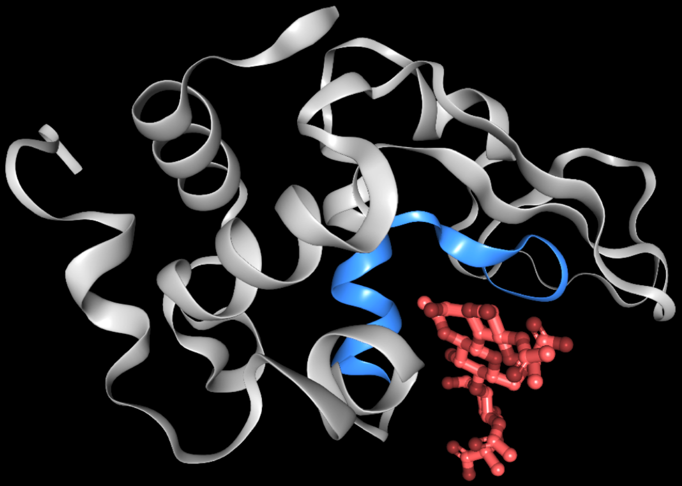
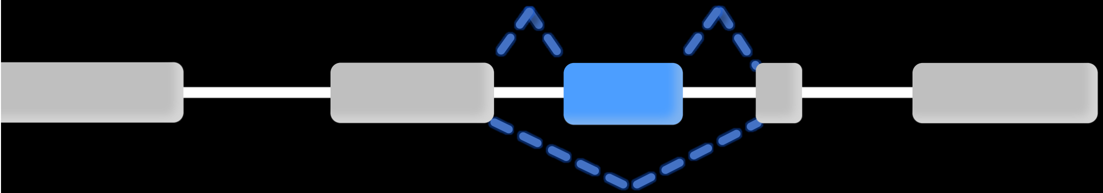
Development of a Bioinformatics Approach for the Functional Analysis of Alternative Splicing

Lorena de la Fuente Lorente

PhD Thesis





Supervisor:

Dr. Ana Conesa Cegarra

May 2019

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Development of a Bioinformatics Approach for the Functional Analysis of Alternative Splicing

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Lorena de la Fuente Lorente

Supervisor:
Dr. Ana Conesa Cegarra

A doctoral thesis submitted to

Department of Biotechnology

May 2019

Abstract

One of the most exciting aspects of transcriptome biology is the contextual adaptability of eukaryotic transcriptomes and proteomes by post-transcriptional regulation (PTR). PTR mechanisms such as alternative splicing (AS) and alternative polyadenylation (APA) have emerged as tightly regulated processes playing a key role in generating transcriptome complexity and coordinating cell differentiation or tissue development. However, how these mechanisms imprint distinct functional characteristics on the resulting set of isoforms to define the observed phenotype remains poorly understood. The number of PTR variants and their resulting range of potentially functional consequences makes their functional validation an impractical task if done on a case-by-case basis. Besides, the lack of isoform-oriented functional profiling approaches has made that much of the computational work done to elucidate transcriptome-wide functional questions has either involved ad hoc computational pipelines applied to specific biological systems or has relied on simple GO-enrichment analysis that are not informative about the PTR impact on isoform properties.

Thus, even though more than 60,000 publications on AS, a few number of existing isoforms have been associated with specific properties while the number of novel AS/APA variants with unknown and even unexplored functions is exponentially increasing thanks to the use of next-generation sequencing (NGS). Due to the technical limitations of NGS to reconstruct the transcript structure, high-throughput sequencing of full-length transcripts using third-generation technologies (TGS) is opening up a new transcriptomics era that enhances the definition of gene models and, for the first time, enables to precisely associate functional events within the RNA molecule.

This thesis addresses three major challenges to the progression of the study of isoform function. First, with the emergence and increasing popularity of TGS, the accurate definition and comprehensive characterisation of de novo transcriptomes is essential to ensure the quality of any conclusions on transcriptome diversity drawn from these data. The lack of long-read oriented quality aware analysis motivated the development of SQANTI (`https://bitbucket.org/ConesaLab/sqanti`), an automated pipeline for the structural characterization and quality assessment of full-length transcriptomes. Secondly, the gene-centric nature of functional resources remained the major limitation to the extended study of functional isoform variability, especially for novel isoforms, which cannot be characterised by static databases. Thus, we designed IsoAnnot, which dynamically constructs an isoform-resolved rich database of functional annotations by using as input transcript sequences and integrating information disseminated across several databases and prediction methods. Finally, because no methods to interrogate the functional impact of PTR were available, we developed novel approaches and user-friendly tools such as tappAS (`http://tappas.org/`), designed to facilitate researchers the transcriptome-wide functional study of context-specific isoform regulation.

Thereby, this thesis describes the development of an analysis framework that tackles the fundamental challenges of the isoform functional analysis by providing a set of novel methods and tools that offer an unique opportunity to explore how the phenotype is specified by altering the functional characteristics of expressed isoforms. Applied to a murine neural differentiation system, our pipeline profiled the effect of isoform regulation on the inclusion of several functional elements within transcripts between motor-neuron and oligodendrocyte differentiation systems and specifically, we discovered isoform-specific transmembrane regions whose modulation by PTR might contribute to control cell type-specific mitochondrial dynamics during neural fate determination.

Resumen

Uno de los aspectos más apasionantes de la transcripción es la plasticidad transcriptómica y proteómica mediada por los procesos de regulación post-transcripcional (PTR). Los mecanismos PTR como el splicing alternativo (AS) y la poliadenilación alternativa (APA) han emergido como procesos estrechamente regulados que juegan un papel clave en la generación de la complejidad transcriptómica y están asociados con la coordinación de la diferenciación celular o el desarrollo de tejidos. Sin embargo nuestro conocimiento sobre cómo estos mecanismos regulan las propiedades de los productos resultantes para definir el fenotipo es aún muy reducido. La cantidad de variantes existentes y el amplio rango de posibles consecuencias funcionales, hacen su validación funcional una tarea impracticable si se realiza caso por caso. Además, la falta de herramientas para la evaluación funcional orientada a isoformas ha provocado que gran parte del trabajo computacional haya empleado pipelines ad-hoc aplicadas a sistemas biológicos específicos o simplemente hayan confiado en análisis de enriquecimiento GO, los cuales no son informativos del impacto en las propiedades de las isoformas que hay detrás de la regulación PTR.

De hecho, a pesar de las más de sesenta mil publicaciones relativas al AS, muy pocas isoformas se han asociado con propiedades específicas, mientras que el número de nuevas variantes AS/APA con function desconocida crece exponencialmente debido a las técnicas de secuenciación de segunda generación (NGS). Además, y debido a limitaciones técnicas de las NGS para reconstruir la estructura de los transcritos, las tecnologías de secuenciación de tercera generación (TGS) están definiendo una nueva era en la que, por primera

vez, es posible conocer la secuencia de elementos estructurales y funcionales en los mRNAs.

En esta tesis se han abordado tres propósitos principales para poder avanzar en el estudio funcional de las isoformas. En primer lugar, con las TGS siendo cada vez más utilizadas, la evaluación de la calidad de los transcriptomas *de novo* es esencial para asegurar la fiabilidad de la diversidad transcriptómica encontrada. La falta de análisis de calidad orientados a secuencias largas ha motivado el desarrollo de SQANTI, una pipeline automatizado para la exhaustiva evaluación de TGS transcriptomas. En segundo lugar, la información a nivel de gen de la mayoría de bases de datos funcionales sigue siendo el principal escollo para el estudio de la variabilidad entre isoformas, especialmente en el caso de las isoformas nuevas, en las que las bases de datos estáticas impiden su caracterización. Así, hemos diseñado IsoAnnot, que construye una base de datos de anotaciones funcionales con resolución a nivel de isoformas integrando información diseminada por múltiples bases de datos y métodos de predicción. Finalmente, la indisponibilidad de métodos para estudiar el impacto funcional de la regulación de isoformas, nos ha motivado a desarrollar tappAS, una herramienta dinámica, flexible y diseñada para facilitar el abordaje de este tipo de estudios.

Por lo tanto, durante esta tesis hemos desarrollado una infraestructura que resuelve los retos principales del análisis funcional de isoformas, proporcionando un conjunto de nuevos métodos y herramientas que ofrecen una oportunidad única para explorar cómo el fenotipo se especifica post-transcripcionalmente, mediante la alteración de las propiedades funcionales de las isoformas expresadas. La aplicación de nuestro análisis a un doble sistema de diferenciación neuronal en ratón definió el efecto de la regulación de isoformas entre la diferenciación de motoneuronas y oligodendrocitos para múltiples elementos funcionales. Entre ellos, hemos descubierto regiones transmembrana que son diferencialmente incluidas en las isoformas expresadas entre ambos tipos celulares y cuya regulación podría estar contribuyendo al control de las dinámica mitocondrial.

Resum

Un dels aspectes més emocionants de la biologia del transcriptoma és l'adaptabilitat contextual de transcriptomes i proteomes eucariotes mitjançant la regulació post-transcripcional (PTR). Els mecanismes PTR, com el splicing alternatiu (AS) i la poliadenilació alternativa (APA), s'han convertit en processos molt regulats que juguen un paper clau en la generació de la complexitat del transcriptoma i en la coordinació de la diferenciació cel·lular o del desenvolupament de teixits. No obstant això, el nostre coneixement de com aquests mecanismes imprimeixen característiques funcionals diferents al conjunt resultant d'isoformes per definir el fenotip observat és encara escàs. El nombre de variants de PTR i les seues conseqüències potencialment funcionals fa que la validació funcional sigui una tasca poc pràctica si es fa cas per cas. A més, la manca d'enfocaments funcionals orientats a isoformes ha fet que gran part del treballs computacionals per esbrinar qüestions funcionals a nivell de transcriptoma siguen estratègies computacionals ad hoc aplicades a sistemes biològics específics o bé basats en un simple anàlisi d'enriquiment GO, que no aporten informació sobre l'impacte de la PTR sobre les propietats de les isoformes.

Així, malgrat les més de 60.000 publicacions existents sobre AS, poques de les isoformes existents s'han associat a propietats específiques, mentre que el nombre de noves variants AS/APA amb funcions desconegudes i fins i tot inexplorades augmenta de manera exponencial gràcies a la seqüenciació de nova generació (NGS). A causa de les limitacions tècniques del NGS per reconstruir l'estructura dels transcrits, la seqüenciació d'alt rendiment de transcrits de longitud completa mitjançant tecnologies de tercera generació (TGS) obre una nova era en la transcriptòmica, ja que millora la definició dels models genètics i, per primera vegada, permet associar amb precisió esdeveniments funcionals dins de la molècula d'ARN.

Aquesta tesi aborda tres grans reptes per a progressar en l'estudi de la funció de les isoformes. En primer lloc, amb l'aparició i la popularitat creixent del TGS, la definició precisa i la caracterització completa dels transcriptomes de novo són essencials per garantir la qualitat de qualsevol conclusió sobre la diversitat del transcriptoma. La manca d'anàlisis de qualitat orientats a lectures llargues va motivar el desenvolupament de SQANTI (https://bitbucket.org/ ConesaLab / sqanti), una estratègia computacional automatitzada per a la caracterització estructural i l'avaluació de la qualitat dels transcriptomes de longitud completa. En segon lloc, els recursos funcionals existents centrats en el gen suposen una gran limitació per a l'estudi extensiu de la variabilitat funcional de les isoformes, especialment en les noves isoformes, que no es poden caracteritzar per bases de dades estàtiques. Per tant, vam dissenyar IsoAnnot, que construeix dinàmicament una base de dades amb anotacions funcionals a nivell d'isoforma, que utilitza com a informació d'entrada les seqüències dels transcrits i integra informació de diverses bases de dades i mètodes de predicció. Finalment, com no hi havia cap mètode per interrogar l'impacte funcional del PTR, vam desenvolupar nous enfocaments i eines fàcils d'utilitzar, com ara tappAS (http://tappas.org/), dissenyada per facilitar als investigadors els estudis funcionals de transcriptoma complet i de regulació d'isoformes en contexts específics.

Per tant, aquesta tesi descriu el desenvolupament d'un marc d'anàlisi que aborda els reptes fonamentals de l'anàlisi funcional d'isoformes, proporcionant un conjunt de nous mètodes i eines que ofereixen una oportunitat única per explorar com l'alteració de les característiques funcionals de les isoformes expressades defineix el fenotip. Aplicada a un sistema de diferenciació neuronal murina, la nostra estratègia va descriure l'efecte de la regulació de les isoformes en la inclusió de diversos elements funcionals en els transcrits al comparar els sistemes de diferenciació a motor-neurona i oligodendròcits i,

concretament, vam descobrir regions transmembrana específiques d'isoformes, la modulació de les quals per PTR podria contribuir a controlar la dinàmica mitocondrial específica del tipus cel·lular durant la determinació del destí neuronal.

A mi familia y a Luisfe

Agradecimientos

Cuando comencé esta tesis nunca pensé que llegar a este momento iba a ser tan difícil. Esta tesis representa un sinfín de bonitos recuerdos pero también aúna etapas complicadas, en las que la incertidumbre y el agotamiento han sido los grandes protagonistas. Cinco años de luces y sombras que al final se traducen en esta tesis de la cual espero estar orgullosa en un futuro.

En primer lugar quería agradecer esta tesis a mi directora Ana Conesa. Sin su tiempo y confianza en mi esta tesis no habría sido posible. Cuando comenzamos esta andadura nunca imaginamos los cambios e imprevistos que íbamos a vivir duante este tiempo. De todo ello me quedo sin ninguna duda con las sabias lecciones que hemos aprendido juntas. Gracias por enseñarme tanto, en lo profesional, pero sobretodo, en lo personal.

Durante los primeros años de tesis tuve a mi alrededor a un grupo de gente excepcional, el I52 festero. Nunca olvidaré las conversaciones en el patio ingles con Cristina, Patricia, Eugenia y Mónica. Muchas gracias escucharme, aconsejarme y darme tantos bonitos momentos. Por supuesto a Rafa y Pedro, mis chicos informáticos, por aguantarme con mis dudas sin perder la sonrisa ni la paciencia. Que decir de nuestra estancia en Florida, grandes compañeros de viaje. Y por supuesto a Sonia, de la cual tanto he aprendido. Gracias por escucharme y ayudarme siempre que lo he necesitado y por darme tantos ánimos en los momentos que más lo necesitaba. Ha sido un placer y un privilegio compartir estos cinco años contigo.

Manu y Hector, mis compañeros de Florida. Juntos creamos un grupo de trabajo que nunca olvidaré. Con ambos disfruté del trabajo en grupo, de los intensos debates científicos y de esta tesis.

Muchas gracias a ambos por acogerme como lo hicísteis en Florida, por aportar tanto a esta tesis, y sobretodo, por hacer más bonito este camino.

Durante los últimos años mi falta de humor se contrarrestó con los increíbles y únicos Manu, Salva, Fran, Ángeles, Teresa, Carlos y Pedro. No imagino la última etapa de esta tesis sin Salva tocándome las narices, Fran haciéndome rabiar o Manu creando esos memes tan geniales que tantas risas nos han dado incluso en los momentos más duros. Gracias por aportarme la luz que necesitaba.

También quería agradecer a todos mis amigos, entre ellos las Golfas, los Erasmus y la Sede, la paciencia que han tenido, por no reprocharme mis ausencias, por entender mi falta de energía y aguantar mis quejas y mala cara cuando me preguntaban por la tesis. Sería dificil nombrar a cada uno de ellos.

Finalmente, pero en el lugar más destacado, se encuentra mi familia. Esta tesis va dedicada a ellos, no por nada en particular, sino por todo. A mis madre y a mi padre por su incondicional confianza, a Susana y Alberto por alegrarme tantos momentos, por visitarme en California y motivarme en todo momento, y a mi abuela por su terqueza en que acabara esta tesis. Ha costado pero aquí está.

Y dentro de esa familia, Luisfe. Esta tesis ha merecido la pena solo por haber hecho que nos encontrásemos. Muchas gracias por estar siempre a mi lado, en lo bueno y en lo malo, por enseñarme a relativizar y mostrarme las cosas que de verdad importan. No podría explicar con palabras lo importante que has sido para mi estos años. Una parte de esta tesis es tuya.

This work was funded by the following grants:

From 2014 to 2018.

FPU: Training programme for Academic Staff.

Spanish Ministry of Education, FPU2013/02348.

From 2016 to 2019.

NOVELSEQ: Novel methods for new challenges in the analysis of high-throughput sequencing data.

MINECO, BIO2015-1658-R.

From 2014 to 2017.

DEANN: Developing a European American NGS Network.

EU Marie Curie IRSES, GA-612583.

Contents

Glossary

a.u	arbitrary units
aa	amino acid
APA	alterantive poliadenylation
AS	alterantive splicing
ATI	alternative transcription initiation
AUC	area under the curve
CC	comment lines [UniprotKb]
CCS	circular consensus sequence
cDNA	complementary DNA
CDS	Coding sequence
CLiP	crosslinking immunoprecipitation
cNLS	classical nuclear localisation signal
CPM	counts per millon
CR-APA	coding-region alternative polyadenylation
DCE	differential coding sequence expression
DCU	differential coding sequence usage
DE	differential expression
DFI	differential feature inclusion
DGE	differential gene expression
DIE	differential isoform expression
DIU	differential isoform usage
DNA	deoxyribonucleic acid
DPA	differential polyadenylation
dPA	distal polyA site
DPAU	distal polyA usage
DS	differential splicing
EBI	European Bioinformatics Institute
ENA	European Nucleotide Archive

ESC	embryonic stem cell
EST	expressed sequence tags
FC	fold change
FD	functional diversity
FDR	false discovery rate
FE	functional enrichment
FET	fisher exact test
FL	full-length [sequence]
FSM	full splice match [isoform]
GEO	Gene Expression Omnibus
GLM	generalised linear model
GMST	GeneMarkS-T [model]
GO	gene ontology
GRE	GU-rich elements
GSE	gene set enrichment
GTF	gene transfer format
GUI	graphical user interface
HMM	hidden Markov model
HQ	high quality
ICE	iterative clustering for error correction
IDP	isoform detection and prediction [pipeline]
IDR	intrinsically disordered region
ISM	incomplete splice match [isoform]
KEGG	Kyoto Encyclopeadia of Genes and Genomes
KW	keyword lines [UniprotKb]
M	million
m.c	manually curated
MCF-7	human breast cancer luminal cell lines
MET	most expressed [gene] transcript
MI	minor Isoform
miRNA	micro RNA
ML	machine learning
MN	motor neuron
mRNA	messenger RNA
NC	non-canonical [splice junction]
NCBI	National Center of Biotechnology
NEURALtc	neural-differentiation time course

nFL	non-full length [sequence]
NGS	next-generation sequencing
NIC	novel in-catalog [isoform]
NLS	nuclear localization signal
NMD	nonsense-mediated decay
NNC	novel not-in catalog [isoform]
NPC	neural precursor cell
ns	not significant
nt	nucleotides
OLG	oligodendrocyte
OLIGd	oligodendrocyte differentiation [transcriptome]
OPC	oligodendrocyte progenitor cell
ORF	open reading frame
PacBio	pacific biosciences
PAS	polyadenylation signal
PbT	PacBio transcriptome
PCA	principal component analysis
PCR	polymerase chain reaction
PI	principal isoform
pPA	proximal polyA site
PSI	percentage spliced in
PSP	PhosphoSitePlus
PTC	premature termination codon
PTM	post-translational modifications
PTR	post-transcriptional regulation
QC	quality control
RBP	RNA binding protein
ReT	RefSeq reference transcriptome
RNA	ribonucleic acid
RNA-seq	RNA sequencing
ROC	receiver operating curve
RT	reverse transcription
RTS	reverse-transcriptase template switching
SJ	splice junction
SMRT	single-molecule real-time
SQANTI	structural and quality annotation of novel transcript isoforms
TF	transcription factors

TGS	third-generation sequencing
TM	transmembrane
TMM	trimmed mean of M values
ToFU	transcript isoforms: full-length and unassembled
TPM	transcripts per millon
TSS	transcription start site
TTS	transcription termination sites
UniProtKb	Uniprot knowledgebase
uORF	upstream open reading frame
UTR	untranslated region
WRS	Wilcoxon rank-sum test
ZMW	zero-mode waveguides

Chapter 1

Introduction

1.1 Alternative Splicing

”DNA makes RNA makes protein.” For many years the central dogma of molecular biology explained the flow of genetic information in this two-step process. However, we now know that, after transcription, ribonucleaic acids (RNAs) undergo a series of intertwining processes that allow the generation of multiple messenger RNA (mRNA) types from only one pre-mRNA molecule. Thereby, post-transcriptional regulation (PTR), which includes the control of splicing and polyadenylation (polyA), provides cells with a mechanism to dramatically diversity and fine-tune transcriptomes and proteomes [288].

The most well-studied PTR mechanism is the alternative splicing (AS). mRNA splicing is a highly conserved biological process in which introns from nascent RNA molecules are removed and exons are ligated to form mature mRNAs [242] [297]. The basic patterns of AS include exon skipping, the use of alternative 5’ and 3’ splice sites, mutually exclusive exons, intron retention, and alternative splicing coupled with alternative first or last exons (Figure 1.1.A). Besides these basic patterns involving the alternative use of single splicing events, eukaryotic transcriptomes can also combine several alternative events, resulting in mRNA variants with complex splicing patterns [332] [289] (Figure 1.1.B). This combinatorial control of AS provides a powerful mechanism for expanding the regulatory and functional complexity of eukaryotic organisms from a reduced number of genes. Genome-wide studies estimate that 90%-95% of multi-exon human genes undergo some level of AS, most of them resulting in mRNA variants with tissue specific expression patterns [348] [237]. However, although AS has been most frequently observed in vertebrates, it is also present in invertebrates (30% in *Drosophila melanogaster* [116][335]), plants (40%-60% in *Arabidopsis thaliana* [170][204]) and fungi (50% in *Verticillium dahliae* [154]), and so represents a widely used mechanism for the generation of molecular diversity in eukaryotes.

