet plots summarising the (C) DEG and (D) DTE overlap between differentiation day comparisons.

4.3.3 Differential isoform usage

To further explore isoform-level regulation following SNORD116 knockout, we performed pairwise isoform switch analysis across each sample collection day. Differential transcript usage (DTU), or isoform switching, refers to changes in the relative contribution of the isoforms to the overall gene expression between conditions. The isoform switch was quantified by the difference in isoform fraction (ΔIF), and switches were considered significant if $\Delta\text{IF} > 0.1$ and $\text{FDR} < 0.05$. In total, 532 significant isoform switches were identified across the three time points (Figure 4.7A, Supplementary Table S7), with Day 6 having the most changes as in previous analyses.

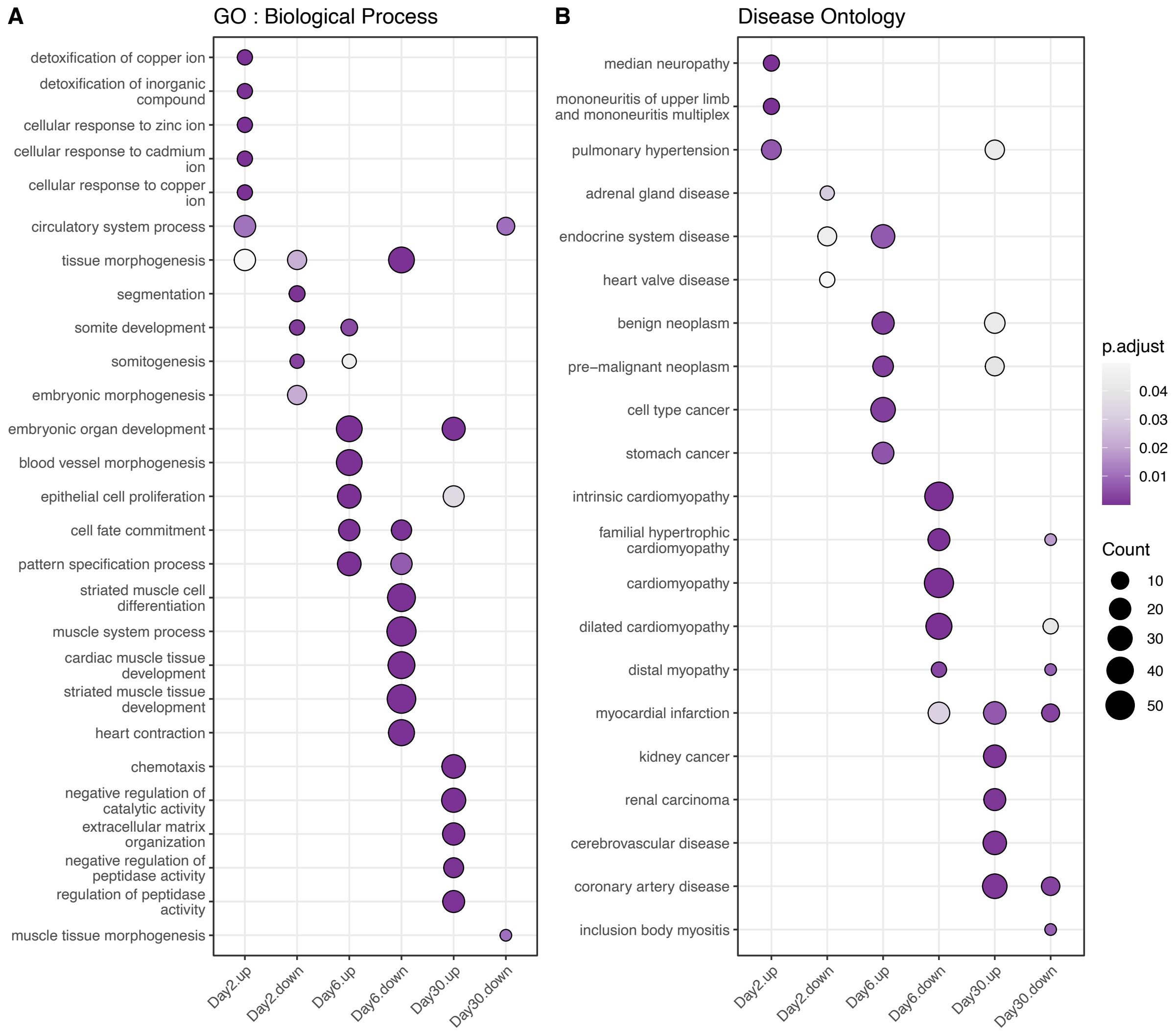


Figure 4.5: **Functional enrichment of differentially expressed genes.** (A) Gene Ontology (GO) biological process enrichment and (B) Disease Ontology term enrichment, summarised by semantic similarity.

Due to there being fewer of the switching events, all genes with significant isoform switches across all three time points were used for GO terms enrichment analysis. These were expectedly enriched in muscle development-related GO-terms, but also in components of the pre-spliceosome complex: PRPF40A, U2AF2 and LUC7L, LUC7L2 (Figure 4.7C). Notably, the same isoform switch in *LUC7L* was observed at both days 2 and 6, with wildtype cells showing higher usage of ENST00000354926.9, while *SNORD116*$^{-/-}$ cells preferentially expressed ENST00000619796.4. Another notable gene identified was *NOP56*, which encodes a core component of box C/D snoRNP complexes (see subsection 4.1.6). We detected a novel isoform, transcript1803.chr20.nic, which was significantly upregulated in Day 6 *SNORD116*$^{-/-}$ samples. This finding contributes to our understanding of the autoregulatory mechanisms governing NOP56 expression. Lykke-Andersen et al. (2018) previously demonstrated that an isoform of *NOP56* retaining SNORD86 prevents the formation of a protein-coding transcript, thereby modulating NOP56 levels through a feedback mechanism. Consistent with this, transcript1803.chr20.nic includes the SNORD86 sequence within one of its exons (4.6B), supporting a model in which non-coding isoform expression contributes to the post-transcriptional regulation of *NOP56*.

Unlike standard differential expression analysis, which focuses on absolute changes in isoform

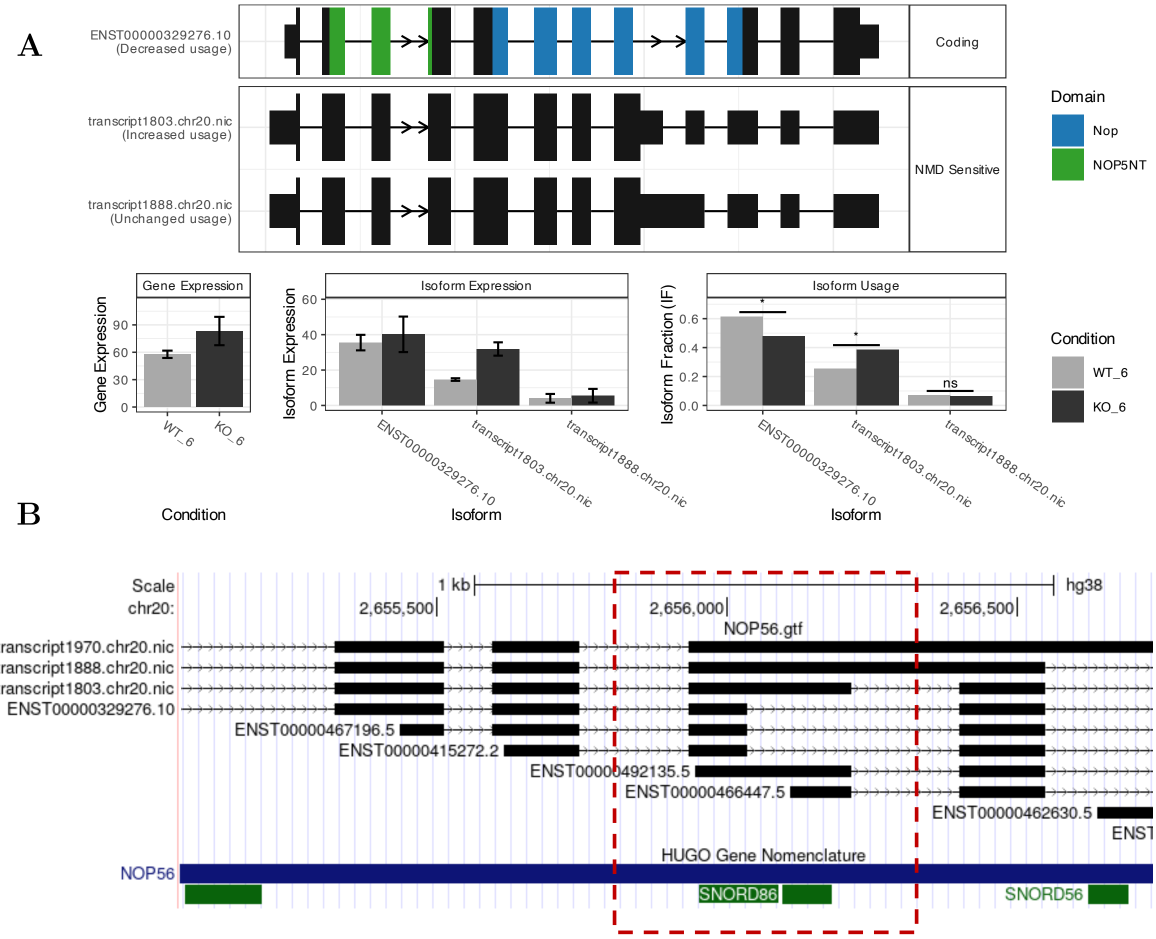
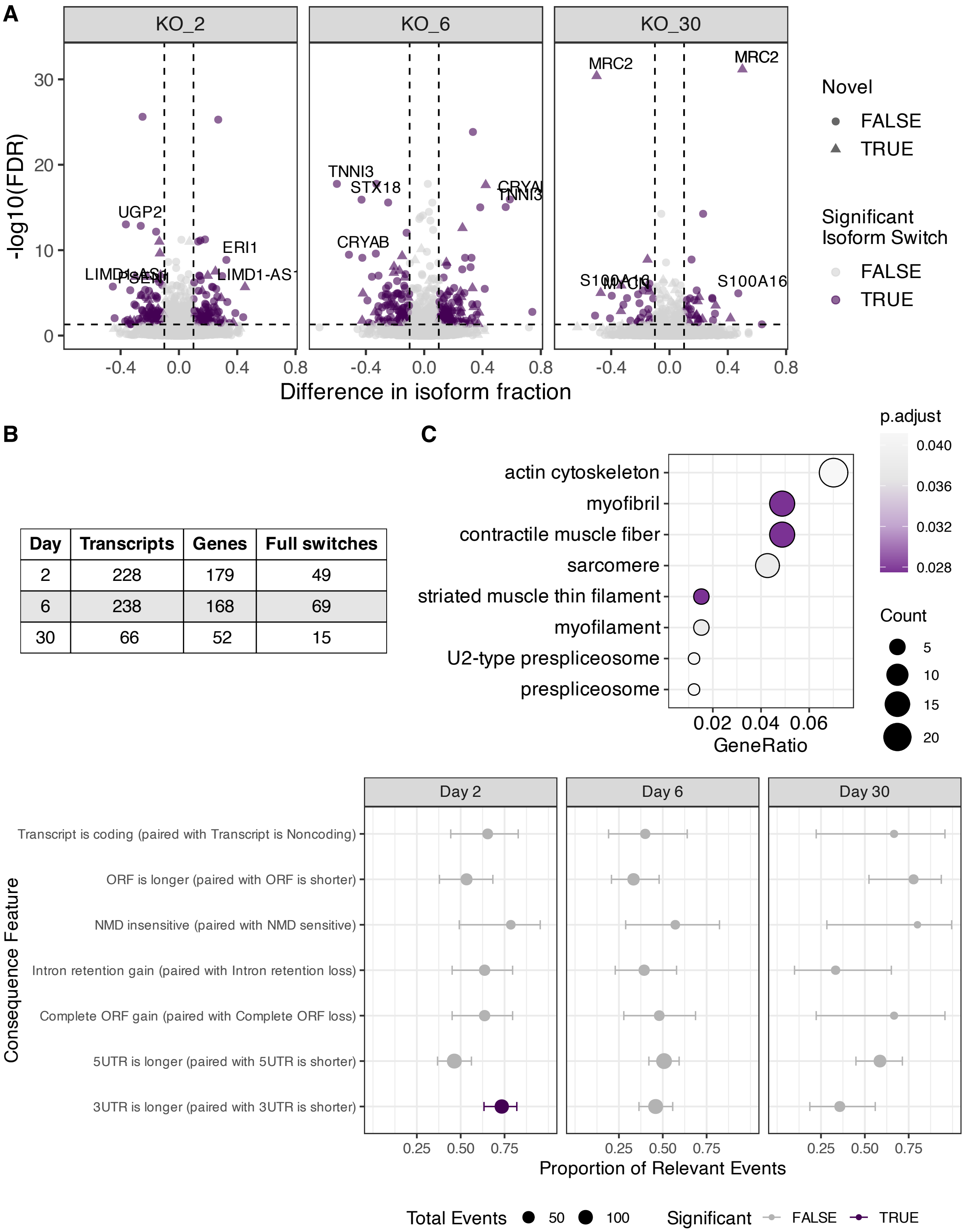


Figure 4.6: **Isoform switching and functional characterisation of NOP56 transcripts.** (A) Isoform switching analysis of the *NOP56* gene, showing significant upregulation of the novel isoform transcript1803.chr20.nic in *SNORD116*$^{-/-}$ samples at Day 6. (B) UCSC Genome Browser view of *NOP56* isoforms. The *SNORD86* sequence is retained in transcript1803.chr20.nic and transcript1888.chr20.nic, but absent from the annotated reference isoform ENST00000329276.10.

abundance, isoform switch analysis identifies reciprocal proportional changes between isoforms. As each switch event involves at least one isoform increasing and another decreasing in relative usage, it enables direct comparison of their structures to infer functional consequences—such as changes in protein domains, or UTR length. This comparative approach is not possible when analysing isoforms in isolation. We used IsoformSwitchAnalyzer's built-in tools to identify the most likely ORF and any PTCs in all the expressed isoforms of genes with at least one differentially used isoform. The amino acid sequences of isoforms not predicted to be subject to NMD and predicted to be protein-coding (CPC2 coding potential value > 0.5; Kang et al. 2017) were extracted; and were then used to predict the protein domains and subcellular localisation of the resulting protein isoforms using external tools PFAM (Mistry et al. 2021), DeepLoc2 (Ødum et al. 2024) and DeepTMHMM (Hallgren et al. 2022), respectively. The functional effect of isoform switching was varied, but one significant effect was enrichment for 3′UTR lengthening in *SNORD116*$^{-/-}$ cells at Day 2. Although no significant enrichment was observed at later stages, we detected a reversal in the direction of 3′UTR changes by Day 30 (Figure 4.7D). These switching isoform-level findings were consistent with results from transcriptome-wide alternative polyadenylation analysis.



Day	Transcripts	Genes	Full switches
2	228	179	49
6	238	168	69
30	66	52	15

Figure 4.7: **Isoform switch analysis across cardiomyocyte differentiation stages.** A) Isoform switching between WT and *SNORD116*$^{-/-}$ samples at each time point. Points represent isoforms. Horizontal and vertical dashed lines indicate thresholds for significance (ΔIF > 0.1 and FDR < 0.05). Novel isoforms are highlighted as triangles. B) Summary table showing the number of transcripts, genes, and full isoform switches detected at each time point. C) GO term enrichment analysis for all genes with significant isoform switches. D) Consequences of isoform switches.

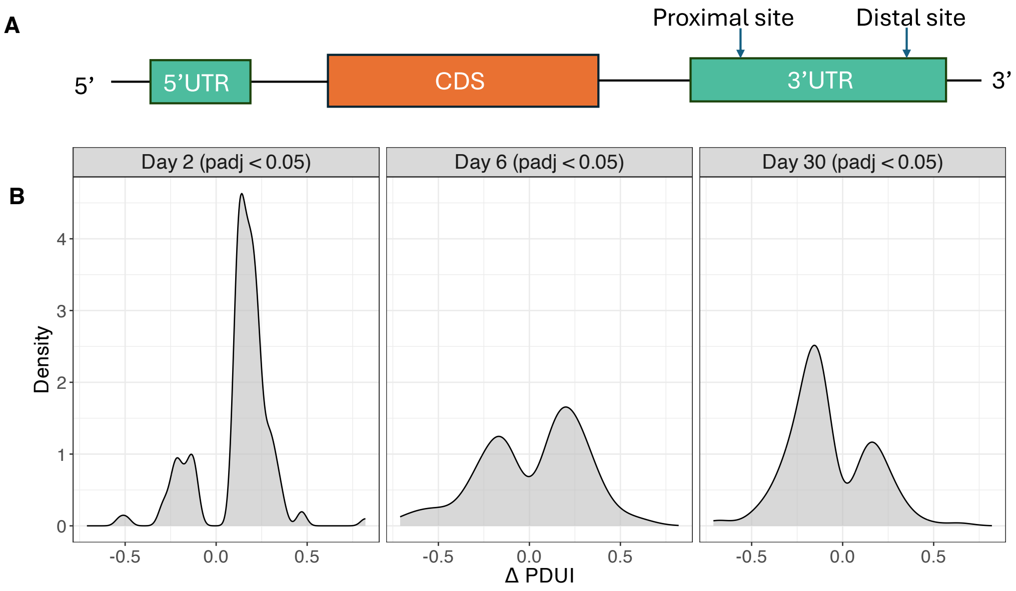


Figure 4.8: **Analysis of alternative polyadenylation with DaPars2.** (A) Schematic representation of proximal (5'-most) and distal (3'-most) poly(A) site usage. (B) Distribution of ΔPDUI values for genes with differential distal site usage (padj < 0.05).

4.3.4 Alternative polyadenylation

To directly investigate changes in 3′UTR regulation, we carried out APA analysis using both DaPars2-LR and LAPA across wildtype and *SNORD116*$^{-/-}$ samples. LAPA reports usage of all detected poly(A) sites, including intronic and coding region sites, while DaPars2-LR, originally developed for short-read data, focuses exclusively on two poly(A) sites per gene within the 3′UTR, inferring APA dynamics based on local changes in read density rather than full-length transcript models. Despite these limitations, DaPars2 provides a valuable framework for quantifying proximal (the 5′-most site) versus distal poly(A) site (the 3′-most site) usage (Figure 4.8A), enabling direct inference of 3′UTR shortening or lengthening. Using a ΔPDUI threshold of 0.1, we identified 433 genes exhibiting significant APA changes on at least one day. DaPars2 is reported to be a high-accuracy but lower-sensitivity method (Shah et al., 2021), which likely accounts for the smaller number of detected events relative to LAPA. APA dynamics varied across differentiation stages (Figure 4.8B). While APA patterns on days 6 and 30 showed a roughly equal distribution of lengthening and shortening events, Day 2 displayed a strong bias toward 3′UTR lengthening (125 genes with positive ΔPDUI vs. 29 with negative ΔPDUI). Consistent with transcriptomic trends observed throughout the study, the majority of APA changes were detected at Day 6 (Figure S2A). However, this result is expected given that isoform changes can be driven by changes in 3′UTR. Consequently, the analyses are not fully independent, and 34 genes were identified by both the APA and isoform usage analyses. Such integrative insights, which link full-length transcript models to alternative 3′ end processing, would not have been possible with solely short-read density-based analysis.

To explore if these 433 SNORD116-regulated genes show a specific pattern of dynamic 3′UTR changes during normal differentiation, we clustered wildtype PDUI values for these genes using k-means clustering (k=4). Among the resulting clusters, only cluster 3 demonstrated significant Gene Ontology enrichment, specifically for the biological process "*RNA decapping*", which comprised five genes: *CNOT7, EDC3, NUDT3, NUDT16* and *NUDT12* (Figure S2B). The strongest APA signal for the cluster was observed at day 2, where *CNOT7, EDC3, NUDT3,* and *NUDT12* showed significantly elevated PDUI values in knockout cells. Since increased PDUI reflects preferential use of distal poly(A) sites, leading to longer 3'UTRs that are generally associated with reduced mRNA stability, these shifts would be expected to correspond to lower protein output. Protein measurements, however, only partly reflected this trend. NUDT12 was significantly upregulated at day 2 ($q = 0.035$) but was not detected at later stages, whereas NUDT16 showed lower abundance in knockout cells at day 30 ($q = 0.07$). Both *NUDT16* and *NUDT12* encode NAD+ 5′cap decapping enzymes: NUDT12 targets mRNAs involved in ribosomal RNA processing and mitochondrial metabolism, while Nudt16 has been shown to decap U8 snoRNA (SNORD118) in *Xenopus laevis* (Ghosh et al., 2004; Sharma et al., 2020; Grudzien-Nogalska et al., 2019).

Although LAPA provided more transcript-aware profiles of APA, interpretation was complicated by variability in the exact poly(A) sites assigned within the same genes between samples, making it difficult to directly compare site usage across conditions. Nevertheless, consistent trends emerged, supporting the conclusion that loss of *SNORD116* perturbs 3′UTR regulation in a dynamic and stage-specific manner. Volcano plots illustrating differential APA detected with lapa at days 2 and 6 are provided in the supplementary material (Figure S3). Functional enrichment of these genes revealed significant over-representation of categories linked to “protein serine kinase activity” and “intellectual disability” (Figure S4). While the latter may appear most relevant to the neuronal phenotypes of PWS, many of the associated genes are also shared across excitable tissues, including the heart.

4.3.5 SNORD116 predicted targets

To investigate whether SNORD116 could mediate these post-transcriptional effects through direct RNA interactions, we turned to snoGlobe (Deschamps-Francoeur et al. (2022)). Although primarily trained on snoRNA–rRNA interactions, the model's positive test set included experimentally detected interactions with mRNAs identified through high-throughput RNA crosslinking methods. As mentioned previously, the authors found predicted sites were enriched in exons and exon boundaries rather than flanking regions. Given this and our focus on post-transcriptional regulation, we restricted our analysis to exonic regions of expressed transcripts. SNORD115 and SNORD96A C/D box snoRNAs were selected as controls. All 30 chromosome 15 cluster paralogs of SNORD116 and 48 paralogue sequences of SNORD115 were used for the analysis. Non-cluster SNORD116 paralogs on other chromosomes (chr1: ENSG00000202498, chr9: ENSG00000252985, chr13:ENSG00000212553) were initially used for enrichment testing together with other SNORD116 sequences, but then split away as extra controls.

To assess whether SNORD116 binding sites are enriched in dysregulated genes, we performed a hypergeometric test. This revealed a significant overrepresentation of SNORD116-predicted target sites among both DE and APA genes versus background genes. Although SNORD116 had the highest total number of predicted binding sites across all expressed genes (D116: `6,119`; D115: `1,394`; D96A: `972`), we also observed enrichment for control snoRNAs within the same set of dysregulated genes. Specifically, SNORD115 sites were significantly enriched in DE genes (Table 4.2), while all snoRNA subgroups, except non-cluster SNORD116s, showed enrichment in DaPars2 detected APA genes.

snoRNA Group	**APA**	**Non-APA**	**Enrichment Ratio**	**p-value**	**FDR**
chr15 SNORD116	166	5458	1.268	0.00016	0.00032
Non-cluster SNORD116	60	2062	1.215	0.06364	0.06364

SNORD115 and SNORD116 exhibited the most significant *p*-values, suggesting a stronger associ-

Table 4.2: Enrichment of differentially expressed genes among targets of specific snoRNA classes.

snoRNA	**N of DEGs**	**N of Background**	**p-value**	**FDR**
SNORD96A	118	1029	0.104673	0.139565
SNORD115	186	1527	0.008151	0.02778
Non-cluster SNORD116	201	1921	0.334755	0.334755
chr15 SNORD116	558	5066	0.013891	0.02778

ation with transcript dysregulation. However, this may instead reflect the higher number of paralog sequences and, therefore, a high number of potential target sites included in the testing, rather than a unique functional role. Similar to miRNAs, which have an even shorter target recognition seed region of 6-8 nt, snoRNAs have many potential binding sites. In our analysis, approximately 36.5% of all expressed genes in the dataset had predicted binding sites for at least one of the tested SNORDs. To determine whether the observed enrichment of control snoRNAs could be explained by sequence similarity and overlapping binding sites with SNORD116, we used BEDTools to compare predicted site coordinates. This analysis revealed minimal overlap, with only 9 shared sites with SNORD115 and 10 with SNORD96A.

We then examined the distribution of predicted binding sites across the transcripts of dysregulated genes. Although previous studies, such as Baldini et al. (2022), classified SNORD116 paralogs into four distinct groups based on evolutionary and interaction patterns, we subdivided SNORD116 into three traditionally defined groups that are based on sequence similarity: Group I (SNORD116-1 to SNORD116-9), Group II (SNORD116-10 to SNORD116-24), and Group III (SNORD116-25 to SNORD116-30) (Good and Kocher, 2017). We adopted this grouping because it is the convention most widely used in the field and was also chosen by our collaborators for preliminary analyses and experimental work, ensuring comparability with previous research. SNORD96A and SNORD115 sites were enriched within 3′ UTRs, which might explain their over-representation in the APA gene set. SNORD116 subgroup III exhibited the most notable positional shift: in DE genes the density of predicted binding sites peaked immediately upstream of the CDS, whereas in APA genes, that peak disappears and most sites are found in the 3'UTR (Figure 4.9).

4.3.6 Post-transcriptional and translational regulation

Matched proteomic and transcriptomic data were generated from the same cell populations, ensuring consistent biological replicates across assays. Proteomic profiling was performed by the Peffers laboratory on samples provided by Terri Holmes. To quantify the global relationship between RNA and protein abundance, we calculated Spearman's rank correlation coefficients (ρ) for wildtype samples at each stage. Correlations were modest (Day 2, $\rho = 0.329$; Day 6, $\rho = 0.308$; Day 30, $\rho = 0.321$), in line with previous reports of weak but reproducible Spearman correlations of 0.35–0.38 across days 6 to 15 of LUHMES neuronal cell differentiation Arad and Geiger (2023); Helwak et al. (2024).

To identify genes where 3'UTR regulation might influence protein output, we compared protein-to-RNA (P/R) ratios between wild-type and SNORD116$^{-/-}$ samples. Using a threshold of $|\log_2(\Delta\text{P/R ratio})| \geq 1.2$, we detected 231, 220, and 180 genes with altered P/R ratios at days 2, 6, and 30, respectively (467–488 transcripts per stage). While most of these changes could arise from multiple alternative post-transcriptional processes, we focused on the subset of 31 genes that also exhibited 3'UTR-associated regulation in the DaPars2 analysis (Table 4.3). This overlap highlights candidates in which SNORD116-dependent changes in 3 ' UTR usage are plausibly linked to altered

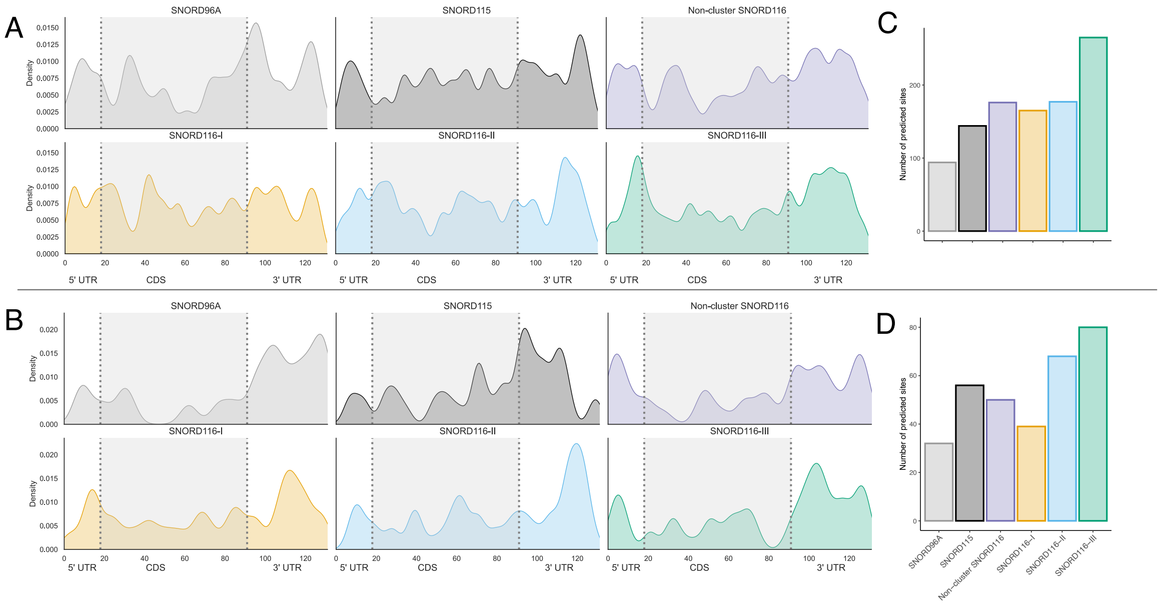


Figure 4.9: **snoRNA target site predictions with snoGlobe.** Distribution of snoGlobe predicted snoRNA target sites across transcripts of dysregulated genes. (A, B) Normalised density plots of snoGlobe-predicted target site distributions across transcript regions (5' UTR, CDS, and 3' UTR) for each snoRNA subgroup. Top panel: Differentially expressed genes (DEGs); Bottom panel: genes with alternative polyadenylation (APA) changes

protein abundance.

To assess whether the 31 candidate genes converge on shared functions, we submitted them to the STRING database (v12.0, combined score ≥ 0.4). This analysis identified three experimentally determined interaction modules: PSME3IP1–CDC5L-WRN, TEX10–IPO5, and USP10–CPSF6 (Figure S5), linking components of multiprotein complexes involved in RNA processing and protein degradation. Among these, the PSME3IP1–CDC5L–WRN interactions had the highest confidence scores (0.92 for PSME3IP1–CDC5L and 0.69 for PSME3IP1–WRN). CPSF6, a core component of the CFIm complex that regulates poly(A) site selection, exhibited altered P/R ratios driven by changes in transcript abundance, while peptide levels remained stable. By contrast, PSME3IP1 showed unchanged RNA levels but almost complete loss of detectable protein in SNORD116$^{-/-}$ samples. CDC5L, in turn, showed more concordant RNA and protein changes, with both upregulated in knockout cells at day 2 and downregulated at days 6 and 30 (all $q < 0.05$).

Correlation analysis showed that absolute RNA and protein abundances were only modestly correlated. However, the protein-to-RNA (P/R) ratios were stable across WT and KO samples (Figure 4.10C–D), consistent with previous studies (Helwak et al., 2024). We therefore fitted a linear model to the median $\log_2$-transformed P/R ratios, calculated residuals per transcript, and extracted the top outliers with the largest deviations from the expected P/R$^{\text{KO}}$ values given P/R$^{\text{WT}}$ levels. Four genes emerged as prominent outliers in both log P/R ratio change and residual analyses (Table 4.3), including *PSME3IP1* and *WRN*.