Alternative splicing regulation

Alternative splicing is regulated in a cell-type and developmental-stage specific manner [374][348][94]. This regulation is dictated by a finely regulated program of protein-RNA interactions that involves cis elements within pre-mRNA sequences and trans-acting factors that bind to these cis-elements [242][348] (Figure 1.1.C). Cis elements include the 5' and 3' splice sites (GU-AG dinucleotides) that define the boundary of an intron with its upstream and downstream exon, respectively, as well as the branch site (A) and polypyrimidine tract (Y(n)), both

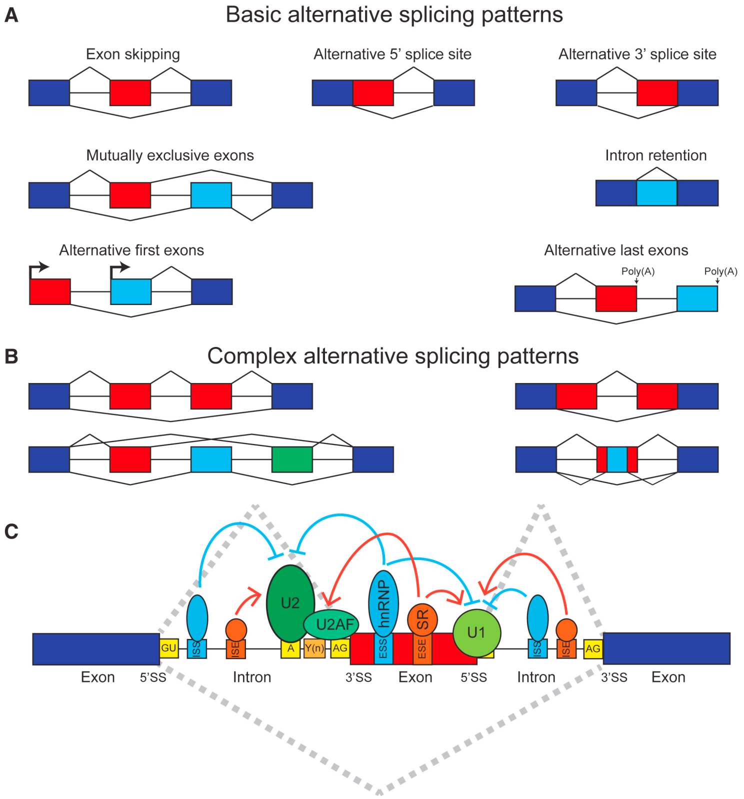


Figure 1.1: Alternative splicing. Figure adapted from Park et al. [242]. A. Basic and B. complex patterns of alternative splicing. Dark-blue boxes represent constitutively-spliced exons. Red, light-blue, and green boxes represent alternatively spliced exons. C. Alternative splicing is regulated by an extensive program of protein-RNA interactions involving cis elements within pre-mRNAs and trans-acting factors that bind to these cis elements.

located upstream of the 3' splice site [242]. These elements are all recognised by the spliceosome (the core splicing mechinery), which plays an essential role in defining exon and intron events [348]. Additionally, auxiliary cis elements in exons or flanking introns can act as splicing enhancer elements (ESEs) or splicing silencer elements (ESSs) to promote or repress exon splicing through their interaction with trans-acting splicing regulators. These include RNA-binding proteins (RBPs), whose combinatorial repertoire within pre-mRNAs determines the splicing-site choice [242][108] and whose coordinated and close regulation is essential to generate context-specific splicing programs such as those seen for the MBNL gene in differentiated cells [132], RBFOX in brain development [115] or NOVA in neurons [86].

Alternative polyadenylation

3' end processing is an essential step of eukaryotic mRNA maturation, which typically involves the cleavage of the 3' end of pre-mRNAs and addition of a poly(A) tail. A large proportion of eukaryotic genes can recognise multiple alternative polyA sites (PAS) within pre-mRNAs, a phenomenon known as alternative polyadenylation (APA) [317]. Together with AS, APA is one of the main sources of transcriptome and proteome diversity in several species. [317][385][386]. In mammalian systems, two different motifs are known to provide the signals that define PASs: (1) the AAUAAA sequence located 20-30 nucleotides (nt) upstream the cleavage site where the poly(A) is added, and (2) a GU-rich region downstream of the previous signal [21] [260]. In all cases, recognition of these sequences by specific proteins leads to mRNA cleavage and subsequent polyadenylation. However, the regulatory mechanisms governing global and gene-specific APA are only starting to be deciphered.

Depending on the location of the PAS, APA can be classified into two major categories. First, coding region-APAs (CR-APAs) are located within internal exons or introns and involve the alteration of the coding region. Thus, CR-APA generates proteins with different C terminals (Figure 5.24.B). Second, untranslated regions (UTR) APAs (UTR-APAs) are located in the 3' UTR and generate transcripts with APA but identical coding regions (Figure 5.24.A) [52][76]. APA has recently emerged as a widespread mechanism to modulate RNA transcription and fate by the generation of transcripts with alternative 3' UTRs and protein-coding potential. Context-specific global profiling studies have also illustrated how APA landscape is tissue-specific [228] and its regulation plays key roles in numerous physiological processes such as neurogenesis or cell differentiation [153][103].

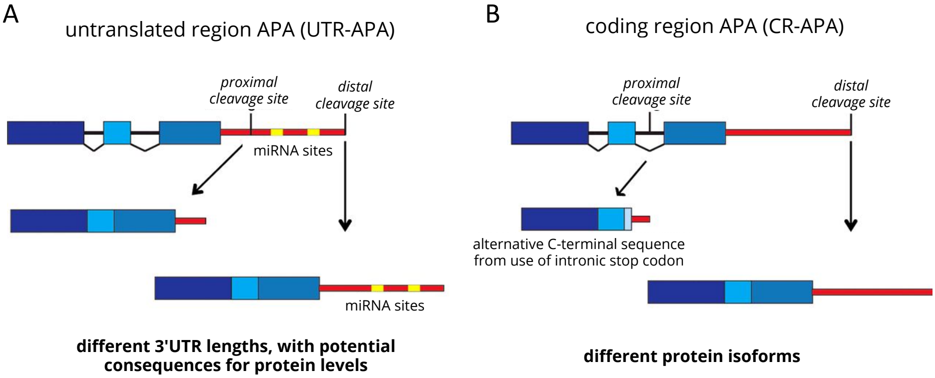

A
untranslated region APA (UTR-APA)
proximal cleavage site
distal cleavage site
miRNA sites
miRNA sites
different 3'UTR lengths, with potential consequences for protein levels
B
coding region APA (CR-APA)
proximal cleavage site
distal cleavage site
alternative C-terminal sequence from use of intronic stop codon
different protein isoforms

Figure 1.2: Types of alternative polyadenylation. Figure adapted from Hardy et al. 2016 [135] A. Untranslated region alternative polyadenylation (UTR-APA) involves the presence of more than one cleavage site within the 3' UTR whose alternative use changes the 3' UTR length. B. In coding-region alternative polyadenylation (CR-APA), the use of polyA sites upstream to the 3' UTR are used, which alters the resulting coding sequence.

1.2 High-throughput technologies to characterise transcriptome complexity

The genome-wide analysis of transcriptomes has been performed using exon microarrays first and more recently RNA-seq. The development of exon microarrays in the early 2000s allowed researchers to overcome the low-throughput constraints of previous approaches such as the reverse transcription polymerase chain reaction [251] (RT-PCR) and the sequencing of expressed sequence tags [212](ESTs) and address the quantification and characterisation of global splicing regulatory programs across many tissue types, organisms and physiological stages [184]. However, since microarrays are based on sequence hybridisation, they generate high levels of noise in expression estimates and cannot be used to discover novel splicing events.

In the late 2000s sequencing experienced a revolution because of the emergence of Next Generation Sequencing (NGS) platforms. Applied to the sequencing of RNA [220], high-throughput sequencing rapidly demonstrated its ability to study transcriptome complexity and accurately quantifying splicing events [348][237] and soon became the standard approach for transcriptome profiling. NGS platforms have evolved and some of them, such as Roche/454 and ABI/SOLiD, became quickly obsolete because of continuous improvements in the technology. The most widely used NGS sequencing platforms are currently those supported by Illumina, which are characterised by their high-throughput and accuracy as well as the short length of sequenced reads (50-200 nt). Today, the new era of Illumina platforms (e.g. the NovaSeq6000 system) can reach up to 10 billion sequenced single-reads per flow cell in a single run.

In transcriptomics, the standard procedure for transcript identification from short-reads is either to map them to the reference genome followed by computational determination of the set expressed transcripts or, to infer them by *de novo* assembly when no reference genome is available. However, even though dozens of isoform reconstruction algorithms for short-read data have been published (e.g., Cufflinks [323] for genome-guided reconstruction or Trinity[125] for de novo assembly), accurate transcript inference remains far from accurate [305]

[319]. This is mainly because short-reads do not usually span multiple splice junctions. This breaks the continuity of the transcript sequence and hinders the resolution of assembly ambiguities (Figure 1.3). Particularly complicated is the analysis of complex transcriptomes where multiple, highly similar isoforms are expressed from the same gene. Thus, despite being extremely valuable to identify and quantify individual splicing events, short-reads have serious limitations in the full-length (FL) reconstruction of expressed transcripts.

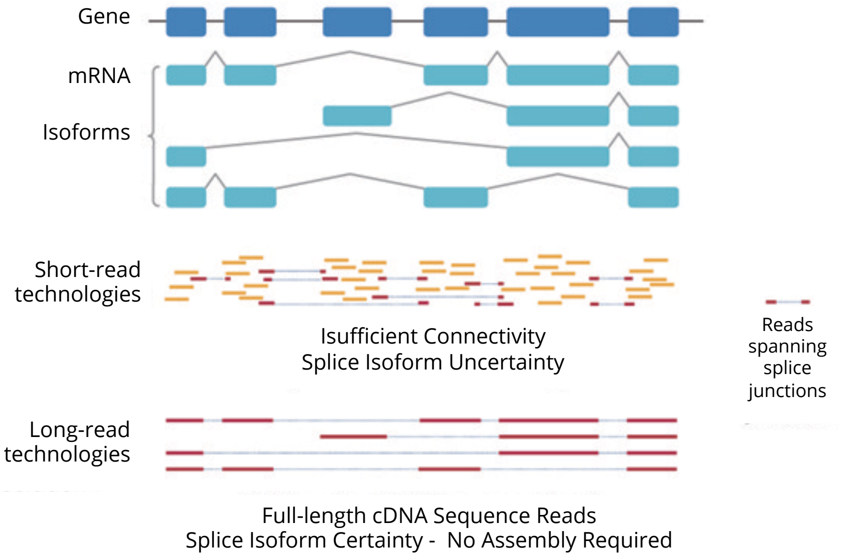

Gene
mRNA
Isoforms
Short-read technologies
Isufficient Connectivity
Splice Isoform Uncertainty
Reads spanning splice junctions
Long-read technologies
Full-length cDNA Sequence Reads
Splice Isoform Certainty - No Assembly Required

Figure 1.3: Differences between short-read and long-read sequencing approaches in the characterisation of transcriptome complexity.

To overcome these short-read limitations in the identification of FL splicing variants, Tilgner et al. developed a novel ”synthetic long-read” RNA-seq approach based on short-read sequencing [320], which became Ilumina’s TruSeq synthetic long-read technology. This protocol divides the sample into small pools containing a limited number of molecules (less than 1,000) which reduces the probability of any one pool containing variants from the same gene. Hence, *de novo* assembly of the short-reads generated from single pools greatly reduces the issues of ambiguity arising from the use of short-reads. However, the assumption that each pool contains only one RNA molecule per gene in each pool cannot be guaranteed, especially for genes with very high expression levels and so, mis-assembly and repetitive-region problems associated with *de novo* assembly algorithms remain [44].

Third Generation Sequencing (TGS) has recently emerged as a technology capable of solving most of the issues of short-read data to define transcriptome complexity. When applied to transcriptome sequencing, TGS provides both the FL combination of splicing events of each expressed molecule without requiring assembly steps, and detects variability at the 3' ends [387][51], thus facilitating a comprenhensive analysis of the alternative PTR mechanisms that generate transcriptome complexity. The most notable TGS platforms are PacBio, which was unveiled in 2010 by Pacific Biosciences [273] and the portable MinION sequencer, presented by Oxford Nanopore Technologies (ONT) in 2014 [234]. PacBio generates reads averaging around 10 kb at the expense of a higher error rate (15%) than short-reads. Nanopore technology produces even longer reads (up to a few hundreds of thousand base pairs long), but with even a lower read accuracy than PacBio.

PacBio RNA-seq (Iso-Seq) has been already used to resolve transcriptome complexity across multiple organisms including human [319], insects [112], animals [178] and plants [84][1][347][53]). However, the MinION nanopore system has been applied just in a few number of transcriptomic studies [113] [39] and, often, has been used to only characterise a bunch of genes of interest [136] [29]. To date, more than 100 publications have reported using PacBio Iso-Seq, making it the most proven and widely used technology to resolve isoform diversity at complex loci and across multiple organisms.

1.2.0.1 PacBio Iso-Seq sequencing

PacBio sequencing is also known as single-molecule real-time (SMRT) sequencing because of its ability to read individual cDNA molecules while replication is taking place. RNA sequencing (Iso-Seq) using PacBio SMRT relies on the FL cDNA sample enrichment by using the the Clontech SMARTer PCR cDNA synthesis kit, which generates around 60% of FL cDNA molecules. The bias of RSII systems towards preferentially loading smaller fragments requires sample size-fractionation by BluePippin^{TM} or SageELF^{TM} systems (Figure 1.4.A). However, the new Sequel system has a lower loading bias, which eliminates the size-fractionation of transcripts smaller than $<$4kb and simplifies the Iso-Seq sample preparation workflow.

Next, complementary DNA (cDNA) is converted into a SMRTbell library by ligation with SMRTbell hairpin adapters (Figure 1.4.C), resulting in single-stranded circular molecules that are subsequently attached to zero-mode waveguides (ZMWs) and sequenced (Figure 1.4.B). Depending on the length of the molecule, the polymerase may pass the sequence several times, generating reads that are frequently longer than the FL cDNA sequence. The sequence generated by each individual molecule pass is termed as subread while the consensus of these passes is called a circular consensus sequence (CCS) read and is computed bioinformatically (Figure 1.4.C). The FL status of a CCS requires both the presence of both 5' and 3' SMARTer primers (ligated during retrotranscription (RT)) and the polyA tail, which indicate the full-length cDNA sequencing and the 3' end mRNA completeness, respectively. Based on these signals, CCS sequences can be catalogued into FL CCSs if the primers and the polyA tail are present and into non-FL (nFL) CCSs when any of these elements is missing (nFL). Despite that, FL reads do not necessarily indicate the FL status of the transcript molecule since thse signals do not guarantee 5'-end completeness. Factors such as 5'-end degradation before RT or incomplete 5'-end RT during library preparation result in the FL classification of reads originated from incomplete 5'-end cDNAs.

The main limitation of PacBio sequencing is its high read-error rate. CCS computation increases the quality of reads by taking advantage of the multiple posible passes of the polymerase over the cDNA molecule. The shorter the insert between the SMRTbell^{TM} adapters, the higher number of full passes through the molecule will be competed, and thus, the more accurate and confident the read will be. Therefore, improvements in PacBio chemistry focus on increasing sequencing lengths (now $>$10 kb in the Sequel System) to produce CCSs with sufficient passes to eventually allow the direct determination of isoform-resolved transcriptomes from single CCS reads without requiring extra correction steps [64].

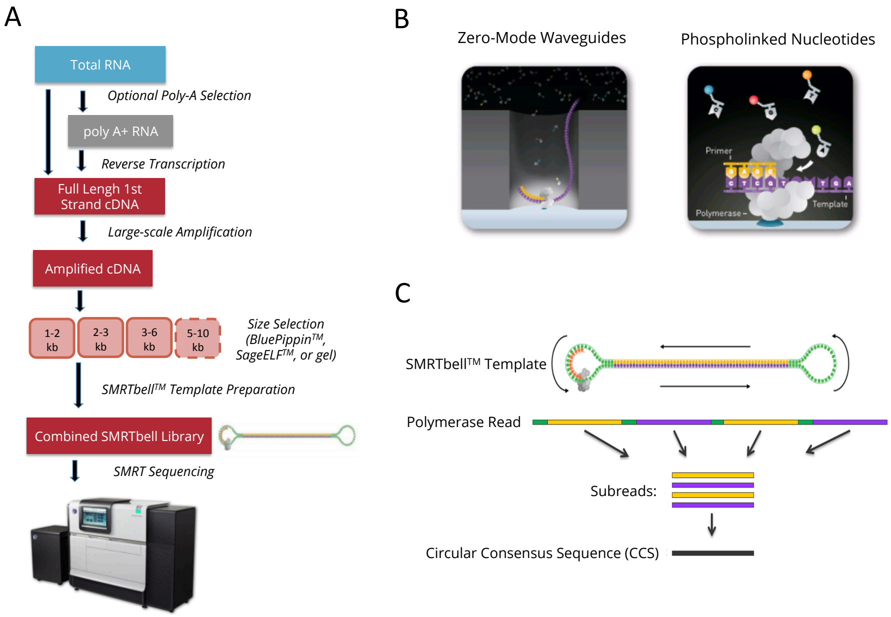

A
Total RNA
Optional Poly-A Selection
poly A+ RNA
Reverse Transcription
Full Lengh 1st Strand cDNA
Large-scale Amplification
Amplified cDNA
1-2 kb
2-3 kb
3-6 kb
5-10 kb
Size Selection (BluePippin™, SageELF™, or gel)
SMRTbell™ Template Preparation
Combined SMRTbell Library
SMRT Sequencing
B
Zero-Mode Waveguides
Phospholinked Nucleotides
Primer
Template
Polymerase
C
SMRTbell™ Template
Polymerase Read
Subreads:
Circular Consensus Sequence (CCS)

Figure 1.4: PacBio Iso-Seq sequencing. A. Iso-Seq sample preparation workflow. B. Single-molecule real-time (SMRT) templates are attached to the zero-mode waveguides (ZMWs). When a nucleotide is incorporated by the DNA polymerase, the fluorescent tag is cleaved off and monitored by the PacBio system in real time. SMRT cells contain up to a million ZMWs. C. Definition of the circular template molecule, the polymerase read and its set of subreads, and the bioinformatically generated circular consensus sequence (CCS). The DNA template is labeleld in yellow and purple and the adapters in green.

Current CCS sequences are still insufficient for the study of transcriptome complexity. This has boost the development of error-correction algorithms for long-reads, most of them taking advantage of highly accurate short-reads, such as LSC [12], proovread [130] and LoRDEC [286]. Moreover, three main pipelines have been recently developed to deliver comprehensive and non-redundant sets of high-quality FL isoforms by Iso-Seq data. These are described in the following sections:

IDP pipeline

The isoform detection and prediction pipeline (IDP) [13] was the first to appear (in 2013). It is known as a “hybrid” approach because it relies on Illumina short-reads to correct long-reads using the large-scale consensus (LSC) algorithm [12] to align short-reads to long-reads, replacing any bases that do not match. Complementary, short-reads are used to detect splice junctions by mapping them to a reference genome with the SpliceMap tool [11]. Finally, IDP module is run both to detect and predict isoforms.

Isoform are detected when an error-corrected long-read spans a reference transcript from the 5’ end to 3’ end and so, the detection step requires a reference transcriptome. Conversely, isoforms are predicted when the combination of spliced junctions is inferred by statistical modelling (using both short-read junction and long-read genome alignment information), which allows the characterisation of long transcript variants that are very difficult for PacBio to fully sequence.

Iso-Seq^{TM} Analysis

Iso-Seq^{TM} RNA isoform sequencing analysis, also known as ToFU (for ‘Transcript isOforms: Full-length and Unassembled’), was presented by PacBio in collaboration with other institutions of USA in 2015 [123]. The pipeline comprises two main steps. First, in the iterative clustering for error correction (ICE) step, FL reads are divided into different clusters by similarity: reads clustered together are highly likely to belong to the same transcript variant and so they are merged to correct randomly-distributed sequencing errors. Thus, ICE provides high-quality, low error-rate consensus sequences and dramatically reduces the number of redundant isoforms. Although ICE filters out non-FL reads because they would otherwise cause the definition of incomplete isoforms, they are used during ’Quiver polishing’ step to increase the coverage of detected isoforms and further improve their quality. Iso-Seq^{TM} is the only PacBio transcriptome-definition approach able to correct sequencing errors and generate a set of high-quality isoforms without using a reference genome.

Even though Iso-Seq^{TM} analysis can remove a high proportion of the redundancy found in the data, some sequences may still be redundant isoforms. The collapsing strategy implemented by PacBio (*Cupcake*) when reference-genome is available groups together sequences with both an identical combination of splice junctions (detected after sequence genome alignment) and 3' and 5' ends. However, while short sequence variations (>100 bp) at the 3' end are considered biological variability and maintained as different isoforms, Cupcake minimise the definition of 5'-degraded sequences by evaluating more conservatively differences at 5' end (merging sequences with missing 5' exons and or less than 5000 bp difference if same 5' end exon). For each group of redundant sequences, the longest one becomes the representative isoform.

TAPIS

TAPIS ('transcriptome analysis pipeline using isoform sequencing') was developed in 2016 as a method for the identification of FL transcript isoforms and APA events without using short-read sequencing [1]. TAPIS uses an iterative process that alternates CCS mapping by GMAP [370] and error correction based on comparison with the reference genome. During iterations, only sites that are mapped with a high level of confidence are corrected. Because mismatches detected close to the splicing sites are left uncorrected during the iterative process, alignments with gaps are eventually evaluated and filtered using SpliceGrapher [279] in order to minimise the number of false-positive splice junctions. Finally, similar to Cupcake, TAPIS eliminates read redundancy by collapsing reads according to the splice junction combination and PAS detected. Reads with identical intron patterns and 3' ends (using 15 nt as a cutoff to define the PAS) are grouped together. The largest CCS is defined as the representative transcript of the redundant group.

An overview of the wide range of existing long-read applications and the set of available bioinformatics tools and opportunities for research is reviewed in Sedlazeck et al. [292] the we refer to this resource for further reading.

1.3 Approaches for isoform quantification and differential isoform usage

Although RNA-seq accurately provides a relatively acurrate estimates of gene expression, the deconvolution of the expression of a gene into the expression of its isoform variants using short-reads is challenging because a high rate of similarity prevents the uniquely association of short-reads with individual isoforms. Several approaches using RNA-seq data have been proposed, including alignment-dependent tools such as RSEM [188], eXpress [275] or Cufflinks [323] as well as alignment-free methods as Sailfish [245], Kallisto [32] and Salmon [246]. Their comprehensive evaluation showed that RSEM and alignment-free methods performed the best, both in terms of accuracy and computational resources [387]. RSEM implements iterations of Expectation-Maximization (EM) algorithms to assign genome-mapped reads to their originating isoforms. In contrast, Salmon, Sailfish and Kallisto rely on the so-called *pseudo-alignment* concept, based on the idea that precise alignments are not required to assign reads to founder isoforms. Pseudo-alignment does compromise the accuracy of expression estimates and considerably reduces computational time and memory requirements.

Additionally, the detection of changes in the use of transcript variants between experimental conditions is key to define context-specific splicing programs. There are generally two main types of strategy for testing differences in the relative abundances of gene isoforms between conditions, depending on whether they require the estimation of isoform expression or not. The *assembly-based* (or isoform deconvolution) strategies integrate isoform expression and gene models to capture genes with differential isoform usage (DIU). Methods such as CuffDIFF2 [323] or UITA [230] use information on isoform structure to estimate the isoform expression levels that best explain the observed reads and subsequently test genes for DIU between two experimental groups. However, this strategy is usually hampered by the intrinsic limitations of short-read sequencing to correctly identify the structure and event combination of expressed isoforms and accurately quantify their absolute expression [64], a factor that remains a challenge in complex models despite the large number of existing tools [387].