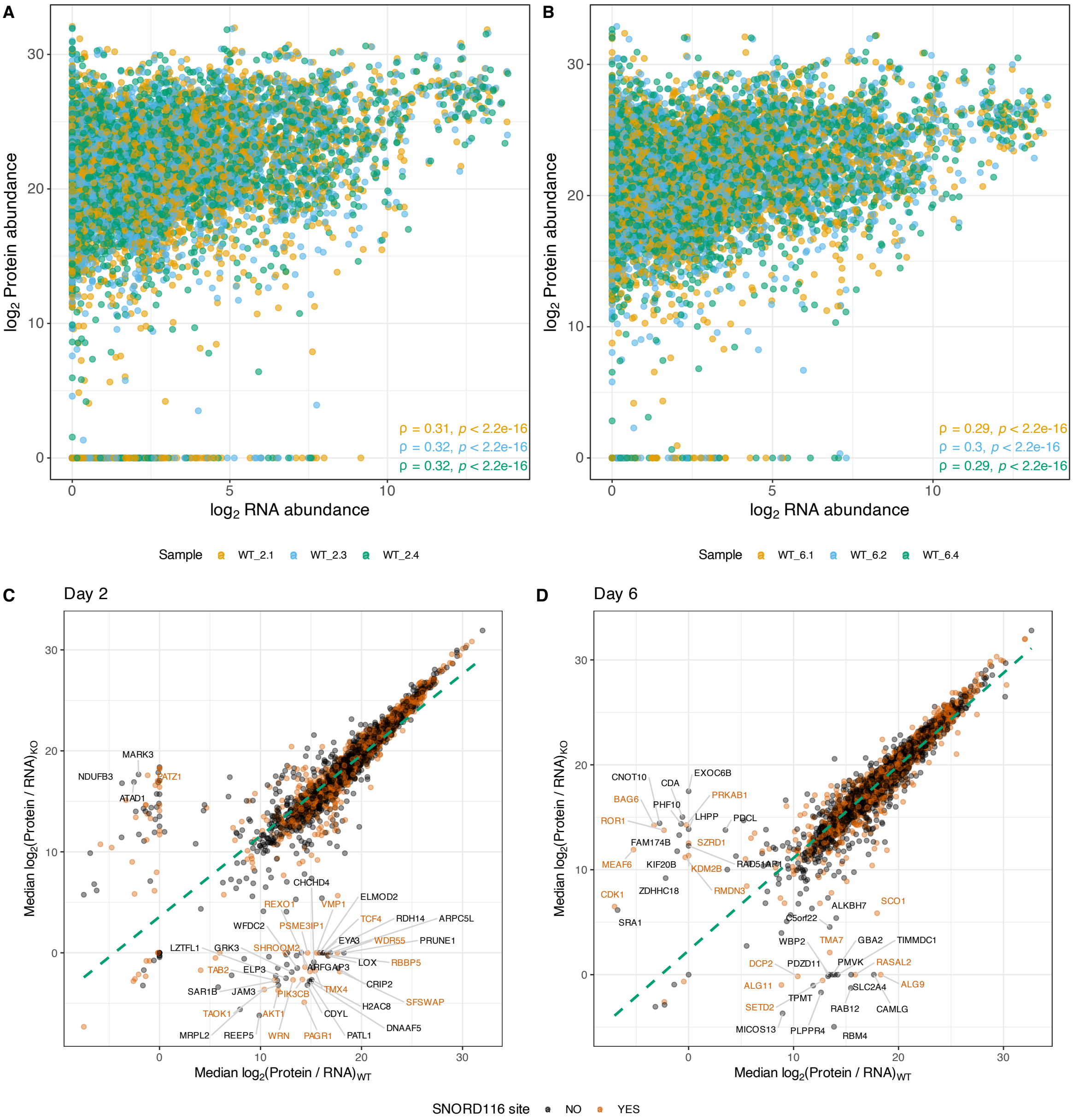


Figure 4.10: **Correlation of transcript and protein abundance in WT and *SNORD116*$^{-/-}$ samples.** (A–B) Correlation between log$_2$-transformed transcript and protein abundances in wildtype (WT) samples at Day 2 and Day 6. Spearman correlation coefficients (ρ) are shown for each biological replicate. (C–D) Correlation of protein-to-RNA (P/R) ratios between WT and *SNORD116*$^{-/-}$ samples at day 2 and day 6. Transcripts with snoGlobe-predicted SNORD116 binding sites are highlighted in orange.

Table 4.3: **Genes with differential 3'UTR length and protein-to-RNA ratio changes across time points.** Genes identified with residual analysis are in bold.

Day	Gene ID	Gene Symbol	ΔPDUI	3'UTR	SNORD116 site	$\log_2$(P/R) change
2	ENSG00000096401.8	CDC5L	0.13	longer		-2.24
2	ENSG00000134014.18	**ELP3**	0.25	longer		-9.19
2	ENSG00000135930.15	EIF4E2	0.18	longer		-1.63
2	ENSG00000165209.19	STRBP	0.20	longer		-1.61
2	ENSG00000165392.11	**WRN**	0.31	longer	✓	-1.76
2	ENSG00000172775.18	**PSME3IP1**	0.10	longer	✓	-24.04
2	ENSG00000179750.16	APOBEC3B	0.12	longer		-2.57
2	ENSG00000184787.19	UBE2G2	0.16	longer	✓	-6.62
2	ENSG00000198791.12	CNOT7	0.26	longer	✓	1.78
6	ENSG00000007923.17	DNAJC11	-0.18	shorter	✓	-1.60
6	ENSG00000095739.11	BAMBI	-0.11	shorter	✓	-1.31
6	ENSG00000103194.16	USP10	0.36	longer	✓	-1.63
6	ENSG00000108946.17	PRKAR1A	0.11	longer		1.73
6	ENSG00000123983.15	ACSL3	0.20	longer	✓	2.36
6	ENSG00000133028.12	**SCO1**	-0.14	shorter	✓	-7.49
6	ENSG00000136891.14	TEX10	0.39	longer		-2.22
6	ENSG00000137500.10	CCDC90B	-0.11	shorter		-1.69
6	ENSG00000145715.15	RASA1	0.21	longer		1.65
30	ENSG00000065150.21	IPO5	0.10	longer	✓	1.63
30	ENSG00000111605.18	CPSF6	0.12	longer	✓	1.73
30	ENSG00000116750.14	UCHL5	0.29	longer	✓	2.22
30	ENSG00000153113.24	CAST	0.21	longer	✓	1.53
30	ENSG00000160049.12	DFFA	0.13	longer	✓	4.41
30	ENSG00000160688.19	FLAD1	0.21	longer	✓	-1.66
30	ENSG00000170275.15	CRTAP	-0.13	shorter	✓	1.62

4.3.7 PolyA tail changes

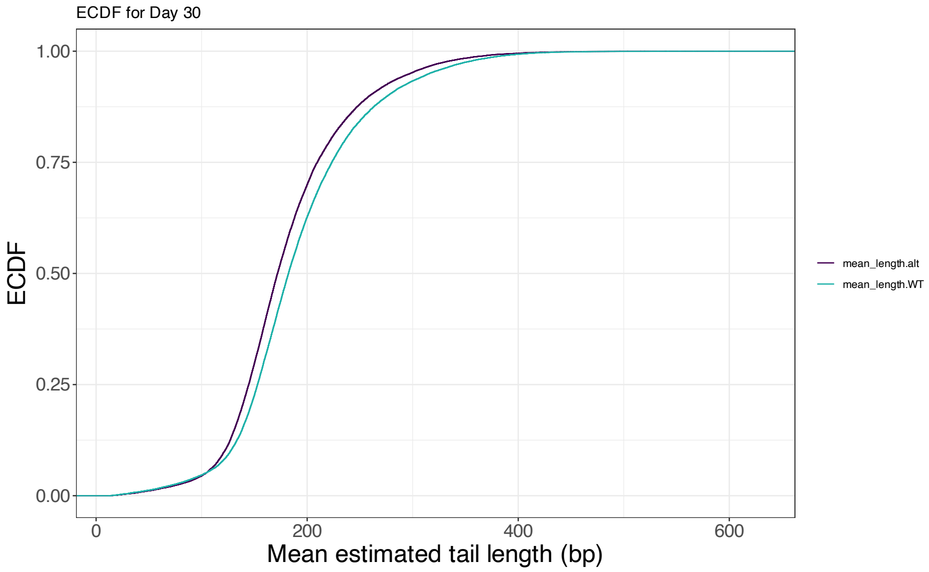


Figure 4.11: Empirical Cumulative Distribution Function (ECDF) plot of tail lengths at Day 30.

cDNA synthesis depends on the binding of the oligo(dT) primer to the poly(A) tail at the 3' end of the mRNA. As a result, while the length of the poly(A) tail may be affected, the cDNA molecules should retain the sequence corresponding to the original poly(A) tail. To assess transcripts with significant changes in tail length, we used Nanoplen (Dar et al., 2024). The Wilcoxon non-parametric test was chosen due to the non-normal distribution of estimated tail lengths, which span a broad range (10-600 bp). The most significant differences were observed at Day 30, with 910 transcripts showing altered tail lengths, of which only 48 had increased tail lengths in *SNORD116*$^{-/-}$ samples. Figure 4.11 illustrates the overall trend for poly(A) tail shortening at this stage. Notably, 65 genes showed overlap between Day 6 and Day 30, with 36 of these being ribosomal protein genes.

4.4 Discussion

Although best known for its association with PWS and its regulatory roles in neuronal differentiation, SNORD116 function in non-neuronal tissues has remained largely unexplored. Here, we provide the first comprehensive long-read transcriptomic characterisation of SNORD116 deletion during human cardiomyocyte differentiation from iPSCs, revealing widespread impacts on gene expression, isoform regulation, polyadenylation, and post-transcriptional control.

4.4.1 Stage-specific effects of SNORD116 loss

Using long-read sequencing across three stages of cardiomyocyte differentiation, we observed temporally distinct outcomes of *SNORD116* deletion. At early stages (Day 2), *SNORD116*$^{-/-}$ cells exhibited increased usage of distal poly(A) sites, whereas by Day 30 this effect was reversed. These findings suggest that SNORD116 may influence APA in a stage-dependent manner, possibly through interaction with the 3' end processing machinery. A similar regulatory function has been reported for

SNORD50A, which can regulate mRNA 3′ end formation by directly interacting with polyadenylation complex components (Shi et al., 2018).

This observation that SNORD116 knockout produces temporally distinct outcomes aligns with recent findings by Pilcher et al. (Pilcher, 2024), who reported that myocardial slices from *Snord116* paternal knockout (*Snord116*$^{+/-\mathrm{P}}$) mice were less sensitive to ischaemia-induced contractile dysfunction, suggesting a cardioprotective effect. This appears at odds with the increased cardiovascular risk observed in PWS patients. One possible explanation lies in the temporal and context-specific nature of SNORD116 function: while its loss during development may disrupt regulatory programmes essential for cardiac maturation, acute deletion in mature tissues may confer stress resilience.

4.4.2 Isoform-level regulation

Transcriptomic profiling revealed isoform switches across differentiation stages. A significant fraction of the differentially expressed (19.2 - 31.0%) and differentially used isoforms were classified as novel (NIC/NNIC), with many supported by independent evidence from refTSS or PolyASite databases. These results reinforce the importance of long-read sequencing for annotating poorly characterised developmental systems, where reliance on reference annotations alone would obscure a substantial proportion of the regulatory landscape.

Several isoforms gained premature termination codons, however, some may evade degradation through cell-type-specific NMD evasion mechanisms. Because many protein isoforms cannot be distinguished by proteomics due to shared peptides, the biological consequences of these isoform changes remain challenging to assess.

4.4.3 RNA–protein integration highlights post-transcriptional regulation

As in previous studies, RNA and protein abundances were only modestly correlated (Cheng et al., 2022; Taggart et al., 2020). However, P/R ratios were stable across wildtype and *SNORD116*$^{-/-}$ samples, providing a reliable baseline for identifying dysregulated genes. Linear modelling of P/R ratios and residual analysis revealed transcripts with disproportionate protein changes relative to RNA levels. Genes with detected *SNORD116*-regulated 3'UTR and P/R changes included genes encoding components of multiprotein complexes involved in RNA processing and protein degradation, like CPSF6 and PSME3IP1. CPSF6 showed altered transcript but stable peptide levels, whereas PSME3IP1 was undetectable at the protein level despite stable RNA expression. The absence of PSME3IP1 protein in *SNORD116*$^{-/-}$ samples is consistent with prior reports that proteasome-related genes often show weak RNA–protein correlations due to strong post-transcriptional and post-translational control (Cheng et al., 2022; Taggart et al., 2020).

Furthermore, STRING predicted an interaction between PSME3IP1 and CDC5L, a component of a non-snRNA spliceosome. Recent cryo-EM structures of the human pre-C*-I spliceosome complex identified PSME3IP1 (FAM192A) as an unexpected pre-mRNA splicing cofactor, with its α-helices contributing to spliceosome maturation (Zhan et al., 2022). This raises the possibility that SNORD116 indirectly influences both proteostasis and RNA splicing through effects on PSME3IP1. Finally, residual analysis of protein-to-RNA ratios provided a complementary lens to identify transcripts with disproportionate protein changes. The identification of PSME3IP1 among the top outliers reinforces its role as a key candidate for post-transcriptional dysregulation in the context of SNORD116 deletion.

Furthermore, STRING predicted an interaction between PSME3IP1 and CDC5L. Importantly, recent cryo-EM structures of the human pre-C*-I spliceosome complex identified PSME3IP1 (also known as FAM192A) as a step II splicing factor, interacting directly with CDC5L as part of the NTC complex (Zhan et al., 2022). In this context, the α-helices of PSME3IP1 contribute to spliceosome maturation, demonstrating that it participates not only in proteasome-associated protein degradation but also in pre-mRNA splicing. STRING also linked this pair to WRN, a DNA helicase/exonuclease required for genome stability that has been shown to interact functionally with the CDC5L/Prp19 splicing complex during DNA repair (Zhang et al., 2005), suggesting that SNORD116 deletion may influence both proteostasis and RNA processing pathways.

4.4.4 Links to metabolic regulation

The overlap in differentially expressed genes and transcripts was greater between days 2 and 6 than between days 6 and 30, even though the latter stages are phenotypically more similar. This suggests that SNORD116 loss exerts direct regulatory effects during early mesodermal progenitor and early cardiomyocyte stages, whereas disruptions observed by Day 30 are more likely to reflect secondary consequences of altered proliferation and metabolism. For this reason, the earlier stages in knockout models provide a more informative window to identify direct targets of SNORD116.

Among the splicing factors affected by SNORD116 loss, LUC7L2 showed consistent isoform switching between SNORD116$^{+/+}$ and SNORD116$^{-/-}$ cells at both days 2 and 6. LUC7L2 is a core component of the U1 snRNP complex and plays a key role in splice site recognition. Jourdain et al. (Jourdain et al., 2021) showed that loss of LUC7L2 in human cell lines induces widespread alternative splicing changes in glycolytic regulators, leading to reduced glycolysis and a compensatory shift toward oxidative phosphorylation (OXPHOS). Complementary work by Daniels et al. (Daniels et al., 2021) further established that LUC7L2 loss is directly linked to metabolic rewiring. In parallel, SNORD116 deletion has been associated with decreased carbohydrate metabolism and increased fatty acid oxidation (FAO) (Holmes et al., 2025). Given the tight coupling of FAO and OXPHOS pathways (Wang et al., 2010), our observation of LUC7L2 isoform switching raises the possibility that SNORD116 may indirectly modulate cardiac metabolic programming through effects on splicing of regulators such as LUC7L2.

At Day 2, upregulated genes were significantly enriched for pathways related to cadmium ion response and transition metal ion homeostasis (adjusted $p < 0.05$). These findings closely mirror those of Gilmore et al. (Gilmore et al., 2023), who reported 207 consistently dysregulated genes across three independent PWS transcriptome datasets from post-mortem brain tissue and iPSC-derived neurons (Bochukova et al., 2018; Huang et al., 2021). The recurrence of these metal ion response signatures in both neuronal and non-neuronal systems suggests that disruption of ion homeostasis represents a conserved molecular consequence of SNORD116 deficiency.

At Day 30, upregulated DE genes were enriched for terms such as “negative regulation of peptidase activity”, consistent with Seahorse assays and proliferation studies that revealed metabolic alterations in SNORD116-deficient cardiomyocytes (performed by Dr Terri Holmes). These results reinforce metabolic regulation as a major axis of SNORD116 function in the heart.

4.4.5 Tissue-specificity and comparison to neuronal models

Previous studies based on neuronal models have proposed multiple putative SNORD116 targets, including *Nhlh2* and *Magel2*. For instance, *Snord116* overexpression in an embryonic hypothalamic cell line led to a 15-fold increase in *Nhlh2* expression, likely mediated through a predicted 20 bp RNA–RNA interaction that enhanced mRNA stability (Kocher et al., 2021). In vivo, *Snord116* deficiency in mice has been linked to obesity and hypogonadism—hallmark phenotypes of PWS (Cogliati et al., 2007). A recent study using LUHMES neuronal cells showed that deletion of the SNORD116 or SNORD115 clusters reduced *MAGEL2* mRNA levels (Helwak et al., 2024). *MAGEL2* lies $\sim$ 1.5Mb upstream of SNORD116 and is typically included in larger 15q11–q13 deletions in PWS. However, both *NHLH2* and *MAGEL2* are predominantly expressed in neuronal tissues and are nearly absent in the heart. According to GTEx, *NHLH2* expression is 0 TPM in the human left ventricle, and *MAGEL2* is 0.03 TPM (`https://www.gtexportal.org`). In line with this, we did not detect differential expression or splicing of *NHLH2* in our cardiomyocyte model. Importantly, even if *MAGEL2* was expressed in our model, it is a single-exon gene, and in our pipeline, single-exon transcripts were filtered out during the transcriptome pre-processing step to reduce artefactual or low-confidence isoforms. While this filtering is commonly applied in long-read datasets to mitigate noise, it may lead to the exclusion of biologically relevant single-exon genes such as *MAGEL2*. Additionally, compared to short-read RNA-seq, long-read datasets remain limited in sensitivity. Other candidate targets, mentioned in the introduction, include the lncRNAs *Malat1* and *Meg3*. Notably, *MALAT1* was identified as differentially expressed in our dataset.

4.4.6 Poly(A) tail dynamics

The enrichment of ribosomal protein genes among transcripts with altered poly(A) tail lengths may partly reflect their very high expression levels. Because the Wilcoxon rank-sum test is sensitive to large sample sizes, even subtle shifts in tail length can reach strong statistical significance for these genes, potentially inflating the effect size. However, this signal may also represent a genuine biological effect as previous research hinted at disrupted protein synthesis (Helwak et al., 2024). Additionally, recent work by the Whipple group has shown that *Snord116* directly interacts with ribosomes in both mouse and human neurons, supporting a specific connection between *Snord116* and the protein synthesis machinery (Whipple and colleagues, 2025). Notably, they did not find evidence that *Snord116* guides 2'-O-methylation of rRNA, suggesting that its role in ribosome biology may occur through a non-canonical mechanism.

4.4.7 SNORD116 and SNORD115 co-regulation

While SNORD116 exhibited the highest number of predicted targets overall, SNORD115 also showed statistically significant enrichment, particularly among differentially expressed genes. Their binding sites showed minimal overlap, which suggests that enrichment for SNORD115 sites is unlikely to be an artefact of binding site redundancy and may instead reflect a coordinated regulatory function. This hypothesis is supported by evolutionary evidence. Guibert et al. (2024) showed that SNORD115 and SNORD116 copy numbers co-vary across mammalian species, suggesting shared selective constraints and co-regulation. Although distinct subfamilies have evolved in humans Baldini et al. (2022), likely enabling functional diversification, their tandem arrangement and shared transcription regulation of the SNHG14 locus imply potential cooperative activity. It has also been suggested that SNORD116 may influence either the production or stability of SNORD115 target mRNAs. Recent work demonstrated that biallelic mutations in *SREK1*, which encode a splicing factor, decrease expression of both SNORD115 and SNORD116, producing PWS-like phenotypes despite an intact PWS locus (Saeed et al., 2025). Together, these findings point to shared upstream regulation and potential cooperative function.

Some of the effects of SNORD116 knockout could be the result of affected splicing of the SNORD116 host gene SNHG14 (116HG). In mice, the spliced transcripts of *Snhg14* retained within the nucleus form an "RNA cloud" that regulates the expression of genes with epigenetic and metabolic functions Powell et al. (2013). Powell et al. used chromatin isolation by RNA purification followed by sequencing (ChIRP-seq), which was complemented by RNA/DNA FISH, which confirmed the physical co-localisation of the 116HG RNA cloud with target genes such as *Mtor, Crebbp* and *Igf2r*.

4.4.8 Conclusions and future directions

Taken together, our findings converge on a model in which SNORD116 influences pre-mRNA processing and proteostasis. However, these signals remain indirect: target predictions were computational, protein-level effects were modest, and RNA–protein correlations were often weak.

Further work is therefore required to establish causality and define the molecular mechanisms of SNORD116 action. Experimental mapping of snoRNA–RNA interactions using approaches such as snoKARR-seq (Liu, 2025) could validate direct binding partners. Proteomic analyses using ribosome profiling (Ribo-seq) would clarify whether observed 3'UTR changes alter translation efficiency. Finally, CRISPR-based perturbations of candidate targets, such as *PSME3IP1* or *LUC7L2*, could test whether their misregulation mediates the metabolic and transcriptional changes associated with SNORD116 loss.

In summary, this work provides the first evidence that SNORD116 regulates RNA processing in human cardiomyocytes, but definitive mechanistic insights will require direct interaction mapping and experimental validation in physiologically relevant models.

Chapter 5

General Discussion

5.1 Project Aims and Contributions

During my PhD project, I set out to develop and apply computational methods for high-confidence detection of novel, low-abundance, and tissue-specific transcript isoforms in human brain and heart models. My work has centred around leveraging long-read technology to characterise events that are difficult to resolve with short-read sequencing, and placed particular emphasis on splicing in long and structurally complex genes. Across three complementary chapters, I advanced long-read RNA-seq bioinformatic methodology and used it to address biological questions surrounding isoform regulation in human development and disease across different scales from tissue to single-cell resolution.

In **Chapter 2**, I investigated whether targeted long-read sequencing of post-mortem human brain could reveal novel isoforms and differential splicing at disease-relevant loci. These data were technically challenging due to degraded RNA and early ONT chemistries, providing an ideal test case for benchmarking artefact filtering and error-aware annotation strategies. To this end, I developed a bioinformatics workflow that balanced permissive isoform discovery with orthogonal support–based filtering. The resulting catalogue of more than 56,000 previously unannotated isoforms demonstrated how targeted long-read sequencing can resolve transcript structures invisible to short-read or transcriptome-wide approaches, while also emphasising the need for rigorous curation of these novel data. Alongside the methodological advances, this work identified extensive differential isoform expression across human brain regions and revealed developmental transcript switching between fetal and adult prefrontal cortex.

Chapter 3 built directly on this framework, extending the analysis to the single-cell level. While bulk CaptureSeq highlighted extensive isoform diversity, the averaging of signals across heterogeneous tissue limited biological interpretability. This chapter, therefore, aimed to (i) construct comprehensive transcriptome references for isogenic hiPSC-derived neurons, astrocytes, and microglia, and (ii) benchmark single-cell long-read platforms and analysis workflows. Benchmarking revealed that standard pipelines discarded large fractions of the data, particularly low-abundance isoforms, necessitating modifications that retained more informative reads. Applying this refined workflow uncovered isoform heterogeneity both between and within cell lineages, showing that even isogenic differentiated cells display distinct post-transcriptional programmes. By cross-referencing these results with those of Chapter 2, I demonstrated that many highly-expressed isoforms are reproducible across *"in vivo"* and *in vitro* models.

In **Chapter 4**, the focus shifted from annotation to biological mechanisms. Here, the aim was to investigate whether the small nucleolar RNA *SNORD116* regulates RNA processing during human cardiac differentiation, and if so, through which mechanisms and of which transcripts. Using long-read RNA-seq data from iPSC-derived cardiomyocytes at three stages of differentiation, I was able to leverage my bioinformatics pipeline from chapters 2 and 3 to quantify alternative splicing, polyadenylation, and poly(A) tail length within a single framework. By integrating these data with binding-site predictions and proteomics, I was able to prioritise candidate *SNORD116* targets and test putative regulatory functions in the cardiac lineage. A key finding was that much of the novelty lies at the 3' end of the molecule, where alternative cleavage and polyadenylation remodel coding potential and 3' UTRs—features that are under-reported by standard "novel intron chain" metrics.

Synthesised together, the results of my thesis demonstrate that long-read sequencing consistently uncovers extensive isoform diversity of biological and disease relevance, but that these data require particularly careful and biologically informed assessment to glean reliable insights. Contrary to the assumption that human transcript annotation is largely complete, systematic reannotation was necessary to capture the true extent of splicing diversity in both brain and heart models. More broadly, these findings argue for a functional genomics framework that moves beyond the gene as the unit of analysis towards transcript-level resolution. This work also highlights the value of increasing resolution to the single-cell level, while also pointing towards future opportunities to integrate isoform discovery with orthogonal approaches such as protein structure prediction (e.g., AlphaFold), ribosome profiling and perturbation experiments to assess functional consequences of isoform variation.

5.2 State of the art at project onset and project challenges

At the onset of this project (2021), long-read transcriptomics was still an emerging field. ONT and PacBio had only recently enabled full-length isoform sequencing, offering major advantages over short-read platforms. Prior to this, most splicing analysis relied on junction counts or computational transcript model inference, which constrained the detection of closely related isoforms and unannotated splicing events.

However, widespread adoption was hindered by several limitations, many of which I encountered in this project. Early ONT chemistries were characterised by high error rates, which complicated both alignment and downstream interpretation. Sequencing throughput was also relatively low, resulting in insufficient read support for novel isoforms. Over the course of this PhD, the field has advanced dramatically. Sequencing platforms have scaled in throughput, moving from the ONT MinION, which supports a single flow cell, to the PromethION, which can operate up to 48 flow cells in parallel, and from the PacBio Sequel II to the Revio platform. Accuracy also improved markedly with the introduction of ONT R10.4 flow cells and PacBio HiFi reads, and will undoubtedly continue to improve (Jain et al., 2017; Sereika et al., 2022). These developments are reflected in my datasets: in Chapter 2, data were generated using ONT R9 flow cells, of which only 8.9% exceeded Q15. By contrast, in Chapter 3, the adoption of R10 chemistry increased this proportion to 54.1%. This step-change in quality had direct consequences for alignment, splice junction detection, and downstream filtering strategies.

To match the rapid developments in sequencing technology, new software tools had to be able to handle larger amounts of data. Early tools such as TAMA, which I used in Chapter 2, were not optimised for high-coverage PromethION datasets and frequently failed due to high memory demands.

As throughput increased, it became necessary to adjust my pipeline, retaining only the more scalable IsoQuant. More generally, increasing dataset sizes are driving a shift in analytical paradigms in long-read transcriptomics. The approach used throughout this thesis, “merge-and-call”, in which reads are merged across samples into a unified transcriptome before isoforms are quantified per sample, maximises sensitivity but will become increasingly impractical at scale. The field is instead moving towards “call-and-merge”, in which isoforms are first called per sample and subsequently merged across datasets. Although these approaches may sacrifice some sensitivity, they are likely to improve reproducibility and reduce the risk of artefactual isoforms being carried forward. This highlights an example of the data itself influencing the balance between research aims and research practicality, and will continue to demand philosophical consideration in the future.

Despite improvements in sequencing chemistry and basecalling, library preparation artefacts remain a persistent source of noise in long-read datasets. Mispriming events, incomplete reverse transcription, and template switching can all generate false isoforms. The pipeline I developed in this thesis incorporated measures to detect and mitigate such artefacts, thereby improving the reliability of isoform-level analyses. Direct RNA sequencing, which avoids PCR artefacts and simultaneously captures base modifications, offers a route to further increase biological accuracy and reduce the number of artefact filtering steps, but comes with its own practical and analytical challenges, particularly at single-cell resolution.

Benchmarking efforts, including those I described in Chapter 3 and by others to which I contributed (Ahlert Scoones et al., 2025), underscore both the promise and the limitations of single-cell long-read sequencing. Transcript length biases and artefacts introduced during reverse transcription reduce the fraction of reads that are informative for isoform quantification. At the computational level, many reads are lost during filtering; indeed, in the comparison by Ahlert-Scoones et al., only 11% of raw ONT reads remained after processing. Statistical frameworks for isoform-level differential expression remain underdeveloped, and data sparsity means that pseudobulking continues to be the most commonly used strategy. This comes at the cost of cell-level resolution and somewhat defeats the initial reasons for employing single-cell resolution. This limitation is becoming increasingly evident as spatial transcriptomic studies reveal that cellular identities often exist along gradients, blurring the boundaries of discrete cell types or states. Integrating long-read sequencing with spatially resolved data will therefore be essential in refining the interpretation of isoform regulation at single-cell resolution.

Across all three datasets analysed in this thesis, iterative refinement of the isoform detection pipeline proved essential to account for shifting error profiles, changing coverage distributions, and increasing throughput. Looking ahead, it is clear that there will be no universal “one-size-fits-all” long-read pipeline. Instead, analytical choices must remain closely aligned with the biological question at hand. Projects focused on basic discovery may justify more permissive isoform detection strategies in order to maximise sensitivity, whereas translational studies aimed at evaluating disease associations or therapeutic targets will require stricter thresholds to prioritise reproducibility and minimise false positives.

5.3 Future directions for functional genomics

Several avenues remain open for future work. Most importantly, there is a need for greater standardisation of isoform annotation, with community-driven efforts required to feed discoveries back into repositories such as GENCODE and RefSeq. However, a longstanding challenge in transcriptomics is deciding when the extensive isoform diversity revealed by high-throughput sequencing represents meaningful biology rather than transcriptional or technical noise. In this thesis, I often used ORF, coding potential, and domain prediction as first-pass indicators of isoform function. While informative for protein-coding transcripts, such criteria cannot be applied to non-coding RNAs, which constitute the majority of the human genome and transcriptome (ENCODE Project Consortium, 2012). The ENCODE project sharpened the debate over how much of the genome is "functional" as it controversially claimed biochemical activity across 80% of the genome. Critics argued that ENCODE conflated biochemical activity with evolutionary function and suggested that much of the genome may still be "junk DNA" in the evolutionary sense, if it is not maintained by purifying selection (Doolittle, 2013; Ponting and Hardison, 2011). Since then, deep and long-read sequencing methods reported more and more non-coding RNAs and their dynamic expression across tissues and conditions (Troskie et al., 2021; Lagarde et al., 2017; Potolitsyna et al., 2022).

Non-coding RNA

Two broad categories matter here: (i) unproductive isoforms from protein-coding loci, which are targeted for degradation by NMD and (ii) isoforms of non-coding RNA genes, including snoRNAs and lncRNAs. While long-regarded as non-functional, unproductive isoforms of protein-coding genes can modulate steady-state gene expression by engaging NMD pathways, modulating the abundance of their protein-coding counterparts and maintaining transcriptional activity (He and Jacobson, 2015; Nasif et al., 2018; Fair et al., 2024). In Chapter 3, for example, stressed microglia expressed intron-retaining isoforms likely targeted for NMD in the cytoplasm. Their increased detection may reflect reduced NMD efficiency under stress, thereby providing a functional insight into cellular adaptation. Quantifying intron retention, however, remains technically difficult due to low-complexity intronic sequences and contamination from pre-mRNA, both of which reduce mapping specificity (Broseus et al., 2020).

Non-coding RNA genes highlight even more clearly that the absence of coding potential does not imply lack of function. SNORD116, a snoRNA cluster studied in Chapter 4, provides a clear example. Deletion or disruption of this locus causes Prader–Willi syndrome (Bieth et al., 2015; Tan et al., 2020), and in our model of cardiac differentiation, it perturbed RNA processing at multiple levels. I detected effects in splicing, polyadenylation and poly(A) tail length, with the direction of polyadenylation changes varying across developmental stages. These findings highlight that functionality cannot be defined in absolute terms but must instead be understood as context-dependent.

lncRNAs represent a more complex and heterogeneous category. Some, such as *XIST* and *H19*, play essential roles in X-chromosome inactivation and cell differentiation, respectively (Bartolomei et al., 1991; Brown et al., 1992). Yet the vast majority are lowly expressed, less evolutionarily conserved, or have no demonstrable phenotypic impact (Mattick et al., 2023). For instance, MALAT1 is highly expressed, ubiquitously transcribed, and localises to nuclear speckles, leading to the assumption that it was essential. However, it produces minimal phenotypic consequences when knocked out in mice (Eißmann et al., 2012; Nakagawa et al., 2012). Recent studies suggest that MALAT1 may have tissue- or disease-specific roles, such as in cancer, highlighting that high expression does not necessarily equate to broad functionality (Ponting and Haerty, 2022; Duan et al., 2023). The GENCODE 47 release annotates 19,433 protein-coding genes and 35,934 lncRNA genes, contributing to a total of 78,046 genes across all categories (including pseudogenes and small RNAs) (Mudge et al., 2025). However, most lncRNAs remain poorly annotated due to low expression and limited experimental validation. Even targeted enrichment approaches are limited by their reliance on existing annotations for probe design, meaning that poorly characterised loci are frequently overlooked (Mercer et al., 2014). As shown in Chapter 2, 372 lncRNAs included on the panel could not be detected, and those that were recovered typically exhibited only minimal enrichment. Many of the annotated lncRNAs would have come from studies on a single sample or tissue type. Nevertheless, long-read sequencing enabled us to identify novel isoforms of lncRNAs at several loci, some of which displayed evidence of condition-specific regulation.

This context-dependence underscores the importance of studying isoform function in specific tissues and developmental windows, as what appears to be transcriptional noise in one context may be a tightly regulated mechanism in another.

Long-read proteogenomics

From a methodological perspective, functional validation is still limited even for protein-coding isoforms. While individual cases of isoform-specific function are well documented, large-scale annotation strategies are still underdeveloped (Pozo et al., 2021b). In Chapter 4, I worked with collaborators to begin addressing this gap by combining long-read transcriptomes with matched mass spectrometry (MS) data from the same hiPSC-derived cardiomyocyte samples. This approach ensures that proteomic searches are guided by empirically observed transcript isoforms rather than incomplete or outdated reference databases and improves peptide mapping (Miller et al., 2022; Kedan et al., 2025; Abood et al., 2024; Paoli-Iseppi et al., 2024). Despite these advances, several challenges remain. Sequence similarity across isoforms limits the number of unique peptides and corresponding enzymatic cleavage sites, which constrains the capacity of MS to resolve isoform-specific protein products (Carlyle et al., 2018). Detection sensitivity also remains an issue, with proteomics often capturing only a single dominant isoform per locus (Ezkurdia et al., 2015; Hall et al., 2021a).

Future strategies will increasingly exploit multi-layered integration. One promising avenue is the use of protein structure prediction tools such as AlphaFold (Jumper et al., 2021) to evaluate whether alternative isoforms encode proteins with altered structural domains or interaction surfaces (Paoli-Iseppi et al., 2024). In ongoing collaborative work, AlphaFold is being applied to novel isoforms identified in Chapter 2 to predict whether differential transcript usage events alter protein topology in ways that might affect stability or function. When combined with targeted MS validation, this approach could offer a powerful framework for prioritising isoforms of functional interest.

Proteogenomic approaches informed by sample-specific long-read transcriptomes improve peptide assignment, and in the future, combining long-read sequencing with spatially resolved proteomics offers a promising avenue for linking isoform diversity to protein-level consequences (Davis et al., 2023).

5.4 Concluding remarks

In summary, this thesis demonstrates the power and promise of long-read sequencing in revealing widespread isoform diversity, much of it invisible to short-read methods. The bioinformatics workflows developed herein were deployed to expanded isoform catalogues in human brain (Chapter 2), dissect cell-type–specific isoform heterogeneity at single-cell resolution (Chapter 3), and reveal a role for *SNORD116* in alternative polyadenylation during cardiomyocyte differentiation (Chapter 4). Together, these findings establish long-read sequencing as an indispensable tool for transcriptome analysis and provide the community with refined bioinformatic workflows for isoform discovery and interpretation, which lay the foundation for future functional studies. This novel work joins an expanding body of evidence of the utility of long-read transcriptomics in answering some of our biggest biological questions, but also demonstrates the growing need for careful, considerate and biologically-informed treatment of these data in the future.

Appendix A

Additional Material for Chapter 2

Table S1: Performance comparison of isoform detection tools on long-read CaptureSeq data

Tool	**Threads (CPUs)**	**Real Time**	**CPU Time**	**Memory Usage (GB)**
FLAIR (v1.6.2)	20	4h 40min	76 h 23 min	163.7
IsoQuant	20	18h 6min	26 h 5 min	317.3
cDNA Cupcake	20	>7 days (failed)	–	>700
TAMA	22	111h 34min	–	max 26 per job
IsoTools	1	7h 6min	7 h 1 min	150

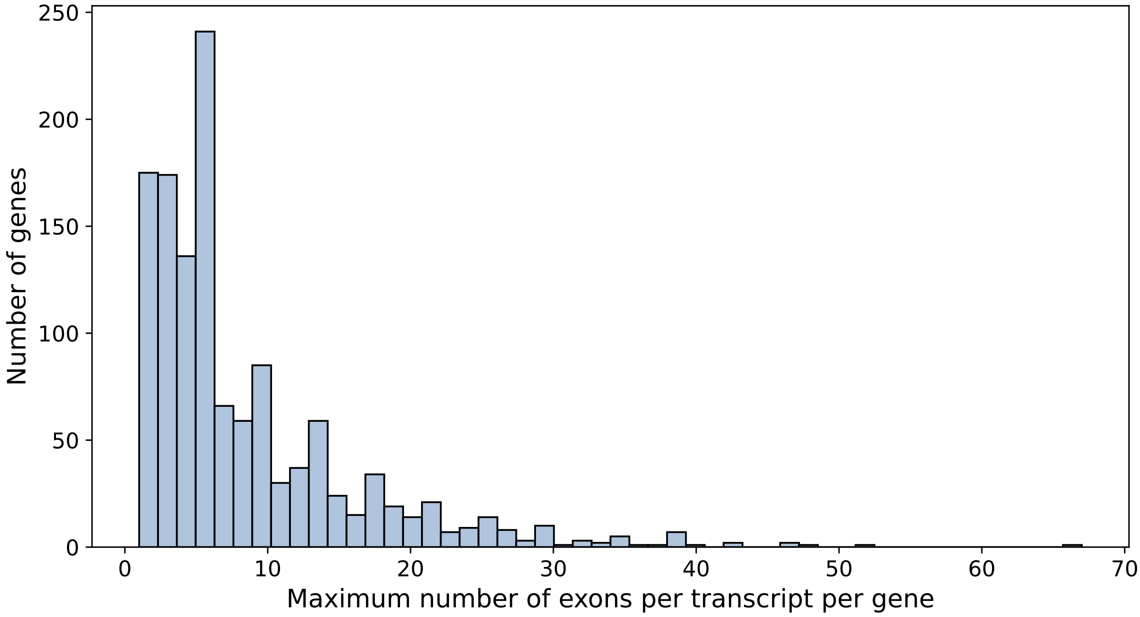
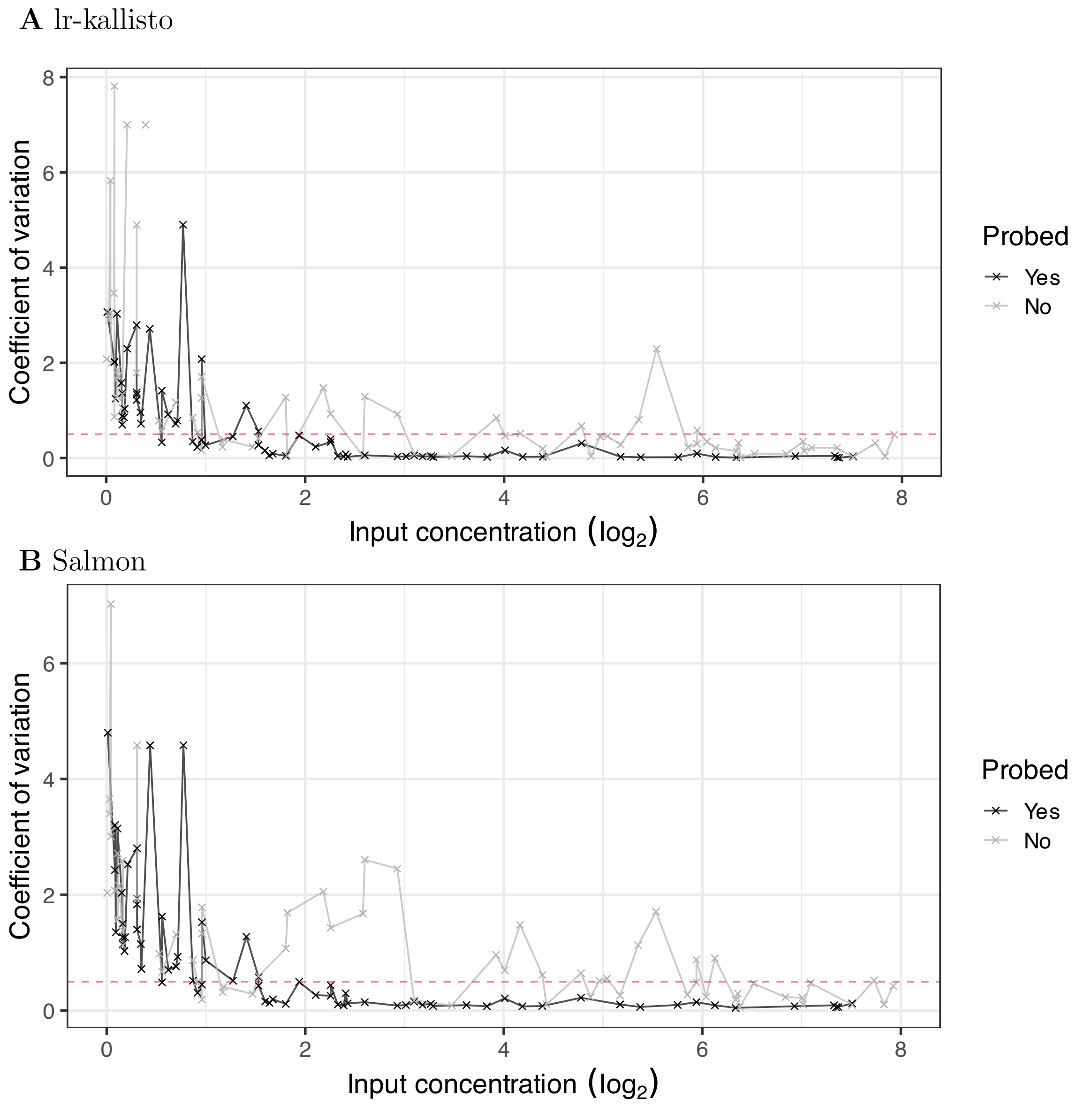


Figure S1: Maximum number of exons per transcript per gene in Capture panel



A lr-kallisto

B Salmon

Figure S2: **Relationship between input concentration and coefficient of variation (CV).** Mean CV of measured expression is shown for probed (black) and non-probed (grey) sequins at increasing input concentrations. The dashed red line indicates a CV of 0.5.

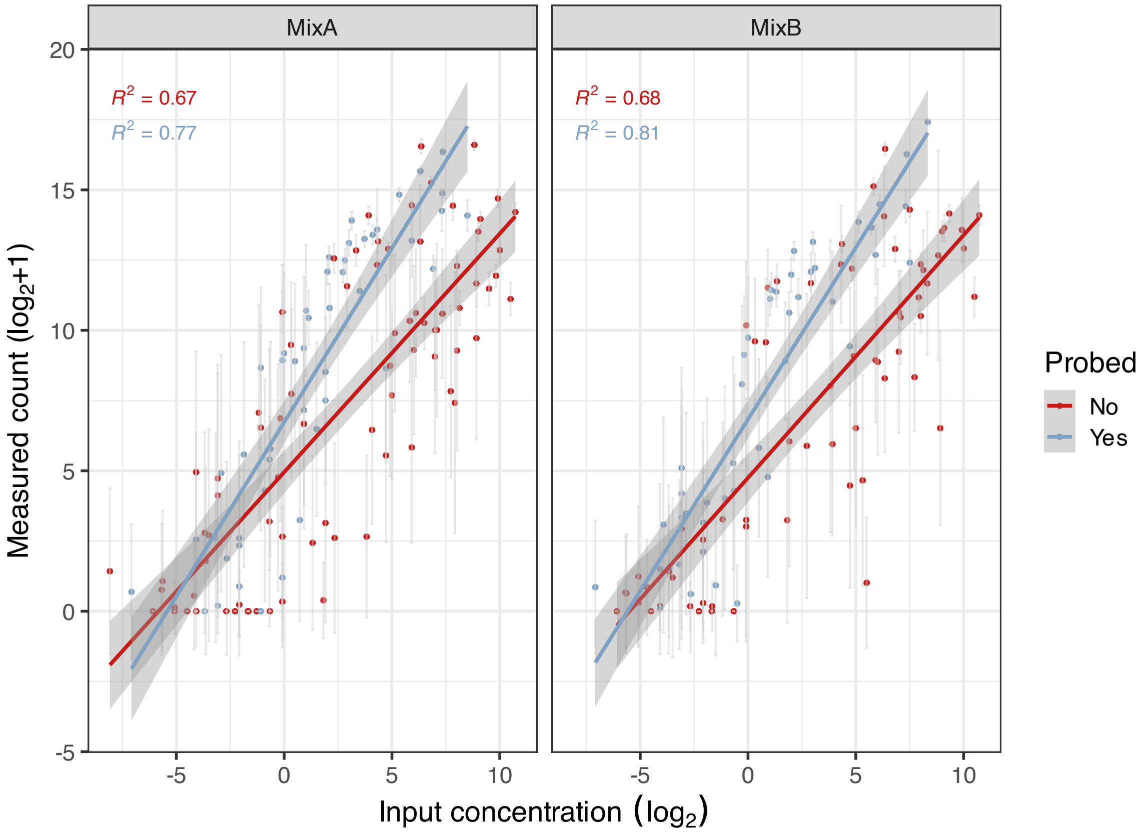
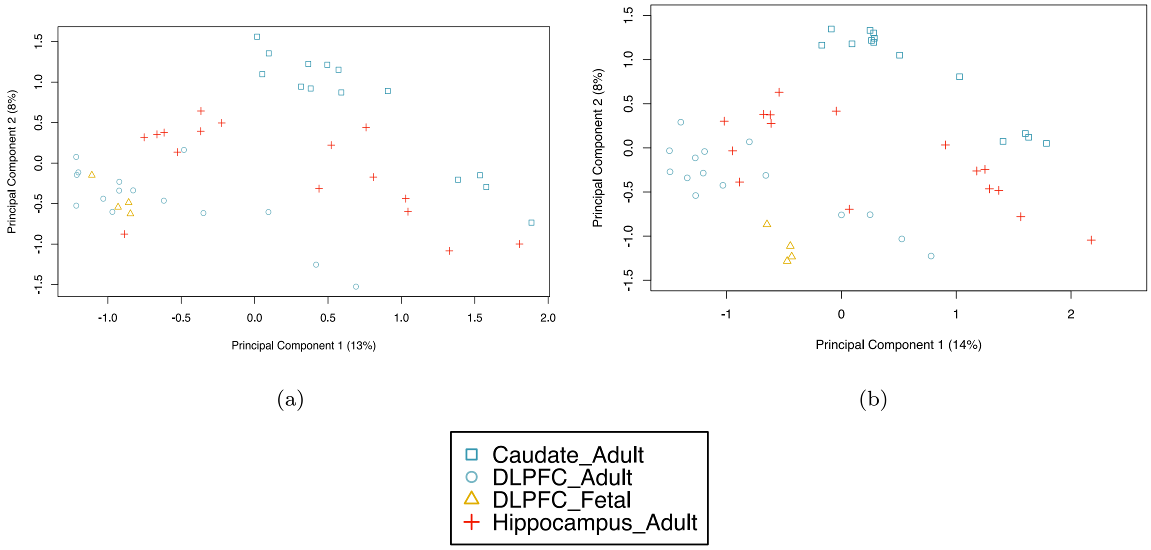


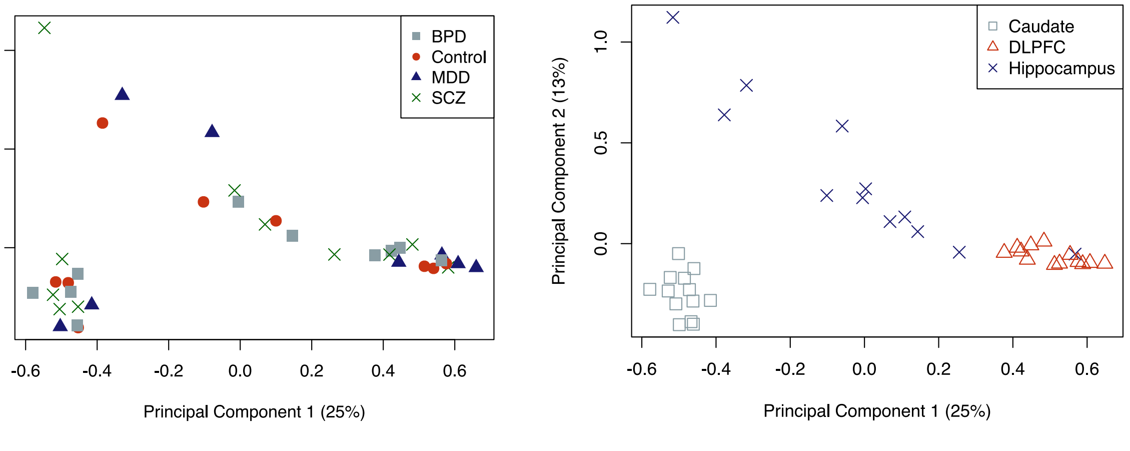
Figure S3: **Correlation of lr-kallisto transcript quantification with known spike-in concentrations.** Correlation between `lr-kallisto`-estimated transcript abundance (TPM) and spike-in concentration and input concentration for 160 synthetic spike-in sequences across MixA and MixB samples.



(a)

(b)

Figure S4: **Transcript-level principal component analysis** using (a) GENCODE annotation and (b) a custom annotation incorporating 39,447 novel isoforms.



(a)

(b)

Figure S5: **Transcript-level principal component analysis of adult brain samples.** Samples visualised according to (a) psychiatric diagnosis and (b) brain region.

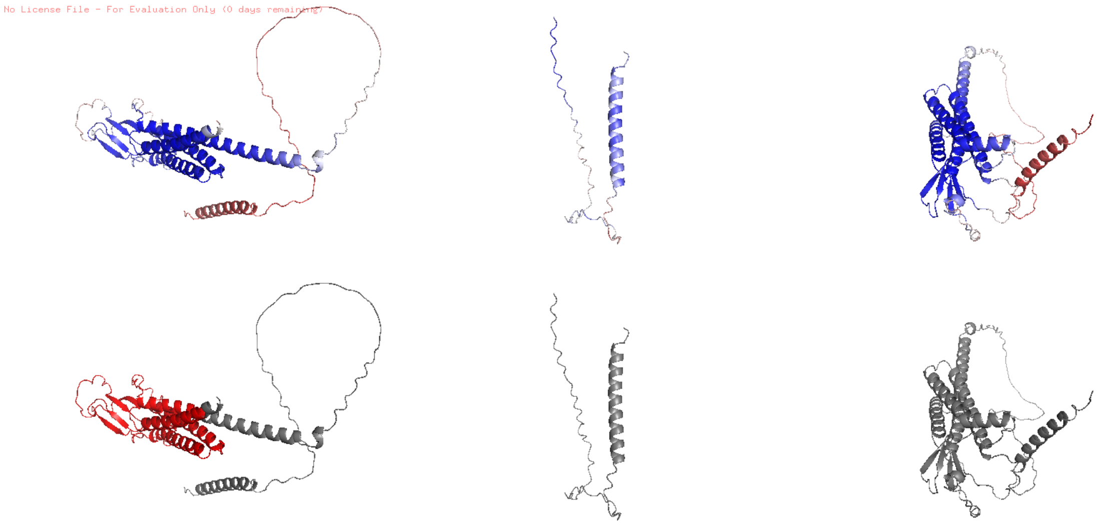


Figure S6: **AlphaFold-predicted structures of novel CACNG4 isoforms.** Structural models were generated by Dr James Wilsenach using AlphaFold (Jumper et al., 2021), based on isoform annotations derived from our CaptureSeq dataset. From left to right: G29432.125.nnc, G29432.127.nnc, and G29432.70_ENST00000262138.4. For each isoform, the top panel shows per-residue confidence scores (pLDDT; blue = high confidence, red = low confidence), and the bottom panel indicates predicted Pfam domains.

Appendix B

Additional Material for Chapter 3

All supplementary tables listed within this chapter are available at: `https://github.com/skudashev/scLR-isoform-analysis/tree/main/supplementary`

Supplementary Table:

S1 - Expression markers provided by collaborators used for cell annotation

S2 - Lineage variable transcript marker genes

S3 - S5 Isoform switches within neuron, astrocyte and microglia lineages

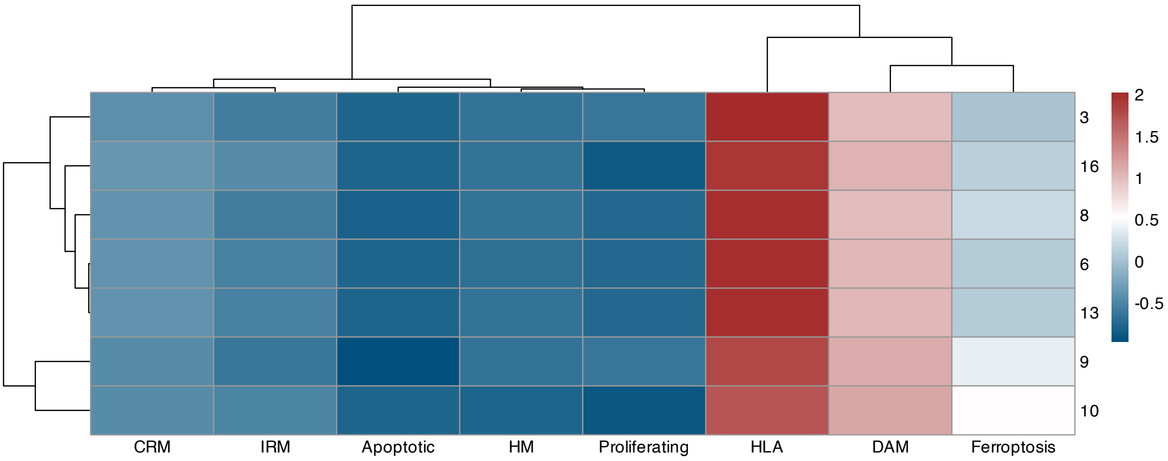


Figure S1: **Heatmap of aggregated module scores (AddModuleScore, z-scored across cells) for microglial subtype marker gene sets.** The plot illustrates relative enrichment of transcriptional signatures associated with distinct microglial subtypes.

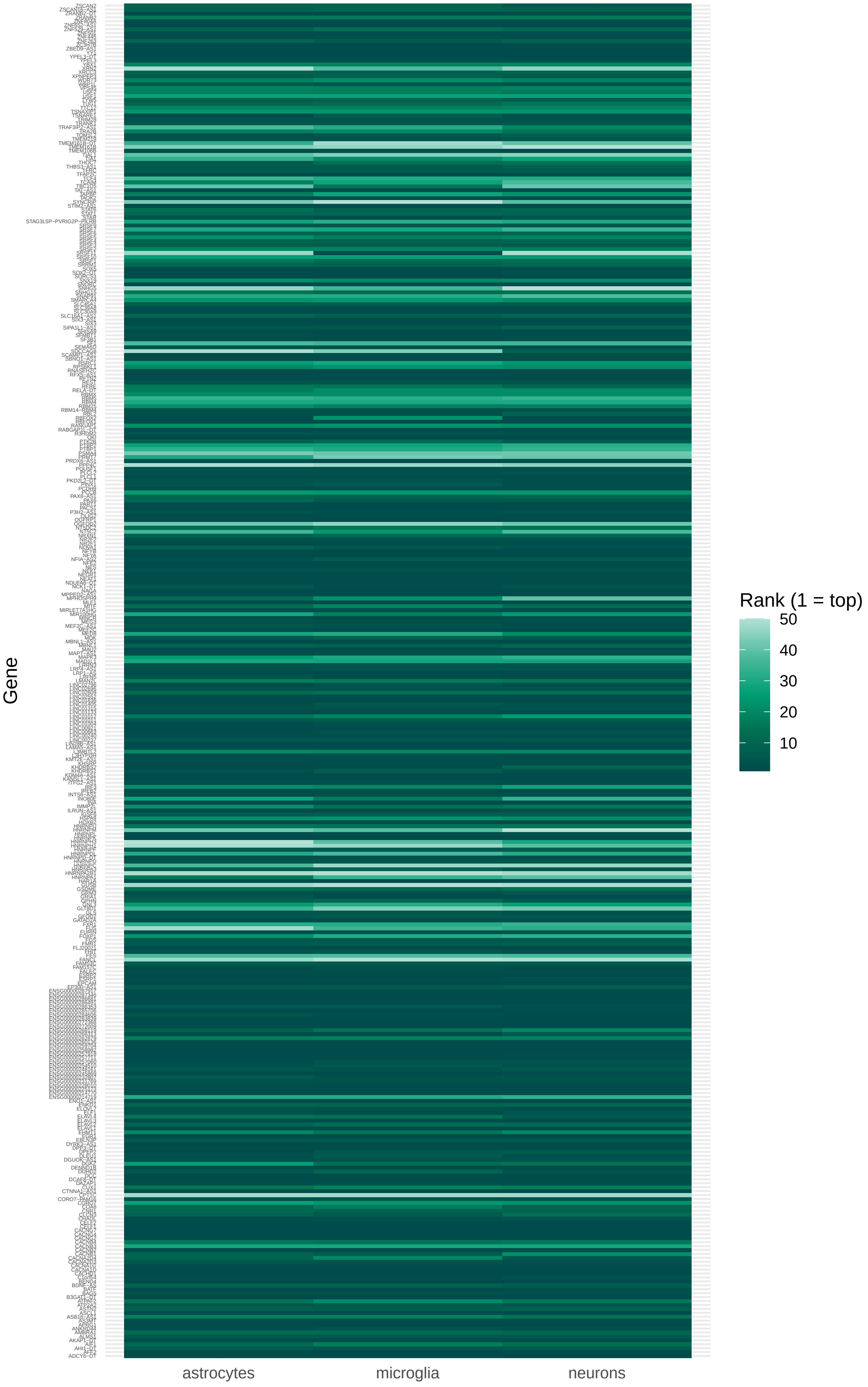


Figure S2: **Lineage-specific ranking of Capture isoforms.** Heatmap showing the rank of CaptureSeq isoforms by mean expression relative to all other isoforms of the same gene across microglia, neuronal, and astrocyte lineages.

Appendix C

Additional Material for Chapter 4

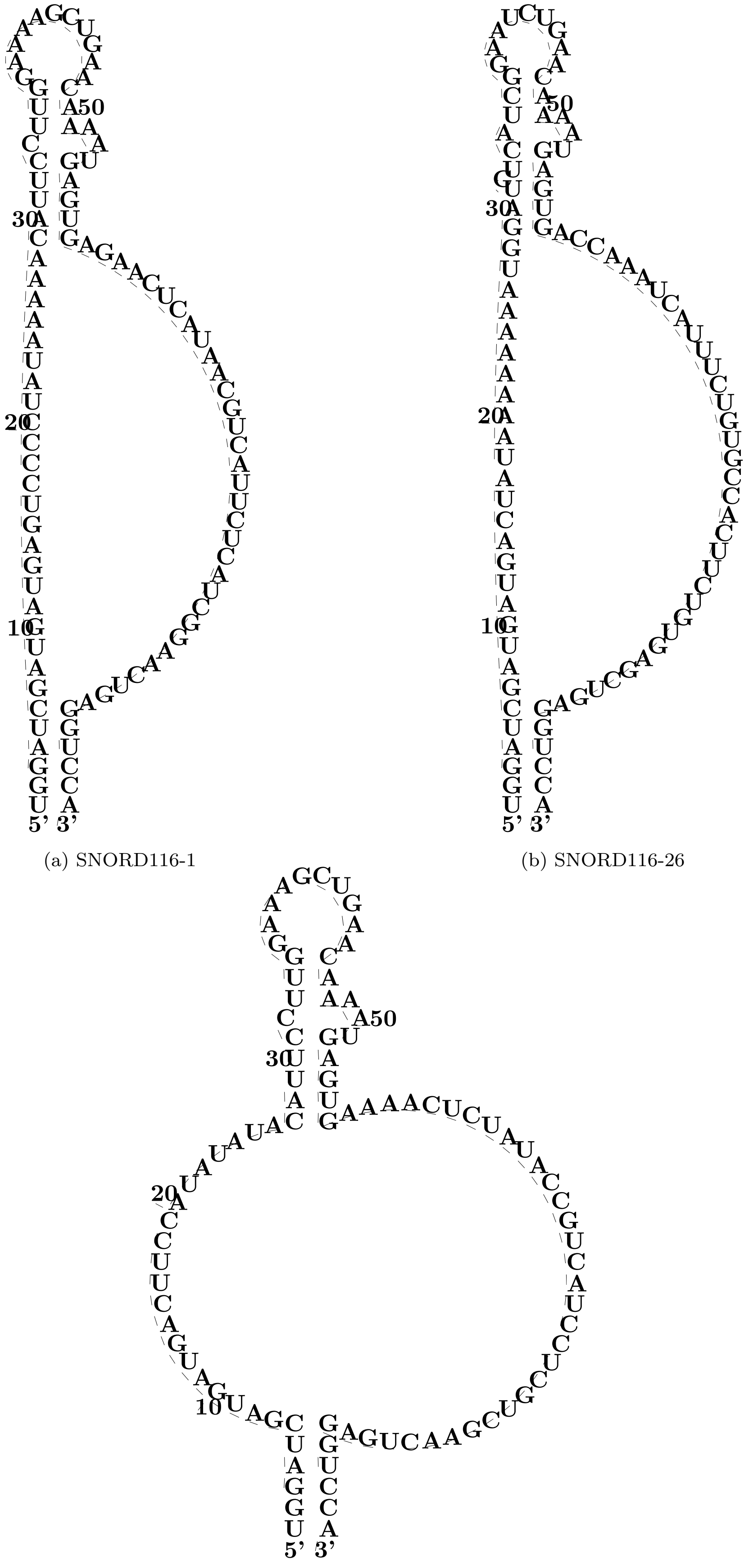
All supplementary tables listed within this chapter are available at: `https://github.com/skudashev/AltPolya_analysis`

Supplementary Tables:

S1–S3 - Differentially expressed genes (DEGs) identified with edgeR at days 2, 6, and 30 of cardiomyocyte differentiation. Each table lists significantly up- or down-regulated genes, $\log_2$ fold changes, false discovery rate (FDR)–adjusted p-values, and expression summaries.