In contrast to the *assembly-based* approaches, *exon-based* methods quantify single splicing events such as exons or junctions and them individually compare their relative abundances across conditions. These approaches skip isoform quantification and take advantage of the greater accuracy of short reads in the quantification of individual events, which can simply be addressed by counting how many reads map to each feature, as performed by tools such as HTSeq-count [262]. The abundance of specific splicing events is generally described as the *percentage splice-in* (PSI), which denotes the percentage of isoforms that include the splicing event (exon or junction) compared to the gene's total isoform population. Differential splicing (DS) is then estimated as the difference of these relative inclusion levels between two given conditions (ΔPSI) [348] [335]. However, this approach neither accounts for biological variability between replicates nor estimates the uncertainty of the difference. To try to tackle this limitation, the SUPPA2 method [325] monitors the uncertainly level of each observed ΔPSI value to infer the biological relevance of splicing changes. Similarly, several methods including DEXSeq [6] and DSGSeq [353] adopt a similar idea to detect differentially spliced genes based on single events but fit regression models directly onto read counts instead of using ΔPSI to determine the significance of DS. Several alternatives such as the *diffSplice* function from the limma R package [274] and rDiff [87] are also available. Even though these *exon-based* approaches accurately quantify single splicing events, they are only appropriate for studying the inclusion and exclusion of specific exons but cannot resolve the actual behaviour of transcript molecules or be used to investigate the cis-acting regulation of events within transcripts.

1.4 Functional impact of isoform regulation

Studies of isoform regulation have progressed from the evaluation of single splicing or APA events to genome-wide analysis describing global post-transcriptional patterns and their context-specific regulation. Published data indicate that post-transcriptional patterns in metazoan organisms constantly change in response to environmental stresses [255] and that the regulation of specific AS or APA programs are essential for lineage determination, cell differentiation and tissue or organ development [18]. Moreover, the functional relevance of these PTR mechanisms is further supported by the large number of human diseases that have been associated with splicing abnormalities such as mutations or dysregulation of cis-acting sequence elements, trans-acting factors or spliceosome components and include neurodegenerative disorders, the autism spectrum disorder, neuromuscular abnormalities, diabetes and cancer [18][291][58][70]. Furthermore, experimental validation of hundreds of isoforms has revealed the wide range of effects arising from the expression of alternative isoforms [165][304]. AS or APA events can modulate transcript expression levels by subjecting mRNAs to nonsense-mediated decay (NMD), impacting the function of gene products by modifying the amino acid (aa) sequence (Figure 1.5.A), or, shortening/lengthening 5' and 3' UTRs, which are essential for the regulating the mRNA fate (Figure 1.5.B).

1.4.0.1 Functional impact on protein properties

Hundreds of experimental validations of isoform variants have shown that the magnitude of AS regulation ranges from subtle functional effects to completely losses of function, as seen in apoptosis genes [333]. Protein isoforms may even acquire novel functions different from the canonical isoform of same gene [165][304]. The functional impact of protein isoforms can alter enzymatic activity by deletion of the active site or loss of the substrate binding region, both preventing product formation (Figure 1.5.A.1). Different transcription factors (TF) components can also undergo AS, producing well-studied effects in TF activity such as DNA-binding modulation (Figure 1.5.A) or transactivation domain structure alterations, influencing the activation of transcription polymerase II (Figure 1.5.A.2).

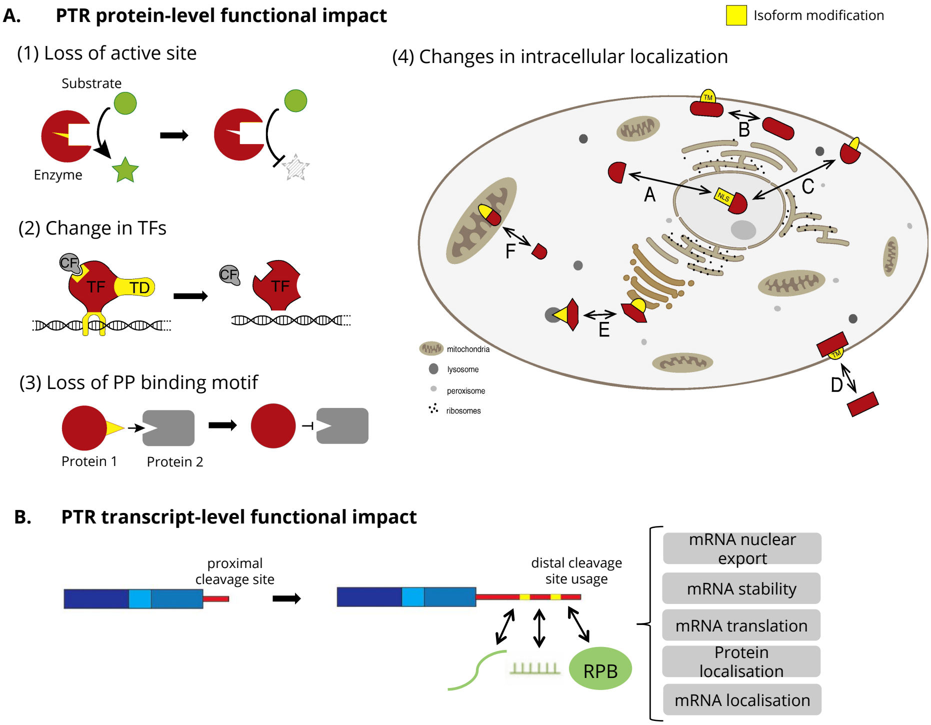


Figure 1.5: Functional impact of isoform modifications. Figure modified from Kelemen et al. [165]. A. Functional divergence of protein isoforms. AS and APA events impact proteins and lead to (1) the loss of active sites, thus altering enzymatic activity, (2) changes in transcription factors affecting the DNA binding domain, the transactivation domain or transcriptional cofactor binding, (3) the regulation of protein-protein interactions and (4) changes in the intracellular localisation of proteins. B. Functional effect of transcript isoforms. UTR regulation of transcripts involves the gain and loss of cis elements which alters mRNA stability, localisation or tranlation rates by interacting with trans elements such as RBPs, miRNAs or long non-coding RNAs.

Moreover, aside from modulating of protein properties and activation levels, intracellular protein localization is also subjected to regulation by APA and AS, what can lead to the acquisition of new functions or new protein interactions. Alteration of nuclear localisation signals (NLSs), post-translational modification (PTMs) or protein interaction sites can lead to the re-distribution of protein isoforms among different cellular compartments (Figure 1.5.A.4). Skipping of transmembrane regions can also result in the loss of protein attachment to cell membranes and the generation of soluble forms which may acquire novel functions and interaction partners (Figure 1.5.A.4).

Finally, alternative exons can encode complete or partial interaction domains, modulating interactions with other proteins. In most cases the binding affinity is modulated but not completely abolished. Similarly, binding of low molecular-weight ligands or hormones can be influenced by splicing events. A classical example of this is the insulin receptor, in which an exon skipping generates a receptor with a higher affinity to IGF-II [22] (Figure 1.5.A.3).

Even though changes caused by individual splicing isoforms interfere with almost every biological function [165][304] (Figure 1.5.A), the contribution of AS to proteome diversity and complexity remains controversial. Tress et al. claim that, although extensive AS is found in higher eukaryotes, currently available proteomics data provides little evidence that most transcript variants are actually translated into functional proteins and suggest that most detected variants are not functional [324]. In contrast, based on evidence from several studies for active translation of variants and their presence in polysome fractions [105][306] or bound to ribosomes [358], others claim that gene isoforms significantly contribute to both proteome composition and diversity [27]. Indeed, recent large-scale proteomic studies suggest that the proteome actually explains a significant proportion of RNA-level diversity [196].

1.4.0.2 Functional impact on UTR properties

Not all transcript variants necessarily result in the production of new protein isoforms. Alternative transcription initiation (ATI), AS and especially APA contribute to transcriptome complexity by generating isoforms with different UTRs which can potentially affecting mRNA metabolism or protein localisation [93] [316] (Figure 1.5.B).

Modulating the availability of post-transcriptional control elements within mRNA UTRs, such as microRNAs or RBP recognition sites, by the shortening or lengthening of UTRs can produce changes in mRNA half-life, translation efficiency, mRNA export, isoform localisation and AS regulation. Moreover, the alternative processing of UTRs can even lead to the alteration of the RNA secondary structure, which is essential for controlling the initiation of translation [131]. While 3' UTRs are preferentially regulated by APA events, 5' UTRs are modulated by AS and ATI. The most well-known effect is the inclusion of uORFs, repressive elements found within the 5' UTR of invertebrate mRNAs that mediate the translational repression of the main coding sequence (CDS) [55]. The role of UTR regulation has been highlighted in neurons. While 5' UTR length modulation influences global translation, transcripts with alternative, extended 3' UTRs accumulate regulatory sequences that are crucial to drive cell-type specific translation [25].

1.4.0.3 Nonsense-mediated decay

Changes caused by individual splicing isoforms can also regulate gene expression by triggering mRNAs to the nonsense-mediated mRNA decay (NMD) pathway [248]. NMD was originally discovered as a cellular surveillance pathway that safeguards the quality of mRNA transcripts in eukaryotic cells. Thus, in abnormal contexts NMD degrades mRNAs with premature termination codons (PTCs), preventing the accumulation of such transcripts and the truncate proteins they encode [201]. However, recent studies have shown that NMD has a much broader role by regulating the stability of many intact transcripts in order to post-transcriptionally modulate gene expression levels by altering the mRNA splicing pattern. Thus, coupled to AS, NMD has recently emerged as an expanded and conserved mechanism of gene expression regulation in natural contexts and across multiple organisms [248][201].

Chapter 2

Motivation, Aims, and Contributions

2.1 Motivation

A wealth of knowledge has been gathered about how AS and APA are regulated and sufficient evidence exists on the importance of these changes for the cellular physiology [18][52]. However, our understanding of how these mechanisms imprint distinct functional characteristics on the resulting set of transcript isoforms and lead to the observed phenotype is still very limited. Only a few transcript variants have been associated with specific properties [304][165] while the number of isoforms with unknown and even unexplored functions is exponentially increasing due to the discovery of hundreds new variants by high-throughput technologies [296][319][1][13]. Thus, even though more than 60,000 publications on AS, we still do not know the functional impact of most alternatively spliced exons or APA sites. Currently, it is clearly impractical for any single research group to individually test the differential function of all known isoforms. Even restricting the study to a particular set of genes (for example TFs or kinases), the remaining number of variants and the range of potentially functional consequences would still make this an experimentally impractical task if done on a case-by-case basis.

At computational level, some recent studies have carried out genome-wide functional studies of isoform regulation. For example, Buljan et al. revealed that the enrichment of spliced exons in disordered regions mediates new protein interactions [37]. Yang et al. and Ellis et al. showed the impact of splicing variants on the rewiring of protein-protein interaction networks in a tissue-specific manner [377][94]. Moreover, APA has been highlighted as an spread mechanism to escape microRNA regulation [206][142][26] and both 3' and 5' UTR regulation has been associated with widespread translation changes in Embryonic Stem Cells [369] and neuronal differentiation [25]. Nevertheless, much of the work done to answer transcriptome-wide questions in the functional impact of AS and APA have either involved *ad hoc* computational pipelines applied to specific organisms, biological systems and functional properties, or rely on simple GO-enrichment analysis of the set of genes regulated by AS or APA. Comparatively, a lot is already known about the functional consequences of differential gene expression patterns thanks to the widely extended use of accurate RNA sequencing technologies to measure gene expression levels and the availability of user-friendly bioinformatics tools that support the functional profiling of deferentially expressed genes for virtually any organism. However, no similar situation exists for the genome-wide functional study of differential isoform usage.

Therefore, although the wealth of data suggest that alternative splicing has important physiological functions and a rapid progress has been made in the development of tools for characterising splicing events and profiling their dynamics, strategies for interrogating alternative isoforms from a functional perspective are still lacking. In consequence, assessing if differential isoform usage is impacting specific functional features such as PTMs or NLSs, or detecting genes modulating mRNA stability by differential availability of UTR AU-rich elements are currently tasks difficult to address. Thus, the development of bioinformatics approaches facing some of the main challenges associated with the isoform analysis becomes essential to dynamically and routinely interrogate the context-specific functional effect of isoform regulation.

In this thesis we develop a new bioinformatics paradigm for studying the potential functional impact of isoform regulation based on three main pillars: the accurate definition of *de novo* isoform-resolved transcriptomes by TGS technologies (Chapter 3), the dynamic annotation of transcript and protein isoforms with rich functional information (Chapter 4) and the development of approaches that, by integration of expression dynamics and functional annotation, provide transcriptome-wide insights into the context-specific effect of AS and APA on isoform properties (Chapter 5). In Chapter 6 we describe the application of our analysis framework to a multiple time-course differentiation system that includes neural precursor cells (NPCs), motor-neurons (MNs) and oligodendrocyte progenitor cells (OPCs) in order to decipher the functional effect of isoform regulation on neural cell fate determination.

2.2 Aims

1) **To accurately define and quantify transcriptomes generated by long-read sequencing.**

   High-throughput sequencing of full-length transcripts using long-reads has paved the way for the discovery of thousands of novel transcripts and the study of transcript co-occurring events. Advances in sequencing technology have created a need for studies and tools that can characterise the isoform diversity generated by long-reads. In particular we address the following aims:

   - Comparatively assess alternative pipelines for the definition of transcriptome complexity using PacBio Iso-Seq data.
   - Develop a strategy to comprehensively characterise and describe the composition and quality of FL transcriptomes.
   - Create a bioiformatics pipeline for quality control of long-read data and curation of full-length transcriptomes.
   - Evaluate the ability of long-reads to accurately estimate isoform expression levels and predict protein-coding potential.

2) **To implement a pipeline for the extensive functional annotation of gene products at isoform-resolution**

   Even though a wide range of sources providing functional information at gene-level are available, the systematic annotation of functional properties at isoform resolution, especially in the case of novel isoforms, is one of the major limitations to address the genome-wide functional profiling of post-transcriptional regulation. Three sub-aims are considered:

   - Rich characterization of transcript and protein isoforms using sequence-based predictors annotating a wide range of functional and regulatory properties.

- Development of a strategy that overcomes disparities between databases and project experimental functional features compiled from gene-centric databases onto query isoforms.
- Adaptation of the approach to be potentially applied to any organism, independently of the transcriptome novelty rate.

3) **To develop an analysis framework to address the functional impact of context-specific isoform regulation.**

   Despite several tools exist for characterizing AS and APA events and for modelling isoform regulation, we still lack genome-wide strategies to interrogate alternative isoforms from a functional perspective. We address this challenge by focusing on the following aspects:

   - Design an approach to measure the functional transcriptome complexity resulting from APA, AS and ATI mechanisms by systematically capturing the functional feature divergence between gene isoforms.
   - Develop new approaches to study the context-specific effect of isoform regulation on gene properties by profiling the inclusion or exclusion of functional features and UTR shortening and lengthening.
   - Adaptation of methods to three different experimental designs: pairwise analysis, single-series time courses and multiple-series time courses.
   - Implementation of this analysis framework in a user-friendly software to facilitate research on isoform function to the broad scientific community.

4) **Understand the functional consequences of isoform usage regulation on neural cell fate determination.**

   Extensive work in the past few decades has focused on understanding the molecular mechanisms of neural cell fate decisions. However, the post-transcriptional landscape underlying fate determination and its functional impact remains poorly understood. In this thesis we use the developed computational approaches to study the process of neural differentiation from Neural Progenitor Cells to Oligodendrocyes and Motor neurons as a proof of principle of the power of functional isoform profiling to understand basic cellular processes. This specific aims consists of the following steps:

- Definition of neural transcriptome complexity using PacBio Iso-Seq sequencing and characterisation of functional isoform divergence.
- Characterise and profile the effect of post-transcriptional regulation on the inclusion of functional elements between the glial and the neuronal differentiation.
- Capture post-transcriptionally regulated events that could potentially generate cell-type specific functional readouts for their subsequent experimental validation.

2.3 Main contributions

During the course of this thesis I have delivered a number of contributions in the form of manuscripts, posters and talks where high-throughput sequencing technologies are used to study fundamental aspects of cell biology. Moreover, I have contributed to teaching of NGS and transcriptomics methods through participation as lecturer in courses and the direct supervison of Master students.

2.3.1 Journal papers

1. de la Fuente L, Conesa A, Lloret A, Badenes ML and Gabino R. *Genome-wide changes in histone H3 lysine 27 trimethylation associated with bud dormancy release in peach*.
   Tree Genetics and Genomes, 11(3), **2015**.

2. Ogando, J, Tardáguila M, Díaz-Alderete A, Usategui A, Miranda-Ramos V, Martínez-Herrera DJ, de la Fuente L, García-León, MJ. and Moreno MC, and Escudero S, Cañete JD, Toribio ML, Cases I, Pascual-Montano A, Pablos JL, Mañes S. *Notch-regulated miR-223 targets the aryl hydrocarbon receptor pathway and increases cytokine production in macrophages*

from rheumatoid arthritis patients.
Scientific Reports, 6:20223, **2016**.

3. Tardáguila M*, de la Fuente L*, Marti C, Pereira C, Pardo-Palacios FJ, 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, Conesa A.
   *Joint first authorship.
   SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification.
   Genome Research,28(7):1096, **2018**.

4. Martín-Expósito M, Gas ME, Mohamad N, Nuño-Cabanes C, Pascual-García P, de la Fuente L, Merran J, Chaves-Arquero B, Corden J, Conesa A, Pérez-Canadillas JM, Bravo J, Rodríguez-Navarro S.
   Mip6 maintains low levels of Msn2/4 dependent mRNAs through its interaction with Mex67 (Submitted)

5. de la Fuente L*, Tardaguila M*, Del Risco H, Tarazona S, Salmeron P, Moreno V and Conesa A.
   *Joint first authorship.
   tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. (In preparation)

2.3.2 Conferences

- HiTSeq14, 22nd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB). Boston, USA. June, 2014. “Functional Alternative Splicing Analysis Using Long Read Technologies” (Oral Communication).

- SMODIA15, Statistical Methods for Omics Data Integration and Analysis workshop. Valencia, Spain. September, 2015. “Functional Annotation of Sequenced Transcripts at Isoform Resolution” (Poster).

- NGS'16, Next Generation Sequencing Conference in Genome Annotation. Barcelona, Spain. April, 2016. "Functional Analysis of Sequenced Transcripts at Isoform Resolution" (Poster).

- JBI2016, XIII Symposium on Bioinformatics. Valencia, Spain. May, 2016. "FAIR: Functional Analysis at Isoform Resolution" (Oral Communication).

- ECCB16, 15th European Conference on Computational Biology. The Hague, Netherlands. September, 2016. "FAIR, Functional Analysis at Isoform Resolution." (Poster).

- PacBio User Group Meeting SMRT Informatics Developers Conference. Barcelona, Spain. Septiembre, 2016. "Decoding the functional and regulatory impact of alternative splicing by using Iso-seq." (Oral Communication).

- I Congreso Nacional Biomedicina Jtóvenes Investigadores en Valencia. Valencia, Spain. Octubre, 2016. "Decoding the functional and regulatory impact of alternative splicing by using long-read sequencing.." (Oral Communication).

- RNA-SIG 2017, 25th Conference on Intelligent Systems for Molecular Biology and the 16th European Conference on Computational Biology (ISMB17). Prague, Czech Republic. July, 2017. "T2GO, deciphering the functional and regulatory impact of differential splicing." (Poster).

- HiTSeq17, 25th Conference on Intelligent Systems for Molecular Biology and the 16th European Conference on Computational Biology (ISMB/ ECCB 17). Prague, Czech Republic. July, 2017. "SQANTI, extensive characterization of long read transcript sequences to remove artifacts in transcriptome identification and quantification." (Oral Communication).

- Florida Genetics Symposium. Florida, USA. October, 2017. "T2GO, deciphering the functional and regulatory impact of differential splicing." (Poster Communication).

- Bioinformatics@Valencia Meeting. Valencia, Spain. July, 2018. “TappAS: Tool for the functional analysis of alternative isoform usage.” (Poster).

- 2nd International Caparica Conference in Splicing. Lisbon, Portugal. July, 2018. “Regulation of 3’ untranslated regions along neural differentiation systems.” (Oral Presentation).

- JBI2018, XIV Symposium on Bioinformatics. Granada, Spain. November, 2018. “Bioinformatics approach to decipher the functional consequences of post- transcriptional regulation in neural differentiation systems.” (Poster).

- JBI2018, XIV Symposium on Bioinformatics. Granada, Spain. November, 2018. “tappAS: a comprenhensive computational framework for the analysis of the funciotnal impact of differential splicing”(Poster).

2.3.3 Awards

- ISCB NGS’16 Conference on Genome Annotation, Barcelona.
  F1000 Presentation Prize.
  2016

- 14th edition of the ISMB RNA-SIG meeting, Prague, Czech Republic.
  Poster winner prize.
  2017

- Florida Genetics Symposium, University of Florida, USA.
  Poster winner prize.
  2017

- 2nd International Caparica Conference in Splicing, Portugal.
  Winners of the Call for conference grants sponsored by the RNA Society and the ProteoMass Scientific Society.
  2018

2.3.4 Software

- de la Fuente L, Tardaguila M, and Conesa A.
  SQANTI, platform-independent tool.
  https://bitbucket.org/ConesaLab/sqanti

- de la Fuente L, Tardaguila M, Del Risco H, Tarazona S, Salmeron P, Moreno V and Conesa A
  TappAS, platform-independent application.
  http://tappas.org/

2.3.5 Master's Thesis Supervisions

- Alberto Manuel Lerma Aguilera.
  Consecuencias funcionales de la regulación del splicing alternativo mediado por RBPs en sistema de diferenciación neuronal.
  Master's degree in Bioinformatics, University of Valencia
  2018

- Francisco Jose Pardo Palacios
  Transcriptome reconstruction with Iso-Seq: a comparison of approaches using public data.
  Master's degree in Integrated Systems Biology, University of Luxembourg.
  2017

2.3.6 Teaching

- Biotechnology BSc (Polytechnic University of Valencia, Valencia). From 2015 to 2018 (120 hours). Lectures on "Genomics and Bioinformatics".

- Bioinformatics MSc (University of Valencia, Valencia). 2018. Lectures on "In Silico studies in Biomedicine".

I have also demonstrated the state of the art in omics analyses to the scientific community by teaching as part of different courses organised at the Centro de Investigación Príncipe Felipe [Príncipe Felipe Research Centre], in Valencia, Spain.:

- International Course of Massive Data Analysis (Centro de Investigación Príncipe Felipe, Valencia). 2014 , lecture on “ChIP-seq Analysis.

- The Genomics of Gene Expression RNA-seq course (Centro de Investigación Príncipe Felipe, Valencia). 2014 and 2015 editions , lectures on “Bedtools Visualisation”, “Transcript Assembly Quantification”, “Count Extraction”, “Full RNA-seq Analysis”.