S4–S6 - Differential transcript expression (DTE) identified with edgeR at days 2, 6, and 30.

S7 - Differential transcript usage (DTU) analysis results.



(a) SNORD116-1

(b) SNORD116-26

(c) SNORD116-17/19

Figure S1: **Predicted secondary structures of selected SNORD116 family members.** Taken from RNA-central (Sweeney et al., 2021)

.

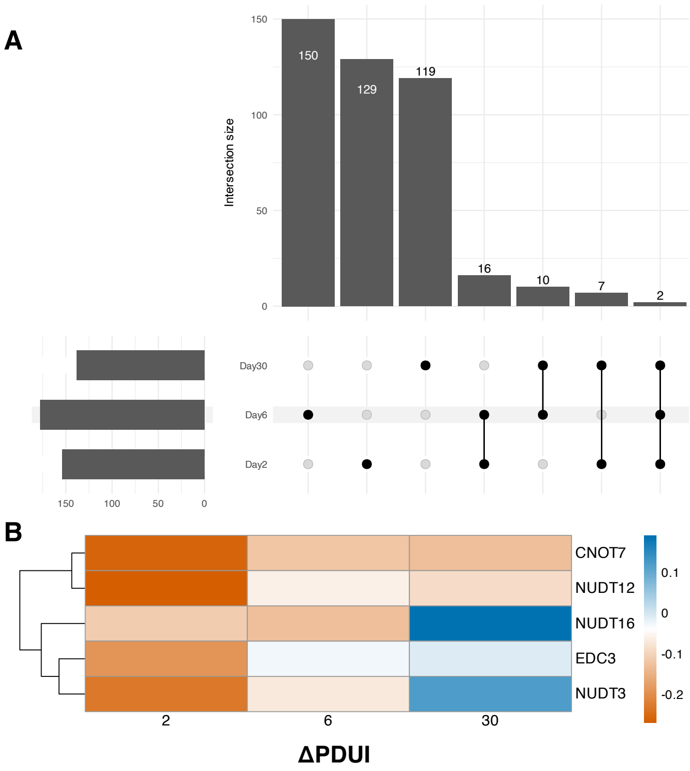


Figure S2: **Alternative polyadenylation dynamics and clustering of SNORD116-regulated genes.** (A) UpSet plot summarising the overlap of alternatively polyadenylated genes across developmental timepoints. (B) Heatmap of ΔPDUI values for cluster 3, the only k-means cluster (k=4) showing significant Gene Ontology enrichment.

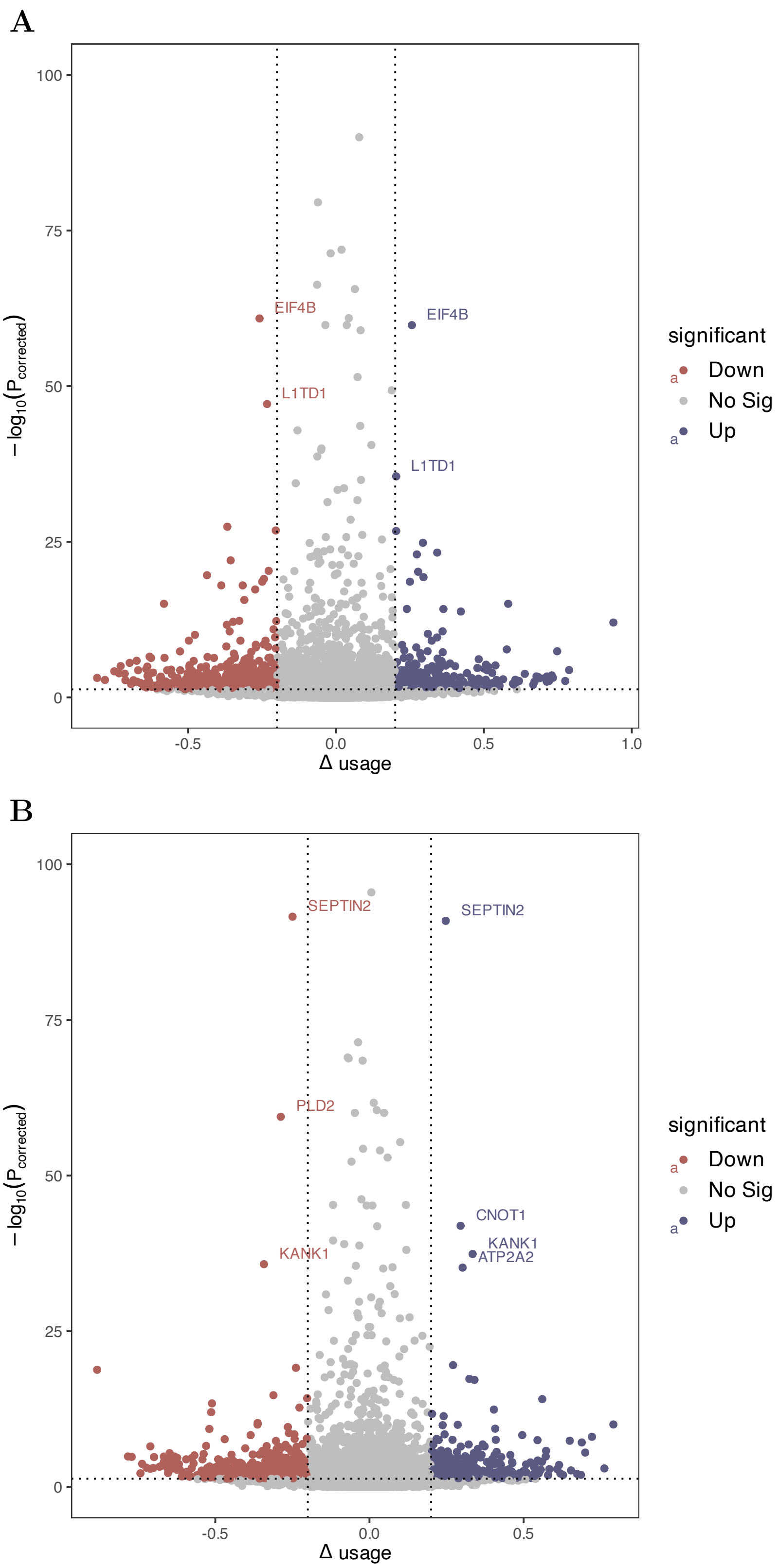


Figure S3: **Differential alternative polyadenylation between *SNORD116*$^{-/-}$ and wild-type cardiomyocytes.** Volcano plots show differential poly(A) site usage detected by `LAPA` at (A) day 2 and (B) day 6 of differentiation. Each point represents an individual poly(A) site, with significant changes in usage highlighted. Top differentially regulated genes are annotated.

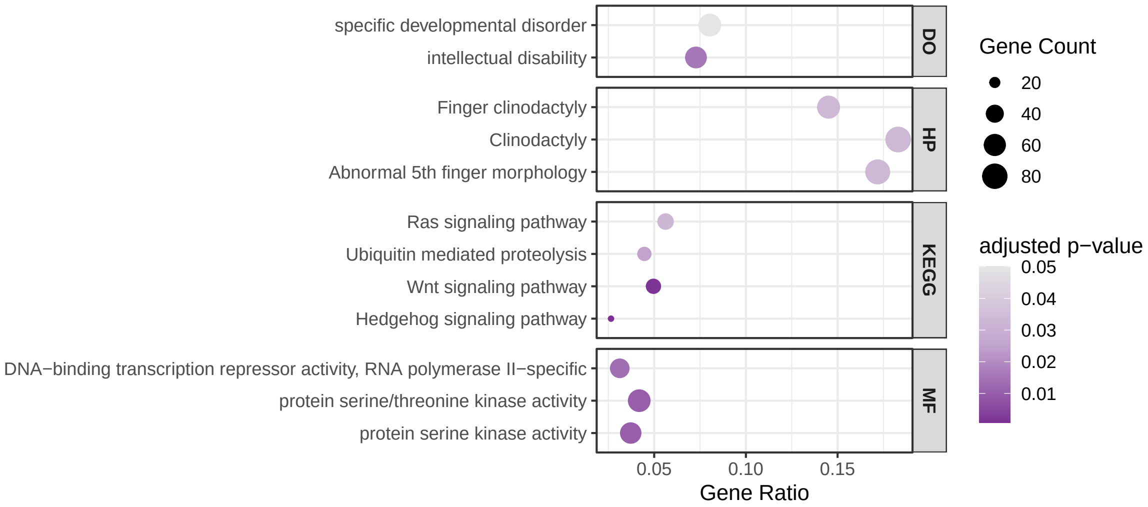


Figure S4: **Functional enrichment of alternatively polyadenylated genes.** Dot plot of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment. Dot size reflects the number of APA genes assigned to each term.

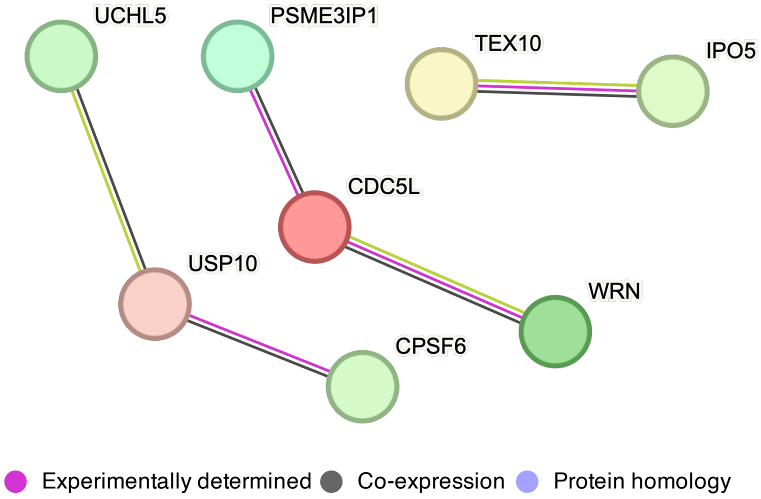


Figure S5: **Human STRING protein-protein interaction (PPI) network for P/R ratio change proteins.**

Bibliography

Abood, A., Mesner, L. D., Jeffery, E. D., Murali, M., Lehe, M. D., Saquing, J., Farber, C. R., and Sheynkman, G. M. (2024). Long-read proteogenomics to connect disease-associated sQTLs to the protein isoform effectors of disease. *Am. J. Hum. Genet.*, 111(9):1914–1931.

Abou Alezz, M., Celli, L., Belotti, G., Lisa, A., and Bione, S. (2020). GC-AG introns features in long non-coding and protein-coding genes suggest their role in gene expression regulation. *Front. Genet.*, 11:488.

Abrahams, L., Savisaar, R., Mordstein, C., Young, B., Kudla, G., and Hurst, L. D. (2021). Evidence in disease and non-disease contexts that nonsense mutations cause altered splicing via motif disruption. *Nucleic Acids Res.*, 49(17):9665–9685.

Abugessaisa, I., Noguchi, S., Hasegawa, A., Kondo, A., Kawaji, H., Carninci, P., and Kasukawa, T. (2019). refTSS: A reference data set for human and mouse transcription start sites. *J. Mol. Biol.*, 431(13):2407–2422.

Agarwal, V., Lopez-Darwin, S., Kelley, D. R., and Shendure, J. (2021). The landscape of alternative polyadenylation in single cells of the developing mouse embryo. *Nat. Commun.*, 12(1):5101.

Ahlert Scoones, A. L., Lan, Y., Utting, C., Pouncey, L., Lister, A., Kudasheva, S., Mehta, N., Irish, N., Swarbreck, D., Gharbi, K., Haerty, W., Cribbs, A. P., Wright, D. J., and Macaulay, I. C. (2025). A comparison of long-read single-cell transcriptomic approaches. *bioRxiv*, page 2025.07.03.662955.

Ake, F., Schilling, M., Fernández-Moya, S. M., Jaya Ganesh, A., Gutiérrez-Franco, A., Li, L., and Plass, M. (2025). Quantification of transcript isoforms at the single-cell level using SCALPEL. *Nat. Commun.*, 16(1):6402.

Al'Khafaji, A. M., Smith, J. T., Garimella, K. V., Babadi, M., Popic, V., Sade-Feldman, M., Gatzen, M., Sarkizova, S., Schwartz, M. A., Blaum, E. M., Day, A., Costello, M., Bowers, T., Gabriel, S., Banks, E., Philippakis, A. A., Boland, G. M., Blainey, P. C., and Hacohen, N. (2024). High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nat. Biotechnol.*, 42(4):582–586.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.*, 21(1):30.

Anczuków, O., Akerman, M., Cléry, A., Wu, J., Shen, C., Shirole, N. H., Raimer, A., Sun, S., Jensen, M. A., Hua, Y., Allain, F. H.-T., and Krainer, A. R. (2015). SRSF1-regulated alternative splicing in breast cancer. *Mol. Cell*, 60(1):105–117.

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Res.*, 22(10):2008–2017.

Arad, G. and Geiger, T. (2023). Functional impact of protein-RNA variation in clinical cancer analyses. *Mol. Cell. Proteomics*, 22(7):100587.

Au, K. F., Sebastiano, V., Afshar, P. T., Durruthy, J. D., Lee, L., Williams, B. A., van Bakel, H., Schadt, E. E., Reijo-Pera, R. A., Underwood, J. G., and Wong, W. H. (2013). Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 110(50):E4821–30.

Bacher, R., Chu, L.-F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., Newton, M., and Kendziorski, C. (2017). SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods*, 14(6):584–586.

Bak, M., van Nimwegen, E., Kouzel, I. U., Gur, T., Schmidt, R., Zavolan, M., and Gruber, A. J. (2024). MAPP unravels frequent co-regulation of splicing and polyadenylation by RNA-binding proteins and their dysregulation in cancer. *Nat. Commun.*, 15(1):4110.

Baldini, L., Charpentier, B., and Labialle, S. (2021). Emerging data on the diversity of molecular mechanisms involving C/D snoRNAs. *Noncoding RNA*, 7(2):30.

Baldini, L., Robert, A., Charpentier, B., and Labialle, S. (2022). Phylogenetic and molecular analyses identify SNORD116 targets involved in the prader-willi syndrome. *Mol. Biol. Evol.*, 39(1):msab348.

Baldoni, P. L., Chen, Y., Hediyeh-Zadeh, S., Liao, Y., Dong, X., Ritchie, M. E., Shi, W., and Smyth, G. K. (2024). Dividing out quantification uncertainty allows efficient assessment of differential transcript expression with edgeR. *Nucleic Acids Res.*, 52(3):e13.

Baralle, F. E. and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.*, 18(7):437–451.

Barry, G., Briggs, J. A., Vanichkina, D. P., Poth, E. M., Beveridge, N. J., Ratnu, V. S., Nayler, S. P., Nones, K., Hu, J., Bredy, T. W., Nakagawa, S., Rigo, F., Taft, R. J., Cairns, M. J., Blackshaw, S., Wolvetang, E. J., and Mattick, J. S. (2014). The long non-coding RNA gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing. *Mol. Psychiatry*, 19(4):486–494.

Bartolomei, M. S., Zemel, S., and Tilghman, S. M. (1991). Parental imprinting of the mouse H19 gene. *Nature*, 351(6322):153–155.

Baumgartner, M., Drake, K., and Kanadia, R. N. (2019). An integrated model of minor intron emergence and conservation. *Front. Genet.*, 10:1113.

Bayraktar, O. A., Bartels, T., Holmqvist, S., Kleshchevnikov, V., Martirosyan, A., Polioudakis, D., Ben Haim, L., Young, A. M. H., Batiuk, M. Y., Prakash, K., Brown, A., Roberts, K., Paredes, M. F., Kawaguchi, R., Stockley, J. H., Sabeur, K., Chang, S. M., Huang, E., Hutchinson, P., Ullian, E. M., Hemberg, M., Coppola, G., Holt, M. G., Geschwind, D. H., and Rowitch, D. H. (2020). Astrocyte layers in the mammalian cerebral cortex revealed by a single-cell in situ transcriptomic map. *Nat. Neurosci.*, 23(4):500–509.

Bello, E., Long, K., Iwama, S., Steer, J., Cooper, S., Alasoo, K., Kumasaka, N., Schwartzentruber, J., Panousis, N. I., and Bassett, A. (2023). An alzheimer's disease-associated common regulatory variant in a PTK2B intron alters microglial function. *bioRxiv*, page 2023.11.04.565613.

Beneventi, G., Munita, R., Cao Thi Ngoc, P., Madej, M., Cieśla, M., Muthukumar, S., Krogh, N., Nielsen, H., Swaminathan, V., and Bellodi, C. (2021). The small cajal body-specific RNA 15 (SCARNA15) directs p53 and redox homeostasis via selective splicing in cancer cells. *NAR Cancer*, 3(3):zcab026.

Bentley, D. L. (2005). Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr. Opin. Cell Biol.*, 17(3):251–256.

Bentley, D. L. (2014). Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.*, 15(3):163–175.

Bergeron, D., Faucher-Giguère, L., Emmerichs, A.-K., Choquet, K., Song, K. S., Deschamps-Francoeur, G., Fafard-Couture, , Rivera, A., Couture, S., Churchman, L. S., Heyd, F., Abou Elela, S., and Scott, M. S. (2023). Intronic small nucleolar RNAs regulate host gene splicing through base pairing with their adjacent intronic sequences. *Genome Biol.*, 24(1):160.

Bhattacharya, A., Vo, D. D., Jops, C., Kim, M., Wen, C., Hervoso, J. L., Pasaniuc, B., and Gandal, M. J. (2023). Isoform-level transcriptome-wide association uncovers genetic risk mechanisms for neuropsychiatric disorders in the human brain. *Nat. Genet.*, 55(12):2117–2128.

Bieth, E., Eddiry, S., Gaston, V., Lorenzini, F., Buffet, A., Conte Auriol, F., Molinas, C., Cailley, D., Rooryck, C., Arveiler, B., Cavaillé, J., Salles, J. P., and Tauber, M. (2015). Highly restricted deletion of the SNORD116 region is implicated in prader-willi syndrome. *Eur. J. Hum. Genet.*, 23(2):252–255.

Biosciences, P. (2023). MAS-seq for single-cell isoform sequencing. Technical report.

Blencowe, B. J. (2000). Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.*, 25(3):106–110.

Blencowe, B. J. (2017). The relationship between alternative splicing and proteomic complexity. *Trends Biochem. Sci.*, 42(6):407–408.

Bochukova, E. G., Lawler, K., Croizier, S., Keogh, J. M., Patel, N., Strohbehn, G., Lo, K. K., Humphrey, J., Hokken-Koelega, A., Damen, L., Donze, S., Bouret, S. G., Plagnol, V., and Farooqi, I. S. (2018). A transcriptomic signature of the hypothalamic response to fasting and BDNF deficiency in prader-willi syndrome. *Cell Rep.*, 22(13):3401–3408.

Boutz, P. L., Stoilov, P., Li, Q., Lin, C.-H., Chawla, G., Ostrow, K., Shiue, L., Ares, Jr, M., and Black, D. L. (2007). A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev.*, 21(13):1636–1652.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527.

Broseus, L., Thomas, A., Oldfield, A. J., Severac, D., Dubois, E., and Ritchie, W. (2020). TALC: Transcript-level aware long-read correction. *Bioinformatics*, 36(20):5000–5006.

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J., and Willard, H. F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71(3):527–542.

Buckanovich, R. J., Posner, J. B., and Darnell, R. B. (1993). Nova, the paraneoplastic ri antigen, is homologous to an RNA-binding protein and is specifically expressed in the developing motor system. *Neuron*, 11(4):657–672.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 36(5):411–420.

Bühler, M., Steiner, S., Mohn, F., Paillusson, A., and Mühlemann, O. (2006). EJC-independent degradation of nonsense immunoglobulin-mu mRNA depends on 3' UTR length. *Nat. Struct. Mol. Biol.*, 13(5):462–464.

Caldas, P., Luz, M., Baseggio, S., Andrade, R., Sobral, D., and Grosso, A. R. (2024). Transcription readthrough is prevalent in healthy human tissues and associated with inherent genomic features. *Commun. Biol.*, 7(1):100.

Calvo-Roitberg, E., Daniels, R. F., and Pai, A. A. (2023). Challenges in identifying mRNA transcript starts and ends from long-read sequencing data. *bioRxiv*, page 2023.07.26.550536.

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C., and Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667.

Cao, J., Routh, A. L., and Kuyumcu-Martinez, M. N. (2021a). Nanopore sequencing reveals full-length tropomyosin 1 isoforms and their regulation by RNA-binding proteins during rat heart development. *J. Cell. Mol. Med.*, 25(17):8352–8362.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., and Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502.

Cao, J., Verma, S. K., Jaworski, E., Mohan, S., Nagasawa, C. K., Rayavara, K., Sooter, A., Miller, S. N., Holcomb, R. J., Powell, M. J., Ji, P., Elrod, N. D., Yildirim, E., Wagner, E. J., Popov, V., Garg, N. J., Routh, A. L., and Kuyumcu-Martinez, M. N. (2021b). RBFOX2 is critical for maintaining alternative polyadenylation patterns and mitochondrial health in rat myoblasts. *Cell Rep.*, 37(5):109910.

Carlyle, B. C., Kitchen, R. R., Zhang, J., Wilson, R. S., Lam, T. T., Rozowsky, J. S., Williams, K. R., Sestan, N., Gerstein, M. B., and Nairn, A. C. (2018). Isoform-level interpretation of high-throughput proteomics data enabled by deep integration with RNA-seq. *J. Proteome Res.*, 17(10):3431–3444.

Cavaillé, J., Buiting, K., Kiefmann, M., Lalande, M., Brannan, C. I., Horsthemke, B., Bachellerie, J. P., Brosius, J., and Hüttenhofer, A. (2000). Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc. Natl. Acad. Sci. U. S. A.*, 97(26):14311–14316.

Chang, K.-T., Cheng, C.-F., King, P.-C., Liu, S.-Y., and Wang, G.-S. (2017). CELF1 mediates connexin 43 mRNA degradation in dilated cardiomyopathy. *Circ. Res.*, 121(10):1140–1152.

Chau, K. K., Zhang, P., Urresti, J., Amar, M., Pramod, A. B., Chen, J., Thomas, A., Corominas, R., Lin, G. N., and Iakoucheva, L. M. (2021). Full-length isoform transcriptome of the developing human brain provides further insights into autism. *Cell Rep.*, 36(9):109631.

Chen, A., Sun, Y., Lei, Y., Li, C., Liao, S., Meng, J., Bai, Y., Liu, Z., Liang, Z., Zhu, Z., Yuan, N., Yang, H., Wu, Z., Lin, F., Wang, K., Li, M., Zhang, S., Yang, M., Fei, T., Zhuang, Z., Huang, Y., Zhang, Y., Xu, Y., Cui, L., Zhang, R., Han, L., Sun, X., Chen, B., Li, W., Huangfu, B., Ma, K., Ma, J., Li, Z., Lin, Y., Wang, H., Zhong, Y., Zhang, H., Yu, Q., Wang, Y., Liu, X., Peng, J., Liu, C., Chen, W., Pan, W., An, Y., Xia, S., Lu, Y., Wang, M., Song, X., Liu, S., Wang, Z., Gong, C., Huang, X., Yuan, Y., Zhao, Y., Chai, Q., Tan, X., Liu, J., Zheng, M., Li, S., Huang, Y., Hong, Y., Huang, Z., Li, M., Jin, M., Li, Y., Zhang, H., Sun, S., Gao, L., Bai, Y., Cheng, M., Hu, G., Liu, S., Wang, B., Xiang, B., Li, S., Li, H., Chen, M., Wang, S., Li, M., Liu, W., Liu, X., Zhao, Q., Lisby, M., Wang, J., Fang, J., Lin, Y., Xie, Q., Liu, Z., He, J., Xu, H., Huang, W., Mulder, J., Yang, H., Sun, Y., Uhlen, M., Poo, M., Wang, J., Yao, J., Wei, W., Li, Y., Shen, Z., Liu, L., Liu, Z., Xu, X., and Li, C. (2023). Single-cell spatial transcriptome reveals cell-type organization in the macaque cortex. *Cell*, 186(17):3726–3743.e24.

Chen, M. and Manley, J. L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.*, 10(11):741–754.

Chen, S., Wang, H., Zhang, D., Chen, R., and Luo, J. (2024a). Readon: a novel algorithm to identify read-through transcripts with long-read sequencing data. *Bioinformatics*, 40(6):btae336.

Chen, X., Birey, F., Li, M.-Y., Revah, O., Levy, R., Thete, M. V., Reis, N., Kaganovsky, K., Onesto, M., Sakai, N., Hudacova, Z., Hao, J., Meng, X., Nishino, S., Huguenard, J., and Pașca, S. P. (2024b). Antisense oligonucleotide therapeutic approach for timothy syndrome. *Nature*, 628(8009):818–825.

Chen, X., Huang, Y., Huang, L., Huang, Z., Hao, Z.-Z., Xu, L., Xu, N., Li, Z., Mou, Y., Ye, M., You, R., Zhang, X., Liu, S., and Miao, Z. (2024c). A brain cell atlas integrating single-cell transcriptomes across human brain regions. *Nat. Med.*, 30(9):2679–2691.

Chen, X., Sun, G., Feng, L., Tian, E., and Shi, Y. (2025a). Human iPSC-derived microglial cells protect neurons from neurodegeneration in long-term cultured adhesion brain organoids. *Commun. Biol.*, 8(1):30.

Chen, Y., Chen, L., Lun, A. T. L., Baldoni, P. L., and Smyth, G. K. (2025b). edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Res.*, 53(2):gkaf018.

Chen, Y., Sim, A., Wan, Y. K., Yeo, K., Lee, J. J. X., Ling, M. H., Love, M. I., and Göke, J. (2022). Context-aware transcript quantification from long read RNA-seq data with bambu. *bioRxiv*, page 2022.11.14.516358.

Cheng, P., Zhao, X., Katsnelson, L., Camacho-Hernandez, E. M., Mermerian, A., Mays, J. C., Lippman, S. M., Rosales-Alvarez, R. E., Moya, R., Shwetar, J., Grun, D., Fenyo, D., and Davoli, T. (2022). Proteogenomic analysis of cancer aneuploidy and normal tissues reveals divergent modes of gene regulation across cellular pathways. *Elife*, 11:e75227.

Chiriboga, C. A. (2017). Nusinersen for the treatment of spinal muscular atrophy. *Expert Rev. Neurother.*, 17(10):955–962.

Cho, V., Mei, Y., Sanny, A., Chan, S., Enders, A., Bertram, E. M., Tan, A., Goodnow, C. C., and Andrews, T. D. (2014). The RNA-binding protein hnRNPLL induces a T cell alternative splicing program delineated by differential intron retention in polyadenylated RNA. *Genome Biol.*, 15(1):R26.

Chorev, M. and Carmel, L. (2012). The function of introns. *Front. Genet.*, 3:55.

Choudhary, S. and Satija, R. (2022). Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biol.*, 23(1):27.

Clark, M. B., Mercer, T. R., Bussotti, G., Leonardi, T., Haynes, K. R., Crawford, J., Brunck, M. E., Cao, K.-A. L., Thomas, G. P., Chen, W. Y., Taft, R. J., Nielsen, L. K., Enright, A. J., Mattick, J. S., and Dinger, M. E. (2015). Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat. Methods*, 12(4):339–342.

Clark, M. B., Wrzesinski, T., Garcia, A. B., Hall, N. A. L., Kleinman, J. E., Hyde, T., Weinberger, D. R., Harrison, P. J., Haerty, W., and Tunbridge, E. M. (2020a). Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain. *Mol. Psychiatry*, 25(1):37–47.

Clark, M. B., Wrzesinski, T., Garcia, A. B., Hall, N. A. L., Kleinman, J. E., Hyde, T., Weinberger, D. R., Harrison, P. J., Haerty, W., and Tunbridge, E. M. (2020b). Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain. *Mol. Psychiatry*, 25(1):37–47.

Cocquet, J., Chong, A., Zhang, G., and Veitia, R. A. (2006). Reverse transcriptase template switching and false alternative transcripts. *Genomics*, 88(1):127–131.

Cogliati, T., Delgado-Romero, P., Norwitz, E. R., Guduric-Fuchs, J., Kaiser, U. B., Wray, S., and Kirsch, I. R. (2007). Pubertal impairment in Nhlh2 null mice is associated with hypothalamic and pituitary deficiencies. *Mol. Endocrinol.*, 21(12):3013–3027.

Cohen, O. S., Weickert, T. W., Hess, J. L., Paish, L. M., McCoy, S. Y., Rothmond, D. A., Galletly, C., Liu, D., Weinberg, D. D., Huang, X.-F., Xu, Q., Shen, Y., Zhang, D., Yue, W., Yan, J., Wang, L., Lu, T., He, L., Shi, Y., Xu, M., Che, R., Tang, W., Chen, C.-H., Chang, W.-H., Hwu, H.-G., Liu, C.-M., Liu, Y.-L., Wen, C.-C., Fann, C. S.-J., Chang, C.-C., Kanazawa, T., Middleton, F. A., Duncan, T. M., Faraone, S. V., Weickert, C. S., Tsuang, M. T., and Glatt, S. J. (2016). A splicing-regulatory polymorphism in DRD2 disrupts ZRANB2 binding, impairs cognitive functioning and increases risk for schizophrenia in six han chinese samples. *Mol. Psychiatry*, 21(7):975–982.

Colbourne, L., Luciano, S., and Harrison, P. J. (2021). Onset and recurrence of psychiatric disorders associated with anti-hypertensive drug classes. *Transl. Psychiatry*, 11(1):319.

Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S., and Yosef, N. (2019). Performance assessment and selection of normalization procedures for single-cell RNA-seq. *Cell Syst*, 8(4):315–328.e8.

Collado-Torres, L., Burke, E. E., Peterson, A., Shin, J., Straub, R. E., Rajpurohit, A., Semick, S. A., Ulrich, W. S., BrainSeq Consortium, Price, A. J., Valencia, C., Tao, R., Deep-Soboslay, A., Hyde, T. M., Kleinman, J. E., Weinberger, D. R., and Jaffe, A. E. (2019). Regional heterogeneity in gene expression, regulation, and coherence in the frontal cortex and hippocampus across development and schizophrenia. *Neuron*, 103(2):203–216.e8.

Connelly, S. and Manley, J. L. (1988). A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev.*, 2(4):440–452.

Corden, J. L. and Patturajan, M. (1997). A CTD function linking transcription to splicing. *Trends Biochem. Sci.*, 22(11):413–416.

Cui, L., Li, S., Wang, S., Wu, X., Liu, Y., Yu, W., Wang, Y., Tang, Y., Xia, M., and Li, B. (2024). Major depressive disorder: hypothesis, mechanism, prevention and treatment. *Signal Transduct. Target. Ther.*, 9(1):30.

Cvetkovic, C., Patel, R., Shetty, A., Hogan, M. K., Anderson, M., Basu, N., Aghlara-Fotovat, S., Ramesh, S., Sardar, D., Veiseh, O., Ward, M. E., Deneen, B., Horner, P. J., and Krencik, R. (2022). Assessing gq-GPCR-induced human astrocyte reactivity using bioengineered neural organoids. *J. Cell Biol.*, 221(4).

Dai, Q., Zhang, L.-S., Sun, H.-L., Pajdzik, K., Yang, L., Ye, C., Ju, C.-W., Liu, S., Wang, Y., Zheng, Z., Zhang, L., Harada, B. T., Dou, X., Irkliyenko, I., Feng, X., Zhang, W., Pan, T., and He, C. (2023). Quantitative sequencing using BID-seq uncovers abundant pseudouridines in mammalian mRNA at base resolution. *Nat. Biotechnol.*, 41(3):344–354.

Dai, W., Zhang, G., and Makeyev, E. V. (2012). RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage. *Nucleic Acids Res.*, 40(2):787–800.

Dainis, A., Tseng, E., Clark, T. A., Hon, T., Wheeler, M., and Ashley, E. (2019). Targeted long-read RNA sequencing demonstrates transcriptional diversity driven by splice-site variation in MYBPC3. *Circ Genom Precis Med*, 12(5):e002464.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2).

Daniels, N. J., Hershberger, C. E., Gu, X., Schueger, C., DiPasquale, W. M., Brick, J., Saunthararajah, Y., Maciejewski, J. P., and Padgett, R. A. (2021). Functional analyses of human LUC7-like proteins involved in splicing regulation and myeloid neoplasms. *Cell Rep.*, 35(2):108989.

Dar, R. D., Razooky, B. S., Singh, A., Trimeloni, T. V., McCollum, J. M., Cox, C. D., Simpson, M. L., and Weinberger, L. S. (2012). Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl. Acad. Sci. U. S. A.*, 109(43):17454–17459.

Dar, S. A., Malla, S., Belair, C., and Maragkakis, M. (2024). Differential poly(a) tail length analysis using nanopore sequencing. *Methods Mol. Biol.*, 2723:267–283.

Dark, N., Cosson, M.-V., Tsansizi, L. I., Owen, T. J., Ferraro, E., Francis, A. J., Tsai, S., Bouissou, C., Weston, A., Collinson, L., Abi-Gerges, N., Miller, P. E., MacLeod, K. T., Ehler, E., Mitter, R., Harding, S. E., Smith, J. C., and Bernardo, A. S. (2023). Generation of left ventricle-like cardiomyocytes with improved structural, functional, and metabolic maturity from human pluripotent stem cells. *Cell Rep. Methods*, 3(4):100456.

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Hayden Gephart, M. G., Barres, B. A., and Quake, S. R. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.*, 112(23):7285–7290.

Das, S. and Krainer, A. R. (2014). Emerging functions of SRSF1, splicing factor and oncoprotein, in RNA metabolism and cancer. *Mol. Cancer Res.*, 12(9):1195–1204.

Davis, S., Scott, C., Oetjen, J., Charles, P. D., Kessler, B. M., Ansorge, O., and Fischer, R. (2023). Deep topographic proteomics of a human brain tumour. *Nat. Commun.*, 14(1):7710.

De Conti, L., Baralle, M., and Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing: Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip. Rev. RNA*, 4(1):49–60.

de la Rubia, I., Srivastava, A., Xue, W., Indi, J. A., Carbonell-Sala, S., Lagarde, J., Albà, M. M., and Eyras, E. (2022). RATTLE: reference-free reconstruction and quantification of transcriptomes from nanopore sequencing. *Genome Biol.*, 23(1):153.

Degener, M. J. F., van Cruchten, R. T. P., Otero, B. A., Wang, E. T., Wansink, D. G., and 't Hoen, P. A. C. (2022). A comprehensive atlas of fetal splicing patterns in the brain of adult myotonic dystrophy type 1 patients. *NAR Genom Bioinform*, 4(1):lqac016.

Deschamps-Francoeur, G., Couture, S., Abou-Elela, S., and Scott, M. S. (2022). The snoGloBe interaction predictor reveals a broad spectrum of C/D snoRNA RNA targets. *Nucleic Acids Res.*, 50(11):6067–6083.

Deveson, I. W., Brunck, M. E., Blackburn, J., Tseng, E., Hon, T., Clark, T. A., Clark, M. B., Crawford, J., Dinger, M. E., Nielsen, L. K., Mattick, J. S., and Mercer, T. R. (2018). Universal alternative splicing of noncoding exons. *Cell Syst*, 6(2):245–255.e5.

Dhillon, S. (2020). Risdiplam: First approval. *Drugs*, 80(17):1853–1858.

Di Giammartino, D. C., Nishida, K., and Manley, J. L. (2011). Mechanisms and consequences of alternative polyadenylation. *Mol. Cell*, 43(6):853–866.

Ding, J., Adiconis, X., Simmons, S. K., Kowalczyk, M. S., Hession, C. C., Marjanovic, N. D., Hughes, T. K., Wadsworth, M. H., Burks, T., Nguyen, L. T., Kwon, J. Y. H., Barak, B., Ge, W., Kedaigle, A. J., Carroll, S., Li, S., Hacohen, N., Rozenblatt-Rosen, O., Shalek, A. K., Villani, A.-C., Regev, A., and Levin, J. Z. (2020). Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.*, 38(6):737–746.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Domínguez Conde, C., Xu, C., Jarvis, L. B., Rainbow, D. B., Wells, S. B., Gomes, T., Howlett, S. K., Suchanek, O., Polanski, K., King, H. W., Mamanova, L., Huang, N., Szabo, P. A., Richardson, L., Bolt, L., Fasouli, E. S., Mahbubani, K. T., Prete, M., Tuck, L., Richoz, N., Tuong, Z. K., Campos, L., Mousa, H. S., Needham, E. J., Pritchard, S., Li, T., Elmentaite, R., Park, J., Rahmani, E., Chen, D., Menon, D. K., Bayraktar, O. A., James, L. K., Meyer, K. B., Yosef, N., Clatworthy, M. R., Sims, P. A., Farber, D. L., Saeb-Parsy, K., Jones, J. L., and Teichmann, S. A. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197.

Dondi, A., Lischetti, U., Jacob, F., Singer, F., Borgsmüller, N., Coelho, R., Tumor Profiler Consortium, Heinzelmann-Schwarz, V., Beisel, C., and Beerenwinkel, N. (2023). Detection of isoforms and genomic alterations by high-throughput full-length single-cell RNA sequencing in ovarian cancer. *Nat. Commun.*, 14(1):7780.

Dong, X., Du, M. R. M., Gouil, Q., Tian, L., Baldoni, P. L., Smyth, G. K., Amarasinghe, S. L., Law, C. W., and Ritchie, M. E. (2022). Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *bioRxiv*, page 2022.07.22.501076.

Doolittle, W. F. (2013). Is junk DNA bunk? a critique of ENCODE. *Proc. Natl. Acad. Sci. U. S. A.*, 110(14):5294–5300.

Dredge, B. K., Stefani, G., Engelhard, C. C., and Darnell, R. B. (2005). Nova autoregulation reveals dual functions in neuronal splicing. *EMBO J.*, 24(8):1608–1620.

Duan, Y., Yue, K., Ye, B., Chen, P., Zhang, J., He, Q., Wu, Y., Lai, Q., Li, H., Wu, Y., Jing, C., and Wang, X. (2023). LncRNA MALAT1 promotes growth and metastasis of head and neck squamous cell carcinoma by repressing VHL through a non-canonical function of EZH2. *Cell Death Dis.*, 14(2):149.

Duncan, L. E., Li, T., Salem, M., Li, W., Mortazavi, L., Senturk, H., Shahverdizadeh, N., Vesuna, S., Shen, H., Yoon, J., Wang, G., Ballon, J., Tan, L., Pruett, B. S., Knutson, B., Deisseroth, K., and Giardino, W. J. (2025). Mapping the cellular etiology of schizophrenia and complex brain phenotypes. *Nat. Neurosci.*, 28(2):248–258.

Dvinge, H. (2018). Regulation of alternative mRNA splicing: old players and new perspectives. *FEBS Lett.*, 592(17):2987–3006.

Edwalds-Gilbert, G., Veraldi, K. L., and Milcarek, C. (1997). Alternative poly(a) site selection in complex transcription units: means to an end? *Nucleic Acids Res.*, 25(13):2547–2561.

Edwards, C. R., Ritchie, W., Wong, J. J.-L., Schmitz, U., Middleton, R., An, X., Mohandas, N., Rasko, J. E. J., and Blobel, G. A. (2016). A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. *Blood*, 127(17):e24–e34.

Eißmann, M., Gutschner, T., Hämmerle, M., Günther, S., Caudron-Herger, M., Groß, M., Schirmacher, P., Rippe, K., Braun, T., Zörnig, M., and Diederichs, S. (2012). Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. *RNA Biol.*, 9(8):1076–1087.

Ellis, J. D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J. A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P. M., Wrana, J. L., and Blencowe, B. J. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol. Cell*, 46(6):884–892.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Engal, E., Zhang, Z., Geminder, O., Jaffe-Herman, S., Kay, G., Ben-Hur, A., and Salton, M. (2024). The spectrum of pre-mRNA splicing in autism. *Wiley Interdiscip. Rev. RNA*, 15(2):e1838.

Eze, U. C., Bhaduri, A., Haeussler, M., Nowakowski, T. J., and Kriegstein, A. R. (2021). Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci.*, 24(4):584–594.

Ezkurdia, I., Rodriguez, J. M., Carrillo-de Santa Pau, E., Vázquez, J., Valencia, A., and Tress, M. L. (2015). Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.*, 14(4):1880–1887.

Fahey, L. and Lopez, L. M. (2024). Pathway-based polygenic score analysis identifies the NRF2-KEAP1 and mRNA splicing - minor pathways as enriched in shared genetic variation between chronotype and bipolar disorder. *bioRxiv*, page 2024.02.16.24302920.

Fair, B., Buen Abad Najar, C. F., Zhao, J., Lozano, S., Reilly, A., Mossian, G., Staley, J. P., Wang, J., and Li, Y. I. (2024). Global impact of unproductive splicing on human gene expression. *Nat. Genet.*, 56(9):1851–1861.

Falaleeva, M., Pages, A., Matuszek, Z., Hidmi, S., Agranat-Tamir, L., Korotkov, K., Nevo, Y., Eyras, E., Sperling, R., and Stamm, S. (2016). Dual function of C/D box small nucleolar RNAs in rRNA modification and alternative pre-mRNA splicing. *Proc. Natl. Acad. Sci. U. S. A.*, 113(12):E1625–34.

Fansler, M. M., Mitschka, S., and Mayr, C. (2024). Quantifying 3'UTR length from scRNA-seq data reveals changes independent of gene expression. *Nat. Commun.*, 15(1):4050.

Farach, L. S., Little, M. E., Duker, A. L., Logan, C. V., Jackson, A., Hecht, J. T., and Bober, M. (2018). The expanding phenotype of RNU4ATAC pathogenic variants to lowry wood syndrome. *Am. J. Med. Genet. A*, 176(2):465–469.

Feng, H., Moakley, D. F., Chen, S., McKenzie, M. G., Menon, V., and Zhang, C. (2021). Complexity and graded regulation of neuronal cell-type–specific alternative splicing revealed by single-cell RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 118(10).

Feng, Q., Lin, Z., Zhao, D., Li, M., Yang, S., Xiang, A. P., Ye, C., and Yao, C. (2025). Functional inhibition of core spliceosomal machinery activates intronic premature cleavage and polyadenylation of pre-mRNAs. *Cell Rep.*, 44(3):115376.

Feng, X., Li, L., Wagner, E. J., and Li, W. (2018). TC3A: The cancer 3' UTR atlas. *Nucleic Acids Res.*, 46(D1):D1027–D1030.

Fernández-Mendívil, C., Luengo, E., Trigo-Alonso, P., García-Magro, N., Negredo, P., and López, M. G. (2021). Protective role of microglial HO-1 blockade in aging: Implication of iron metabolism. *Redox Biol.*, 38(101789):101789.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., and Gottardo, R. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, 16:278.

Fong, N., Brannan, K., Erickson, B., Kim, H., Cortazar, M. A., Sheridan, R. M., Nguyen, T., Karp, S., and Bentley, D. L. (2015). Effects of transcription elongation rate and Xrn2 exonuclease activity on RNA polymerase II termination suggest widespread kinetic competition. *Mol. Cell*, 60(2):256–267.

Foord, C., Prjibelski, A. D., Hu, W., Michielsen, L., Vandelli, A., Narykov, O., Evans, B., Hsu, J., Belchikov, N., Jarroux, J., He, Y., Ross, M. E., Hajirasouliha, I., Tartaglia, G. G., Korkin, D., Tomescu, A. I., and Tilgner, H. U. (2025). A spatial long-read approach at near-single-cell resolution reveals developmental regulation of splicing and polyadenylation sites in distinct cortical layers and cell types. *bioRxivorg*.

Gandal, M. J., Haney, J. R., Wamsley, B., Yap, C. X., Parhami, S., Emani, P. S., Chang, N., Chen, G. T., Hoftman, G. D., de Alba, D., Ramaswami, G., Hartl, C. L., Bhattacharya, A., Luo, C., Jin, T., Wang, D., Kawaguchi, R., Quintero, D., Ou, J., Wu, Y. E., Parikshak, N. N., Swarup, V., Belgard, T. G., Gerstein, M., Pasaniuc, B., and Geschwind, D. H. (2022). Broad transcriptomic dysregulation occurs across the cerebral cortex in ASD. *Nature*, 611(7936):532–539.

Gandal, M. J., Zhang, P., Hadjimichael, E., Walker, R. L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., Shieh, A. W., Haney, J., Parhami, S., Belmont, J., Kim, M., Moran Losada, P., Khan, Z., Mleczko, J., Xia, Y., Dai, R., Wang, D., Yang, Y. T., Xu, M., Fish, K., Hof, P. R., Warrell, J., Fitzgerald, D., White, K., Jaffe, A. E., PsychENCODE Consortium, Peters, M. A., Gerstein, M., Liu, C., Iakoucheva, L. M., Pinto, D., and Geschwind, D. H. (2018). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science*, 362(6420).

Gao, L., Wang, X., Guo, S., Xiao, L., Liang, C., Wang, Z., Li, Y., Liu, Y., Yao, R., Liu, Y., and Zhang, Y. (2019). LncRNA HOTAIR functions as a competing endogenous RNA to upregulate SIRT1 by sponging miR-34a in diabetic cardiomyopathy: GAOet al. *J. Cell. Physiol.*, 234(4):4944–4958.

Gao, Y., Li, L., Amos, C. I., and Li, W. (2021). Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression. *Genome Res.*, 31(10):1856–1866.

Garcia-Cabau, C., Bartomeu, A., Tesei, G., Cheung, K. C., Pose-Utrilla, J., Picó, S., Balaceanu, A., Duran-Arqué, B., Fernández-Alfara, M., Martín, J., De Pace, C., Ruiz-Pérez, L., García, J., Battaglia, G., Lucas, J. J., Hervás, R., Lindorff-Larsen, K., Méndez, R., and Salvatella, X. (2025). Mis-splicing of a neuronal microexon promotes CPEB4 aggregation in ASD. *Nature*, 637(8045):496–503.

García-Bea, A., Bermudez, I., Harrison, P. J., and Lane, T. A. (2017). A group II metabotropic glutamate receptor 3 (mGlu3, GRM3) isoform implicated in schizophrenia interacts with canonical mGlu3 and reduces ligand binding. *J. Psychopharmacol.*, 31(12):1519–1526.

Gauthier, J., Spiegelman, D., Piton, A., Lafrenière, R. G., Laurent, S., St-Onge, J., Lapointe, L., Hamdan, F. F., Cossette, P., Mottron, L., Fombonne, E., Joober, R., Marineau, C., Drapeau, P., and Rouleau, G. A. (2009). Novel de novo SHANK3 mutation in autistic patients. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 150B(3):421–424.

Geisberg, J. V., Moqtaderi, Z., Fong, N., Erickson, B., Bentley, D. L., and Struhl, K. (2022). Nucleotide-level linkage of transcriptional elongation and polyadenylation. *Elife*, 11.

Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W., and Robinson, M. D. (2021). Doublet identification in single-cell sequencing data using scDblFinder. *F1000Res.*, 10:979.

Ghosh, T., Peterson, B., Tomasevic, N., and Peculis, B. A. (2004). Xenopus U8 snoRNA binding protein is a conserved nuclear decapping enzyme. *Mol. Cell*, 13(6):817–828.

Gilmore, R. B., Liu, Y., Stoddard, C. E., Chung, M. S., Carmichael, G. G., and Cotney, J. (2023). Identifying key underlying regulatory networks and predicting targets of orphan C/D box SNORD116 snoRNAs in prader-willi syndrome. *bioRxivorg*, page 2023.10.03.560773.

Giudice, J., Xia, Z., Li, W., and Cooper, T. A. (2016). Neonatal cardiac dysfunction and transcriptome changes caused by the absence of Celf1. *Sci. Rep.*, 6(1):35550.

Gleeson, J., Leger, A., Prawer, Y. D. J., Lane, T. A., Harrison, P. J., Haerty, W., and Clark, M. B. (2022). Accurate expression quantification from nanopore direct RNA sequencing with NanoCount. *Nucleic Acids Res.*, 50(4):e19.

Glinos, D. A., Garborcauskas, G., Hoffman, P., Ehsan, N., Jiang, L., Gokden, A., Dai, X., Aguet, F., Brown, K. L., Garimella, K., Bowers, T., Costello, M., Ardlie, K., Jian, R., Tucker, N. R., Ellinor, P. T., Harrington, E. D., Tang, H., Snyder, M., Juul, S., Mohammadi, P., MacArthur, D. G., Lappalainen, T., and Cummings, B. (2021). Transcriptome variation in human tissues revealed by long-read sequencing. *bioRxiv*, page 2021.01.22.427687.

Glover-Cutter, K., Kim, S., Espinosa, J., and Bentley, D. L. (2008). RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat. Struct. Mol. Biol.*, 15(1):71–78.

Goes, F. S., Collado-Torres, L., Zandi, P. P., Huuki-Myers, L., Tao, R., Jaffe, A. E., Pertea, G., Shin, J. H., Weinberger, D. R., Kleinman, J. E., and Hyde, T. M. (2025). Large-scale transcriptomic analyses of major depressive disorder reveal convergent dysregulation of synaptic pathways in excitatory neurons. *Nat. Commun.*, 16(1):3981.

Goetz, A. E. and Wilkinson, M. (2017). Stress and the nonsense-mediated RNA decay pathway. *Cell. Mol. Life Sci.*, 74(19):3509–3531.

Gonatopoulos-Pournatzis, T. and Blencowe, B. J. (2020). Microexons: at the nexus of nervous system development, behaviour and autism spectrum disorder. *Curr. Opin. Genet. Dev.*, 65:22–33.

González-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.*, 14(7):R70.

Good, D. J. and Kocher, M. A. (2017). Phylogenetic analysis of the SNORD116 locus. *Genes (Basel)*, 8(12):358.

Grudzien-Nogalska, E., Wu, Y., Jiao, X., Cui, H., Mateyak, M. K., Hart, R. P., Tong, L., and Kiledjian, M. (2019). Structural and mechanistic basis of mammalian Nudt12 RNA deNADding. *Nat. Chem. Biol.*, 15(6):575–582.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nat. Methods*, 11(6):637–640.

GTEx Consortium (2020). The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330.

Guibert, M., Marty-Capelle, H., Robert, A., Charpentier, B., and Labialle, S. (2024). Coordinated evolution of the SNORD115 and SNORD116 tandem repeats at the imprinted Prader–Willi/angelman locus. *NAR Mol. Med.*, 1(1):ugad003.

Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A. B., Sloan, S. A., Luo, W., Fedrigo, O., Ross, M. E., and Tilgner, H. U. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.*

Guttikonda, S. R., Sikkema, L., Tchieu, J., Saurat, N., Walsh, R. M., Harschnitz, O., Ciceri, G., Sneeboer, M., Mazutis, L., Setty, M., Zumbo, P., Betel, D., de Witte, L. D., Pe'er, D., and Studer, L. (2021). Fully defined human pluripotent stem cell-derived microglia and tri-culture system model C3 production in alzheimer's disease. *Nat. Neurosci.*, 24(3):343–354.

Haberman, N., Digby, H., Faraway, R., Cheung, R., Jobbins, A. M., Parr, C., Yasuzawa, K., Kasukawa, T., Yip, C. W., Kato, M., Takahashi, H., Carninci, P., Vernia, S., Ule, J., Sibley, C. R., Martinez-Sanchez, A., and Lenhard, B. (2023). Abundant capped RNAs are derived from mRNA cleavage at 3'UTR G-quadruplexes. *bioRxiv*, page 2023.04.27.538568.

Hadar, S., Meller, A., Saida, N., and Shalgi, R. (2022). Stress-induced transcriptional readthrough into neighboring genes is linked to intron retention. *iScience*, 25(12).

Hafemeister, C. and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, 20(1):296.

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, 36(5):421–427.

Hall, N. A. L., Carlyle, B. C., Haerty, W., and Tunbridge, E. M. (2021a). Roadblock: improved annotations do not necessarily translate into new functional insights. *Genome Biol.*, 22(1):320.

Hall, N. A. L., Husain, S. M., Lee, H., and Tunbridge, E. M. (2021b). Chapter fourteen - long read transcript profiling of ion channel splice isoforms. In Minor, D. L. and Colecraft, H. M., editors, *Methods in Enzymology*, volume 654, pages 345–364. Academic Press.

Hallgren, J., Tsirigos, K. D., Pedersen, M. D., Armenteros, J. J. A., Marcatili, P., Nielsen, H., Krogh, A., and Winther, O. (2022). DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv*, page 2022.04.08.487609.

Hansel-Frose, A. F. F., Allmer, J., Friedrichs, M., Dos Santos, H. G., Dallagiovanna, B., and Spangenberg, L. (2024). Alternative polyadenylation and dynamic 3' UTR length is associated with polysome recruitment throughout the cardiomyogenic differentiation of hESCs. *Front. Mol. Biosci.*, 11:1336336.

Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., and Satija, R. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.*, 42(2):293–304.

Haque, A., Engel, J., Teichmann, S. A., and Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.*, 9(1):75.

Hardwick, S. A., Chen, W. Y., Wong, T., Deveson, I. W., Blackburn, J., Andersen, S. B., Nielsen, L. K., Mattick, J. S., and Mercer, T. R. (2016). Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods*, 13(9):792–798.

Hardwick, S. A., Hu, W., Joglekar, A., Fan, L., Collier, P. G., Foord, C., Balacco, J., Lanjewar, S., Sampson, M. M., Koopmans, F., Prjibelski, A. D., Mikheenko, A., Belchikov, N., Jarroux, J., Lucas, A. B., Palkovits, M., Luo, W., Milner, T. A., Ndhlovu, L. C., Smit, A. B., Trojanowski, J. Q., Lee, V. M. Y., Fedrigo, O., Sloan, S. A., Tombácz, D., Ross, M. E., Jarvis, E., Boldogkői, Z., Gan, L., and Tilgner, H. U. (2022). Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. *Nat. Biotechnol.*

Harrison, P. J. (1997). Schizophrenia: a disorder of neurodevelopment? *Curr. Opin. Neurobiol.*, 7(2):285–289.

Harrison, P. J., Geddes, J. R., and Tunbridge, E. M. (2018). The emerging neurobiology of bipolar disorder. *Trends Neurosci.*, 41(1):18–30.

Harrison, P. J., Husain, S. M., Lee, H., Los Angeles, A. D., Colbourne, L., Mould, A., Hall, N. A. L., Haerty, W., and Tunbridge, E. M. (2022). CACNA1C (CaV1.2) and other L-type calcium channels in the pathophysiology and treatment of psychiatric disorders: Advances from functional genomics and pharmacoepidemiology. *Neuropharmacology*, 220(109262):109262.

Harrison, P. J. and Weinberger, D. R. (2005). Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. *Mol. Psychiatry*, 10(1):40–68; image 5.

Hasel, P., Rose, I. V. L., Sadick, J. S., Kim, R. D., and Liddelow, S. A. (2021). Neuroinflammatory astrocyte subtypes in the mouse brain. *Nat. Neurosci.*, 24(10):1475–1487.

Hayano, T., Yanagida, M., Yamauchi, Y., Shinkawa, T., Isobe, T., and Takahashi, N. (2003). Proteomic analysis of human Nop56p-associated pre-ribosomal ribonucleoprotein complexes. possible link between Nop56p and the nucleolar protein treacle responsible for treacher collins syndrome. *J. Biol. Chem.*, 278(36):34309–34319.

Hayes, J. F., Lundin, A., Wicks, S., Lewis, G., Wong, I. C. K., Osborn, D. P. J., and Dalman, C. (2019). Association of hydroxylmethyl glutaryl coenzyme a reductase inhibitors, L-type calcium channel antagonists, and biguanides with rates of psychiatric hospitalization and self-harm in individuals with serious mental illness. *JAMA Psychiatry*, 76(4):382–390.

He, F. and Jacobson, A. (2015). Nonsense-mediated mRNA decay: Degradation of defective transcripts is only part of the story. *Annu. Rev. Genet.*, 49:339–366.

Head, S. A., Hernandez-Alias, X., Yang, J.-S., Ciampi, L., Beltran-Sastre, V., Torres-Méndez, A., Irimia, M., Schaefer, M. H., and Serrano, L. (2021). Silencing of SRRM4 suppresses microexon inclusion and promotes tumor growth across cancers. *PLoS Biol.*, 19(2):e3001138.

Hebras, J., Marty, V., Personnaz, J., Mercier, P., Krogh, N., Nielsen, H., Aguirrebengoa, M., Seitz, H., Pradere, J.-P., Guiard, B. P., and Cavaille, J. (2020). Reassessment of the involvement of Snord115 in the serotonin 2c receptor pathway in a genetically relevant mouse model. *Elife*, 9.

Helwak, A., Turowski, T., Spanos, C., and Tollervey, D. (2024). Roles of SNORD115 and SNORD116 ncRNA clusters during neuronal differentiation. *Nat. Commun.*, 15(1):10427.

Herrmann, C. J., Schmidt, R., Kanitz, A., Artimo, P., Gruber, A. J., and Zavolan, M. (2020). PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic Acids Res.*, 48(D1):D174–D179.

Herzel, L., Ottoz, D. S. M., Alpert, T., and Neugebauer, K. M. (2017). Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nat. Rev. Mol. Cell Biol.*, 18(10):637–650.

Herzel, L., Straube, K., and Neugebauer, K. M. (2018). Long-read sequencing of nascent RNA reveals coupling among RNA processing events. *Genome Res.*, 28(7):1008–1019.

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578.

Hilgers, V., Perry, M. W., Hendrix, D., Stark, A., Levine, M., and Haley, B. (2011). Neural-specific elongation of 3' UTRs during drosophila development. *Proc. Natl. Acad. Sci. U. S. A.*, 108(38):15864–15869.

Hirose, Y. and Manley, J. L. (1998). RNA polymerase II is an essential mRNA polyadenylation factor. *Nature*, 395(6697):93–96.

Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., Close, J. L., Long, B., Johansen, N., Penn, O., Yao, Z., Eggermont, J., Höllt, T., Levi, B. P., Shehata, S. I., Aevermann, B., Beller, A., Bertagnolli, D., Brouner, K., Casper, T., Cobbs, C., Dalley, R., Dee, N., Ding, S.-L., Ellenbogen, R. G., Fong, O., Garren, E., Goldy, J., Gwinn, R. P., Hirschstein, D., Keene, C. D., Keshk, M., Ko, A. L., Lathia, K., Mahfouz, A., Maltzer, Z., McGraw, M., Nguyen, T. N., Nyhus, J., Ojemann, J. G., Oldre, A., Parry, S., Reynolds, S., Rimorin, C., Shapovalova, N. V., Somasundaram, S., Szafer, A., Thomsen, E. R., Tieu, M., Quon, G., Scheuermann, R. H., Yuste, R., Sunkin, S. M., Lelieveldt, B., Feng, D., Ng, L., Bernard, A., Hawrylycz, M., Phillips, J. W., Tasic, B., Zeng, H., Jones, A. R., Koch, C., and Lein, E. S. (2019). Conserved cell types with divergent features in human versus mouse cortex. *Nature*, 573(7772):61–68.

Hogg, J. R. and Goff, S. P. (2010). Upf1 senses 3' UTR length to potentiate mRNA decay. *Cell*, 143(3):379–389.

Holmes, T. L., Chabronova, A., Denning, C., James, V., Peffers, M. J., and Smith, J. G. W. (2025). Footprints in the sno: investigating the cellular and molecular mechanisms of SNORD116. *Open Biol.*, 15(3):240371.

Holmgren, A., Hansson, L., Bjerkaas-Kjeldal, K., Impellizzeri, A. A. R., Gilfillan, G. D., Djurovic, S., and Hughes, T. (2022). Mapping the expression of an ANK3 isoform associated with bipolar disorder in the human brain. *Transl. Psychiatry*, 12(1):45.

Houseley, J. and Tollervey, D. (2010). Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One*, 5(8):e12271.

Howard, D. M., Adams, M. J., Clarke, T.-K., Hafferty, J. D., Gibson, J., Shirali, M., Coleman, J. R. I., Hagenaars, S. P., Ward, J., Wigmore, E. M., Alloza, C., Shen, X., Barbu, M. C., Xu, E. Y., Whalley, H. C., Marioni, R. E., Porteous, D. J., Davies, G., Deary, I. J., Hemani, G., Berger, K., Teismann, H., Rawal, R., Arolt, V., Baune, B. T., Dannlowski, U., Domschke, K., Tian, C., Hinds, D. A., 23andMe Research Team, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Trzaskowski, M., Byrne, E. M., Ripke, S., Smith, D. J., Sullivan, P. F., Wray, N. R., Breen, G., Lewis, C. M., and McIntosh, A. M. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.*, 22(3):343–352.

Hua, Y., Sahashi, K., Hung, G., Rigo, F., Passini, M. A., Bennett, C. F., and Krainer, A. R. (2010). Antisense correction of SMN2 splicing in the CNS rescues necrosis in a type III SMA mouse model. *Genes Dev.*, 24(15):1634–1644.

Huang, C., Shi, J., Guo, Y., Huang, W., Huang, S., Ming, S., Wu, X., Zhang, R., Ding, J., Zhao, W., Jia, J., Huang, X., Xiang, A. P., Shi, Y., and Yao, C. (2017). A snoRNA modulates mRNA 3' end processing and regulates the expression of a subset of mRNAs. *Nucleic Acids Res.*, 45(15):8647–8660.

Huang, W.-K., Wong, S. Z. H., Pather, S. R., Nguyen, P. T. T., Zhang, F., Zhang, D. Y., Zhang, Z., Lu, L., Fang, W., Chen, L., Fernandes, A., Su, Y., Song, H., and Ming, G.-L. (2021). Generation of hypothalamic arcuate organoids from human induced pluripotent stem cells. *Cell Stem Cell*, 28(9):1657–1670.e10.