- Multi-omic Integrative Anaylysis of Gene Expression (Centro de Investigación Príncipe Felipe, Valencia). 2017 and 2018 editions , lectures on “NGS pipelines”, “Proteomics”, “Matching omics”, “RNA-seq ChIP-seq Omics Integration”, “Hands on Multiomics Integration”.

Chapter 3

Extensive characterization and quality control of long-read sequencing transcriptomes

Chapter adapted from Tardáguila M*, de la Fuente L*, Marti C, Pereira C, Pardo-Palacios FJ, 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, Conesa A. *Joint first authorship.
SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. **Genome Research**, 28(7):1096, **2018**.

3.1 Introduction

During last years, there has been increasing interest in the use of single-molecule sequencing to characterise the transcriptome diversity generated by AS in animals and plants as this it allows direct sequencing of full-length splicing variants, eliminating the need for short-read assembly and transcript reconstruction. Two different long read transcriptome sequencing platforms are currently available: PacBio [319][296], and Nanopore [234], being PacBio RNA-seq (Iso-Seq) the technology with the highest number of publications so far.

Although PacBio Iso-Seq technology has proven useful for unravelling isoform diversity at complex loci, it suffers from a relatively high raw error rate (∼15% [42]) and has a lower throughput compared to Illumina. Different methods for transcriptome definition using Iso-Seq have recently been developed, each using different strategies and combining different sources of data to overcome the limitations of single-molecule sequencing, while leveraging its capacity to generate full-length transcripts (Further details into alternative methods in Section 1.2.0.1).

During last years, several studies have reported thousands of new transcripts accumulated in known genes by long-read technologies [296][320][13] [1] [347]. Sequencing the transcriptome of hESCs with long reads followed by IDP analysis identified over 2,000 novel transcripts (∼30%), and discovered new genes that were proven to be functional [13]. Tilgner et al. used PacBio to sequence the GM12878 cell line, and found about 12,000 novel transcripts fully supported either by previous splice-site annotations or by Illumina reads, although they did not study detected novel junctions in detail [319]. From nearly 1M sorghum PacBio long-reads, 11,342 novel transcripts (∼40% of detected isoforms) were found in combination with the application of a splice-junction quality filter (SpliceGrapher [279]); of these, 6/6 random transcripts were confirmed by PCR. Finally, a maize multi-tissue transcriptome analysis identified over 111,151 transcripts from among 3.7M CCS, most of which were novel and tissue-specific [347]. Despite the authors found that between 10% and 20% of the PacBio junctions didn't show coverage by Illumina reads and around 1% were non-canonical, they did not report the number of affected transcripts or carried out any further experimental validation. Despite most of these long-read transcriptome papers propose classification strategies to call novel genes by comparing defined isoforms to reference annotations in a intron-based mode, they lack in the description and sub-classification of the type of novelties introduced by transcripts not matching the splice pattern of annotated references. None one of these studies performed any in-depth characterisation of these novel transcripts and junctions that could have revealed any potential biases and would have justified their analysis strategies. Thus, implementing a comprehensive, quality aware analysis of single molecule sequencing is fundamental at a time when long read methods are becoming more popular and important conclusions on transcriptome diversity can be drawn from these data.

In this chapter, we compare different strategies for transcritpome definition using long-read technologies and define an analysis framework for generating curated transcriptomes at isoforms resolution (Figure 3.1). Moreover, we develop SQANTI (Structural and Quality Annotation of Novel Transcript Isoforms), a pipeline that maximise the analytical outcome of long-read technologies by providing the tools which can deliver quality-evaluated and curated full-length transcriptomes. SQANTI was implemented as a open source software, and is available at https://bitbucket.org/ConesaLab/sqanti. SQANTI has been applied to multiple organisms and long-read sequencing platforms that will be presented and briefly discussed during this chapter.

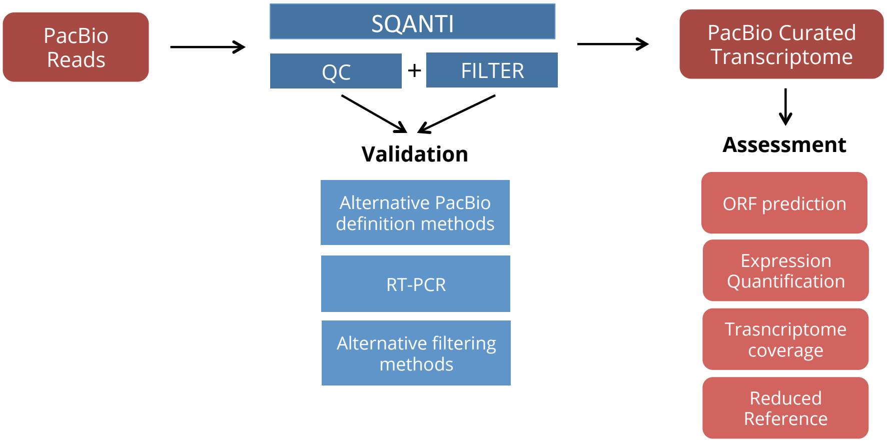


Figure 3.1: Chapter 1 analysis workflow. An strategy of analysis and quality control of PacBio Iso-Seq data was implemented to characterise and asses the results provided by alternative PacBio definition pipelines. Based on several quality attributes, we designed an approach to filter artefactual isoforms. Isoform validations by RT-PCR were performed in order to evaluate filtering perfomance and compare it to alternative strategies described in literature. We evaluated open reading frame prediction on PacBio curated transcriptomes as well as the accuracy of PacBio data to quantify and capture low expressed isoforms. Finally, we assessed the impact of using a reduced and specific transcriptome on the performace of short-read based isoform quantification methods. The SQANTI tool was implemented to automatically carry out the quality control and curation of long-read based transcriptomes.

3.2 Data

A murine neural differentiation system was used to assess long-read sequencing transcriptomes and to develop pipelines for the quality-control and characterisation of transcripts. We chose this system due to the extensive splicing program repeatly identified during brain development in mammals. In addition, we also used public datasets from long-read technologies and from different organisms to validate the use of the pipelines we developed in subsequent analyses.

3.2.1 Neural System in mouse

Experimental design

Neural precursor cells (NPCs) were isolated from the subventricular zone of 4-day old c57/BL6 mice and were cultured as neurospheres in media supplemented with EGF and bFG. Oligodendrocyte progenitor cells (OPCs) were generated *in vitro* from NPCs by adding all-trans retinoic acid (ATRA) to the culture medium, as previously described by Keirstead et al. [164]. To account for biological variability, these In vitro differentiation assays were performed in 2 biological replicates at the Neuronal and Tissue Regeneration Lab headed by Victoria Moreno.

Library preparation and sequencing

Total RNA was transcribed using the Clontech SMARTer cDNA synthesis kit which, unlike commonly used cDNA synthesis methods, enriches the full-length cDNAs contained in the final sample. Full-length cDNA from NPCs and OPCs, two biological replicates each, was obtained and split to prepare Illumina and PacBio sequencing libraries.

Iso-Seq libraries were sequenced on the PacBio RS II platform using the P4-C2 chemistry. To prevent the preferential sequencing of shorter transcripts caused by loading bias, we used a BluePippin device to produce three transcript-size fractions (1–2 kb, 2–3 kb, 3–6 kb). A total of 8 SMRT cells per sample were sequenced (1–2 kb: 3 SMRT cells, 2–3 kb: 3 SMRT cells, 3–6 kb: 2 SMRT cells) following the Iso-Seq PacBio sequencing protocol.

We also sequenced the same samples using the Illumina Nextseq instrument in combination with Nextera tagging and 250 paired-end sequencing, which yielded around 60M single-end reads per sample. PacBio sequencing was performed at the ICBR sequencing facility at the University of Florida and Illumina data was generated at the University of California Irvine.

3.2.2 Public datasets

3.2.2.1 Maize PacBio

Maize, specifically the B73 cultivar, is a well-studied crop of agricultural relevance. Wang et al. performed deep-sequencing of in six different maize tissues: root, pollen, embryo, endosperm, immature ear, and immature tassel [347]. PacBio sequencing was performed using RS II platform with P6-C4 chemistry and 47 SMRT cells. Tissue-specific barcodes were added before pooling for amplification and size-fractionation ($<$1, 1-2, 2-3, 3-5, 4-6 and $>$5 kb) was performed using a SageELF device before sequencing. We selected ear tissue for the purposes of this Chapter as Wang et al. reported a high level of ear-specific splicing variants. SRP067440 and E-MTAB-3826 are the accession numbers for Iso-Seq and Ilumina data, respectively.

3.2.2.2 MCF-7 Human PacBio

Pacific Bioscience has made different datasets available to the scientific community (https://github.com/PacificBiosciences/DevNet/wiki/IsoSeq-Human-MCF7-Transcriptome). Among them, the MCF-7 human breast cancer cell line has one of the highest sequencing depths. We used the most recent release (from 2015) of this dataset consisting of 28 SMRT cells. MCF-7 was sequenced by using P5-C3 chemistry and sizing was performed by using the SageELF platform (fractions: 1-2 kb, 2-3 kb, 3-5 kb, and 5-10 kb).

Illumina reads were not available from the same biological original material. However, numerous Illumina datasets have been generated for this stable immortal cell line and we used the short-read datasets (SRX426377) published by Schueler et al. [290]. Of note, Weirather et al. [359] used this combination of PacBio and Illumina MCF-7 datasets to develop a PacBio-based fusion transcript discovery tool.

3.2.2.3 B-cell mouse nanopore

To evaluate the ability of Nanopore technology to identify and quantify isoforms in complex gene models, Byrne et al. performed single-cell sequencing using the MinION Nanopore sequencer [39]. Libraries from seven FACS-sorted B1a cells were generated in a multiplexed manner following the ONT library preparation protocol. ONT reads were processed using the Metrichor cloud platform 2D wokflow and subsequently aligned to the genome using BLAT software [166]. Reads from cell number 1 (SRA accession number SRR4048177) were downloaded and used to assess out pipeline on Nanopore data.

3.3 Methods

3.3.1 Transcriptome definition using Iso-Seq PacBio long-reads

In this work we comprehensively evaluated and compared three alternative pipeli– nes for the definion of Iso-Seq transcriptomes in order to pinpoint their strengths and weaknesses. Before running these alternative tools, we performed several raw PacBio data preprocessing steps using functions in the PacBio Iso-Seq bioinformatics toolkik to trim primers, generate Consensus Circular Sequences (CCS) and evaluate CCS full-lengthness.

Regarding CCS generation, we set a minimum predicted accuracy of 0.8 (Iso-Seq Analysis predicted accuracy of a read - ranging from 0 to 1) and a minimum number of 0 full-passes as parameters. This latter setting meant that all the ZMWs produce a CCS, even if the polymerase did not replicate the entire insert located between the two SMRTbell adapters. However, in specific subsequent evaluations, we increased the number of full-passes to 1, retaining consensus sequences only for the molecules that were entirely sequenced. We used default parameters to classify CCSs into full-length (FL) and non-full length (nFL) sets, remove chimeral sequences and trim the SMARTer primers. Moreover, we also implemented proovread [130] to correct PacBio read errors by short-read data. The error rate decrease was assessed by identifying the number of mistmaches and indels contained in the proovread output sequences after reference genome alignment by GMAP [370]. We used preprocessed Iso-Seq data as the input for the three different isoform-definition pipelines considered in this work: IDP [13], ToFU [123] and TAPIS [1].

The ToFU pipeline [123] was used to generate the set of consensus isoforms, specifiying the Quiver polishing option. high-quality (HQ) polished isoform sequences were aligned to the reference genome (mm10) and subsequently collapsed with the cDNA_Cupcake collapse function (`https://github.com/Magdoll/cDNA_Cupcake/wiki`) to remove isoform redundancy. Because 5' end completeness cannot be verified, stringent parameters (1000 bp) and a 5' merge option were used to avoid the definition of false-positive alternative TSSs. We maintained the 100 bp cutoff at the 3' end to define isoforms with the same polyA site. This resulted in the ”ToFU” set of non-redundant full-length transcripts.

For IDP pipeline [13], we used the SpliceMap aligner [279] with the default mammalian parameters to map short reads to the reference genome, thus allowing splice junction detection. In addition, we error corrected the CCS long-reads with short-read data by using the LSC algorithm [12], following the IDP guidelines. Finally, we input splice-junction coverage information, short-read corrected long-reads and both the reference genome and the murine RefSeq [235] gene models into the IDP tool. We used the following parameters to obtain the set of predicted and detected isoforms: GMAP as the aligner, maximum posteriori probability as the MAP expression-estimator approach, and a minimum isoform fraction of 0.05.

Finally, we ran TAPIS using default parameters except that only the set of full-length classified long-reads were provided in order to minimize the number of detected incomplete isoforms. Murine RefSeq gene models were the input for the TAPIS collapsing step using the default parameters. In all cases, the mm10 reference genome assembly was used.

3.3.2 Iso-Seq PacBio evaluation of isoform quantification and detection

In addition to accurate identification of transcript sequences, accurate expression level estimation of splicing variants is essential to study their role and relevance. Since transcript isoform quantification by Illumina is limited by the high levels of ambiguity generated by short reads during isoform identification, we evaluated whether PacBio reads could be used to quantify the expression of transcripts. Isoform expression using long and short read data was evaluated by computing pairwise correlations between sequencing replicates. We also separated the transcripts into high, medium, and low expression levels to account for the influence that different noise levels (associated with high and low isoform expression) could have on this correlation. Isoforms estimations on short-reads were computed using the ENCODE3 pipeline consisting in the combination of

STAR [83] as mapper and RSEM as quantification algorithm [188]. Isoform expression estimations using PacBio reads alone were estimated by extracting the number of full-length reads associated with each defined isoform, and normalizing the values with the total number of FLs in the sample.

Secondly, to evaluate how the magnitude and nature of the transcriptome affects quantification, we compared expression levels obtained using as reference either the complete mouse transcriptome (ReT) or the set of transcripts identified by PacBio (PbT). We defined the most expressed gene transcript (MET) as the gene transcript with the highest average TPM value across all the samples and compared METs between quantification results using the ReT and the PbT transcriptomes. This analysis evaluates if quantification on PacBio transcripts would have a minor or a significant impact in the redistribution of gene expression across isoforms.

3.3.3 Classification of transcripts to describe long-read captured novelty

To characterize the nature and magnitude of the novelty found by long-read sequencing, we developed a classification scheme that capture the range and main characteristics of novel calls. This scheme compares identified transcripts to reference annotations in a splice-junction (SJ) based mode (Figure 3.2).

PacBio transcripts matching a reference transcript at all splice junctions were labelled as *full splice match* (FSM), while transcripts matching consecutive, but not all, splice junctions of the reference transcripts were designated as *incomplete splice match* (ISM). Besides, ISM isoforms were divided into different subcategories depending on their type of incompleteness (*3' end incomplete, 5' end incomplete, internal fragment*). Moreover, ISM transcripts with 95% or more of their sequence within the UTR3 sequence of their cognate reference transcript are labeled *UTR3 Fragment*. Monoexonic transcripts matching a monoexonic reference were included in the FSM category whereas those matching a multiexonic reference were placed into the ISM group (Figure 3.2).

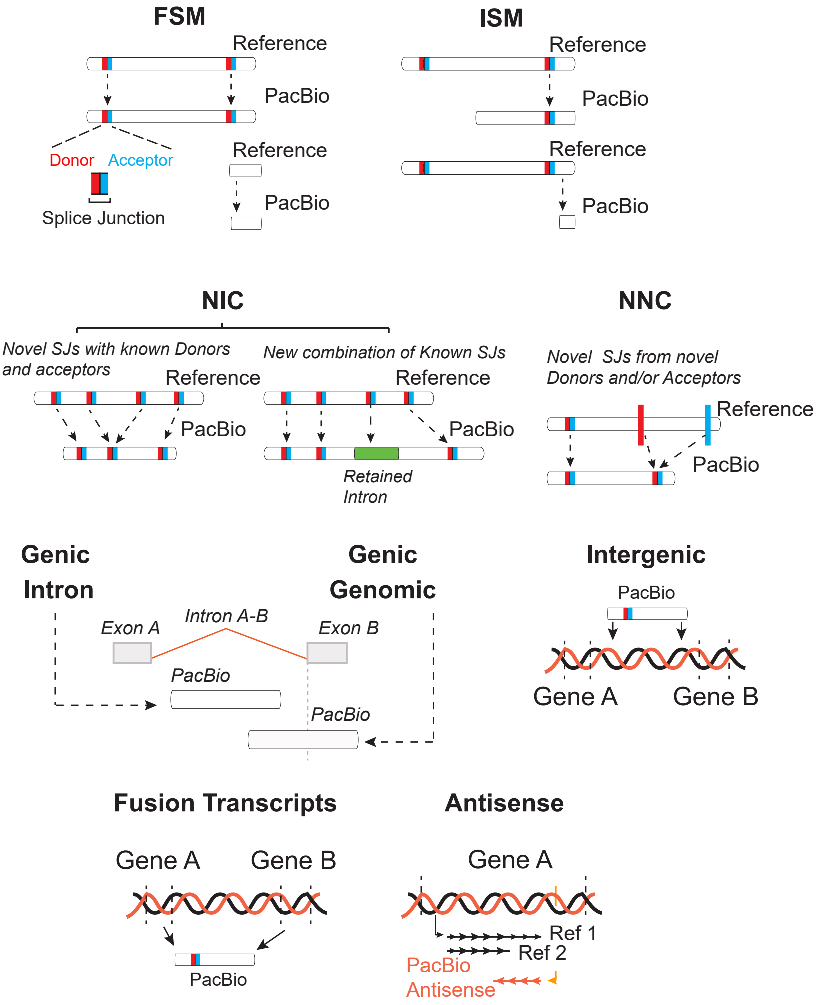


Figure 3.2: Structural classification of transcripts according to their SJs and donor and acceptor sites relative to the reference transcripts. Splice donors and acceptors are indicated in red and blue, respectively.

Furthermore, novel transcripts overlapping reference genes were classified into two categories: *novel in catalogue* (NIC) and *novel not in catalogue* (NNC). NIC transcripts contain new combinations of already annotated JSs in the associtated reference gene or novel SJs formed from already annotated donors and acceptors (NIC subcategory). On they contrary, NNC transcripts contain donors and/or acceptors not previously seen in the reference-gene annotations (Figure 3.2).
Finally, transcripts in novel genes were classified as *intergenic*, if lying outside the boundaries of an annotated gene, and as *genic intron* if lying entirely within the boundaries of an annotated intron. In addition, the *genic genomic* category encompasses transcripts with partial exon and intron/intergenic overlap in a known gene (Figure 3.2). Finally, we labeled transcripts as *fusion* if they span two annotated loci, and as *antisense* when poly(A)-containing transcripts overlap the complementary strand of an annotated transcript (Figure 3.2). In addition to classification, which is based on SJs, we also added other features to facilitate the identification of novel alternative polyadenylation sites (PASs) or fragmentation levels at the ends of transcripts. Hence, the minimum distance of transcript 3'/5' ends to the reference transcript 3'/5' ends were annotated.

3.3.4 Extensive isoform characterisation as a means for quality control

To capture different sources of artefacts, from library preparation to data preprocessing, we defined a quality-evaluation strategy for long-read transcripts based on the definition of a wide range of attributes to characterise several aspects of isoforms and their associated SJs. Appendixes 1 and 2 show the total set of defined features. Among them we highlight:

1. **SJ status:** We categorized *canonical junctions* as those with the combination of GT at the beginning and AG at the end of the intron as well as GC-AG and AT-AC pairs, which together represent more than 99.9% of all human introns [240]. Any other possible combination is labelled as *non-canonical splicing*.

2. SJ support: The number of uniquely-mapping short-reads at each defined junction was estimated using the STAR aligner [83]. Furthermore, to account for the relative expression level of transcripts, we defined *the relative coverage of a splice junction* as its short-read coverage normalised to the total expression of all the transcripts in which it is present. We summarised the supporting junction information for each transcript by defining the *minimum sample coverage*, as the lowest number of samples showing the presence of a SJ from among all the junctions present in a transcript; the *minimum coverage* was the lowest number of short reads supporting the presence of a SJ within the transcript and the *minimum coverage position* was the position of the junction in the transcript with the minimum coverage.

3. Reverse transcriptase template switching: An RNA with two direct repeats is a potential template for reverse transcriptase swithching (RTS) from one repeat to other. This causes gaps during cDNA synthesis [60][145], that when sequenced result in false transcript detections. These gaps are enriched for non-canonical junctions. To detect this problem, we developed an algorithm that identifies junctions that are likely to derive from a RTS event. The algorithm analyses all the junctions for possible RTS event and checks for a direct repeat pattern at the end of the SJ 5' exon which must match the pattern at its 3' end SJ intron. Three parameters control pattern matching: (1) the minimum number of matching nucleotides (4-10 nucleotides); (2) the amount of wiggle room allowed from the ideal pattern location (0-3 nucleotides); (3) the allowance of single mismatch or indels or not. In this Chapter, we used repeat sequences at least 8-bases long, allowed a maximum wiggle of 1, and did not permit any mismatches. We assumed that the FSM transcripts with the highest mean expression in each gene would act as templates for RTS and were therefore excluded from the analysis.

4. Bite junction: We defined junctions whose associated intron completely overlaps an annotated intron and that partially overlaps the 3' and 5' annotated exons as bite junctions.

5. FL isoform coverage: The amount of FL support is representative of the confidence level in an isoform. FL-count information is provided through the ToFU pipeline and therefore, are only available for ToFU-defined isoforms.

6. Intra-priming/off-priming: We also evaluated possible off-priming of the oligo(dTs)in A-rich regions of the mRNA template to account for internal poly(A) priming during reverse transcription [221]. To investigate these events, we calculate the percentage of adenine (A) within a established window downstream of the genomic coordinates corresponding to the 3' end of the long-read defined transcripts. We set a window size of 20 nucleotides.

3.3.5 Using quality control features to build a filter of isoform artefacts

We developed a machine learning (ML)-based filtering method to discriminate potential transcript artefacts from true novel transcripts. The approach takes advantage of the total set of long-read quality control (QC) attributes previously defined for quality evaluation. To make the classifier generally applicable and independent from the availability of validation results, we defined a "best guess" of true (positive set) and artefact (negative set) transcripts based on the information obtained from the classification of the long-read sequencing transcriptome. FSMs (whose splicing patterns are identical to the reference ones) were used to define the set of positive transcripts while NNC-NC transcripts (which contain novel and non-canonical SJs) were used as the negative set. It is important to mention that the labelled sets (FSM and NNC–NC) only applied to multi-exonic transcripts and hence, this classifier cannot be applied to single-exon isoforms.

The labeled data was then separated into to sets: the training set for classifier training (80% of the data) and the test set for subsequent evaluation of the classifiers (20% of the data). Algorithms were ran using down-sampling to balance the positive and negative sets and applying a 10x cross-validation. From the total set of transcript descriptors previously defined, we selected 16 variables. Attributes related to reference transcripts or those given a structural classification were removed because they discriminate between novel and known transcripts and consequently, are irrelevant to the classifier. Variables related to canonical junction status were also excluded because they were used to define the positive and negative transcript sets. Finally, variables with near-zero variance or with a correlation coefficient higher than 0.9 in the labeled sets were also removed.