Hughes, T., Hansson, L., Sønderby, I. E., Athanasiu, L., Zuber, V., Tesli, M., Song, J., Hultman, C. M., Bergen, S. E., Landén, M., Melle, I., Andreassen, O. A., and Djurovic, S. (2016). A loss-of-function variant in a minor isoform of ANK3 protects against bipolar disorder and schizophrenia. *Biol. Psychiatry*, 80(4):323–330.

Ip, J. Y., Sone, M., Nashiki, C., Pan, Q., Kitaichi, K., Yanaka, K., Abe, T., Takao, K., Miyakawa, T., Blencowe, B. J., and Nakagawa, S. (2016). Gomafu lncRNA knockout mice exhibit mild hyperactivity with enhanced responsiveness to the psychostimulant methamphetamine. *Sci. Rep.*, 6(1):27204.

Irimia, M. and Roy, S. W. (2008). Evolutionary convergence on highly-conserved 3' intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS Genet.*, 4(8):e1000148.

Irimia, M., Weatheritt, R. J., Ellis, J. D., Parikshak, N. N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., Barrios-Rodiles, M., Sternberg, M. J. E., Cordes, S. P., Roth, F. P., Wrana, J. L., Geschwind, D. H., and Blencowe, B. J. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*, 159(7):1511–1523.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, 11(2):163–166.

Jacob, A. G. and Smith, C. W. J. (2017). Intron retention as a component of regulated gene expression programs. *Hum. Genet.*, 136(9):1043–1057.

Jaffe, A. E., Straub, R. E., Shin, J. H., Tao, R., Gao, Y., Collado-Torres, L., Kam-Thong, T., Xi, H. S., Quan, J., Chen, Q., Colantuoni, C., Ulrich, W. S., Maher, B. J., Deep-Soboslay, A., BrainSeq Consortium, Cross, A. J., Brandon, N. J., Leek, J. T., Hyde, T. M., Kleinman, J. E., and Weinberger, D. R. (2018). Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.*, 21(8):1117–1125.

Jain, M., Tyson, J. R., Loose, M., Ip, C. L. C., Eccles, D. A., O'Grady, J., Malla, S., Leggett, R. M., Wallerman, O., Jansen, H. J., Zalunin, V., Birney, E., Brown, B. L., Snutch, T. P., Olsen, H. E., and MinION Analysis and Reference Consortium (2017). MinION analysis and reference consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Res.*, 6(760):760.

Jangi, M., Boutz, P. L., Paul, P., and Sharp, P. A. (2014). Rbfox2 controls autoregulation in RNA-binding protein networks. *Genes Dev.*, 28(6):637–651.

Jaramillo Oquendo, C., Wai, H. A., Rich, W. I., Bunyan, D. J., Thomas, N. S., Hunt, D., Lord, J., Douglas, A. G. L., and Baralle, D. (2024). Identification of diagnostic candidates in mendelian disorders using an RNA sequencing-centric approach. *Genome Med.*, 16(1):110.

Jenal, M., Elkon, R., Loayza-Puch, F., van Haaften, G., Kühn, U., Menzies, F. M., Oude Vrielink, J. A. F., Bos, A. J., Drost, J., Rooijers, K., Rubinsztein, D. C., and Agami, R. (2012). The poly(a)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell*, 149(3):538–553.

Jenkins, A. K., Paterson, C., Wang, Y., Hyde, T. M., Kleinman, J. E., and Law, A. J. (2016). Neurexin 1 (NRXN1) splice isoform expression during human neocortical development and aging. *Mol. Psychiatry*, 21(5):701–706.

Ji, X., Wan, J., Vishnu, M., Xing, Y., and Liebhaber, S. A. (2013). αcp poly(C) binding proteins act as global regulators of alternative polyadenylation. *Mol. Cell. Biol.*, 33(13):2560–2573.

Ji, Z., Lee, J. Y., Pan, Z., Jiang, B., and Tian, B. (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. U. S. A.*, 106(17):7028–7033.

Jiang, R., Sun, T., Song, D., and Li, J. J. (2022). Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biol.*, 23(1):31.

Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., and Inoue, K. (2003). A vertebrate RNA-binding protein fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *EMBO J.*, 22(4):905–912.

Jin, Z., Huang, W., Shen, N., Li, J., Wang, X., Dong, J., Park, P. J., and Xi, R. (2022). Single-cell gene fusion detection by scFusion. *Nat. Commun.*, 13(1):1084.

Joglekar, A., Hu, W., Zhang, B., Narykov, O., Diekhans, M., Balacco, J., Ndhlovu, L. C., Milner, T. A., Fedrigo, O., Jarvis, E. D., Sheynkman, G., Korkin, D., Elizabeth Ross, M., and Tilgner, H. U. (2023). Single-cell long-read mRNA isoform regulation is pervasive across mammalian brain regions, cell types, and development. *bioRxiv*, page 2023.04.02.535281.

Joglekar, A., Hu, W., Zhang, B., Narykov, O., Diekhans, M., Marrocco, J., Balacco, J., Ndhlovu, L. C., Milner, T. A., Fedrigo, O., Jarvis, E. D., Sheynkman, G., Korkin, D., Ross, M. E., and Tilgner, H. U. (2024). Single-cell long-read sequencing-based mapping reveals specialized splicing patterns in developing and adult mouse and human brain. *Nat. Neurosci.*, 27(6):1051–1063.

Joglekar, A., Prjibelski, A., Mahfouz, A., Collier, P., Lin, S., Schlusche, A. K., Marrocco, J., Williams, S. R., Haase, B., Hayes, A., Chew, J. G., Weisenfeld, N. I., Wong, M. Y., Stein, A. N., Hardwick, S. A., Hunt, T., Wang, Q., Dieterich, C., Bent, Z., Fedrigo, O., Sloan, S. A., Risso, D., Jarvis, E. D., Flicek, P., Luo, W., Pitt, G. S., Frankish, A., Smit, A. B., Ross, M. E., and Tilgner, H. U. (2021). A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat. Commun.*, 12(1):463.

Johnson, B. B., Cosson, M.-V., Tsansizi, L. I., Holmes, T. L., Gilmore, T., Hampton, K., Song, O.-R., Vo, N. T. N., Nasir, A., Chabronova, A., Denning, C., Peffers, M. J., Merry, C. L. R., Whitelock, J., Troeberg, L., Rushworth, S. A., Bernardo, A. S., and Smith, J. G. W. (2024). Perlecan (HSPG2) promotes structural, contractile, and metabolic development of human cardiomyocytes. *Cell Rep.*, 43(1):113668.

Jourdain, A. A., Begg, B. E., Mick, E., Shah, H., Calvo, S. E., Skinner, O. S., Sharma, R., Blue, S. M., Yeo, G. W., Burge, C. B., and Mootha, V. K. (2021). Loss of LUC7L2 and U1 snRNP subunits shifts energy metabolism from glycolysis to OXPHOS. *Mol. Cell*, 81(9):1905–1919.e12.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Jurica, M. S. and Roybal, G. A. (2013). RNA splicing. In Lennarz, W. J. and Lane, M. D., editors, *Encyclopedia of Biological Chemistry (Second Edition)*, pages 185–190. Academic Press, Waltham.

Kabza, M., Ritter, A., Byrne, A., Sereti, K., Le, D., Stephenson, W., and Sterne-Weiler, T. (2024). Accurate long-read transcript discovery and quantification at single-cell, pseudo-bulk and bulk resolution with isosceles. *Nat. Commun.*, 15(1):7316.

Kafali, H. Y., Bildik, T., Bora, E., Yuncu, Z., and Erermis, H. S. (2019). Distinguishing prodromal stage of bipolar disorder and early onset schizophrenia spectrum disorders during adolescence. *Psychiatry Res.*, 275:315–325.

Kalsotra, A., Xiao, X., Ward, A. J., Castle, J. C., Johnson, J. M., Burge, C. B., and Cooper, T. A. (2008). A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *Proc. Natl. Acad. Sci. U. S. A.*, 105(51):20333–20338.

Kang, B., Yang, Y., Hu, K., Ruan, X., Liu, Y.-L., Lee, P., Lee, J., Wang, J., and Zhang, X. (2023). Infernape uncovers cell type-specific and spatially resolved alternative polyadenylation in the brain. *Genome Res.*, 33(10):1774–1787.

Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., and Gao, G. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*, 45(W1):W12–W16.

Karousis, E. D., Gypas, F., Zavolan, M., and Mühlemann, O. (2021). Nanopore sequencing reveals endogenous NMD-targeted isoforms in human cells. *Genome Biol.*, 22(1):223.

Kedan, A., Zauber, H., Wang, M.-R., Zhu, Q., Fang, L., Chen, W., and Selbach, M. (2025). An integrated landscape of mRNA and protein isoforms. *bioRxiv*.

Keil, N., Monzó, C., McIntyre, L., and Conesa, A. (2024). SQANTI-reads: a tool for the quality assessment of long read data in multi-sample lrRNA-seq experiments. *bioRxivorg*, page 2024.08.23.609463.

Kelly, S. T. (2023). leiden: R implementation of the leiden algorithm.

Khajuria, D. K., Nowak, I., Leung, M., Karuppagounder, V., Imamura, Y., Norbury, C. C., Kamal, F., and Elbarbary, R. A. (2023). Transcript shortening via alternative polyadenylation promotes gene expression during fracture healing. *Bone Res.*, 11(1):5.

Kiltschewskij, D. J., Harrison, P. F., Fitzsimmons, C., Beilharz, T. H., and Cairns, M. J. (2023). Extension of mRNA poly(a) tails and 3'UTRs during neuronal differentiation exhibits variable association with post-transcriptional dynamics. *Nucleic Acids Res.*, 51(15):8181–8198.

Kishore, S., Khanna, A., Zhang, Z., Hui, J., Balwierz, P. J., Stefan, M., Beach, C., Nicholls, R. D., Zavolan, M., and Stamm, S. (2010). The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. *Hum. Mol. Genet.*, 19(7):1153–1164.

Kiss, T. (2001). Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J.*, 20(14):3617–3622.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201.

Kocher, M. A., Huang, F. W., Le, E., and Good, D. J. (2021). Snord116 post-transcriptionally increases Nhlh2 mRNA stability: Implications for human prader-willi syndrome. *Hum. Mol. Genet.*, 30(12):1101–1110.

Konermann, S., Lotfy, P., Brideau, N. J., Oki, J., Shokhirev, M. N., and Hsu, P. D. (2018). Transcriptome engineering with RNA-targeting type VI-D CRISPR effectors. *Cell*, 173(3):665–676.e14.

Kornblihtt, A. R., Schor, I. E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M. J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.*, 14(3):153–165.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. (2019a). Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, 16(12):1289–1296.

Korsunsky, I., Nathan, A., Millard, N., and Raychaudhuri, S. (2019b). Presto scales wilcoxon and auROC analyses to millions of observations. *bioRxiv*, page 653253.

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.*, 20(1):278.

Kralovicova, J., Hwang, G., Asplund, A. C., Churbanov, A., Smith, C. I. E., and Vorechovsky, I. (2011). Compensatory signals associated with the activation of human GC 5' splice sites. *Nucleic Acids Res.*, 39(16):7077–7091.

Krueger, F. (2018). TrimGalore: A wrapper around cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data.

Kuo, R. I., Cheng, Y., Zhang, R., Brown, J. W. S., Smith, J., Archibald, A. L., and Burt, D. W. (2020). Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics*, 21(1):751.

Kwon, S.-K., Sando, 3rd, R., Lewis, T. L., Hirabayashi, Y., Maximov, A., and Polleux, F. (2016). LKB1 regulates mitochondria-dependent presynaptic calcium clearance and neurotransmitter release properties at excitatory synapses along cortical axons. *PLoS Biol.*, 14(7):e1002516.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., and Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, 560(7719):494–498.

Ladd, A. N. (2016). New insights into the role of RNA-binding proteins in the regulation of heart development. *Int. Rev. Cell Mol. Biol.*, 324:125–185.

Lagarde, J., Uszczynska-Ratajczak, B., Carbonell, S., Pérez-Lluch, S., Abad, A., Davis, C., Gingeras, T. R., Frankish, A., Harrow, J., Guigo, R., and Johnson, R. (2017). High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.*, 49(12):1731–1740.

Lake, B. B., Ai, R., Kaeser, G. E., Salathia, N. S., Yung, Y. C., Liu, R., Wildberg, A., Gao, D., Fung, H.-L., Chen, S., Vijayaraghavan, R., Wong, J., Chen, A., Sheng, X., Kaper, F., Shen, R., Ronaghi, M., Fan, J.-B., Wang, W., Chun, J., and Zhang, K. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, 352(6293):1586–1590.

Lamichhane Rajan, Daubner Gerrit M., Thomas-Crusells Judith, Auweter Sigrid D., Manatschal Cristina, Austin Keyunna S., Valniuk Oksana, Allain Frédéric H.-T., and Rueda David (2010). RNA looping by PTB: Evidence using FRET and NMR spectroscopy for a role in splicing repression. *Proceedings of the National Academy of Sciences*, 107(9):4105–4110.

Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C., and Brenner, S. E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, 446(7138):926–929.

Law, A. J., Kleinman, J. E., Weinberger, D. R., and Weickert, C. S. (2007). Disease-associated intronic variants in the ErbB4 gene are related to altered ErbB4 splice-variant expression in the brain in schizophrenia. *Hum. Mol. Genet.*, 16(2):129–141.

Lebrigand, K., Magnone, V., Barbry, P., and Waldmann, R. (2020). High throughput error corrected nanopore single cell transcriptome sequencing. *Nat. Commun.*, 11(1):4025.

Lee, J.-S., Lamarche-Vane, N., and Richard, S. (2021). Microexon alternative splicing of small GTPase regulators: Implication in central nervous system diseases. *Wiley Interdiscip. Rev. RNA*, page e1678.

Legendre, M. and Gautheret, D. (2003). Sequence determinants in human polyadenylation site selection. *BMC Genomics*, 4(1):7.

Leng, K., Rose, I. V. L., Kim, H., Xia, W., Romero-Fernandez, W., Rooney, B., Koontz, M., Li, E., Ao, Y., Wang, S., Krawczyk, M., Tcw, J., Goate, A., Zhang, Y., Ullian, E. M., Sofroniew, M. V., Fancy, S. P. J., Schrag, M. S., Lippmann, E. S., and Kampmann, M. (2022). CRISPRi screens in human iPSC-derived astrocytes elucidate regulators of distinct inflammatory reactive states. *Nat. Neurosci.*, 25(11):1528–1542.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100.

Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., Susztak, K., Reilly, M. P., Hu, G., and Li, M. (2020). Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat. Commun.*, 11(1):2338.

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., and Pritchard, J. K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.*, 50(1):151–158.

Li, Y. I., Sanchez-Pulido, L., Haerty, W., and Ponting, C. P. (2015). RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Res.*, 25(1):1–13.

Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S., and Mayr, C. (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.*, 27(21):2380–2396.

Lipscombe, D., Andrade, A., and Allen, S. E. (2013). Alternative splicing: functional diversity among voltage-gated calcium channels and behavioral consequences. *Biochim. Biophys. Acta*, 1828(7):1522–1529.

Liu, B. (2025). Mapping snoRNA targets transcriptome-wide with snoKARR-seq. *ACS Chem. Biol.*, 20(2):242–244.

Liu, B., Liu, Y., Li, J., Guo, H., Zang, T., and Wang, Y. (2019). deSALT: fast and accurate long transcriptomic read alignment with de bruijn graph-based index. *Genome Biol.*, 20(1):274.

Lopes, I., Altab, G., Raina, P., and de Magalhães, J. P. (2021). Gene size matters: An analysis of gene length in the human genome. *Front. Genet.*, 12:559998.

Loving, R., Sullivan, D. K., Reese, F., Rebboah, E., Sakr, J., Rezaie, N., Liang, H. Y., Filimban, G., Kawauchi, S., Oakes, C., Trout, D., Williams, B. A., MacGregor, G., Wold, B. J., Mortazavi, A., and Pachter, L. (2024a). Long-read sequencing transcriptome quantification with lr-kallisto. *bioRxiv*, page 2024.07.19.604364.

Loving, R. K., Sullivan, D. K., Reese, F., Rebboah, E., Sakr, J., Rezaie, N., Liang, H. Y., Filimban, G., Kawauchi, S., Oakes, C., Trout, D., Williams, B. A., MacGregor, G., Wold, B. J., Mortazavi, A., and Pachter, L. (2024b). Long-read sequencing transcriptome quantification with lr-kallisto. *bioRxivorg*.

Lowther, C., Speevak, M., Armour, C. M., Goh, E. S., Graham, G. E., Li, C., Zeesman, S., Nowaczyk, M. J. M., Schultz, L.-A., Morra, A., Nicolson, R., Bikangaga, P., Samdup, D., Zaazou, M., Boyd, K., Jung, J. H., Siu, V., Rajguru, M., Goobie, S., Tarnopolsky, M. A., Prasad, C., Dick, P. T., Hussain, A. S., Walinga, M., Reijenga, R. G., Gazzellone, M., Lionel, A. C., Marshall, C. R., Scherer, S. W., Stavropoulos, D. J., McCready, E., and Bassett, A. S. (2017). Molecular characterization of NRXN1 deletions from 19,263 clinical microarray cases identifies exons important for neurodevelopmental disease expression. *Genet. Med.*, 19(1):53–61.

Luco, R. F., Pan, Q., Tominaga, K., Blencowe, B. J., Pereira-Smith, O. M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. *Science*, 327(5968):996–1000.

Lykke-Andersen, S., Ardal, B. K., Hollensen, A. K., Damgaard, C. K., and Jensen, T. H. (2018). Box C/D snoRNP autoregulation by a cis-acting snoRNA in the NOP56 pre-mRNA. *Mol. Cell*, 72(1):99–111.e5.

Makeyev, E. V., Zhang, J., Carrasco, M. A., and Maniatis, T. (2007). The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol. Cell*, 27(3):435–448.

Mancuso, R., Fattorelli, N., Martinez-Muriana, A., Davis, E., Wolfs, L., Van Den Daele, J., Geric, I., Premereur, J., Polanco, P., Bijnens, B., Preman, P., Serneels, L., Poovathingal, S., Balusu, S., Verfaillie, C., Fiers, M., and De Strooper, B. (2024). Xenografted human microglia display diverse transcriptomic states in response to alzheimer's disease-related amyloid-β pathology. *Nat. Neurosci.*, 27(5):886–900.

Maquat, L. E. (1995). When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells. *RNA*, 1(5):453–465.

Marquez, Y., Höpfler, M., Ayatollahi, Z., Barta, A., and Kalyna, M. (2015). Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity. *Genome Res.*, 25(7):995–1007.

Matera, A. G. and Wang, Z. (2014). A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.*, 15(2):108–121.

Mathur, M., Kim, C. M., Munro, S. A., Rudina, S. S., Sawyer, E. M., and Smolke, C. D. (2019). Programmable mutually exclusive alternative splicing for generating RNA and protein diversity. *Nat. Commun.*, 10(1):2673.

Matsudaira, T. and Prinz, M. (2022). Life and death of microglia: Mechanisms governing microglial states and fates. *Immunol. Lett.*, 245:51–60.

Mattick, J. S., Amaral, P. P., Carninci, P., Carpenter, S., Chang, H. Y., Chen, L.-L., Chen, R., Dean, C., Dinger, M. E., Fitzgerald, K. A., Gingeras, T. R., Guttman, M., Hirose, T., Huarte, M., Johnson, R., Kanduri, C., Kapranov, P., Lawrence, J. B., Lee, J. T., Mendell, J. T., Mercer, T. R., Moore, K. J., Nakagawa, S., Rinn, J. L., Spector, D. L., Ulitsky, I., Wan, Y., Wilusz, J. E., and Wu, M. (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat. Rev. Mol. Cell Biol.*, 24(6):430–447.

Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., Catallini, 2nd, J. L., Tran, M. N., Besich, Z., Tippani, M., Chew, J., Yin, Y., Kleinman, J. E., Hyde, T. M., Rao, N., Hicks, S. C., Martinowich, K., and Jaffe, A. E. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.*, 24(3):425–436.

Mazin, P., Xiong, J., Liu, X., Yan, Z., Zhang, X., Li, M., He, L., Somel, M., Yuan, Y., Phoebe Chen, Y.-P., Li, N., Hu, Y., Fu, N., Ning, Z., Zeng, R., Yang, H., Chen, W., Gelfand, M., and Khaitovich, P. (2013). Widespread splicing changes in human brain development and aging. *Mol. Syst. Biol.*, 9(1):633.

McManus, C. J. and Graveley, B. R. (2011). RNA structure and the mechanisms of alternative splicing. *Curr. Opin. Genet. Dev.*, 21(4):373–379.

Mehlferber, M. M., Jeffery, E. D., Saquing, J., Jordan, B. T., Sheynkman, L., Murali, M., Genet, G., Acharya, B. R., Hirschi, K. K., and Sheynkman, G. M. (2022). Characterization of protein isoform diversity in human umbilical vein endothelial cells via long-read proteogenomics. *RNA Biol.*, 19(1):1228–1243.

Meier, U. T. (2017). RNA modification in cajal bodies. *RNA Biol.*, 14(6):693–700.

Meldolesi, J. (2020). Alternative splicing by NOVA factors: From gene expression to cell physiology and pathology. *Int. J. Mol. Sci.*, 21(11).

Melsted, P., Booeshaghi, A. S., Liu, L., Gao, F., Lu, L., Min, K. H. J., da Veiga Beltrame, E., Hjörleifsson, K. E., Gehring, J., and Pachter, L. (2021). Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat. Biotechnol.*, 39(7):813–818.

Melville, J. (2022). uwot: The uniform manifold approximation and projection (UMAP) method for dimensionality reduction.

Mercer, T. R., Clark, M. B., Crawford, J., Brunck, M. E., Gerhardt, D. J., Taft, R. J., Nielsen, L. K., Dinger, M. E., and Mattick, J. S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.*, 9(5):989–1009.

Mercer, T. R., Gerhardt, D. J., Dinger, M. E., Crawford, J., Trapnell, C., Jeddeloh, J. A., Mattick, J. S., and Rinn, J. L. (2011). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.*, 30(1):99–104.

Merico, D., Roifman, M., Braunschweig, U., Yuen, R. K. C., Alexandrova, R., Bates, A., Reid, B., Nalpathamkalam, T., Wang, Z., Thiruvahindrapuram, B., Gray, P., Kakakios, A., Peake, J., Hogarth, S., Manson, D., Buncic, R., Pereira, S. L., Herbrick, J.-A., Blencowe, B. J., Roifman, C. M., and Scherer, S. W. (2015). Compound heterozygous mutations in the noncoding RNU4ATAC cause roifman syndrome by disrupting minor intron splicing. *Nat. Commun.*, 6(1):8718.

Michielsen, L., Prjibelski, A., Foord, C., Hu, W., Jarroux, J., Hsu, J., Tomescu, A., Hajirasouliha, I., and Tilgner, H. (2025). Spatial isoform sequencing at sub-micrometer single-cell resolution reveals novel patterns of spatial isoform variability in brain cell types. *bioRxivorg*, page 2025.06.25.661563.

Mikheenko, A., Prjibelski, A. D., Joglekar, A., and Tilgner, H. U. (2022). Sequencing of individual barcoded cDNAs using pacific biosciences and oxford nanopore technologies reveals platform-specific error patterns. *Genome Res.*, 32(4):726–737.

Miller, R. M., Jordan, B. T., Mehlferber, M. M., Jeffery, E. D., Chatzipantsiou, C., Kaur, S., Millikin, R. J., Dai, Y., Tiberi, S., Castaldi, P. J., Shortreed, M. R., Luckey, C. J., Conesa, A., Smith, L. M., Deslattes Mays, A., and Sheynkman, G. M. (2022). Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biol.*, 23(1):69.

Mincarelli, L., Uzun, V., Wright, D., Scoones, A., Rushworth, S. A., Haerty, W., and Macaulay, I. C. (2023). Single-cell gene and isoform expression analysis reveals signatures of ageing in haematopoietic stem and progenitor cells. *Commun. Biol.*, 6(1):558.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., and Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.*, 49(D1):D412–D419.

Mitschka, S. and Mayr, C. (2022). Context-specific regulation and function of mRNA alternative polyadenylation. *Nat. Rev. Mol. Cell Biol.*, 23(12):779–796.

Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J. O., and Lai, E. C. (2013). Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Res.*, 23(5):812–825.

Montañés-Agudo, P., Casini, S., Aufiero, S., Ernault, A. C., van der Made, I., Pinto, Y. M., Remme, C. A., and Creemers, E. E. (2021). Inhibition of minor intron splicing reduces na+ and Ca2+ channel expression and function in cardiomyocytes. *J. Cell Sci.*

Moraes, K. C. M., Wilusz, C. J., and Wilusz, J. (2007). CUG-BP and 3'UTR sequences influence PARN-mediated deadenylation in mammalian cell extracts. *Genet. Mol. Biol.*, 30(3):646–655.

Mourelatos, Z., Abel, L., Yong, J., Kataoka, N., and Dreyfuss, G. (2001). SMN interacts with a novel family of hnRNP and spliceosomal proteins. *EMBO J.*, 20(19):5443–5452.

Mudge, J. M., Carbonell-Sala, S., Diekhans, M., Martinez, J. G., Hunt, T., Jungreis, I., Loveland, J. E., Arnan, C., Barnes, I., Bennett, R., Berry, A., Bignell, A., Cerdán-Vélez, D., Cochran, K., Cortés, L. T., Davidson, C., Donaldson, S., Dursun, C., Fatima, R., Hardy, M., Hebbar, P., Hollis, Z., James, B. T., Jiang, Y., Johnson, R., Kaur, G., Kay, M., Mangan, R. J., Maquedano, M., Gómez, L. M., Mathlouthi, N., Merritt, R., Ni, P., Palumbo, E., Perteghella, T., Pozo, F., Raj, S., Sisu, C., Steed, E., Sumathipala, D., Suner, M.-M., Uszczynska-Ratajczak, B., Wass, E., Yang, Y. T., Zhang, D., Finn, R. D., Gerstein, M., Guigó, R., Hubbard, T. J. P., Kellis, M., Kundaje, A., Paten, B., Tress, M. L., Birney, E., Martin, F. J., and Frankish, A. (2025). GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res.*, 53(D1):D966–D975.

Mulica, P., Venegas, C., Landoulsi, Z., Badanjak, K., Delcambre, S., Tziortziou, M., Hezzaz, S., Ghelfi, J., Smajic, S., Schwamborn, J., Krüger, R., Antony, P., May, P., Glaab, E., Grünewald, A., and Pereira, S. L. (2023). Comparison of two protocols for the generation of iPSC-derived human astrocytes. *Biol. Proced. Online*, 25(1):26.

Nachtigall, P. G., Kashiwabara, A. Y., and Durham, A. M. (2021). CodAn: predictive models for precise identification of coding regions in eukaryotic transcripts. *Brief. Bioinform.*, 22(3):bbaa045.

Nagasawa, C., Ogren, A., Kibiryeva, N., Marshall, J., O'Brien, J. E., Kenmochi, N., and Bittel, D. C. (2018). The role of scaRNAs in adjusting alternative mRNA splicing in heart development. *J. Cardiovasc. Dev. Dis.*, 5(2):26.

Nagasawa, C. K., Kibiryeva, N., Marshall, J., O'Brien, Jr, J. E., and Bittel, D. C. (2020). ScaRNA1 levels alter pseudouridylation in spliceosomal RNA U2 affecting alternative mRNA splicing and embryonic development. *Pediatr. Cardiol.*, 41(2):341–349.

Nagy, C., Maitra, M., Tanti, A., Suderman, M., Théroux, J.-F., Davoli, M. A., Perlman, K., Yerko, V., Wang, Y. C., Tripathy, S. J., Pavlidis, P., Mechawar, N., Ragoussis, J., and Turecki, G. (2020). Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.*, 23(6):771–781.

Nakagawa, S., Ip, J. Y., Shioi, G., Tripathi, V., Zong, X., Hirose, T., and Prasanth, K. V. (2012). Malat1 is not an essential component of nuclear speckles in mice. *RNA*, 18(8):1487–1499.

Nanni, A., Titus-McQuillan, J., Bankole, K. S., Pardo-Palacios, F., Signor, S., Vlaho, S., Moskalenko, O., Morse, A. M., Rogers, R. L., Conesa, A., and McIntyre, L. M. (2024). Nucleotide-level distance metrics to quantify alternative splicing implemented in TranD. *Nucleic Acids Res.*, 52(5):e28.

Naro, C. and Sette, C. (2013). Phosphorylation-mediated regulation of alternative splicing in cancer. *Int. J. Cell Biol.*, 2013:151839.

Nascimento, J. M., Saia-Cereda, V. M., Zuccoli, G. S., Reis-de Oliveira, G., Carregari, V. C., Smith, B. J., Rehen, S. K., and Martins-de Souza, D. (2022). Proteomic signatures of schizophrenia-sourced iPSC-derived neural cells and brain organoids are similar to patients' postmortem brains. *Cell Biosci.*, 12(1):189.

Nasif, S., Contu, L., and Mühlemann, O. (2018). Beyond quality control: The role of nonsense-mediated mRNA decay (NMD) in regulating gene expression. *Semin. Cell Dev. Biol.*, 75:78–87.

Nazim, M., Masuda, A., Rahman, M. A., Nasrin, F., Takeda, J.-I., Ohe, K., Ohkawara, B., Ito, M., and Ohno, K. (2017). Competitive regulation of alternative splicing and alternative polyadenylation by hnRNP H and CstF64 determines acetylcholinesterase isoforms. *Nucleic Acids Res.*, 45(3):1455–1468.

Nguyen, T.-M., Schreiner, D., Xiao, L., Traunmüller, L., Bornmann, C., and Scheiffele, P. (2016). An alternative splicing switch shapes neurexin repertoires in principal neurons versus interneurons in the mouse hippocampus. *Elife*, 5.

Ni, J. Z., Grate, L., Donohue, J. P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T. A., Blume, J. E., and Ares, Jr, M. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.*, 21(6):708–718.

Nimura, K., Yamamoto, M., Takeichi, M., Saga, K., Takaoka, K., Kawamura, N., Nitta, H., Nagano, H., Ishino, S., Tanaka, T., Schwartz, R. J., Aburatani, H., and Kaneda, Y. (2016). Regulation of alternative polyadenylation by Nkx2-5 and Xrn2 during mouse heart development. *Elife*, 5:e16030.

Novak, G., Kyriakis, D., Grzyb, K., Bernini, M., Rodius, S., Dittmar, G., Finkbeiner, S., and Skupin, A. (2022). Single-cell transcriptomics of human iPSC differentiation dynamics reveal a core molecular network of parkinson's disease. *Commun. Biol.*, 5(1):49.

Ochiai, H., Hayashi, T., Umeda, M., Yoshimura, M., Harada, A., Shimizu, Y., Nakano, K., Saitoh, N., Liu, Z., Yamamoto, T., Okamura, T., Ohkawa, Y., Kimura, H., and Nikaido, I. (2020). Genome-wide kinetic properties of transcriptional bursting in mouse embryonic stem cells. *Sci. Adv.*, 6(25):eaaz6699.

O'Dea, M. R. and Hasel, P. (2025). Are we there yet? exploring astrocyte heterogeneity one cell at a time. *Glia*, 73(3):619–631.