After evaluating different ML methods, we selected the tree-based Random Forest learning method because it performed the best in our pilot tests (data not shown). We ran 500 trees and artefacts were defined as those transcripts with a probability for positive classification exceeding 0.75. The performance of this ML classifier was evaluated in the test set by using ROC analysis.

Moreover, we evaluated alternative single-feature filtering methods described in the literature and our ML classifier using a set of 67 PacBio-defined transcripts that were validated by reverse transcription PCR (Further details in Section 3.4.4). The confusion matrix shown below 3.1 illustrates the potential classification errors.

		Actual values	
		Positives P	Negatives N
Predicted values	Positives P'	True Positives TP	False Positives FP
	Negatives N'	False Negatives FN	True Negatives TN

Table 3.1: Confusion matrix for two-class classifications

The following performance indicators were used:

- **Recall**, also known as the sensitivity or the true positive rate (TPR):
  $$recall = \frac{TP}{TP + FN}$$
- **Precision**, also known as the positive predicted value:
  $$precision = \frac{TP}{TP + FP}$$
- **The false discovery rate** (FDR), the percentage of FPs from the total number of detections:

$$FDR = \frac{FP}{TP + FP} = \frac{FP}{P'}$$

- **The F1-score**, the harmonic mean of the recall and precision parameters. $$F1score = \frac{2TP}{2TP + FP + FN} = 2.\frac{recall.precision}{recall + precision}$$
- **The receiver operating characteristic curve** (ROC) is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold varies. The area under the curve (AUC) measures the classifier's ability to distinguish classes. AUC ranges between 0 and 1 [101].

We excluded the transcripts assessed by RT-PCR from the training set to prevent biases during the evaluation of the ML classifier.

3.3.6 Open reading frame prediction benchmarking and assessment of UTR/ORF variability in PacBio-defined transcriptomes

The GeneMarkS-T algorithm (GMST) [30] was used to predict the open reading frame (ORF) in PacBio transcripts by using AUGs as the initial codon. Because GMST can predict ORFs in incomplete transcripts, incomplete 5' transcripts may produce some truncated ORFs. In these cases, the first in-frame downstream methionine was detected and identified as the start codon.

We defined different event that characterise the changes between ORF sequences in order to benchmark the ORF prediction algorithms and to study upstream ORF (uORF) variability in PacBio-defined transcriptomes. Microexon definition was restricted to novel amino-acid (aa) stretches obtained by in-frame indels or to substitutions up to 27 nt (9aas) according to previously published work [149]. N-terminal or C-terminal deletions were labeled as *N-Ter deletion* and *C-Ter deletion*, respectively. Indels and substitutions greater than 9 aas, whether combined with N-Ter and C-Ter deletions or not, were labelled as *major changes*. Finally, ORFs without sequence overlapping were deemed as *no align ORFs*.

To assess the coding prediction results, we performed comparisons between the predicted sequence and its reference ORF sequences for isoforms belonguing to reference-associated catergories (FSM, ISM and 3'UTR fragment). To study to what extent alternative splicing affects the coding region of novel transcripts, we compared the predicted ORFs with the ORF encoded by the principal isoform (PI) of each gene. PIs were defined based on the information retrieved from the APPRIS database [277] - which defines the PI ORF as the ORF isoform with the highest functional load and cross-species conservation. A non-redundant ORF database encompassing the set of predicted proteins from our neural transcriptome was generated for subsequent classification into three groups: (1) Principal Isoform ORF if annotated as such by APPRIS, (2) Alternative ORF if found in Ensembl or RefSeq databases without being the PI, and (3) Novel ORF if the protein was present only in our mouse PacBio data. Alternative and Novel ORFs sets were then compared based on the gene PI ORF. Finally, UTR variability was also evaluated by considering UTRs to be different if they started in different genomic coordinates or if they shared a common start point but had a difference longer than 30 nt.

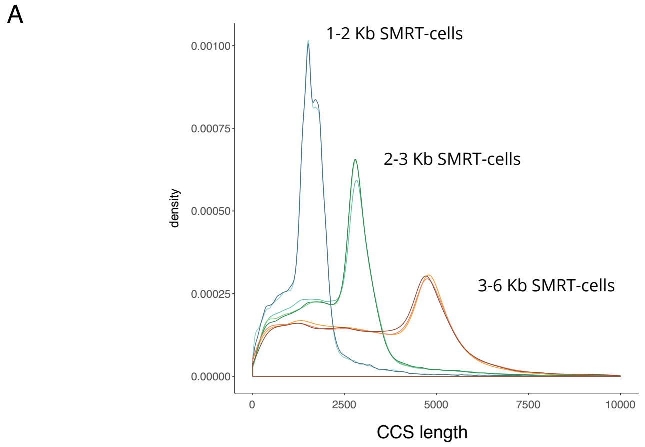
3.4 Results

3.4.1 PacBio Iso-Seq sequencing quality

We used PacBio to sequence both Neural Progenitor Cells (NPCs) and Oligodendrocyte Precursors Cells (OPCs), two biological replicates per condition, in 8 SMRT cells each. SMRT cells were pre-processed independently obtaining about half a million of circular consensus sequences (CCS) per sample, where length distribution matches the expected fractionation pattern (Figure 3.3.A). FL classification yielded 544,184 FL-catalogued CCSs, 25% of the total CCSs that decreased to 11% in 3-6 kb fraction (Figure 3.3.B). This low rate of FL PacBio-catalogued transcripts could be the result of degradation of cDNA during library preparation and pre-processing or due to incomplete sequencing. To further investigate this, we ran CCS computing and FL classification tools by changing the minimum number of passes to 1, thereby preventing the generation of CCSs when the sequencing did not reach a full pass of the molecule. We found a dramatic decrease on the number of CCSs (30% of total CCSs obtained with a 0-passes setting, Table 3.3.B) and a concomitant increase in the proportion of FL reads (74% of FL reads, Figure 3.3.B). Hence, these results revealed that the high rate of nFL reads in our samples was a consequence of incomplete molecule sequencing instead of caused by cDNA degradation during library preparation. To further check the quality of RNA before RT and library preparation as well as confirming the true full-length status of isoforms, we run Blast [4] against RefSeq reference. We found about 60% of FL CCSs with at least a 90% coverage hit, that agrees agreement with the expected FL enrichment levels of Clontech protocol in FL cDNA and verifies the high quality of our FL reads.

3.4.2 Transcriptome complexity and transcript full-lengthness assessment across alternative pipelines

Once the quality of the raw Iso-Seq data was evaluated, the PacBio CCSs were pooled together to obtain a total of 544,184 FL and 1,427,731 nFL reads. The PacBio reads were then processed by three alternative Iso-Seq transcriptome



B

Sample	# Passes	# CCS	# FL non Chimeric	# FLnc with 90 % coverage hit
NSC1	0	380307	98989 (26.03 %)	54431 (55 %)
NSC2	0	602546	157558 (26.15 %)	82062 (53 %)
OPC1	0	553031	167320 (30.26 %)	96124 (59 %)
OPC1	1	178619	131986 (74 %)	-
OPC2	0	511642	120317 (23.52 %)	69632 (59 %)

Figure 3.3: Raw Iso-Seq data quality. A. CCS length distribution in function of the fractionation size of each sequenced SMRT cell. B. Number of CCS, FL non chimeric yield and proportion with Blast identity above 90% for each of samples being studied. One OPC replicate was analysed twice using alternative values for the number of passes required to generate a CCS.

definition pipelines: ToFU, IDP and TAPIS, which identified 16,106, 13,525 and 91,428 isoforms, respectively, indicating that a huge difference in the magnitude of the transcriptomes is obtained by the alternative strategies.

To understand the nature of these differences, we performed an extensive characterisation of transcriptome composition, as described in methods section.

ToFU transcritome characterization

The ToFU pipeline generated 33,635 high-quality, but redundant isoforms. Subsequent alignment against the reference genome (assembly mm10) and collapsing in order to remove redundancy generated a final murine neural transcriptome, which included 16,106 unique transcripts resulting from the expression of 7,704 different genes. Classification of ToFU isoforms showed that a small proportion of isoforms fell outside the boundaries of known genes (640 isoforms belonging to 511 novel genes, 6% of the total number of expressed genes). Moreover, in terms of splicing diversity, we found remarkable differences between known and novel gene models. While novel genes showed low levels of alternative splicing, 50% of known genes had at least two splicing variant (Figure 3.4.A). Furthermore, only 13.8% of the novel isoforms had SJs, indicating that mono-exon isoforms were clearly enriched in novel genes.

Based on the structural classification of isoforms, 49% of them were classified as FSM (Figure 3.4.B) and the total transcripts mapping a known reference (FSM, ISM and UTR3-fragment) accounted for 60% of the transcriptome. ISM transcripts might be a combination of biological shorter versions of long reference transcripts and partial fragments resulting either from incomplete retrotranscription or mRNA decay. Our analysis showed that PacBio transcripts classified as ISM or UTR3-fragment matched reference transcripts that were longer (t-test, p = 0, Figure 3.4.C) and had more exons (t-test, p = 0, data not shown) than FSM sequences, suggesting that they are enriched in retrotranscription fragments. Novel transcripts assigned to known genes (NIC, NNC) made up 35.6% of our sequences (Figure 3.4.B), a proportion higher than expected in well-studied model organisms as mouse. Transcripts from novel genes (Intergenic and Genic Intron categories) represented about 2.3%. Other categories as antisense and fusion isoforms only accounted for 1.1% and 0.4%, respectively (Figure 3.4.B). Although most isoform categories had a similar median length (Figure 3.4.D), genic intron transcripts were found significantly shorter (t-test p = 1.421e-15), and almost entirely composed of single exons (data not shown) without any predicted coding sequence (Figure 3.4.B), suggesting the accumulation of non-coding transcriptional events.

Regarding expression features across the transcript categories, transcript expression levels were significantly lower in ISM, NIC and NNC categories than in the FSM set (t-test $p < 2.2e\text{-}16$ for all comparisons) and were significantly lower for novel genes compared to annotated genes (t-test $p < 2.2e\text{-}16$ for both comparisons), confirming that the novel isoforms discovered in model organ-

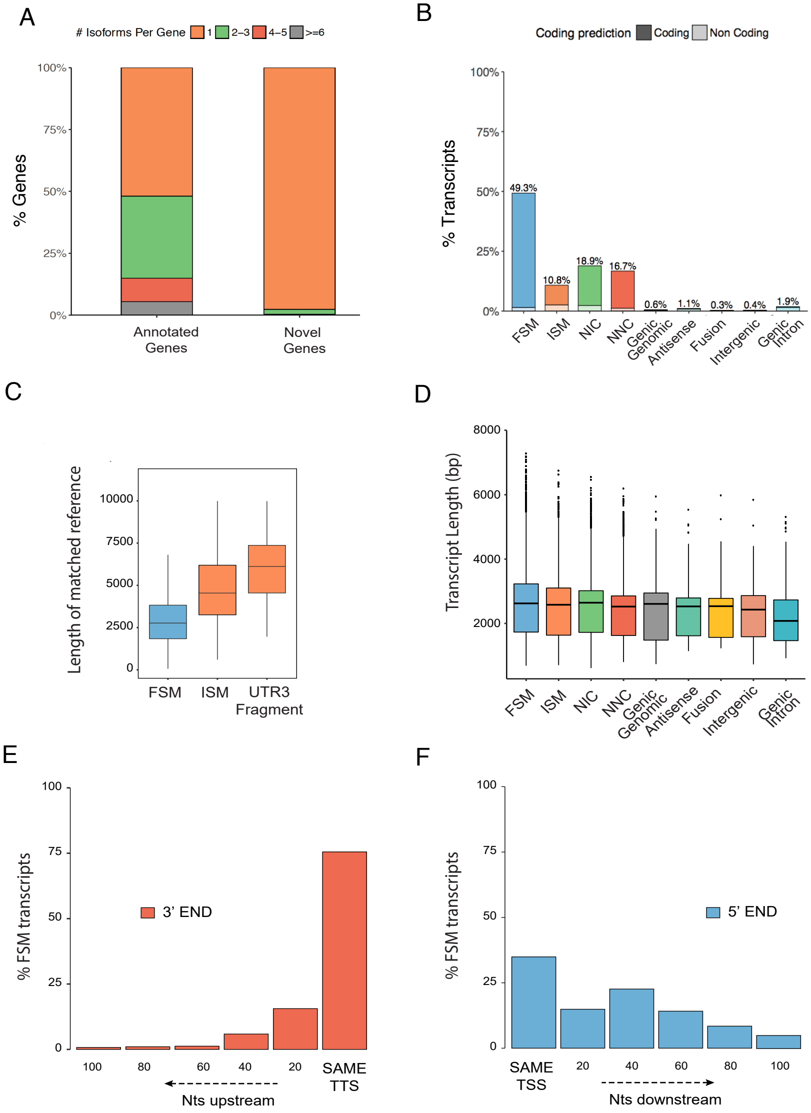


Figure 3.4: Characterisation of the ToFU-defined transcriptome. A. Distribution of the number of variants derived from annotated and novel genes. B. Distribution of transcripts among the set of defined structural categories. C. Length of the reference transcripts to which FSM, ISM, and UTR3 Fragment PacBio transcripts matched. D. Length of the PacBio transcripts by structural categories. E-F. Overlap at 3' and 5' ends between the FSM transcripts and their respective matched reference transcripts.

isms such as mouse are usually minor isoforms of genes already described in reference databases.

In terms of transcript full-lengthness, the majority of our FSM transcripts showed a complete or close to complete 3' end overlap with the 3' end of the matched reference transcript: 76% had an exact 3' end match and 16% were within 20 nt upstream of the annotated 3' end (Figure 3.4.E). This contrasts with the lower proportion of FSM transcripts showing a complete overlap with their reference 5' ends (35%) and the higher number of transcripts falling short by 40 to 100 nts (50%, Figure 3.4.F). This finding concurs with the strategy used during cDNA library preparation and with the Iso-Seq^{TM} analysis pipeline because both steps have less control over completeness at 5' ends (Oligo(dT) primming and polyA tail identification to control 3' completeness). Interestingly, 851 and 1,361 FSM transcripts had 3' end and 5' end positions that extended beyond the matched reference transcript, while 1,610 and 1,439 of our FSM sequences were shorter by more than 100 nt at their 3' and 5' ends, respectively. These cases could potentially represent alternative polyadenylation/alternative TSS events.

IDP transcritome characterization

Error correction of long-reads by LSC yielded 99.2% of CCSs corrected by short-reads. This set of LSC corrected long-reads and the set of splice junctions detected by SpliceMap were fed to the IDP tool, resulting in the detection of 12,521 isoforms and the prediction of 4,387 isoforms. After removing redundancy between the detection and prediction IDP modules, we obtained 13,525 unique transcripts.

In contrast with the results from ToFU, most of the isoforms we identified by IDP were classified as FSMs (96%, Figure 3.5.A). Only 4% of the transcripts were catalogued as novel isoforms from known genes (Figure 3.5.A). Remarkably, the majority of them were categorized as NICs and 97% of them were composed of known junctions in a novel combination. Moreover, the IDP did not identify any isoform outside the boundaries of kwown genes and consequently, all the isoforms belonged to previously annotated genes. Surprisingly, all of the 3' and 5' ends of the IDP-defined transcripts perfectly matched the reference ends (data not shown).

TAPIS transcritome characterization

Out the aprroximately 204,984 (94,5%) FL CCSs that were properly mapped to the reference genome, a total of 57,776 transcripts originated from 14,775 expressed genes were defined by TAPIS. Regarding transcript classification, a small proportion of the total defined transcripts belonged to novel genes (6.5%, Figure 3.5.B). Similar to the ToFU strategy, the novel genes were characterized by including mono-exon transcripts (data not shown) and a reduced number of splicing variants compared to the high number of transcript variants detected in annotated genes (median = 4 isoforms, Figure 3.5.C). Surprisingly, just 9% of the isoforms fell into the FSM category (Figure 3.5.B) and almost 70% were classified as NNC category, as characterized by using novel splicing sites.

Moreover, it should be noted that novel transcript categories as NIC, NNC and Genic Genomic showed a higher length distribution compared to the FSM category (Figure 3.5.D, t-test $p < 2.2e\text{-}16$), which contrasts with the results obtained using the ToFU strategy. Moreover, the FSM category showed drastic higher gene expression distribution in comparison to the rest of categories (t-test $p < 2.2e\text{-}16$ for all comparisons), again revealing the minor expression status of transcripts not yet annotated in public databases.

Finally, as seen for the ToFU pipeline and in agreement with our cDNA library preparation, the 3' end overlaps for the FSM transcripts was almost complete but was less so for the 5' ends (35% of the transcripts overlapped the annotated TSS).

Comparative Overview

We systematically compared the strategies in order to highlight the advantages and disadvantages of each transcriptome-definition strategy. For comparative purposes, two isoforms were considered identical across pipelines if they shared the exact splicing junction linkage which enables the association of known and

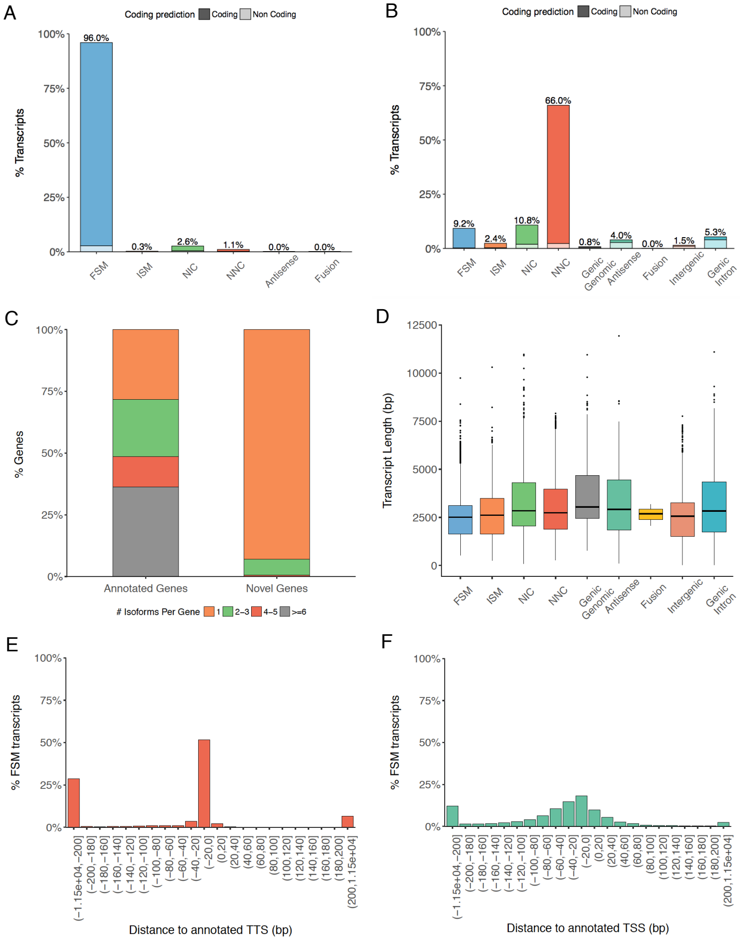


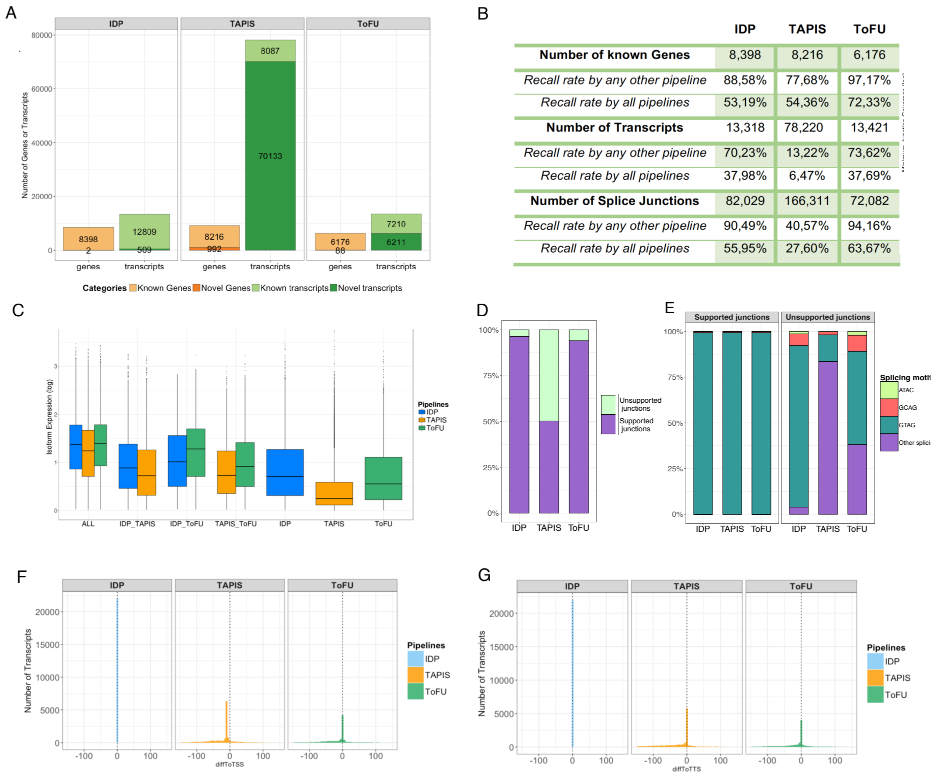
Figure 3.5: Characterisation of transcriptomes defined by IDP and TAPIS. A-B. Distribution of transcripts among the set of defined structural categories for IDP and TAPIS transcriptomes, respectively. C. Distribution of the number of variants derived from the annotated and novel genes present in the TAPIS transcriptome. D. Length of the PacBio transcripts based on the structural categories in the TAPIS transcriptome. E-F. Overlap at 3' and 5' ends between the FSM transcripts and their respective matched reference transcripts in the TAPIS transcriptome.

novel isoforms across pipelines, thus excluding mono-exonic genes from this comparison.

First, we compared the approximate number of genes and transcripts (known and novel) identified by the three different approaches ((Figure 3.6.A). Even though around 7,000 genes were identified by each of the different pipelines, the very high number of transcripts found by TAPIS, as well as its high proportion of novel isoforms (89.7% novel transcripts) stood out compared to the other pipelines (Figure 3.6.A). Moreover, while a roughly equal proportion of novel and known isoforms were identified by ToFU, IDP identified fewer novel transcripts (4% novel isoforms, Figure 3.6.A).

To further investigate the differences between pipelines and their nature, we compared the sets of detected genes, transcripts and SJs. Recall rates were calculated (Figure 3.6.B) to evaluate the ability of each pipeline to identify the same genes, transcripts or junctions described by other strategies. When known genes detected by each pipeline were compared, the recall rate by at at least one other pipeline reached a mean value close to 90% (Figure 3.6.B), revealing that the three pipelines mostly agree in the set of detected genes. Among them, ToFU shows the highest number of detected known genes that are also captured by the other two pipelines (72.33%), suggesting a higher sensitivity compared to the other two strategies. However, this consensus was no longer achieved at the transcript level, mainly because of the incredibly high discovery rate of TAPIS (Figure 3.6.B). Thus, although IDP and ToFU reached more than 70% recall between pipelines, the recall rate of TAPIS was only 13%, meaning that only 13% of transcripts defined by TAPIS were identified by another strategy. When all the novel isoforms were removed from the analysis and the recall rate was calculated considering only known transcripts, the recall rates of ToFU and TAPIS reached 94% and 89%, respectively (data not shown), revealing the high intersection of known isoforms. Meanwhile, IDP pipeline detected mamy more known transcripts that any other strategy, but only 67% of them were also found by TAPIS and/or ToFU.

Next, we analysed differences in the expression levels for different groups of intersection isoforms. Interestingly, the expression of isoforms found by TAPIS were lower than that of the other pipelines. Because TAPIS generated a higher transcript-per-gene ratio, the same short reads were assigned to a more complex gene model with an increased number of alternatively-spliced isoforms, likely resulting in a significantly decreased expression at the transcript level. In contrast, the expression levels of isoforms found by more than one pipeline were higher than the those identified by just one approach (Figure 3.6.C).



	IDP	TAPIS	ToFU
Number of known Genes	8,398	8,216	6,176
Recall rate by any other pipeline	88,58%	77,68%	97,17%
Recall rate by all pipelines	53,19%	54,36%	72,33%
Number of Transcripts	13,318	78,220	13,421
Recall rate by any other pipeline	70,23%	13,22%	73,62%
Recall rate by all pipelines	37,98%	6,47%	37,69%
Number of Splice Junctions	82,029	166,311	72,082
Recall rate by any other pipeline	90,49%	40,57%	94,16%
Recall rate by all pipelines	55,95%	27,60%	63,67%

Figure 3.6: Comparison of alternative methods for defining Iso-Seq PacBio transcriptomes. A. Distribution of the number of genes and transcripts, novel and known obtained with the different approaches. B. Recall summary between approaches at the gene, transcript and splice-junction levels. C. Transcript expression distribution for the different intersected sets. D. Illumina short-read splice-junction support across these different pipelines. E. Distribution of canonical and non-canonical splicing motifs across the junctions which are supported or not supported by short reads. F–G. Histograms representing the differences in the detected and reference ends for each method in the TSS and the TTS, respectively.

We characterised each SJ based on their short-read support and splicing motif and found important differences in the fraction of supported SJs across pipelines, with the TAPIS pipeline standing out because almost 50% of the SJs in this pipeline were not supported by Illumina data (Figure 3.5.D). Moreover, almost 60% of the SJs detected by TAPIS were not detected by the other pipelines, in contrast to the 10% of junctions specific only to the ToFU and IDP pipelines (Figure 3.5.B). Furthermore, splicing junction categorisation revealed that in all cases, more than 98% of the supported junctions presented the most common canonical motif (GT-AG; Figure 3.5.E). In contrast, unsupported SJ splicing motifs were more diverse; this was especially the case in TAPIS, in which 80% of all unsupported junctions present a non-canonical splicing motif (Figure 3.6.E). Thereby, TAPIS detected more than 68,000 non-canonical and unsupported junctions, which represents the 40% of the total detected junctions. In contrast, IDP did not retrieve any non-canonical sites either in supported or non-supported categories because the aligner discards them before the isoforms are defined.

Because one of the most important aspects of Iso-Seq is the theoretical ability to capture entire transcripts from end to end, the full-lengthness status of isoforms was compared among pipelines using only the set of FSM transcripts. As we have already mentioned, FSM isoforms identified by IDP perfectly matched both the TSS and TTS reference sites (Figure 3.6.F-G), suggesting that the reference information was preferentially used and may have biased results and hidden potential APAs and alternative TSS events. In contrast, both Tappis and ToFU did find transcript end variability between discovered and annotated transcripts (Figure 3.6.F-G), which might represent alternative TSSs and TTSs or be a consequence of mRNA degradation prior to sequencing.

In conclusion, the loci detected by these alternative approaches for Iso-Seq transcriptome reconstruction almost always coincide but differ in the magnitude and nature of the transcript isoforms they define. This reveals that the computational pipeline chosen strongly impacts transcriptome reconstruction. The IDP strategy is highly dependent on reference information resulting on transcript calls tha faithfully match annotated transcripts (96%) and have very little 3'/5' end variability. IDP was also unable to detect any of the 16 novel PCR-validated transcripts, suggesting that this method is highly restrictive for novel isoform calling. In contrast, TAPIS works without short-read data. The tool returns significantly more transcripts (91,428) most of which are NNCs (66%). It also has the lowest recall rate (13%) of all the pipelines and identified a high proportion of unsupported non-canonical junctions (40%). We conclude that TAPIS might have a high rate of false calls. Finally, ToFU defined a balanced proportion of novel/known transcripts and had the best recall rates at the gene, transcript, and splice-junction level, without relying on short-read sequencing or high-quality transcriptome annotation.

These results indicate that, ToFU provides high flexibility to generate full-length transcriptomes without requiring prior knowledge or the need for short-read sequencing data. Hence, ToFU appears to be the best existing pipeline for the purposes of general transcriptome definition using Iso-Seq PacBio data and so we chose this pipeline as baseline for further analyses.

3.4.3 Characterisation of ToFU-defined novel calls reveals enrichment in artefacts

The descriptive analysis framework provided in previous section for ToFU transcriptome readily indicates that our neural mouse transcriptome, obtained by PacBio single molecule sequencing, recovered full-length transcripts and had an important level of novelty ($\sim$ 40%). Alignment of ToFU transcripts to the reference genome showed an average percentage of identity above 99.8%, indicating that most sequencing errors were corrected by the ToFU clustering approach. However, small indels (average size $\sim$ 1.2 nts) were still detected in 56.2% of the transcripts. To tackle this problem, we first attempted to correct indels with matching Illumina short reads using Proovreads [130]. Although the number of transcripts with at least one indel decreased to 2,550 (16% of transcripts), this was still unsatisfactory for posterior ORF prediction since small indels can provoke a frame shift and consequently a false prediction. Instead, transcripts were corrected using the reference genome sequence. By virtue of this strategy, all indels inside exons were removed, and we obtained what we called a corrected PacBio transcriptome.

To assess the quality and nature of novel calls, canonical status of SJ was evaluated in first place. In our ToFU-defined transcriptome, the ratio of canonical versus non-canonical splicing events fitted the expected genome proportions when looking at known splice junctions: out of 141,332 known splice junctions, 99.9% were canonical. However, deep inspection revealed that novel splice junctions showed a very different distribution: out of 3,837 novel splice junctions 31% (1,188) were non-canonical. When analysed across the different transcript categories, non-canonical splicing was maintained at low rates in FSM (0.1%) and ISM (0.25%) transcripts, which was expected as both are formed entirely by known splicing events (Figure 3.7.A). In NIC transcripts, comprising novel combinations of known splice junctions or novel splice junctions deriving from annotated donors or acceptors, the percentage of non-canonical splicing was 0.15% (Figure 3.7.A). However, in NNC transcripts, characterized by the introduction of alternative donors and/or acceptors, we found 1,155 novel non-canonical junctions, which represented 4.5% of total. Moreover, genic genomic, intergenic, genic intron and Antisense transcripts, despite rarely being multi-exonic, showed relatively high percentages of non-canonical splice junctions with 2.3%, 7.28%, 21.57% and 32.65% respectively (Figure 3.7.A). This unusually high level of non-canonical junctions suggests that experimental artifacts might be accumulating in these categories. Furthermore, when the percentage of transcripts showing at least one non-canonical splice junction was considered, the proportion of NNC affected compared to NIC transcripts became more evident, 41.5% vs. 1.47%, respectively, strongly indicating that this NNC category of transcripts needed deeper inspection.

Positional analysis of junctions along the transcript showed that, although novel junctions could appear at any position in novel transcripts, there was a higher concentration of occurrences towards 5' ends, pattern which is not observed for known - whether canonical or not - junctions (Figure 3.7.B, FET p $<$ 2.2e-16). This could either be the consequence of unannotated variability at

5’ ends or higher accumulation of errors due to lower sequence support. The ToFU pipeline is more permissive with clustering conditions at transcript ends (E. Tseng, personal communication), which accounts for a higher probability of errors at these areas.

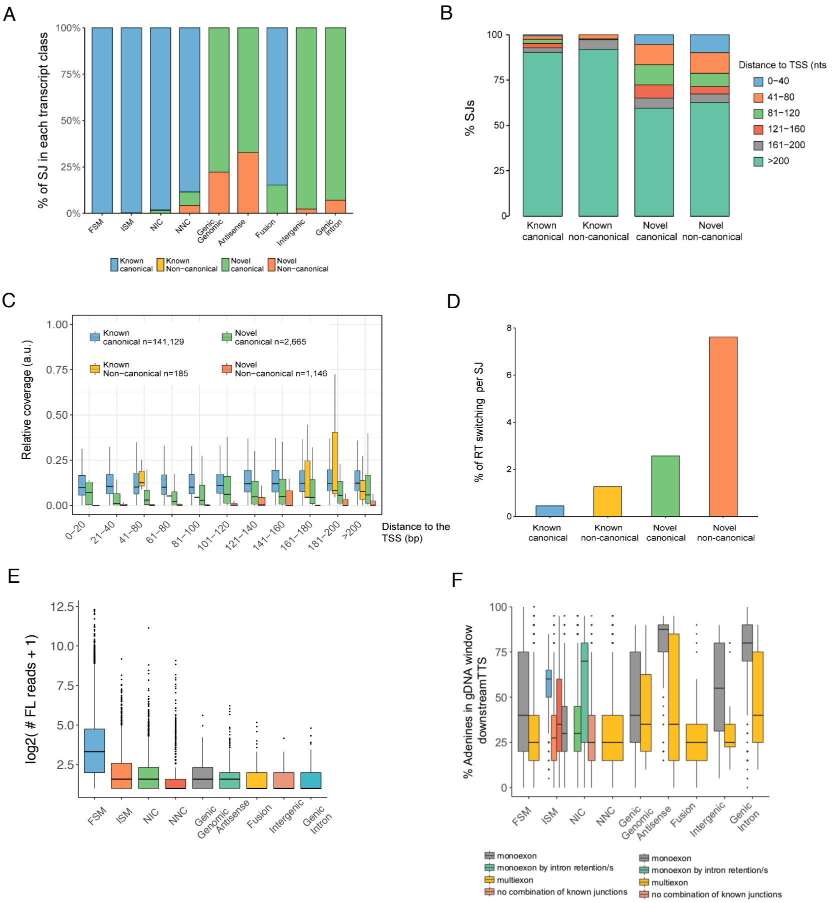


Figure 3.7: Splice junction characterisation in the corrected PacBio transcriptome. A. Distribution of splice junction (SJ) types across structural categories. B. Distribution of the SJs according to their distance to the transcription start site. C. Relative coverage by short reads of SJs as a function of their class and distance to the TSS. (a.u.) arbitrary units. D. Detection of RT switching direct repetitions across SJ types. E. Distribution of Iso-Seq FL reads associated to each isoform across structural categories. F. Intrapriming evaluation across structural subcategories.

Short-read junction coverage computed by STAR was used to calculate the support level for novel junctions called by PacBio. Note that Illumina reads are not always equally distributed along the transcript length and are often less abundant towards the 5' ends, thus providing less support for junction validation. We found that, as suspected, splice junction support by short reads decreased towards the 5' end of the transcripts, but was significantly higher for known junctions (Figure 3.7.C, Wilcoxon Rank Sum test (WRS) $p < 2.2e\text{-}16$). Novel canonical junctions were in general less frequently covered but still significantly more supported than novel non-canonical junctions, which had hardly any supporting reads if located within the first 120 nts of the transcript 5' end (Figure 3.7.C, WRS $p < 2.2e\text{-}16$).

Moreover, events which occur during RNA RT and library generation could also explain this accumulation of non-canonical junctions. Prediction of RT switching events confirmed the enrichment of RT switching in novel splice junctions (Figure 3.7.D, FET $p < 2.2e\text{-}16$) and in NNC compared to NIC transcripts (7.24% versus 1.98%, FET $p < 2.2e\text{-}16$). The described RT switching events affect minor isoforms of genes co-expressed with a major isoform that serves as the template for the intra-molecular switching. Accordingly, we found that NNC transcripts are enriched for being minor transcripts of highly expressed genes (data not shown).

The number of supporting FL reads (number of raw FL reads used to generate a given isoform) affects the capacity of the ToFU pipelines to correct errors and consequently affects the quality level of the final defined isoform. Results showed that FSM transcripts contain significant higher number of FL reads than any other isoform category (Figure 3.7.E, t-test $p < 2.2e\text{-}16$ for all comparisons). Nevertheless, although ISMs and NICs show similar distributions, NNCs present a clear lower FL distribution and hence, a low chance to be error-corrected by the ToFU pipeline.

Finally, A-rich genomic DNA regions downstream of the TTS were concentrated in the relatively minor transcript categories (Figure 3.7.F). Using a cutoff of 60% adenines, a total of 601 transcripts were found to be intra-priming candidates, which affected the antisense and genic intron categories in particular (50% and 30% of their transcripts were flagged). Remarkably, Incomplete

Splice Match transcripts that were versions of the reference transcripts shortened at the 3 end and monoexon NIC transcripts with intron retention events were also significantly enriched in intra-priming candidates (WRS p $<$ 2.2e-16 for all tests).

Altogether, out results suggest that a fraction of the novel transcripts found by ToFU pipeline could be technical artifacts that originated at the cDNA library construction step or via less confident correction by ToFU at the 5' ends of transcripts.

3.4.4 Experimental validation of ToFU results verifies the presence of novel-isoform artifacts

To shed light on whether the transcripts detected by the ToFU analysis were correct or not we performed RT-PCR amplifications for a total of 67 mRNAs encompassing different categories: 23 FSM (3 with non-canonical splice sites), 12 NIC, 30 NNC canonical (11 of them containing at least one non-canonical splice junction) and 3 Fusion. Importantly, we performed RT-PCRs both on the ClonTech oligo(dT) enriched full-length cDNAs used for PacBio sequencing and, for positive NIC/NNC/Fusion and 4 FSM transcripts, on new cDNA retrotranscribed using random hexamers rather than oligo(dT) at both 42 °C and 50 °C. The rationale behind this approach was to test whether novel transcripts could have been spuriously generated by RT switching-like mechanisms at the retrotranscription step of the PacBio protocol. Since higher temperature and/or the use of random hexamers would complicate the formation of secondary structures in the RNA template, retrotranscription artifacts would be less favored in these conditions.

We validated by RT-PCR for all of the 23 FSM, including the 3 cases with non-canonical junctions, (Figure 3.8.A) highlighting the high level of confidence supporting these transcripts. Novel transcripts showed lower validation rates: 8/12 NIC, 1/3 Fusion and 6/30 NNC, highlighting the low detection rate within NNC category (Figure 3.8.B). Importantly, 9 of these non-validated NNC transcripts were amplified by oligo(dT) PCR but were lost when random hexamers and higher temperatures were used (Figure 3.8.C), suggesting the possible occurrence of retrotranscription artifacts. Table 3.2 summarizes the results of the

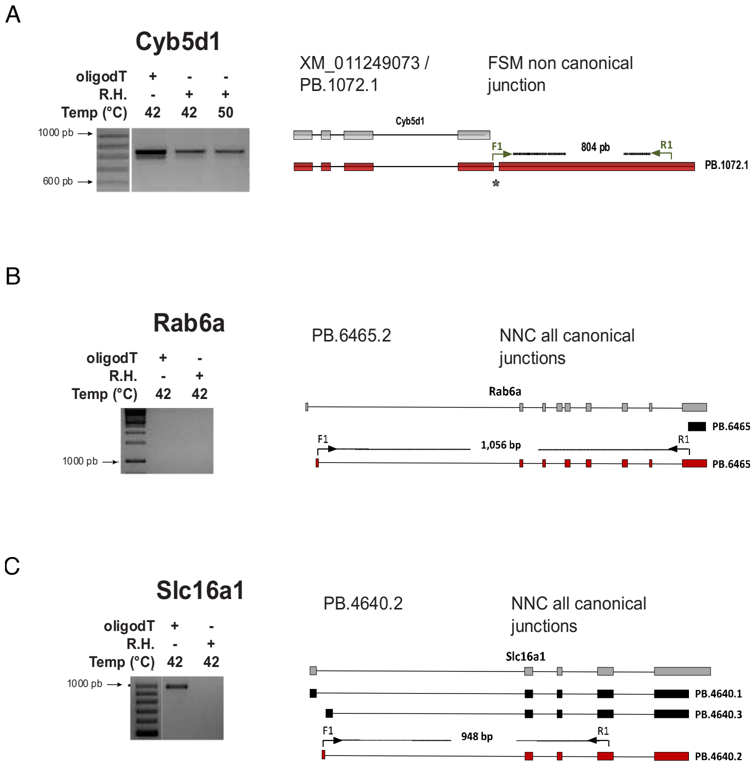


Figure 3.8: Representative examples of RT-PCR validation experiments. A. FSM transcript with a noncanonical SJ successfully amplified at each PCR condition. B. Example of a NNC transcript with a noncanonical SJ that failed to be amplified in the oligo(dT) condition. C. Example of NNC transcript with noncanonical SJ amplified at oligo(dT) but not when using Random Hexamers conditions.

	oligo (dT)			Random hexamers			
Transcript type	**Positive**	**Negative**	**Total**	**Positive**	**Negative**	**Total**	**Overall validation**
FSM	23 (3 nc)	0	23	4 (3 nc)	0	4	100%
NIC	10	1	11	8	2	10	67%
NNC	15 (3 nc)	15 (8 nc)	30	6	9 (3 nc)	15	20%
Fusion	1	2	3	1	0	1	33%

(nc) Trasncript with non-canonical junctions.

Table 3.2: Summary of RT-PCR isoform validation across different structural categories.

PCR validation experiment. Our PCR results indicated that an additional filtering strategy would be useful to remove artifactual transcripts from the ToFU transcriptome output.

3.4.5 Machine learning enables accurate filtering of novel-isoform artefacts

Previous work applied different criteria to discard artifacts from transcriptome sequencing, including support by short reads [12], removal of transcripts with non-canonical splicing [318] or filtering based on sequence features [279]. However, we found that these approaches do not fully capture the complexity of the data. For example, a few known and NIC transcript junctions lack Illumina coverage (148 out of 67,610, and 20 out of 437 respectively), while most of the novel non-canonical junctions did had supporting Illumina reads (543 out of 597). We found that additional features such as RT switching direct repeats and low expression values accumulated in NNC transcripts, but were not exclusive to them. Moreover, our RT-PCR analysis revealed an important number of transcripts (16) having a full set of canonical junctions but failing validation. We hypothesized that the set of quality control attributes and descriptors previously used to evaluate and characterize isoforms ought to be informative of transcript quality and could be used to define a composite filter to remove artifact transcripts efficiently.

Thus, we decided to train a ML classifier based on these features. As previously described in Methods section, we defined the FSM transcripts as the positive set (n = 7,774) and the NNC transcripts with at least one non-canonical SJ as the negative set (NNC-NC, n=1,110). Figure 3.9.A. shows the features selected by the classifier, with flags bite transcripts ranking first in order of importance which we interpret as an indication of the presence of novel SJs caused by secondary RNA structures. Interestingly, five out of the eight top variables were associated with junction expression, suggesting that junction coverage patterns are some of the most useful characteristics for calling *bona fide* novel transcripts.

Application of the Random Forest classifier to our test set using a probability for positive classification higher than 0.75 led to AUC of 99.54% for the receiver operating curve (Figure 3.9.B, blue line), indicating that the created classifier

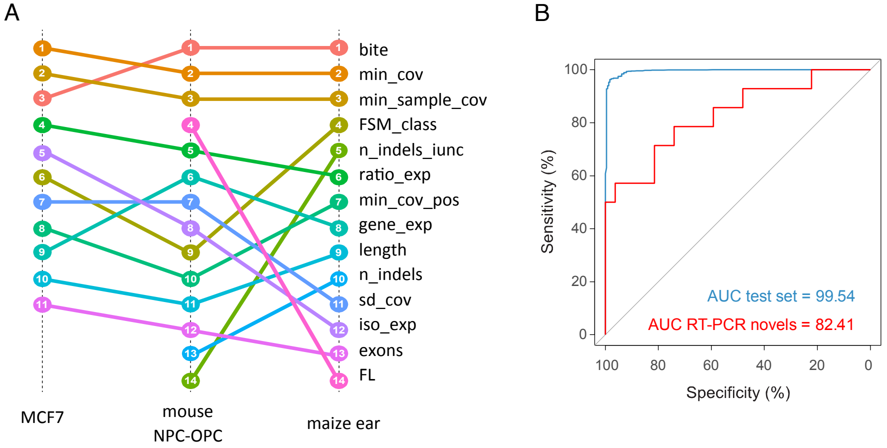


Figure 3.9: Machine Learning (ML) filter performance. A. Ranking of variables according to their relative importance for the ML classifier across different PacBio-defined transcriptomes. In MCF7 dataset no full-length data were available. B. ROC curves of the ML filter in our mouse transcriptome for the test set (blue line) and for the set of novel isoforms assayed by RT-PCRs (red line).

performed well. We then applied this Random Forest classifier to our ToFU-defined transcriptome using the same classification parameters we used in the test set. Evaluation of the classifier on the set of 41 novel NNI/NNC isoforms we had previously assayed by RT-PCR, gave an AUC 82.41% (Figure 3.9.B, red line). This indicates that our classifier faithfully captured differences between our baseline set of positive and negative transcripts, and thus it can be applied to efficiently discriminate true transcripts from artefacts within the set of long-read novel sequences defined by ToFU. Additionally, we used RT-PCR data to compare the performance of our ML method to two previous methods: the non-canonical SJ filter (nc Filter) and SpliceGrapher. Our results indicate that the classifier approach has a higher F1 score (71.7 versus 57.9 and 41.1 respectively), and a lower FDR (11% versus 53.3% and 58.8% respectively) than alternative methods (Table 3.3). These notable FDR differences are mostly due to a high rate of false canonical junction transcripts that are not discarded by the prior approaches. Moreover, the ML filtering strategy was the only one that succeeded in lowering both the non-canonical SJ and the no short-read coverage quality features in NNC transcripts to levels similar to those of the high-confidence FSM category (Figure 3.10.A).

	TP	TN	FP	FN	F-score	FDR
ToFU + SQANTI Filter	8	26	1	6	71.7	11.1%
ToFU + nc Filter	14	11	16	0	57.9	53.3%
ToFU + SpliceGrapher	14	7	20	0	41.1	58.8%

(nc) Non-canonical junctions.