O'Donovan, M. C. and Owen, M. J. (2016). The implications of the shared genetics of psychiatric disorders. *Nat. Med.*, 22(11):1214–1219.

Olthof, A. M., Hyatt, K. C., and Kanadia, R. N. (2019). Minor intron splicing revisited: identification of new minor intron-containing genes and tissue-dependent retention and alternative splicing of minor introns. *BMC Genomics*, 20(1):686.

Ono, M., Scott, M. S., Yamada, K., Avolio, F., Barton, G. J., and Lamond, A. I. (2011). Identification of human miRNA precursors that resemble box C/D snoRNAs. *Nucleic Acids Res.*, 39(9):3879–3891.

OpenAI (2025). ChatGPT (version GPT-5). `https://chat.openai.com/`.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40(12):1413–1415.

Panagiotakos, G., Haveles, C., Arjun, A., Petrova, R., Rana, A., Portmann, T., Paşca, S. P., Palmer, T. D., and Dolmetsch, R. E. (2019). Aberrant calcium channel splicing drives defects in cortical differentiation in timothy syndrome. *Elife*, 8.

Pandey, A. C., Bezney, J., DeAscanis, D., Kirsch, E., Ahmed, F., Crinklaw, A., Choudhary, K. S., Mandala, T., Deason, J., Hamdi, J., Siddique, A., Ranganathan, S., Ordoukhanian, P., Brown, K., Armstrong, J., Head, S., and Topol, E. J. (2022). A CRISPR/Cas9-based enhancement of high-throughput single-cell transcriptomics. *bioRxiv*, page 2022.09.06.506867.

Pandit, S., Zhou, Y., Shiue, L., Coutinho-Mansfield, G., Li, H., Qiu, J., Huang, J., Yeo, G. W., Ares, Jr, M., and Fu, X.-D. (2013). Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol. Cell*, 50(2):223–235.

Paoli-Iseppi, R. D., Joshi, S., Gleeson, J., David, Y., Prawer, J., Yu, Y., Agarwal, R., Li, A., Hull, A., Whitehead, E. M., Seo, Y., Kujawa, R., Chang, R., Dutt, M., McLean, C., Parker, B. L., and Clark, M. B. (2024). Long-read sequencing reveals the RNA isoform repertoire of neuropsychiatric risk genes in human brain. *medRxiv*, 26(1):1–30.

Pardo-Palacios, F. J., Arzalluz-Luque, A., Kondratova, L., Salguero, P., Mestre-Tomás, J., Amorín, R., Estevan-Morió, E., Liu, T., Nanni, A., McIntyre, L., Tseng, E., and Conesa, A. (2024a). SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat. Methods*, 21(5):793–797.

Pardo-Palacios, F. J., Wang, D., Reese, F., Diekhans, M., Carbonell-Sala, S., Williams, B., Loveland, J. E., De María, M., Adams, M. S., Balderrama-Gutierrez, G., Behera, A. K., Gonzalez Martinez, J. M., Hunt, T., Lagarde, J., Liang, C. E., Li, H., Meade, M. J., Moraga Amador, D. A., Prjibelski, A. D., Birol, I., Bostan, H., Brooks, A. M., Çelik, M. H., Chen, Y., Du, M. R. M., Felton, C., Göke, J., Hafezqorani, S., Herwig, R., Kawaji, H., Lee, J., Li, J.-L., Lienhard, M., Mikheenko, A., Mulligan, D., Nip, K. M., Pertea, M., Ritchie, M. E., Sim, A. D., Tang, A. D., Wan, Y. K., Wang, C., Wong, B. Y., Yang, C., Barnes, I., Berry, A. E., Capella-Gutierrez, S., Cousineau, A., Dhillon, N., Fernandez-Gonzalez, J. M., Ferrández-Peral, L., Garcia-Reyero, N., Götz, S., Hernández-Ferrer, C., Kondratova, L., Liu, T., Martinez-Martin, A., Menor, C., Mestre-Tomás, J., Mudge, J. M., Panayotova, N. G., Paniagua, A., Repchevsky, D., Ren, X., Rouchka, E., Saint-John, B., Sapena, E., Sheynkman, L., Smith, M. L., Suner, M.-M., Takahashi, H., Youngworth, I. A., Carninci, P., Denslow, N. D., Guigó, R., Hunter, M. E., Maehr, R., Shen, Y., Tilgner, H. U., Wold, B. J., Vollmers, C., Frankish, A., Au, K. F., Sheynkman, G. M., Mortazavi, A., Conesa, A., and Brooks, A. N. (2024b). Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat. Methods*, 21(7):1349–1363.

Parikshak, N. N., Swarup, V., Belgard, T. G., Irimia, M., Ramaswami, G., Gandal, M. J., Hartl, C., Leppa, V., Ubieta, L. d. l. T., Huang, J., Lowe, J. K., Blencowe, B. J., Horvath, S., and Geschwind, D. H. (2016). Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature*, 540(7633):423–427.

Park, E., Iaccarino, C., Lee, J., Kwon, I., Baik, S. M., Kim, M., Seong, J. Y., Son, G. H., Borrelli, E., and Kim, K. (2011a). Regulatory roles of heterogeneous nuclear ribonucleoprotein M and nova-1 protein in alternative splicing of dopamine D2 receptor pre-mRNA. *J. Biol. Chem.*, 286(28):25301–25308.

Park, J., Wetzel, I., Marriott, I., Dréau, D., D'Avanzo, C., Kim, D. Y., Tanzi, R. E., and Cho, H. (2018). A 3D human triculture system modeling neurodegeneration and neuroinflammation in alzheimer's disease. *Nat. Neurosci.*, 21(7):941–951.

Park, J. Y., Li, W., Zheng, D., Zhai, P., Zhao, Y., Matsuda, T., Vatner, S. F., Sadoshima, J., and Tian, B. (2011b). Comparative analysis of mRNA isoform expression in cardiac hypertrophy and development reveals multiple post-transcriptional regulatory modules. *PLoS One*, 6(7):e22391.

Parker, M. T., Knop, K., Barton, G. J., and Simpson, G. G. (2021). 2passtools: two-pass alignment using machine-learning-filtered splice junctions increases the accuracy of intron detection in long-read RNA sequencing. *Genome Biol.*, 22(1):72.

Paterson, C., Wang, Y., Hyde, T. M., Weinberger, D. R., Kleinman, J. E., and Law, A. J. (2017). Temporal, diagnostic, and tissue-specific regulation of NRG3 isoform expression in human brain development and affective disorders. *Am. J. Psychiatry*, 174(3):256–265.

Patil, P., Kibiryeva, N., Uechi, T., Marshall, J., O'Brien, Jr, J. E., Artman, M., Kenmochi, N., and Bittel, D. C. (2015). scaRNAs regulate splicing and vertebrate heart development. *Biochim. Biophys. Acta*, 1852(8):1619–1629.

Patowary, A., Zhang, P., Jops, C., Vuong, C. K., Ge, X., Hou, K., Kim, M., Gong, N., Margolis, M., Vo, D., Wang, X., Liu, C., Pasaniuc, B., Li, J. J., Gandal, M. J., and de la Torre-Ubieta, L. (2024). Developmental isoform diversity in the human neocortex informs neuropsychiatric risk mechanisms. *Science*, 384(6698):eadh7688.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14(4):417–419.

Perez-Alcantara, M., Washer, S., Chen, Y., Steer, J., Gonzalez-Padilla, D., McWilliam, J., Willé, D., Panousis, N., Kolberg, P., Guerrero, E. N., Alasoo, K., Hall-Roberts, H., Williams, J., Cowley, S. A., Trynka, G., and Bassett, A. (2025). Integrated QTL mapping and CRISPR screening in pooled iPSC-derived microglia reveals genetic drivers of neurodegenerative risk. *bioRxiv*, page 2025.08.18.670767.

Petri, A. J. and Sahlin, K. (2023). isONform: reference-free transcriptome reconstruction from oxford nanopore data. *Bioinformatics*, 39(39 Suppl 1):i222–i231.

Philpott, M., Watson, J., Thakurta, A., Brown, Jr, T., Brown, Sr, T., Oppermann, U., and Cribbs, A. P. (2021). Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq. *Nat. Biotechnol.*, 39(12):1517–1520.

Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.*, 9(1):171–181.

Pickrell, J. K., Pai, A. A., Gilad, Y., and Pritchard, J. K. (2010). Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.*, 6(12):e1001236.

Pilcher, L. (2024). *Loss Of Non-Coding Rna, Snord116, Attenuates Chronic Pathological Cardiac Remodeling And Protects Cardiomyocyte Contraction And Relaxation Kinetics During Ischemia.* PhD thesis, University of Vermont.

Pimentel, H., Parra, M., Gee, S. L., Mohandas, N., Pachter, L., and Conboy, J. G. (2016). A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res.*, 44(2):838–851.

Pohl, M., Bortfeldt, R. H., Grützmann, K., and Schuster, S. (2013). Alternative splicing of mutually exclusive exons–a review. *Biosystems.*, 114(1):31–38.

Ponting, C. P. and Haerty, W. (2022). Genome-wide analysis of human long noncoding RNAs: A provocative review. *Annu. Rev. Genomics Hum. Genet.*, 23(1):153–172.

Ponting, C. P. and Hardison, R. C. (2011). What fraction of the human genome is functional? *Genome Res.*, 21(11):1769–1776.

Potolitsyna, E., Hazell Pickering, S., Tooming-Klunderud, A., Collas, P., and Briand, N. (2022). De novo annotation of lncRNA HOTAIR transcripts by long-read RNA capture-seq reveals a differentiation-driven isoform switch. *BMC Genomics*, 23(1):658.

Powell, W. T., Coulson, R. L., Crary, F. K., Wong, S. S., Ach, R. A., Tsang, P., Alice Yamada, N., Yasui, D. H., and Lasalle, J. M. (2013). A prader-willi locus lncRNA cloud modulates diurnal genes and energy expenditure. *Hum. Mol. Genet.*, 22(21):4318–4328.

Pozo, F., Martinez-Gomez, L., Walsh, T. A., Rodriguez, J. M., Di Domenico, T., Abascal, F., Vazquez, J., and Tress, M. L. (2021a). Assessing the functional relevance of splice isoforms. *NAR Genom Bioinform*, 3(2):lqab044.

Pozo, F., Martinez-Gomez, L., Walsh, T. A., Rodriguez, J. M., Di Domenico, T., Abascal, F., Vazquez, J., and Tress, M. L. (2021b). Assessing the functional relevance of splice isoforms. *NAR Genom Bioinform*, 3(2):lqab044.

Prjibelski, A. D., Mikheenko, A., Joglekar, A., Smetanin, A., Jarroux, J., Lapidus, A. L., and Tilgner, H. U. (2023). Accurate isoform discovery with IsoQuant using long reads. *Nat. Biotechnol.*

PsychENCODE Consortium, Akbarian, S., Liu, C., Knowles, J. A., Vaccarino, F. M., Farnham, P. J., Crawford, G. E., Jaffe, A. E., Pinto, D., Dracheva, S., Geschwind, D. H., Mill, J., Nairn, A. C., Abyzov, A., Pochareddy, S., Prabhakar, S., Weissman, S., Sullivan, P. F., State, M. W., Weng, Z., Peters, M. A., White, K. P., Gerstein, M. B., Amiri, A., Armoskus, C., Ashley-Koch, A. E., Bae, T., Beckel-Mitchener, A., Berman, B. P., Coetzee, G. A., Coppola, G., Francoeur, N., Fromer, M., Gao, R., Grennan, K., Herstein, J., Kavanagh, D. H., Ivanov, N. A., Jiang, Y., Kitchen, R. R., Kozlenkov, A., Kundakovic, M., Li, M., Li, Z., Liu, S., Mangravite, L. M., Mattei, E., Markenscoff-Papadimitriou, E., Navarro, F. C. P., North, N., Omberg, L., Panchision, D., Parikshak, N., Poschmann, J., Price, A. J., Purcaro, M., Reddy, T. E., Roussos, P., Schreiner, S., Scuderi, S., Sebra, R., Shibata, M., Shieh, A. W., Skarica, M., Sun, W., Swarup, V., Thomas, A., Tsuji, J., van Bakel, H., Wang, D., Wang, Y., Wang, K., Werling, D. M., Willsey, A. J., Witt, H., Won, H., Wong, C. C. Y., Wray, G. A., Wu, E. Y., Xu, X., Yao, L., Senthil, G., Lehner, T., Sklar, P., and Sestan, N. (2015). The PsychENCODE project. *Nat. Neurosci.*, 18(12):1707–1712.

Puvogel, S., Alsema, A., Kracht, L., Webster, M. J., Weickert, C. S., Sommer, I. E. C., and Eggen, B. J. L. (2022). Single-nucleus RNA sequencing of midbrain blood-brain barrier cells in schizophrenia reveals subtle transcriptional changes with overall preservation of cellular proportions and phenotypes. *Mol. Psychiatry*, 27(11):4731–4740.

Qi, T., Wu, Y., Fang, H., Zhang, F., Liu, S., Zeng, J., and Yang, J. (2022). Genetic control of RNA splicing and its distinct role in complex trait variation. *Nat. Genet.*, 54(9):1355–1363.

Quesnel-Vallières, M., Irimia, M., Cordes, S. P., and Blencowe, B. J. (2015). Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development. *Genes Dev.*, 29(7):746–759.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Raj, B. and Blencowe, B. j. (2015). Alternative splicing in the mammalian nervous system: Recent insights into mechanisms and functional roles. *Neuron*, 87(1):14–27.

Rasetti, R., Mattay, V. S., White, M. G., Sambataro, F., Podell, J. E., Zoltick, B., Chen, Q., Berman, K. F., Callicott, J. H., and Weinberger, D. R. (2014). Altered hippocampal-parahippocampal function during stimulus encoding: a potential indicator of genetic liability for schizophrenia: A potential indicator of genetic liability for schizophrenia. *JAMA Psychiatry*, 71(3):236–247.

Ratni, H., Ebeling, M., Baird, J., Bendels, S., Bylund, J., Chen, K. S., Denk, N., Feng, Z., Green, L., Guerard, M., Jablonski, P., Jacobsen, B., Khwaja, O., Kletzl, H., Ko, C.-P., Kustermann, S., Marquet, A., Metzger, F., Mueller, B., Naryshkin, N. A., Paushkin, S. V., Pinard, E., Poirier, A., Reutlinger, M., Weetall, M., Zeller, A., Zhao, X., and Mueller, L. (2018). Discovery of risdiplam, a selective survival of motor neuron-2 (SMN2) gene splicing modifier for the treatment of spinal muscular atrophy (SMA). *J. Med. Chem.*, 61(15):6501–6517.

Ray, T. A., Cochran, K., Kozlowski, C., Wang, J., Alexander, G., Cady, M. A., Spencer, W. J., Ruzycki, P. A., Clark, B. S., Laeremans, A., He, M.-X., Wang, X., Park, E., Hao, Y., Iannaccone, A., Hu, G., Fedrigo, O., Skiba, N. P., Arshavsky, V. Y., and Kay, J. N. (2020). Comprehensive identification of mRNA isoforms reveals the diversity of neural cell-surface molecules with roles in retinal development and disease. *Nat. Commun.*, 11(1):3328.

Reed, X., Cobb, M. M., Skinbinski, G., Roosen, D., Kaganovich, A., Ding, J., Finkbeiner, S., and Cookson, M. R. (2021). Transcriptional signatures in iPSC-derived neurons are reproducible across labs when differentiation protocols are closely matched. *Stem Cell Res.*, 56:102558.

Rees, E., Creeth, H. D. J., Hwu, H.-G., Chen, W. J., Tsuang, M., Glatt, S. J., Rey, R., Kirov, G., Walters, J. T. R., Holmans, P., Owen, M. J., and O'Donovan, M. C. (2021). Schizophrenia, autism spectrum disorders and developmental disorders share specific disruptive coding mutations. *Nat. Commun.*, 12(1):5353.

Reixachs-Solé, M., Ruiz-Orera, J., Albà, M. M., and Eyras, E. (2020). Ribosome profiling at isoform level reveals evolutionary conserved impacts of differential splicing on the proteome. *Nat. Commun.*, 11(1):1768.

Rogalska, M. E., Vivori, C., and Valcárcel, J. (2022). Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. *Nat. Rev. Genet.*

Runte, M., Varon, R., Horn, D., Horsthemke, B., and Buiting, K. (2005). Exclusion of the C/D box snoRNA gene cluster HBII-52 from a major role in prader-willi syndrome. *Hum. Genet.*, 116(3):228–230.

Saeed, S., Siegert, A.-M., Tung, Y. C. L., Khanam, R., Janjua, Q. M., Manzoor, J., Derhourhi, M., Toussaint, B., Lam, B. Y. H., Mahmoud, S. A., Vaillant, E., Falay, E. B., Amanzougarene, S., Ayesha, H., Khan, W. I., Ramazan, N., Saudek, V., O'Rahilly, S., Goldstone, A. P., Arslan, M., Bonnefond, A., Froguel, P., and Yeo, G. S. H. (2025). Biallelic variants in*SREK1*downregulating*SNORD115*and*SNORD116*cause a novel prader-willi-like syndrome. *medRxiv*, page 2025.02.26.24313254.

Sahakyan, A. B. and Balasubramanian, S. (2016). Long genes and genes with multiple splice variants are enriched in pathways linked to cancer and other multigenic diseases. *BMC Genomics*, 17:225.

Samsom, J. N. and Wong, A. H. C. (2015). Schizophrenia and depression co-morbidity: What we have learned from animal models. *Front. Psychiatry*, 6:13.

Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., and Burge, C. B. (2008). Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*, 320(5883):1643–1647.

Sartorius, L. J., Weinberger, D. R., Hyde, T. M., Harrison, P. J., Kleinman, J. E., and Lipska, B. K. (2008). Expression of a GRM3 splice variant is increased in the dorsolateral prefrontal cortex of individuals carrying a schizophrenia risk SNP. *Neuropsychopharmacology*, 33(11):2626–2634.

Saudemont, B., Popa, A., Parmley, J. L., Rocher, V., Blugeon, C., Necsulea, A., Meyer, E., and Duret, L. (2017). The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biol.*, 18(1):208.

Schertzer, M. D., Stirn, A., Isaev, K., Pereira, L., Das, A., Harbison, C., Park, S. H., Wessels, H.-H., Sanjana, N. E., and Knowles, D. A. (2023). Cas13d-mediated isoform-specific RNA knockdown with a unified computational and experimental toolbox. *bioRxiv*, page 2023.09.12.557474.

Schuster, J., Ritchie, M. E., and Gouil, Q. (2023). Restrander: rapid orientation and artefact removal for long-read cDNA data. *NAR Genom. Bioinform.*, 5(4):lqad108.

Schwenk, V., Leal Silva, R. M., Scharf, F., Knaust, K., Wendlandt, M., Häusser, T., Pickl, J. M. A., Steinke-Lange, V., Laner, A., Morak, M., Holinski-Feder, E., and Wolf, D. A. (2023). Transcript capture and ultradeep long-read RNA sequencing (CAPLRseq) to diagnose HNPCC/lynch syndrome. *J. Med. Genet.*

Scott, M. S., Ono, M., Yamada, K., Endo, A., Barton, G. J., and Lamond, A. I. (2012). Human box C/D snoRNA processing conservation across multiple cell types. *Nucleic Acids Res.*, 40(8):3676–3688.

Sebastian, R., Jin, K., Pavon, N., Bansal, R., Potter, A., Song, Y., Babu, J., Gabriel, R., Sun, Y., Aronow, B., and Pak, C. (2023). Schizophrenia-associated NRXN1 deletions induce developmental-timing- and cell-type-specific vulnerabilities in human brain organoids. *Nat. Commun.*, 14(1):3770.

Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sørensen, E. A., Wollenberg, R. D., and Albertsen, M. (2022). Oxford nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods*, 19(7):823–826.

Shah, A., Mittleman, B. E., Gilad, Y., and Li, Y. I. (2021). Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. *Genome Biol.*, 22(1):291.

Sharma, S., Grudzien-Nogalska, E., Hamilton, K., Jiao, X., Yang, J., Tong, L., and Kiledjian, M. (2020). Mammalian nudix proteins cleave nucleotide metabolite caps on RNAs. *Nucleic Acids Res.*, 48(12):6788–6798.

Shaul, O. (2017). How introns enhance gene expression. *Int. J. Biochem. Cell Biol.*, 91(Pt B):145–155.

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One*, 11(10):e0163962.

Sheynkman, G. M., Tuttle, K. S., Laval, F., Tseng, E., Underwood, J. G., Yu, L., Dong, D., Smith, M. L., Sebra, R., Willems, L., Hao, T., Calderwood, M. A., Hill, D. E., and Vidal, M. (2020). ORF capture-seq as a versatile method for targeted identification of full-length isoforms. *Nat. Commun.*, 11(1):2326.

Shi, J., Huang, C., Huang, S., and Yao, C. (2018). snoRNAs associate with mRNA 3' processing complex: New wine in old bottles. *RNA Biol.*, 15(2):194–197.

Shi, Y., Di Giammartino, D. C., Taylor, D., Sarkeshik, A., Rice, W. J., Yates, 3rd, J. R., Frank, J., and Manley, J. L. (2009). Molecular architecture of the human pre-mRNA 3' processing complex. *Mol. Cell*, 33(3):365–376.

Shi, Y. and Manley, J. L. (2015). The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. *Genes Dev.*, 29(9):889–897.

Shiau, C.-K., Lu, L., Kieser, R., Fukumura, K., Pan, T., Lin, H.-Y., Yang, J., Tong, E. L., Lee, G., Yan, Y., Huse, J. T., and Gao, R. (2023). High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. *Nat. Commun.*, 14(1):4124.

Shimada, M., Omae, Y., Kakita, A., Gabdulkhaev, R., Hitomi, Y., Miyagawa, T., Honda, M., Fujimoto, A., and Tokunaga, K. (2024). Identification of region-specific gene isoforms in the human brain using long-read transcriptome sequencing. *Sci. Adv.*, 10(4):eadj5279.

Shulman, E. D. and Elkon, R. (2019). Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic Acids Res.*, 47(19):10027–10039.

Sibley, C. R., Blazquez, L., and Ule, J. (2016). Lessons from non-canonical splicing. *Nat. Rev. Genet.*, 17(7):407–421.

Singh, R. (2002). RNA-protein interactions that regulate pre-mRNA splicing. *Gene Expr.*, 10(1-2):79–92.

Smith, M. A., Ersavas, T., Ferguson, J. M., Liu, H., Lucas, M. C., Begik, O., Bojarski, L., Barton, K., and Novoa, E. M. (2020). Molecular barcoding of native RNAs using nanopore sequencing and deep learning. *Genome Res.*, 30(9):1345–1353.

Smith, R. M. and Sadee, W. (2011). Synaptic signaling and aberrant RNA splicing in autism spectrum disorders. *Front. Synaptic Neurosci.*, 3:1.

Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.*, 27(3):491–499.

Soetanto, R., Hynes, C. J., Patel, H. R., Humphreys, D. T., Evers, M., Duan, G., Parker, B. J., Archer, S. K., Clancy, J. L., Graham, R. M., Beilharz, T. H., Smith, N. J., and Preiss, T. (2016). Role of miRNAs and alternative mRNA 3'-end cleavage and polyadenylation of their mRNA targets in cardiomyocyte hypertrophy. *Biochim. Biophys. Acta*, 1859(5):744–756.

Song, Z., Bae, B., Schnabl, S., Yuan, F., De Zoysa, T., Akinyi, M., Le Roux, C., Choquet, K., Whipple, A., and Van Nostrand, E. (2024). Mapping snoRNA-target RNA interactions in an RNA binding protein-dependent manner with chimeric eCLIP. *bioRxivorg*, page 2024.09.19.613955.

Spellman, R., Llorian, M., and Smith, C. W. J. (2007). Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. *Mol. Cell*, 27(3):420–434.

Splawski, I., Timothy, K. W., Sharpe, L. M., Decher, N., Kumar, P., Bloise, R., Napolitano, C., Schwartz, P. J., Joseph, R. M., Condouris, K., Tager-Flusberg, H., Priori, S. G., Sanguinetti, M. C., and Keating, M. T. (2004). Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell*, 119(1):19–31.

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.*, 20(11):631–656.

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., RGASP Consortium, Hubbard, T. J., Guigó, R., Harrow, J., and Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, 10(12):1177–1184.

Steyn, C., Mishi, R., Fillmore, S., Verhoog, M. B., More, J., Rohlwink, U. K., Melvill, R., Butler, J., Enslin, J. M. N., Jacobs, M., Sauka-Spengler, T., Greco, M., Quiñones, S., Dulla, C. G., Raimondo, J. V., Figaji, A., and Hockman, D. (2024). A temporal cortex cell atlas highlights gene expression dynamics during human brain maturation. *Nat. Genet.*, 56(12):2718–2730.

Stroup, E. K. and Ji, Z. (2023). Deep learning of human polyadenylation sites at nucleotide resolution reveals molecular determinants of site usage and relevance in disease. *Nat. Commun.*, 14(1):7378.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, 3rd, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21.

Su, Y., Yu, Z., Jin, S., Ai, Z., Yuan, R., Chen, X., Xue, Z., Guo, Y., and others (2023). Comprehensive assessment of isoform detection methods for third-generation sequencing data. *bioRxiv*.

Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.*, 38(2):147–150.

Svensson, V., Gayoso, A., Yosef, N., and Pachter, L. (2020). Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421.

Sweeney, B. A., Hoksza, D., Nawrocki, E. P., Ribas, C. E., Madeira, F., Cannone, J. J., Gutell, R., Maddala, A., Meade, C. D., Williams, L. D., Petrov, A. S., Chan, P. P., Lowe, T. M., Finn, R. D., and Petrov, A. I. (2021). R2DT is a framework for predicting and visualising RNA secondary structure using templates. *Nat. Commun.*, 12(1):3494.

Tabrez, S. S., Sharma, R. D., Jain, V., Siddiqui, A. A., and Mukhopadhyay, A. (2017). Differential alternative splicing coupled to nonsense-mediated decay of mRNA ensures dietary restriction-induced longevity. *Nat. Commun.*, 8(1):306.

Taggart, J. C., Zauber, H., Selbach, M., Li, G.-W., and McShane, E. (2020). Keeping the proportions of protein complex components in check. *Cell Syst.*, 10(2):125–132.

Takata, A., Matsumoto, N., and Kato, T. (2017). Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat. Commun.*, 8(1):14519.

Tan, Q., Potter, K. J., Burnett, L. C., Orsso, C. E., Inman, M., Ryman, D. C., and Haqq, A. M. (2020). Prader-willi-like phenotype caused by an atypical 15q11.2 microdeletion. *Genes (Basel)*, 11(2):128.

Tang, A. D., Soulette, C. M., van Baren, M. J., Hart, K., Hrabeta-Robinson, E., Wu, C. J., and Brooks, A. N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.*, 11(1):1438.

Tang, P., Yang, Y., Li, G., Huang, L., Wen, M., Ruan, W., Guo, X., Zhang, C., Zuo, X., Luo, D., Xu, Y., Fu, X.-D., and Zhou, Y. (2022). Alternative polyadenylation by sequential activation of distal and proximal PolyA sites. *Nat. Struct. Mol. Biol.*, 29(1):21–31.

Tang, Z. Z., Sharma, S., Zheng, S., Chawla, G., Nikolic, J., and Black, D. L. (2011). Regulation of the mutually exclusive exons 8a and 8 in the CaV1.2 calcium channel transcript by polypyrimidine tract-binding protein. *J. Biol. Chem.*, 286(12):10007–10016.

Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., Edelmann, M., Ezkurdia, I., Vazquez, J., Tress, M., Mortazavi, A., Martens, L., Rodriguez-Navarro, S., Moreno-Manzano, V., and Conesa, A. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.*

Teng, P., Li, Y., Ku, L., Wang, F., Goldsmith, D. R., Wen, Z., Yao, B., and Feng, Y. (2023). The human lncRNA GOMAFU suppresses neuronal interferon response pathways affected in neuropsychiatric diseases. *Brain Behav. Immun.*, 112:175–187.

Tesfaye, M., Jaholkowski, P., Shadrin, A. A., van der Meer, D., Hindley, G. F. L., Holen, B., Parker, N., Parekh, P., Birkenæs, V., Rahman, Z., Bahrami, S., Kutrolli, G., Frei, O., Djurovic, S., Dale, A. M., Smeland, O. B., O'Connell, K. S., and Andreassen, O. A. (2024). Identification of novel genomic loci for anxiety symptoms and extensive genetic overlap with psychiatric disorders. *Psychiatry Clin. Neurosci.*, 78(12):783–791.

Tian, B. and Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.*, 18(1):18–30.

Tian, L., Jabbari, J. S., Thijssen, R., Gouil, Q., Amarasinghe, S. L., Voogd, O., Kariyawasam, H., Du, M. R. M., Schuster, J., Wang, C., Su, S., Dong, X., Law, C. W., Lucattini, A., Prawer, Y. D. J., Collar-Fernández, C., Chung, J. D., Naim, T., Chan, A., Ly, C. H., Lynch, G. S., Ryall, J. G., Anttila, C. J. A., Peng, H., Anderson, M. A., Flensburg, C., Majewski, I., Roberts, A. W., Huang, D. C. S., Clark, M. B., and Ritchie, M. E. (2021). Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol.*, 22(1):310.

Tilgner, H., Grubert, F., Sharon, D., and Snyder, M. P. (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. U. S. A.*, 111(27):9869–9874.

Torres-Méndez, A., Pop, S., Bonnal, S., Almudi, I., Roberts, R. J. V., Paolantoni, C., Alcaina, A., Avola, A., Martín-Anduaga, A., Haussmann, I. U., Morin, V., Casares, F., Soller, M., Kadener, S., Roignant, J.-Y., Prieto-Godino, L., and Irimia, M. (2021). Parallel evolution of a splicing program controlling neuronal excitability in flies and mammals. *Cold Spring Harbor Laboratory*, page 2021.02.24.432780.

Tress, M. L., Abascal, F., and Valencia, A. (2017). Most alternative isoforms are not functionally important. *Trends Biochem. Sci.*, 42(6):408–410.

Treutlein, B., Gokce, O., Quake, S. R., and Südhof, T. C. (2014). Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 111(13):E1291–9.

Troskie, R.-L., Jafrani, Y., Mercer, T. R., Ewing, A. D., Faulkner, G. J., and Cheetham, S. W. (2021). Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome. *Genome Biol.*, 22(1):146.

Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., Bryois, J., Chen, C.-Y., Dennison, C. A., Hall, L. S., Lam, M., Watanabe, K., Frei, O., Ge, T., Harwood, J. C., Koopmans, F., Magnusson, S., Richards, A. L., Sidorenko, J., Wu, Y., Zeng, J., Grove, J., Kim, M., Li, Z., Voloudakis, G., Zhang, W., Adams, M., Agartz, I., Atkinson, E. G., Agerbo, E., Al Eissa, M., Albus, M., Alexander, M., Alizadeh, B. Z., Alptekin, K., Als, T. D., Amin, F., Arolt, V., Arrojo, M., Athanasiu, L., Azevedo, M. H., Bacanu, S. A., Bass, N. J., Begemann, M., Belliveau, R. A., Bene, J., Benyamin, B., Bergen, S. E., Blasi, G., Bobes, J., Bonassi, S., Braun, A., Bressan, R. A., Bromet, E. J., Bruggeman, R., Buckley, P. F., Buckner, R. L., Bybjerg-Grauholm, J., Cahn, W., Cairns, M. J., Calkins, M. E., Carr, V. J., Castle, D., Catts, S. V., Chambert, K. D., Chan, R. C. K., Chaumette, B., Cheng, W., Cheung, E. F. C., Chong, S. A., Cohen, D., Consoli, A., Cordeiro, Q., Costas, J., Curtis, C., Davidson, M., Davis, K. L., de Haan, L., Degenhardt, F., DeLisi, L. E., Demontis, D., Dickerson, F., Dikeos, D., Dinan, T., Djurovic, S., Duan, J., Ducci, G., Dudbridge, F., Eriksson, J. G., Fañanás, L., Faraone, S. V., Fiorentino, A., Forstner, A., Frank, J., Freimer, N. B., Fromer, M., Frustaci, A., Gadelha, A., Genovese, G., Gershon, E. S., Giannitelli, M., Giegling, I., Giusti-Rodríguez, P., Godard, S., Goldstein, J. I., González Peñas, J., González-Pinto, A., Gopal, S., Gratten, J., Green, M. F., Greenwood, T. A., Guillin, O., Gülöksüz, S., Gur,

R. E., Gur, R. C., Gutiérrez, B., Hahn, E., Hakonarson, H., Haroutunian, V., Hartmann, A. M., Harvey, C., Hayward, C., Henskens, F. A., Herms, S., Hoffmann, P., Howrigan, D. P., Ikeda, M., Iyegbe, C., Joa, I., Julià, A., Kähler, A. K., Kam-Thong, T., Kamatani, Y., Karachanak-Yankova, S., Kebir, O., Keller, M. C., Kelly, B. J., Khrunin, A., Kim, S.-W., Klovins, J., Kondratiev, N., Konte, B., Kraft, J., Kubo, M., Kučinskas, V., Kučinskiene, Z. A., Kusumawardhani, A., Kuzelova-Ptackova, H., Landi, S., Lazzeroni, L. C., Lee, P. H., Legge, S. E., Lehrer, D. S., Lencer, R., Lerer, B., Li, M., Lieberman, J., Light, G. A., Limborska, S., Liu, C.-M., Lönnqvist, J., Loughland, C. M., Lubinski, J., Luykx, J. J., Lynham, A., Macek, Jr, M., Mackinnon, A., Magnusson, P. K. E., Maher, B. S., Maier, W., Malaspina, D., Mallet, J., Marder, S. R., Marsal, S., Martin, A. R., Martorell, L., Mattheisen, M., McCarley, R. W., McDonald, C., McGrath, J. J., Medeiros, H., Meier, S., Melegh, B., Melle, I., Mesholam-Gately, R. I., Metspalu, A., Michie, P. T., Milani, L., Milanova, V., Mitjans, M., Molden, E., Molina, E., Molto, M. D., Mondelli, V., Moreno, C., Morley, C. P., Muntané, G., Murphy, K. C., Myin-Germeys, I., Nenadić, I., Nestadt, G., Nikitina-Zake, L., Noto, C., Nuechterlein, K. H., O'Brien, N. L., O'Neill, F. A., Oh, S.-Y., Olincy, A., Ota, V. K., Pantelis, C., Papadimitriou, G. N., Parellada, M., Paunio, T., Pellegrino, R., Periyasamy, S., Perkins, D. O., Pfuhlmann, B., Pietiläinen, O., Pimm, J., Porteous, D., Powell, J., Quattrone, D., Quested, D., Radant, A. D., Rampino, A., Rapaport, M. H., Rautanen, A., Reichenberg, A., Roe, C., Roffman, J. L., Roth, J., Rothermundt, M., Rutten, B. P. F., Saker-Delye, S., Salomaa, V., Sanjuan, J., Santoro, M. L., Savitz, A., Schall, U., Scott, R. J., Seidman, L. J., Sharp, S. I., Shi, J., Siever, L. J., Sigurdsson, E., Sim, K., Skarabis, N., Slominsky, P., So, H.-C., Sobell, J. L., Söderman, E., Stain, H. J., Steen, N. E., Steixner-Kumar, A. A., Stögmann, E., Stone, W. S., Straub, R. E., Streit, F., Strengman, E., Stroup, T. S., Subramaniam, M., Sugar, C. A., Suvisaari, J., Svrakic, D. M., Swerdlow, N. R., Szatkiewicz, J. P., Ta, T. M. T., Takahashi, A., Terao, C., Thibaut, F., Toncheva, D., Tooney, P. A., Torretta, S., Tosato, S., Tura, G. B., Turetsky, B. I., Üçok, A., Vaaler, A., van Amelsvoort, T., van Winkel, R., Veijola, J., Waddington, J., Walter, H., Waterreus, A., Webb, B. T., Weiser, M., Williams, N. M., Witt, S. H., Wormley, B. K., Wu, J. Q., Xu, Z., Yolken, R., Zai, C. C., Zhou, W., Zhu, F., Zimprich, F., Atbaşoğlu, E. C., Ayub, M., Benner, C., Bertolino, A., Black, D. W., Bray, N. J., Breen, G., Buccola, N. G., Byerley, W. F., Chen, W. J., Cloninger, C. R., Crespo-Facorro, B., Donohoe, G., Freedman, R., Galletly, C., Gandal, M. J., Gennarelli, M., Hougaard, D. M., Hwu, H.-G., Jablensky, A. V., McCarroll, S. A., Moran, J. L., Mors, O., Mortensen, P. B., Müller-Myhsok, B., Neil, A. L., Nordentoft, M., Pato, M. T., Petryshen, T. L., Pirinen, M., Pulver, A. E., Schulze, T. G., Silverman, J. M., Smoller, J. W., Stahl, E. A., Tsuang, D. W., Vilella, E., Wang, S.-H., Xu, S., Indonesia Schizophrenia Consortium, PsychENCODE, Psychosis Endophenotypes International Consortium, SynGO Consortium, Adolfsson, R., Arango, C., Baune, B. T., Belangero, S. I., Børglum, A. D., Braff, D., Bramon, E., Buxbaum, J. D., Campion, D., Cervilla, J. A., Cichon, S., Collier, D. A., Corvin, A., Curtis, D., Forti, M. D., Domenici, E., Ehrenreich, H., Escott-Price, V., Esko, T., Fanous, A. H., Gareeva, A., Gawlik, M., Gejman, P. V., Gill, M., Glatt, S. J., Golimbet, V., Hong, K. S., Hultman, C. M., Hyman, S. E., Iwata, N., Jönsson, E. G., Kahn, R. S., Kennedy, J. L., Khusnutdinova, E., Kirov, G., Knowles, J. A., Krebs, M.-O., Laurent-Levinson, C., Lee, J., Lencz, T., Levinson, D. F., Li, Q. S., Liu, J., Malhotra, A. K., Malhotra, D., McIntosh, A., McQuillin, A., Menezes, P. R., Morgan, V. A., Morris, D. W., Mowry, B. J., Murray, R. M., Nimgaonkar, V., Nöthen, M. M., Ophoff, R. A., Paciga, S. A., Palotie, A., Pato, C. N., Qin, S., Rietschel, M., Riley, B. P., Rivera, M., Rujescu, D., Saka, M. C., Sanders, A. R., Schwab, S. G., Serretti, A., Sham, P. C., Shi, Y., St Clair, D., Stefánsson, H., Stefansson, K., Tsuang, M. T., van Os, J., Vawter, M. P., Weinberger, D. R., Werge, T., Wildenauer, D. B., Yu, X., Yue, W., Holmans, P. A., Pocklington, A. J., Roussos, P., Vassos, E., Verhage, M., Visscher, P. M., Yang, J., Posthuma, D., Andreassen, O. A., Kendler, K. S., Owen, M. J., Wray, N. R., Daly, M. J., Huang, H., Neale, B. M., Sullivan, P. F., Ripke, S., Walters, J. T. R., O'Donovan, M. C., and

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, 604(7906):502–508.

Trujillo, C. A., Rice, E. S., Schaefer, N. K., Chaim, I. A., Wheeler, E. C., Madrigal, A. A., Buchanan, J., Preissl, S., Wang, A., Negraes, P. D., Szeto, R. A., Herai, R. H., Huseynov, A., Ferraz, M. S. A., Borges, F. S., Kihara, A. H., Byrne, A., Marin, M., Vollmers, C., Brooks, A. N., Lautz, J. D., Semendeferi, K., Shapiro, B., Yeo, G. W., Smith, S. E. P., Green, R. E., and Muotri, A. R. (2021). Reintroduction of the archaic variant of NOVA1 in cortical organoids alters neurodevelopment. *Science*, 371(6530).

Tsuang, M. T., Taylor, L., and Faraone, S. V. (2004). An overview of the genetics of psychotic mood disorders. *J. Psychiatr. Res.*, 38(1):3–15.

Tume, C. E., Chick, S. L., Holmans, P. A., Rees, E., O'Donovan, M. C., Cameron, D., and Bray, N. J. (2024). Genetic implication of specific glutamatergic neurons of the prefrontal cortex in the pathophysiology of schizophrenia. *Biol. Psychiatry Glob. Open Sci.*, 4(5):100345.

Tung, K.-F., Pan, C.-Y., and Lin, W.-C. (2022). Dominant transcript expression profiles of human protein-coding genes interrogated with GTEx dataset. *Sci. Rep.*, 12(1):6969.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B. J., and Darnell, R. B. (2006). An RNA map predicting nova-dependent splicing regulation. *Nature*, 444(7119):580–586.

Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.-S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., Zeeberg, B. R., Kane, D., Weinstein, J. N., Blume, J., and Darnell, R. B. (2005). Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet.*, 37(8):844–852.

van der Toorn, W., Bohn, P., Liu-Wei, W., Olguin-Nava, M., Gribling-Burrer, A.-S., Smyth, R. P., and von Kleist, M. (2025). Demultiplexing and barcode-specific adaptive sampling for nanopore direct RNA sequencing. *Nat. Commun.*, 16(1):3742.

van Wilgenburg, B., Browne, C., Vowles, J., and Cowley, S. A. (2013). Efficient, long term production of monocyte-derived macrophages from human pluripotent stem cells under partly-defined and fully-defined conditions. *PLoS One*, 8(8):e71098.

Veikkolainen, V., Vaparanta, K., Halkilahti, K., Iljin, K., Sundvall, M., and Elenius, K. (2011). Function of ERBB4 is determined by alternative splicing. *Cell Cycle*, 10(16):2647–2657.

Velmeshev, D., Magistri, M., Mazza, E. M. C., Lally, P., Khoury, N., D'Elia, E. R., Bicciato, S., and Faghihi, M. A. (2020). Cell-type-specific analysis of molecular pathology in autism identifies common genes and pathways affected across neocortical regions. *Mol. Neurobiol.*, 57(5):2279–2289.

Velmeshev, D., Perez, Y., Yan, Z., Valencia, J. E., Castaneda-Castellanos, D. R., Wang, L., Schirmer, L., Mayer, S., Wick, B., Wang, S., Nowakowski, T. J., Paredes, M., Huang, E. J., and Kriegstein, A. R. (2023). Single-cell analysis of prenatal and postnatal human cortical development. *Science*, 382(6667):eadf0834.

Vitting-Seerup, K. and Sandelin, A. (2019). IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics*, 35(21):4469–4471.

Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R. J., Green, R. E., and Vollmers, C. (2018). Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U. S. A.*, 115(39):9726–9731.

Volden, R. and Vollmers, C. (2022). Single-cell isoform analysis in human immune cells. *Genome Biol.*, 23(1):47.

Walker, R. L., Ramaswami, G., Hartl, C., Mancuso, N., Gandal, M. J., de la Torre-Ubieta, L., Pasaniuc, B., Stein, J. L., and Geschwind, D. H. (2019). Genetic control of expression and splicing in developing human brain informs disease mechanisms. *Cell*, 179(3):750–771.e22.

Wamsley, B., Bicks, L., Cheng, Y., Kawaguchi, R., Quintero, D., Margolis, M., Grundman, J., Liu, J., Xiao, S., Hawken, N., Mazariegos, S., and Geschwind, D. H. (2024). Molecular cascades and cell type-specific signatures in ASD revealed by single-cell genomics. *Science*, 384(6698):eadh2602.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C. P., Clarke, D., Gu, M., Emani, P., Yang, Y. T., Xu, M., Gandal, M. J., Lou, S., Zhang, J., Park, J. J., Yan, C., Rhie, S. K., Manakongtreecheep, K., Zhou, H., Nathan, A., Peters, M., Mattei, E., Fitzgerald, D., Brunetti, T., Moore, J., Jiang, Y., Girdhar, K., Hoffman, G. E., Kalayci, S., Gümüş, Z. H., Crawford, G. E., PsychENCODE Consortium, Roussos, P., Akbarian, S., Jaffe, A. E., White, K. P., Weng, Z., Sestan, N., Geschwind, D. H., Knowles, J. A., and Gerstein, M. B. (2018). Comprehensive functional genomic resource and integrative model for the human brain. *Science*, 362(6420):eaat8464.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476.

Wang, J., Chang, Y. F., Hamilton, J. I., and Wilkinson, M. F. (2002). Nonsense-associated altered splicing: a frame-dependent response distinct from nonsense-mediated decay. *Mol. Cell*, 10(4):951–957.

Wang, W., Yang, X., Cristofalo, V. J., Holbrook, N. J., and Gorospe, M. (2001). Loss of HuR is linked to reduced expression of proliferative genes during replicative senescence. *Mol. Cell. Biol.*, 21(17):5889–5898.

Wang, X., Park, J., Susztak, K., Zhang, N. R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.*, 10(1):380.

Wang, Y., Liu, J., Huang, B. O., Xu, Y.-M., Li, J., Huang, L.-F., Lin, J., Zhang, J., Min, Q.-H., Yang, W.-M., and Wang, X.-Z. (2015). Mechanism of alternative splicing and its regulation. *Biomed Rep*, 3(2):152–158.

Wang, Y., Mohsen, A.-W., Mihalik, S. J., Goetzman, E. S., and Vockley, J. (2010). Evidence for physical association of mitochondrial fatty acid oxidation and oxidative phosphorylation complexes. *J. Biol. Chem.*, 285(39):29834–29841.

Wang, Z., Zhou, K., Yuan, Q., Chen, D., Hu, X., Xie, F., Liu, Y., and Xing, J. (2023). A high-efficiency capture-based NGS approach for comprehensive analysis of mitochondrial transcriptome. *Anal. Chem.*, 95(46):17046–17053.

Washer, S. J., Perez-Alcantara, M., Chen, Y., Steer, J., James, W. S., Trynka, G., Bassett, A. R., and Cowley, S. A. (2022). Single-cell transcriptomics defines an improved, validated monoculture protocol for differentiation of human iPSC to microglia. *Sci. Rep.*, 12(1):19454.

Weißbach, S., Todorov, H., Schlichtholz, L., Mühlbauer, S., Zografidou, L., Soliman, A., Lor-Zade, S., Hartwich, D., Strand, D., Strand, S., Vogel, T., Heine, M., Gerber, S., and Winter, J. (2024). Premature upregulation of miR-92a's target RBFOX2 hijacks PTBP splicing and impairs cortical neuronal differentiation. *bioRxiv*, page 2024.09.20.614071.

Wen, C., Margolis, M., Dai, R., Zhang, P., Przytycki, P. F., Vo, D. D., Bhattacharya, A., Kim, M., Matoba, N., Tsai, E., Hoh, C., Jiao, C., Aygun, N., Walker, R. L., Chatzinakos, C., Clarke, D., Pratt, H., PsychENCODE Consortium, Peters, M. A., Gerstein, M., Daskalakis, N. P., Weng, Z., Jaffe, A. E., Kleinman, J. E., Hyde, T. M., Weinberger, D. R., Bray, N. J., Sestan, N., Geschwind, D. H., Roeder, K., Gusev, A., Pasaniuc, B., Stein, J. L., Love, M. I., Pollard, K. S., Liu, C., and Gandal, M. J. (2023). Cross-ancestry, cell-type-informed atlas of gene, isoform, and splicing regulation in the developing human brain. *medRxiv*.

West, S., Gromak, N., and Proudfoot, N. J. (2004). Human 5' –¿ 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature*, 432(7016):522–525.

Whipple, A. and colleagues (2025). Investigating the role of Snord116 in ribosome biology (year 2). `https://www.fpwr.org/fpwr-funded-projects/investigating-the-role-of-snord116-in-ribosome-biology-year-2`. Accessed: 2025-9-25.

White, E. J. F., Brewer, G., and Wilson, G. M. (2013). Post-transcriptional control of gene expression by AUF1: mechanisms, physiological targets, and regulation. *Biochim. Biophys. Acta*, 1829(6-7):680–688.

Wissel, D., Mehlferber, M. M., Nguyen, K. M., Pavelko, V., Tseng, E., Robinson, M. D., and Sheynkman, G. M. (2025). A systematic benchmark of high-accuracy PacBio long-read RNA sequencing for transcript-level quantification. *bioRxivorg*, page 2025.05.30.656561.

Wojtowicz, W. M., Flanagan, J. J., Millard, S. S., Zipursky, S. L., and Clemens, J. C. (2004). Alternative splicing of drosophila dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell*, 118(5):619–633.

Wong, G. K.-S., Passey, D. A., Huang, Y.-Z., Yang, Z., and Yu, J. (2000). Is “junk” DNA mostly intron DNA? *Genome Res.*, 10(11):1672–1678.

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., and Rasko, J. E. J. (2017). Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment. *Nat. Commun.*, 8:15134.

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., Agerbo, E., Air, T. M., Andlauer, T. M. F., Bacanu, S.-A., Bækvad-Hansen, M., Beekman, A. F. T., Bigdeli, T. B., Binder, E. B., Blackwood, D. R. H., Bryois, J., Buttenschøn, H. N., Bybjerg-Grauholm, J., Cai, N., Castelao, E., Christensen, J. H., Clarke, T.-K., Coleman, J. I. R., Colodro-Conde, L., Couvy-Duchesne, B., Craddock, N., Crawford, G. E., Crowley, C. A., Dashti, H. S., Davies, G., Deary, I. J., Degenhardt, F., Derks, E. M., Direk, N., Dolan, C. V., Dunn, E. C., Eley, T. C., Eriksson, N., Escott-Price, V., Kiadeh, F. H. F., Finucane, H. K., Forstner, A. J., Frank, J., Gaspar, H. A., Gill, M., Giusti-Rodríguez, P., Goes, F. S., Gordon, S. D., Grove, J., Hall, L. S., Hannon, E., Hansen, C. S., Hansen, T. F., Herms, S., Hickie, I. B., Hoffmann, P., Homuth, G., Horn, C., Hottenga, J.-J., Hougaard, D. M., Hu, M., Hyde, C. L., Ising, M., Jansen, R., Jin, F., Jorgenson, E., Knowles, J. A., Kohane, I. S., Kraft, J., Kretzschmar, W. W., Krogh, J., Kutalik, Z., Lane, J. M., Li, Y., Li, Y., Lind, P. A., Liu, X., Lu, L., MacIntyre, D. J., MacKinnon, D. F., Maier, R. M., Maier, W., Marchini, J., Mbarek, H., McGrath, P., McGuffin, P., Medland, S. E., Mehta, D., Middeldorp, C. M., Mihailov, E., Milaneschi, Y., Milani, L., Mill, J., Mondimore, F. M., Montgomery, G. W., Mostafavi, S., Mullins, N., Nauck, M., Ng, B., Nivard, M. G., Nyholt, D. R., O'Reilly, P. F., Oskarsson, H., Owen, M. J., Painter, J. N., Pedersen, C. B., Pedersen, M. G., Peterson, R. E., Pettersson, E., Peyrot, W. J., Pistis, G., Posthuma, D., Purcell, S. M., Quiroz, J. A., Qvist, P., Rice, J. P., Riley, B. P., Rivera, M., Saeed Mirza, S., Saxena, R., Schoevers, R., Schulte, E. C., Shen, L., Shi, J., Shyn, S. I., Sigurdsson, E., Sinnamon, G. B. C., Smit, J. H., Smith, D. J., Stefansson, H., Steinberg, S., Stockmeier, C. A., Streit, F., Strohmaier, J., Tansey, K. E., Teismann, H., Teumer, A., Thompson, W., Thomson, P. A., Thorgeirsson, T. E., Tian, C., Traylor, M., Treutlein, J., Trubetskoy, V., Uitterlinden, A. G., Umbricht, D., Van der Auwera, S., van Hemert, A. M., Viktorin, A., Visscher, P. M., Wang, Y., Webb, B. T., Weinsheimer, S. M., Wellmann, J., Willemsen, G., Witt, S. H., Wu, Y., Xi, H. S., Yang, J., Zhang, F., eQTLGen, 23andMe, Arolt, V., Baune, B. T., Berger, K., Boomsma, D. I., Cichon, S., Dannlowski, U., de Geus, E. C. J., DePaulo, J. R., Domenici, E., Domschke, K., Esko, T., Grabe, H. J., Hamilton, S. P., Hayward, C., Heath, A. C., Hinds, D. A., Kendler, K. S., Kloiber, S., Lewis, G., Li, Q. S., Lucae, S., Madden, P. F. A., Magnusson, P. K., Martin, N. G., McIntosh, A. M., Metspalu, A., Mors, O., Mortensen, P. B., Müller-Myhsok, B., Nordentoft, M., Nöthen, M. M., O'Donovan, M. C., Paciga, S. A., Pedersen, N. L., Penninx, B. W. J. H., Perlis, R. H., Porteous, D. J., Potash, J. B., Preisig, M., Rietschel, M., Schaefer, C., Schulze, T. G., Smoller, J. W., Stefansson, K., Tiemeier, H., Uher, R., Völzke, H., Weissman, M. M., Werge, T., Winslow, A. R., Lewis, C. M., Levinson, D. F., Breen, G., Børglum, A. D., Sullivan, P. F., and Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.*, 50(5):668–681.

Wright, D. J., Hall, N., Irish, N., Man, A. L., Glynn, W., Mould, A., De Los Angeles, A., Angiolini, E., Swarbreck, D., Gharbi, K., Tunbridge, E. M., and Haerty, W. (2021). Long read sequencing reveals novel isoforms and insights into splicing regulation during cell state changes. *bioRxiv*, page 2021.04.27.441628.

Wu, H., Lu, Y., Duan, Z., Wu, J., Lin, M., Wu, Y., Han, S., Li, T., Fan, Y., Hu, X., Xiao, H., Feng, J., Lu, Z., Kong, D., and Li, S. (2023a). Nanopore long-read RNA sequencing reveals functional alternative splicing variants in human vascular smooth muscle cells. *Commun Biol*, 6(1):1104.

Wu, H., Yin, Q.-F., Luo, Z., Yao, R.-W., Zheng, C.-C., Zhang, J., Xiang, J.-F., Yang, L., and Chen, L.-L. (2016). Unusual processing generates SPA LncRNAs that sequester multiple RNA binding proteins. *Mol. Cell*, 64(3):534–548.

Wu, Q., Wu, J., Karim, K., Chen, X., Wang, T., Iwama, S., Carobbio, S., Keen, P., Vidal-Puig, A., Kotter, M. R., and Bassett, A. (2023b). Massively parallel characterization of CRISPR activator efficacy in human induced pluripotent stem cells and neurons. *Mol. Cell*, 83(7):1125–1139.e8.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., and Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb.)*, 2(3):100141.

Wyman, D., Balderrama-Gutierrez, G., Reese, F., Jiang, S., Rahmanian, S., Forner, S., Matheos, D., Zeng, W., Williams, B., Trout, D., England, W., Chu, S.-H., Spitale, R. C., Tenner, A. J., Wold, B. J., and Mortazavi, A. (2020). A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv*, page 672931.

Yang, S., Li, M., and Yao, C. (2025). Splicing inhibits premature cleavage and polyadenylation. *Trends Genet.*, 0(0).

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y. A., Murray, R. R., Spirohn, K., Begg, B. E., Duran-Frigola, M., MacWilliams, A., Pevzner, S. J., Zhong, Q., Wanamaker, S. A., Tam, S., Ghamsari, L., Sahni, N., Yi, S., Rodriguez, M. D., Balcha, D., Tan, G., Costanzo, M., Andrews, B., Boone, C., Zhou, X. J., Salehi-Ashtiani, K., Charloteaux, B., Chen, A. A., Calderwood, M. A., Aloy, P., Roth, F. P., Hill, D. E., Iakoucheva,

L. M., Xia, Y., and Vidal, M. (2016). Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, 164(4):805–817.

Yang, Y., Guo, T., Zhao, Q., Li, Y., Cheung, T., Zhang, L., Zhu, X., Jackson, T., Li, X.-H., and Xiang, Y.-T. (2024). Mapping prodromal symptoms in patients with bipolar disorder: A network perspective. *Psychiatry Res.*, 335(115842):115842.

Yang, Y., Wu, X., Yang, W., Jin, W., Wang, D., Yang, J., Jiang, G., Zhang, W., Niu, X., and Gong, J. (2022). Dynamic alternative polyadenylation during iPSC differentiation into cardiomyocytes. *Comput. Struct. Biotechnol. J.*, 20:5859–5869.

Yang, Y., Yang, R., Kang, B., Qian, S., He, X., and Zhang, X. (2023). Single-cell long-read sequencing in human cerebral organoids uncovers cell-type-specific and autism-associated exons. *Cell Rep.*, 42(11):113335.

Yang, Y. Y., Yin, G. L., and Darnell, R. B. (1998). The neuronal RNA-binding protein nova-2 is implicated as the autoantigen targeted in POMA patients with dementia. *Proc. Natl. Acad. Sci. U. S. A.*, 95(22):13254–13259.

Yao, J., Ding, D., Li, X., Shen, T., Fu, H., Zhong, H., Wei, G., and Ni, T. (2020). Prevalent intron retention fine-tunes gene expression and contributes to cellular senescence. *Aging Cell*, 19(12):e13276.

Yap, K., Lim, Z. Q., Khandelia, P., Friedman, B., and Makeyev, E. V. (2012). Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. *Genes Dev.*, 26(11):1209–1223.

Yeo, G., Holste, D., Kreiman, G., and Burge, C. B. (2004). Variation in alternative splicing across human tissues. *Genome Biol.*, 5(10):R74.

Yeom, K.-H., Mitchell, S., Linares, A. J., Zheng, S., Lin, C.-H., Wang, X.-J., Hoffmann, A., and Black, D. L. (2018). Polypyrimidine tract-binding protein blocks miRNA-124 biogenesis to enforce its neuronal-specific expression in the mouse. *Proc. Natl. Acad. Sci. U. S. A.*, 115(47):E11061–E11070.

Yin, Q.-F., Yang, L., Zhang, Y., Xiang, J.-F., Wu, Y.-W., Carmichael, G. G., and Chen, L.-L. (2012). Long noncoding RNAs with snoRNA ends. *Mol. Cell*, 48(2):219–230.

Yoon, J.-H., De, S., Srikantan, S., Abdelmohsen, K., Grammatikakis, I., Kim, J., Kim, K. M., Noh, J. H., White, E. J. F., Martindale, J. L., Yang, X., Kang, M.-J., Wood, 3rd, W. H., Noren Hooten, N., Evans, M. K., Becker, K. G., Tripathi, V., Prasanth, K. V., Wilson, G. M., Tuschl, T., Ingolia, N. T., Hafner, M., and Gorospe, M. (2014). PAR-CLIP analysis uncovers AUF1 impact on target RNA fate and genome integrity. *Nat. Commun.*, 5(1):5248.

You, Y., Prawer, Y. D. J., De Paoli-Iseppi, R., Hunt, C. P. J., Parish, C. L., Shim, H., and Clark, M. B. (2023). Identification of cell barcodes from long-read single-cell RNA-seq with BLAZE. *Genome Biol.*, 24(1):66.

Zakutansky, P. M. and Feng, Y. (2022). The long non-coding RNA GOMAFU in schizophrenia: Function, disease risk, and beyond. *Cells*, 11(12):1949.

Zeinelabdeen, Y., Abaza, T., Yasser, M. B., Elemam, N. M., and Youness, R. A. (2024). MIAT LncRNA: A multifunctional key player in non-oncological pathological conditions. *Noncoding RNA Res.*, 9(2):447–462.

Zeng, C. and Hamada, M. (2020). RNA-seq analysis reveals localization-associated alternative splicing across 13 cell lines. *Genes*, 11(7).

Zhan, X., Lu, Y., Zhang, X., Yan, C., and Shi, Y. (2022). Mechanism of exon ligation by human spliceosome. *Mol. Cell*, 82(15):2769–2778.e4.

Zhang, C., Frias, M. A., Mele, A., Ruggiu, M., Eom, T., Marney, C. B., Wang, H., Licatalosi, D. D., Fak, J. J., and Darnell, R. B. (2010). Integrative modeling defines the nova splicing-regulatory network and its combinatorial controls. *Science*, 329(5990):439–443.

Zhang, C.-Y., Xiao, X., Zhang, Z., Hu, Z., and Li, M. (2021). An alternative splicing hypothesis for neuropathology of schizophrenia: evidence from studies on historical candidate genes and multi-omics data. *Mol. Psychiatry*.

Zhang, F., Chen, L., Li, W., Yang, C., Xiong, M., Zhou, M., Kazobinka, G., Zhao, J., and Hou, T. (2023). Lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation is associated with tumor progression and poor prognosis of clear cell renal cell carcinoma. *Lab. Invest.*, 103(6):100125.

Zhang, N., Kaur, R., Lu, X., Shen, X., Li, L., and Legerski, R. J. (2005). The Pso4 mRNA splicing and DNA repair complex interacts with WRN for processing of DNA interstrand cross-links. *J. Biol. Chem.*, 280(49):40559–40567.

Zhang, Z., Zabaikina, I., Nieto, C., Vahdat, Z., Bokes, P., and Singh, A. (2024). Stochastic gene expression in proliferating cells: Differing noise intensity in single-cell and population perspectives. *bioRxivorg*.

Zheng, C. L., Fu, X.-D., and Gribskov, M. (2005). Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA*, 11(12):1777–1787.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049.

Zhou, X., Zhang, W., Jin, M., Chen, J., Xu, W., and Kong, X. (2017). lncRNA MIAT functions as a competing endogenous RNA to upregulate DAPK2 by sponging miR-22-3p in diabetic cardiomyopathy. *Cell Death Dis.*, 8(7):e2929.

Zhuravskaya, A., Yap, K., Hamid, F., and Makeyev, E. V. (2024). Alternative splicing coupled to nonsense-mediated decay coordinates downregulation of non-neuronal genes in developing mouse neurons. *Genome Biol.*, 25(1):162.

Ásgrímsdóttir, E. S. and Arenas, E. (2020). Midbrain dopaminergic neuron development at the single cell level: In vivo and in stem cells. *Front. Cell Dev. Biol.*, 8:463.

Çelik, M. H. and Mortazavi, A. (2022). Analysis of alternative polyadenylation from long-read or short-read RNA-seq with LAPA. *bioRxiv*, page 2022.11.08.515683.

Ødum, M. T., Teufel, F., Thumuluri, V., Almagro Armenteros, J. J., Johansen, A. R., Winther, O., and Nielsen, H. (2024). DeepLoc 2.1: multi-label membrane protein type prediction using protein language models. *Nucleic Acids Res.*, 52(W1):W215–W220.