Table 3.3: Performance summary for alterantive artefact filtering methods. The ML-based filter, the non-canonical filter, and SpliceGrapher were evaluated using the set of novel isoforms assayed by RT-PCR. (TP) True Positive, (TN) True Negative, (FP) False Positive, (FN) False Negative, (FDR) False Discovery Rate.

Based on these results, we curated our ToFU-defined transcriptome using our classifier. In addition, we added intra-priming filters to discard transcripts that could have undergone polyA intra-priming, which, as described in section 3.4.3, were enriched in the antisense and ISM categories. When we applied this approach to the mouse neural transcriptome, this combination of ML and intra-priming filters removed 4,134 novel transcripts (2,462 NNC, 1,281 NIC, 32 genic genomic, 36 fusion, 116 antisense, 25 intergenic, 129 genic intron and 53 ISM). The adjusted percentages of each category in our final curated transcriptome were: 66.3% FSM, 14.1% ISM, 15.7% NIC, 2% NNC, 0.5% genic genomic, 0.5% sntisense, 0.2% fusion, 0.3% intergenic and 1.4% genic intron (Figure 3.10.B). Our filter had the strongest impact in the NNC transcript category, which considerably diminished from 14% to 2%, while FSM increased consequently from 49% to 66% in the curated transcriptome (Figure 3.10.B). In our final dataset 9,626 transcripts (80.4%) are in known categories and 2,058 (17.1%) are novel transcripts of which 286 (2.3%) fall within novel genes.

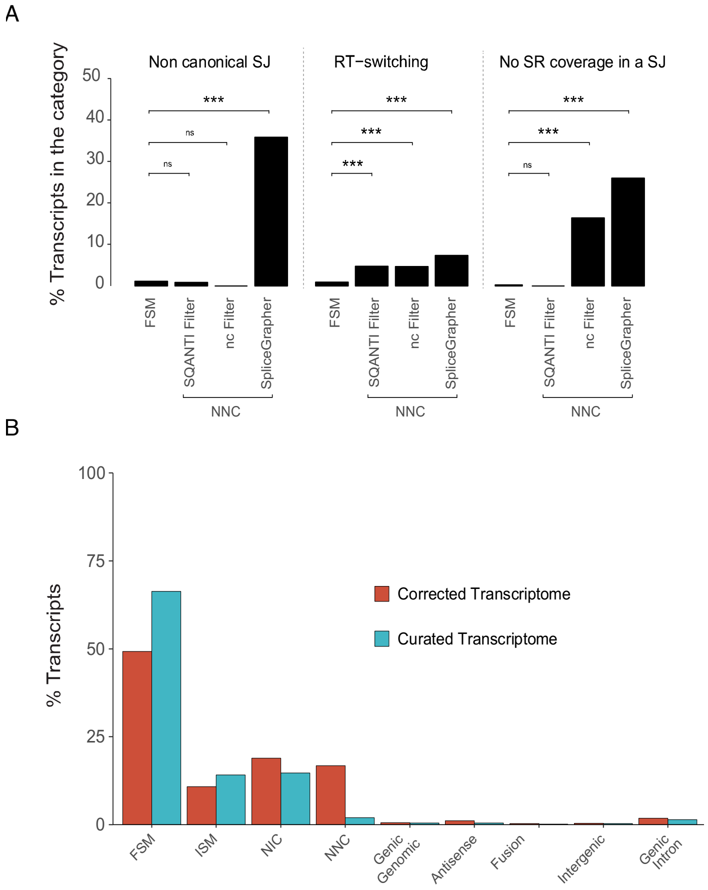


Figure 3.10: Comparison of results between alternative artifact filtering methods. A. Evaluation of quality features in the FSM and NNC categories after the ML-based method, the non-canonical filter, and SpliceGrapher. Statistical differences were test using Fisher's exact tests (FET), (***) $P < 0.001$, (ns) not significant. B. Structural classification of transcripts before and after application of our ML-based filter.

3.4.6 PacBio sequencing unable to accurately quantify low-medium expressed isoforms but capturing most transcriptional signal

Although Iso-Seq PacBio sequencing was effective in defining full-length transcripts and discovering new splicing variants, its low throughput (compared to short-read based sequencing technologies) may be insufficient to compute accurate transcript level expression estimates. Transcript quantification evaluation-revealed that replicate correlation was significantly lower when quantifiying with PacBio FL reads (Figure 3.11.A) compared to short-read quantification, especially at the mid and lower expression ranges, where the correlation dropped to nearly zero (Figure 3.11.B). Thus, our results show that the number of reads would be insufficient for an accurate quantification of transcript expression quantification at the PacBio's current sequencing depth (0.5 M per sample),

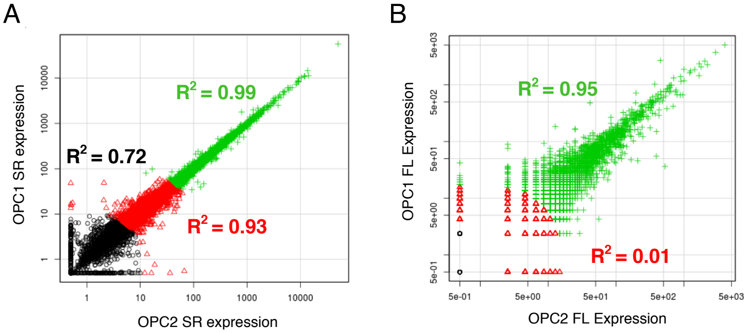


Figure 3.11: Correlation of expression estimates between OPC replicates at three levels: low (black), medium (red) and high (green). A. Isoform expression quantification using short-read data. B. Isoform estimations using PacBio FL data.

Another possible limitation of PacBio sequencing is that transcripts with low expression levels could be difficult to detect because of their low-throughput compared to short-read RNA-seq that generates millions of reads at a fraction of the cost. To test this possibily, we mapped short-reads to both the RefSeq reference transcriptome (ReT, around 160,000 transcripts) and the curated PacBio-defined transcriptome (PbT, 12,408 transcripts) ad evaluated the portion of the signal hidden by PacBio. On average, 87% of our Illumina reads mapped to the mouse genome. Transcriptome mapping results showed that 81.7% of reads had a hit to the ReT and 70.7% to our PbT respectively, indicating that only an 11% in transcriptional signal was missed when considering the PbT alone.

However, this difference in the number of mapped reads translates into a much bigger difference in the number of detected transcripts, equating to 30,071 versus 11,921 transcripts at a 1 count threshold (Figure 3.12.A). This suggests that ReT exclusive transcripts had lower expression than PbT, which we confirmed after analyzing transcript expression levels (Figure 3.12.B). At the gene level, ReT-based quantification totally overlapped PbT except for 357 genes that were a combination of novel, fusion and other reference genes. Further characterisation of PacBio exclusive trancritps revealed that from a total of 3,447 transcripts absent from the ReT, 20.8% of them belong to Ensembl and RefSeq

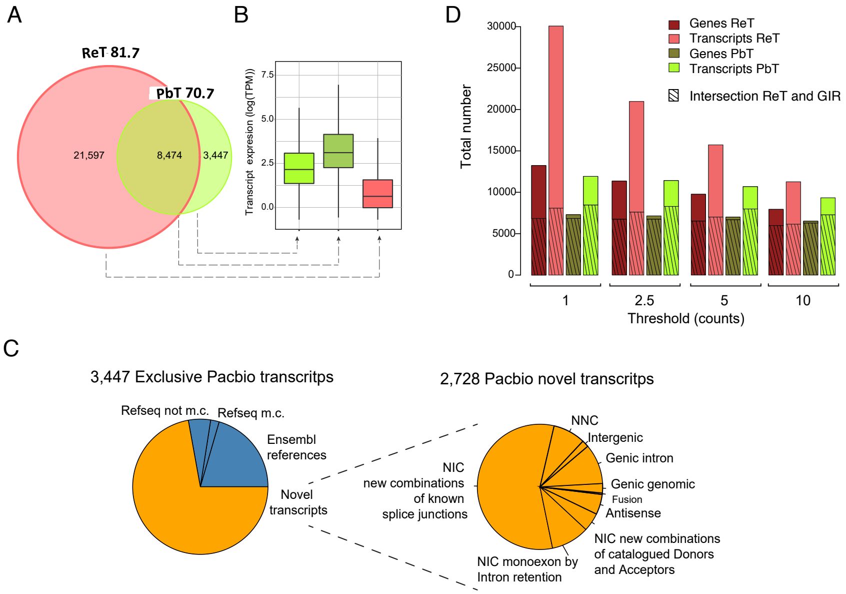


Figure 3.12: Isoform detection by Iso-Seq sequencing. A. Venn diagram of the number of reference transcripts captured by short-read mapping (red, ReT) or Iso-Seq long-read sequencing (green, PbT). The upper part of the Venn diagram indicates the percentage of short reads mapped to each defined transcriptome is indicated. B. Expression level distribution for ReT exclusive, PbT exclusive and shared transcripts. C. Classification of transcripts exclusively detected by PacBio sequencing based on their presence in reference databases and characterization of PacBio novelty . D. Number of genes and transcripts detected by short-read mapping to the PbT and ReT, at increasing expression level cut-offs. m.c., manually curated transcript.

transcriptome references (Figure 3.12.C). However, the great majority of PbT exclusive transcripts were catalogued as novel transcripts (n=2,728, 79%), most of which were NIC transcripts generated by new combinations of already known splice junctions (61%).

In addition, imposing a filter of 10 counts, we eliminate most of the ReT-exclusive transcripts and made the number of transcripts and genes detected by the two mapping approaches similar (Figure 3.12.D). Note that a minimum of 10 counts is required by popular differential expression algorithms such as edgeR [276] to remove transcriptional noise. Furthermore, the proportion of genes with multiple transcripts was almost identical for the PbT and the ReT at this 10 count threshold (Data not shown). We concluded that, at reasonable sequencing depths for long and short-reads technologies (2M and 60M, respectively), the PacBio transcriptome still captures nearly 90% of the transcriptional signal that Illumina would find, is able to rescue transcriptional diversity not yet annotated by the reference databases, and dramatically reduces the calls of transcripts with very low expression levels that could be at the limits of accurate quantification.

3.4.7 Novel transcripts have a major impact on the accuracy of transcriptome quantification by short reads

In order to investigate how the magnitude and nature of the transcriptome affects quantification, we compared the quantification results when using a reduced transcriptome (our curated PacBio Transcriptome, PbT), and the total set of Reference Isoforms (ReT). As explained in Section 3.3.2, we addressed the evaluation of short-read quantification results by comparing the METs in each transcriptome. The MET was the same for 3,976 genes when quantifying with PbT and ReT. Interestingly, this was not the case for 1,433 genes, 996 of them showing a PbT MET transcript already present in ReT but not quantified as MET.

For example, the signal peptidase complex subunit 2 gene (*SPCS2*) was expressed as one transcript in our PacBio neural transcriptome (PB.6460.1) and had two transcripts in ReT quantification (NM_025668 and XM_006508117) (Figure 3.13.A1). PB.6460.1 is a FSM transcript of NM_025668 and both codify for the PI-ORF of the gene (ORF associated to the transcript defined as principal isoform by APPRIS [277] based on its functional load) but the 3' exon of PB.6460.1 is smaller, resulting in a 3' UTR shorter by 1,340 nucleotides, (Figure 3.13.A1, red dashed box). This shorter 3' exon is actually the annotated exon of the RefSeq transcript, XM_006508117, which also uses two alternative 5' exons. XM_006508117 was the MET in the ReT quantification while NM_025668 was estimated as poorly expressed (Figure 3.13.A2). Upon RT-PCR amplification with transcript discriminating primers we confirmed the PbT and not the ReT based quantification scheme (Figure 3.13.A3). When inspecting read coverage at this locus we observed that neither the unique 5' junctions of XM_006508117 nor the extra exonic sequence at the 3'exon of NM_025668 were covered by Illumina short reads, while the short-read pattern nicely fits the PacBio transcript model (Figure 3.13.A1). We speculate that this variability at the 3'UTRs creates a conflict when resolving transcript quantification in the RefSeq gene model that was decided in favour of transcript XM_006508117 by RSEM, as this transcript has a more consistent 3' end coverage. In summary, the transcript quantification error of the SPCS2 gene when using a reference transcriptome as mapping template was due to a discrepancy in the 3' end annotation between the reference and the actual expressed transcripts. Similar disagreement patterns were observed for two additional genes, DHRS7B and BDKRB2 with similar outcomes in terms of MET selection (data not shown).

To estimate how general this pattern was, for all the Mayor Expressed Transcript (MET) discrepant genes, we investigated the RefSeq curation status. The majority of the discrepant genes (57.2%, n = 470 genes) corresponded to situations where the PbT MET was a FSM of a manually curated RefSeq transcript and the ReT MET was not manually curated, as in the case of SPC2 gene. Furthermore, in these cases, the RefSeq-based MET had significantly worse lowest splice junction coverage and lowest mean exon coverage than the MET called by the PbT quantification (Figure 3.13.B-C). Similarly to SPCS2, we found that, for these 470 genes, the differences in the length at the 3 end between the MET selected at PbT quantification and their matched RefSeq transcripts were significantly higher than in genes where both quantifications selected equivalent METs (Figure 3.13.D). Moreover, these differences were also observed for transcripts codifying for the PI-ORF of the genes, indicating that the extensive variability in the 3' ends that is not annotated in a global reference such as RefSeq is not only restricted to secondary/alternative transcripts. These results demonstrate the relevance of using a full-length reference transcriptome updated with novel expressed transcripts for correct quantification estimates.

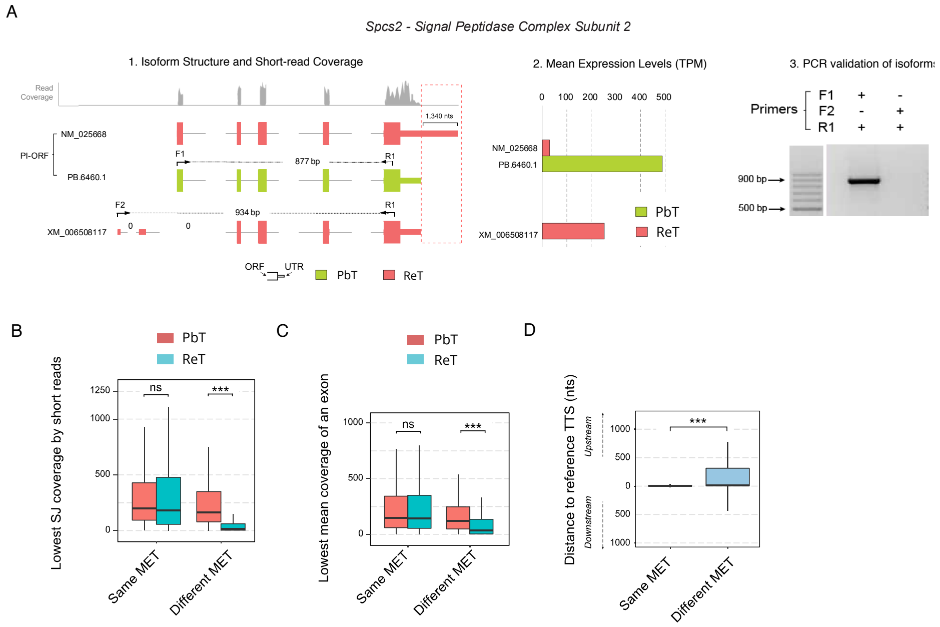


Figure 3.13: Quantification problems caused by 3' UTR variability. A. SPCS2 gene as an example of how 3' UTR variability in a PI-ORF leads to quantification errors. A1.Transcripts associated with the SPCS2 according to PacBio sequencing (green), RefSeq quantification (red), and the short-reads mapping profile at the SPCS2 locus (grey) are shown. The positions of transcript-specific primers are indicated by arrows and differences at the transcription termination sites are highlighted by a red dashed box; 0 indicates splice junctions lacking any short-read support. A2. Expression level of the SPCS2 variant in the OPC condition. A3. Validation of SPCS2 transcript expression by RT-PCR. PB.6460.1/NM_025668 were amplificated but XM_006508117 was not. B-D. Characterisation of genes expressing different METs in the PbT and ReT transcriptomes. B. Lowest SJ coverage by short reads in MET genes. C. Lowest mean exon coverage by short reads in MET genes. D. Distance between the TTS of the MET genes and their FSM references. (***) P $<$ 0.001, (ns) not significant.

3.4.8 Open reading frame prediction in long-read defined transcriptomes

The availability of a full-length corrected and curated transcriptome allows us to predict ORFs with high confidence while also annotating the 3' and 5' UTRs. GMST ORF prediction in our curated long-read neural transcriptome generated 9,269 non-redundant ORFs in a total of 10,813 coding transcripts (90.3% of the total transcripts). Most FSM, NIC and NNC transcripts were predicted to have ORFs (97%, 90%, 87.8% and 92.8%, respectively), while the remaining categories were mostly non-coding. To evaluate ORF prediction results, we selected FSM, ISM and 3'UTR fragment subsets representing variants already annotated in reference transcriptomes. The comparison between predicted and reference coding status revealed a very high true-positive rate and a low true-negative rate for the FSM and ISM subset (Table 3.4). These results are in agreement with the protein-coding transcript enrichment performed in the polyA purification step and demonstrates the capacity of GMST to predict the coding region when the isoform has coding potential. Besides, the high FN ratio observed in the 3' UTR fragment subset also suggests that ORF predictors are specific because ORFs in regions devoid of coding potential such as 3' UTRs cannot be predicted (Table 3.4).

		FSM (7,899 Iso)		**ISM (1,392 Iso)**		**3'UTR Fragment (339 Iso)**	
		Predicted		**Predicted**		**Predicted**	
		Coding	*Non Coding*	*Coding*	*Non Coding*	*Coding*	*Non Coding*
Reference	*Coding*	93.7%	1.2%	86.1%	8.4%	18.3 %	80.2%
	Non Coding	3.3%	1.8%	3.9%	1.6%	0%	1.5%

Table 3.4: Confusion matrices evaluating coding prediction across FSM, ISM and UTR3 fragment transcripts.

Moreover, to assess the quality of predictions we compared the protein sequence of true positive cases against their cognate references (Figure 3.14.A): for the FSM subset we found 90.5% of ORFs identical to the reference protein and 7.8% showing a shorter N-Terminus. Instead, for ISM category, just 14.8% of ORFs were identical to the reference protein and remarkably 55%

showed a shorter N-Terminus and 21.6% had a shorter C-Terminus, certifying their fragmented status. Actually, comparison of the size of the N-Ter deletions between FSM and ISM ORFs classified as N-Ter Deletion reflectes that, when present, the shortening of the N-Terminus was much smaller in FSM ORFs (Figure 3.14.B). Finally, the 3'UTR fragment subset were highly enriched in predicted ORFs with major changes (exceeding 70%) or which do not aligning with the matched reference ORFs (30%), indicating that they are enriched in non-coding retrotranscription/degradation fragments. Our results demonstrate that the GMST algorithm can accurately predict the coding sequence in full-length sequenced isoforms and provide true partial coding sequences in cases of partially-sequenced isoforms.

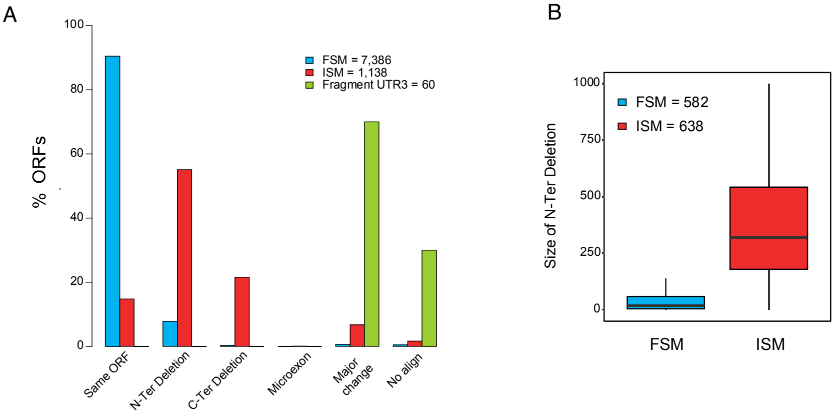


Figure 3.14: ORF predictor benchmarking. A. Types of differences between predicted ORFs and matched reference ORFs. B. Size of N-Ter deletions for FSM and ISM transcripts.

3.4.9 Open reading frame diversity generated by novel long-read defined isoforms

Most of the novel transcripts from the mouse neural transcriptome belong to existing genes (98%). We studied to what extent alternative splicing modifies both coding and non-coding regions of transcripts, and how it impacts the PI of the gene. Approximately, 36% of the genes expressed in our system were multi-isoform genes; of these, 1836 expressed the transcript corresponding to the gene PI and in 592 cases (32%), the PI isoform (57%) was expressed with multiple, distinct UTR regions. Transcripts corresponding to predicted alternative ORFs were expressed in 1,429 genes and in contrast, the UTRs of these non-PI transcripts were much less variable, with only 9% of them showing multiple 3' or 5' UTR variants. Hence, our neural transcriptome shows a significantly higher regulation of UTRs in PI ORFs than in alterantive ORFs, suggesting that further transcriptional regulation of alternative forms might not be required to modulate their functionality.

Finally, evaluation of protein differences in our set of curated transcripts regarding the PI of the gene showed that most of the predicted alternative (Alt-ORF; n = 2127) and novel ORFs (Novel-ORF; n = 1194) are distributed between N-terminal truncations (around 37% for both categories) and major changes (around 53% for both categories) with an enrichment of microexons differences in Alt-ORFs regarding to Novel ORFs (12% vs 3%, respectively).

3.4.10 SQANTI tool

Given the success of the strategies we followed in this work for the in-depth characterisation and curation of long-read transcriptomes, we decided to implement our analysis into an easy-to-use python tool called *Structural and Quality Assessment of Novel Transcript Isoforms* (SQANTI). SQANTI is implemented in Python with calls to R for statistical analyses and to generate descriptive plots. SQANTI has two major functions: sqanti quality control and sqanti filtering (Figure 3.15).

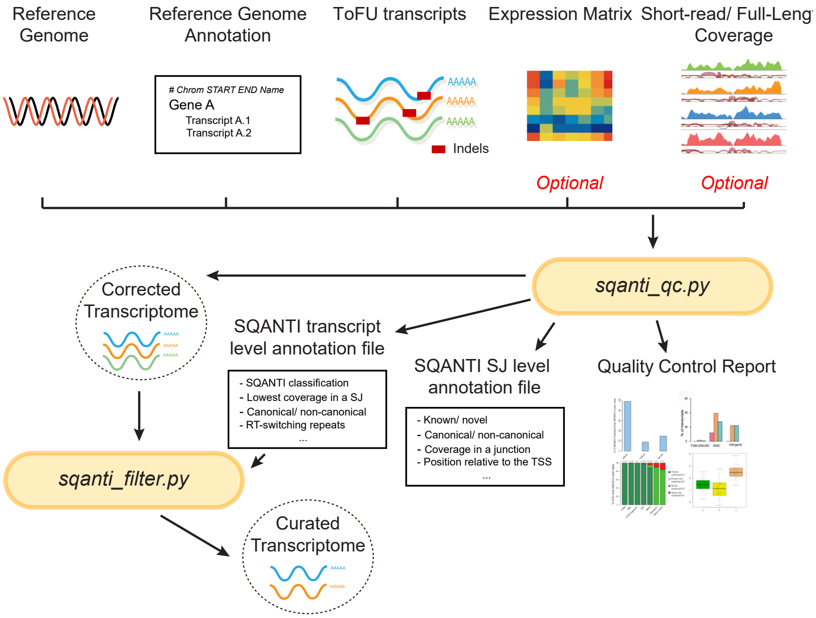


Figure 3.15: SQANTI workflow. SQANTI comprises two main functions: *sqanti_qc.py* uses a FASTA file with transcript sequences, the reference genome in FASTA format, a GTF annotation file, and optionally, full-length and short-read coverage files as inputs. It returns a reference-corrected transcriptome, two characterisation files containing structural classification, transcript and junction-level quality descriptors, and a final QC graphical report. *sqanti_filter.py* uses the reference-corrected transcriptome and the transcript-level attributes file and returns a matching learning-curated transcriptome from which artefacts have been removed.

3.4.10.1 SQANTI quality control

The SQANTI inputs are: a transcript dataset (in gtf or fasta file formats), a genome annotation, and a genome sequence, and it returns a reference-corrected transcriptome and a wide set of transcript and junction attributes in two tabulated text files.To correct transcriptomes, SQANTI aligns sequences to genome references using the GMAP algorithm [370] to obtain a gtf file and a subsequent fasta file and avoiding indels along the exons defined by the alignment step.

A fundamental goal of long-read transcriptome sequencing is to capture the extent of transcriptome complexity while still obtaining full-length transcripts. Thus, SQANTI includes metrics to readily study these aspects as well as to provide a deep classification of transcripts by comparing input isoforms to reference-gene models (Figure 3.2).

Furthermore, QC evaluation of transcripts is sometimes essential to the detection of anomalies in the data or reconstruction pipelines. Thus, all the QC descriptors mentioned and described during this chapter are evaluated and provided along with the SQANTI output files. Appendixes 1 and 2 list the set of descriptors computed by SQANTI at the transcript and junction levels. These files contain 33 and 20 fields, respectively; the first three fields identify the transcript in the reference genome and the remaining fields describe different transcript/junction properties, making a total of 47 SQANTI descriptors.

Moreover, different options can be set to allow users to adjust the evaluation to suit their needs; examples of these options include: the SJ sequences that SQANTI considers as canonical, the size of the genomic DNA window screened for adenine content downstream of TTSs, and alignment parameters. Moreover, extra data can be provided as input so that SQANTI provides an even more extensive QC analysis. These optional inputs that can be useful in the assessment of quality of the tested transcriptome include FL number, junction coverage, and expression quantification data (Figure 3.15).

Thus, in summary, SQANTI QC is implemented in a function called *sqant_qc.py* which performs the following tasks: (1) transcript sequence correction based on the provided reference; (2) comparison of sequenced transcripts with the current genome annotation to generate gene models and classify transcripts according to SJs; (3) ORF prediction using GMST [30]; (4) prediction of RT-switching events; (5) QC characterisation via the analysis of several transcript and junction-level attributes.

3.4.10.2 SQANTI filter

After reference-guided error correction, artefacts may still be present in the resulting transcriptome. SQANTI removes potential artefact transcripts by applying a ML classifier based on the SQANTI features generated beforehand (Figure 3.15). The definition of the true and the artefact sets can be provided by the user when a set of reliable sequences known to be true isoforms and another set of sequences known to be artefacts are available. If not, SQANTI infers these sets by defining FSM transcripts as the positive set and NNC transcripts with at least one non-canonical junction as the negative set. The ML filter trains a Random Forest classifier based on the user's data and following the strategy described above. Hence, SQANTI returns a curated transcriptome from which artefact transcripts have been removed. The SQANTI filter also includes an option to discard transcripts flagged as intra-priming candidates (60% adenines at the genomic 3' end of isoforms) and the curated transcriptome obtained can be evaluated by using the SQANTI QC function to verify the improvement in quality parameters. The SQANTI filter has been implemented in a function named *sqanti_filter.py* which uses *sqanti_qc.py* output information to perform filtering of potential artefacts.

3.4.10.3 Diagnostic plots

SQANTI provides a graphical report generated by R that shows the different evaluated attributes which helps the user to understand the quality and characteristics of the transcriptome, including the distribution of transcript lengths, expression levels, number of exons, the position of junctions, full-lengthiness, and other quality features such as the proportion of non-canonical junctions, evidence of RT switching, and junction coverage by short reads. In addition, SQANTI provides most of these graphs with a transcript category breakdown in order to facilitate quality assessment of the transcriptome obtained by single molecule sequencing. SQANTI is available at `https://bitbucket.org/ConesaLab/sqanti`.

3.4.11 Generalization of the SQANTI approach

To assess the general usefulness of SQANTI, we applied our approach to public datasets from alternative organisms and long-read sequencing technologies (Section 3.2.2).

First, SQANTI was used to analyse human (MCF7 cells) and maize (ear tissue) PacBio datasets (Figure 3.16, A-B). Results indicated that the transcriptome composition in these datasets was substantially similar to our observation for the mouse transcriptome: a significant number of novel transcripts in known genes (50.4% and 38% for MCF7 and ear tissue, respectively, Figure 3.16, A-B) and enriched in low quality features (Figure 3.16.C-D). In each case, we applied the SQANTI filtering approach by training our ML classifier in each case with their sets of FSM and NNC-NC transcripts using default values to remove intra-priming events. As with the mouse data, we obtained high AUC values in the test sets (99.3% for maize ear and 99.7% for MCF-7) and we succeeded in removing a considerable amount of low quality novel transcripts while controlling their enrichment in low quality features (Figure 3.16, C-D). Furthermore, analysis of the importance of SQANTI descriptors for the ML classifier in these datasets with respect to the mouse data revealed noticeable differences (Figure 3.9.A), although in general the top-ranked classification features coincided (i.e. the top three variables were shared among datasets). For example, the number of FL reads was not a highly ranked feature for the maize ear data, probably because the sequencing depth of this dataset was lower and it was absent from the MCF-7 dataset because these values were not available.

Additionally, we assessed the performance of SQANTI when using alternative long-read technologies such as Nanopore. As expected, its higher error rate is probably the cause of the high number of NNC isoforms and exacerbated levels of non-canonical junctions (28.5%) we observed with this technology (Figure

3.17, A). In particular, unlike TSS bias distribution for ToFU PacBio isoforms, we found a similar non-canonical distribution along transcript sequences (Data not shown). We should remember that we hypothesize that the clustering conditions at transcript ends in the ToFU pipeline is more permissive, thus generating this biased positional pattern. Moreover, Nanopore data showed an exacerbated representation of genes with more than 6 isoforms, even in novel genes (Figure 3.17, B). This is probably because there is no collapsing step to remove redundancy in the Nanopore processing pipeline. Finally, the levels of intra-priming detected for this dataset were low (Figure 3.17, C), possibly because of the different cDNA synthesis conditions or preprocessing/filtering read steps in PacBio. Thus, the results of the quality assessment performed by SQANTI greatly help to reveal the characteristics of each particular dataset.

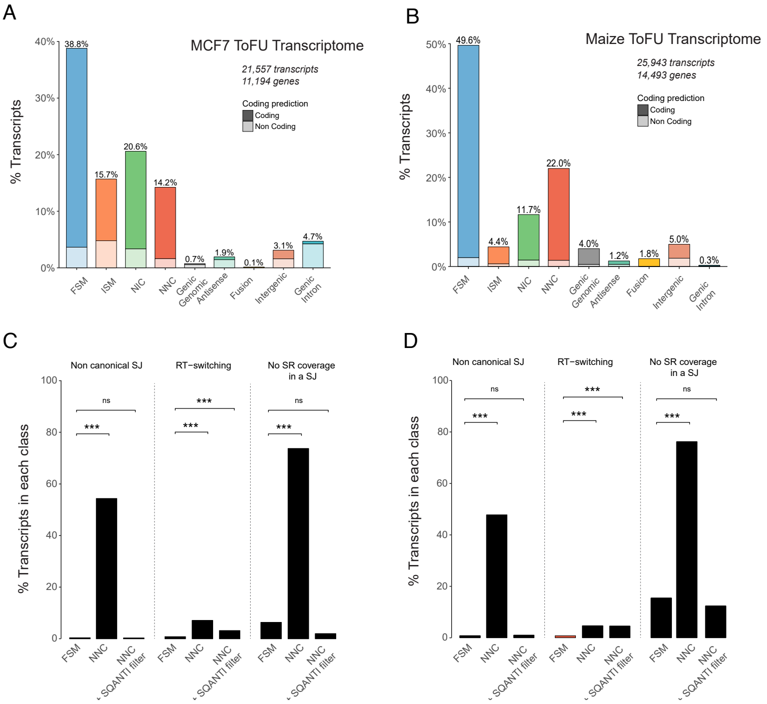


Figure 3.16: SQANTI performance on alternative PacBio datasets. A. SQANTI characterisation of the ToFU-defined transcriptome for the MCF7 human PacBio dataset. B. SQANTI characterisation of the ToFU-defined transcriptome for the maize ear PacBio dataset. C. SQANTI filter results on the MCF7 transcriptome. D. SQANTI filter results on the maize ear transcriptome. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. ns = not significant.

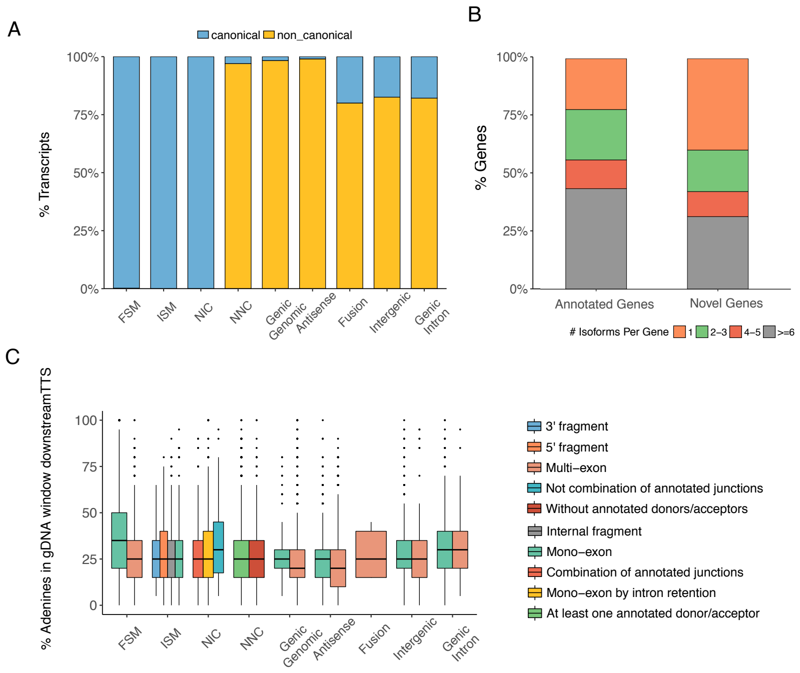


Figure 3.17: Quality control performance by SQANTI on Nanopore sequencing data. A. Rate of transcripts containing non-canonical junctions across SQANTI structural categories. B. Distribution in the number of isoforms per gene across novel and reference genes. C. Intra-priming characterization across SQANTI structural subcategories.

Altogether, this section shows that the SQANTI QC framework is a very useful tool for revealing the structural composition of transcriptomes obtained from long-read sequencing and for evaluating quality of novel calls across different organisms and sequencing technologies.

3.5 Discussion

Long-read sequencing technologies, such as the PacBio platforms or Oxford Nanopore, have brought novel excitement into the challenge of describing the complexity of the transcriptome of higher eukaryotes by providing new means for sequencing full-length transcript models. While early papers concentrated on demonstrating the dramatic enrichment in full-length transcripts achieved by long reads [296] [326], there is an increasing number of publications that describe thousands of new transcripts discovered by this technology. Accordingly, we found that, when sequencing the mouse neural transcriptome using PacBio, a large number of novel transcripts could be detected. However, close inspection of these new transcripts revealed signs of potential errors that required a thorough and systematic analysis of these sequences before making any new transcript calls. This motivated the development of SQANTI, a new software for the structural and quality analysis of transcripts obtained by long-read sequencing.

The three basic aspects of the SQANTI QC pipeline are (1) the classification of transcripts according to the comparison of their junctions to a reference annotation in order to dissect the origin of transcript diversity, (2) the computation of a wide range of descriptors to chart transcript characteristics, and (3) the generation of graphs from descriptors data, frequently with a transcript-type breakdown, to facilitate interpretation of the sequencing output and reveal potential biases in the novel sequences. Using this analysis framework, we were able to show that, at least in our mouse experiment, novel transcripts - especially those in the NNC category - are typically poorly expressed transcripts of known genes, consistent with previous reports [296][319][320]. We also observed that novel junctions accumulate at the 5' end of transcripts, have lower coverage by Illumina reads, and are enriched in non-canonical splicing and direct repeats typical of RT switching.

However, none of these features are exclusive of any of the novel transcripts categories, which invites the question on how best to remove transcript artifacts. This has been solved in the past by either eliminating all novel transcripts with at least one junction not supported by short reads [296], by systematically discarding transcripts with noncanonical splicing [13], or by developing models to estimate the likelihood of a certain splicing event [1]. In our case, we performed an extensive PCR validation of transcripts belonging to different known and novel types. We found a significant number of transcripts, both with canonical and noncanonical junctions, that had complete junction support by Illumina and that were amplified by RT-PCR of the sequenced cDNA library but that failed to be validated when PCR conditions were adjusted to avoid secondary RNA structures. We concluded that these might be cases of retrotranscription artifacts, which would have escaped filtering solely based on short-read support. This result may suggest that a revision of library preparation protocols is needed, which goes beyond the scope of this study. As an alternative, we were able to combine our set of SQANTI descriptors with a machine learning strategy to build a filter that discards poor quality transcripts with better performance than alternative existing approaches. Moreover, the SQANTI filter is data-adaptive, and we showed that it can be successfully applied to other long-read transcriptomics datasets and technologies.

Thereby, SQANTI is designed to leverage genome annotation data to characterize and filter long-read transcriptomes. Where no genome is available or the assembly is low-quality, reference-guided correction of transcript sequences will be compromised and therefore also the accurate translation into ORFs. If, additionally, the gene content annotation is poor, this will impact SQANTI transcript classification, leading to enrichment in novel isoforms and genes. In these conditions, it might be difficult to define robust FSM positive and NNC-NC negative training sets for the SQANTI classifier: the first set, because of the low number of known transcripts, and the second, because of poor correction of PacBio sequences. Subsampling experiments showed that 150-200 training set transcripts would be sufficient to obtain comparable performance to that in 6.3.B, indicating that the SQANTI filter can be used confidently even when reduced training sets are available. Furthermore, the SQANTI set of quality descriptors will be extremely useful in these cases, as they will provide a comprehensive characterization of the quality of the transcript calls in situations where little additional data is available. Finally, note that SQANTI is agnostic to the sequencing technology that generated the transcripts and can accept transcript sequences from other long-read approaches such as Nanopore and Moleculo (Illumina synthetic long-read technology). Obviously, the results of the quality assessment will vary as a function of the characteristics of the underlying technology.

The fundamental advantage of single-molecule, long-read technologies over short reads is their direct detection of full-length isoform diversity including novel transcripts. The availability of a curated full-length transcriptome data set of our mouse neural tissue allowed us to explore these aspects confidently. We found 2,058 novel transcript isoforms, representing 17% of total transcriptome and most of them falling within reference genes, revealing the relevant transcriptional signal hidden by reference gene models and highlighting the need of long-read sequencing for whole transcriptome definition in tissues with extensive post-transcriptional programs as brain.

We also show how high variability at transcript ends is a source of quantification errors that can be alleviated when an expressed full-length reference transcriptome is used. Our data suggests that unannotated alternative polyadenylation events are frequent in mammalian genomes, which, in turn, induce incorrect quantification estimates. Full-length sequencing of the expressed transcriptome readily identifies this 3-end diversity to provide the correct templates for transcript quantification. On the other hand, variability at the 5 end is still an issue for full-length transcriptome sequencing, as biological variability cannot be unequivocally differentiated from technical artifacts in cDNA library preparation protocols. The SMARTer protocol typically used in PacBio sequencing may not always capture the full extension of the 5’ ends due to transcript degradation or incomplete retrotranscription. This may account for the lack of 5’-end coverage observed in FSM and ISM transcripts. Trapping of the 5’ CAP prior to the synthesis of the secondary cDNA strand has been shown to increase the overlap of the 5 end without seriously compromising the yield of long reads [43] and in the future may represent the preferred form of library preparation to study 5-end diversity.

In conclusion, the results presented in this chapter indicate that long-read technologies, as any other large-scale genome technology, are subjected to the accumulation of false positives if proper quality evaluations are not established. However, provided adequate quality control, long-read technologies are effective tools for the characterization of isoform-resolved transcriptomes, the accurate estimation of isoform expression as well as enhancing the study of the biological significance of isoform diversity. Due to effectiveness of SQANTI to maximize the analytical outcome of long-read technologies and deliver quality-evaluated transcriptomes, PacBio recommends it as the standard quality control tool for best practices to analyse Iso-Seq data.

Chapter 4

In-silico annotation of isoforms with functional and regulatory features

4.1 Introduction

Functional profiling is by far the most widely adopted genome-wide approach for those interested in characterising the functional relevance of gene expression regulation [64]. This has been possible by the availability of structured and precisely vocabulary describing the functional properties of gene products such as Gene Ontology (GO) [10], the Kyoto Encyclopedia of Genes and Genome (KEGG) [158][159][160] or Reactome [98]. Despite these resources currently include rich annotation for most model species, functional entries are still recorded at the gene level. Thus, even though both AS and APA mechanisms have emerged as central mechanisms of proteome and transcriptome diversity and playing a key role in lineage determination, cell differentiation or tissue development [18][52], current gene-centric annotation information impedes to study the functional consequences of differential splicing in specific contexts and conditions of interest. Therefore, there is a great need for databases and methods that provide isoform-resolved functional information of gene products.

Trying to cope with functional annotation at isoform resolution, a series of prediction methods have recently appeared [191], [199] [239] [91]. Unlike gene function prediction, computational methods for predicting isoform function are limited by the little functional information available at the isoform level, which makes traditional supervised learning algorithms such as Bayensian networks not directly applicable [190]. Thus, alternative approaches such as the Multiple Instance Learning (MIL) [41] have been recently adopted to deliver isoform-resolved function based on GO terms by mining multiple mouse and human RNA-seq datasets. However, these methods are limited by: (1) the lack of isoform-level gold-standard functional annotation, hindering the evaluation of prediction results and limiting the accuracy of these methods [192]; (2) their high sensitivity to the initial isoform labels inherited from their host genes [199]; (3) the need of large datasets to increase the applicability and reliability of these methods, given that most isoforms are known to be tissue- or developmental-stage specific [94] [374] and their number is steadily growing as the result of application of long-read technologies; (4) the high dependency of these methods on the complexity of the gene annotation reference being used (e.g. RefSeq, GENCODE, Ensembl, etc).

Moreover, genome-wide studies indicate that alternative exons do not radically change or disrupt the function of gene products [351][280] but they do introduce elements that modulate properties such as its enzymatic activity, binding or stability. Hundreds of experimental validations have demonstrated that almost all aspects of protein functions are influenced by alternative splicing [165][304], making the traditional GO annotations not granular enough to characterise the functional properties that differentiate isoforms.

In that direction, UniprotKB [66] annotates a large set of functional domains, motifs and topological regions along protein sequences. UniprotKB is biased towards the annotation of canonical isoforms and fails to capture the divergence between isoform proteins. In contrast, the APPRIS [278][277], ASPicDB [205], VastDB [311] databases provide important resources of isoform-resolved functional information by annotating structural information, domains, transmembrane regions or intracellular location for the different protein isoform sequences defined in reference databases. Nevertheless, the annotation in both these resources is highly biased towards capturing protein-level differences and ignore how AS/APA mechanisms impact the accumulation of regulatory elements in mRNA UTRs to control essential properties as stability, translational efficiency or localization of mRNAs [206][287][151][182][7]. Moreover, the applicability of all these databases, including ASpedia [147] is limited because they are restricted to the annotation of reference gene products in a few range of organisms as human or mouse. Thereby, they cannot cope with the dynamics of the transcript novelty being identified by current sequencing technologies, both in model and non-model organisms [52] [347] [1].

In this chapter we detail the development of IsoAnnot, a new pipeline for the functional characterization of isoforms which relies on the annotation of isoforms as a combination of domains, motifs and functionally relevant sites. IsoAnnot considers an extensive variety of functional properties, both at RNA and protein level, to capture the widest possible range of the functional divergence originating by AS and APA mechanisms. The main advantage of IsoAnnot is its ability to annotate splicing variants based on sequence, thus eliminating the need for large experimental datasets and increasing the scope of the method to annotate novel isoforms obtained from long-read sequencing, both from model and non-model organisms. IsoAnnot provides a dynamic pipeline to extensively annotate isoforms by integrating multiple state of the art sequence predictive tools and are complemented with functional information collected from databases such as UniprotKB and PhosphositePlus, adding experimentally validated elements to our annotation.

4.2 IsoAnnot pipeline

The IsoAnnot pipeline (Figure 4.1) was developed to populate splicing isoforms with rich functional information at the RNA and protein level. Importantly, all the functional labels annotated at isoform variants are defined by protein/RNA coordinates which enables the direct mapping of splicing events to functional elements. IsoAnnot is a pipeline comprising several modules that integrate annotations derived from experimentally validated information stored in public databases and tools based on sequence-prediction.

4.2.1 Input data

IsoAnnot requires three unique pieces of information as input data (Figure 4.1):

1. Isoform sequences, either de novo or from reference databases.

2. The predicted or reference ORF sequence associated with query isoforms in order to annotate functional labels at the coding part of isoforms.

3. Gene Models in GTF format so that functional information from public databases can be transferred and for the prediction of nonsense-mediated decay (NMD).

4.2.2 Functional annotation at transcript isoform resolution

Untranslated regions (UTR) of mRNAs play crucial roles in the fate of mRNAs by the presence of cis-regulatory elements and the availability o binding sites for RNA-binding proteins (RBP) and microRNAs [206][287][151][182][7]. Hence, annotation of isoform-especific regulatory elements becomes essential to characterize the functional properties of UTRs and facilitate the understanding of the influence of APA/AS mechanisms in mRNA fate. IsoAnnot implements different approaches to generate extensive annotation of RNA regulatory features at isoform resolution (Figure 4.1).

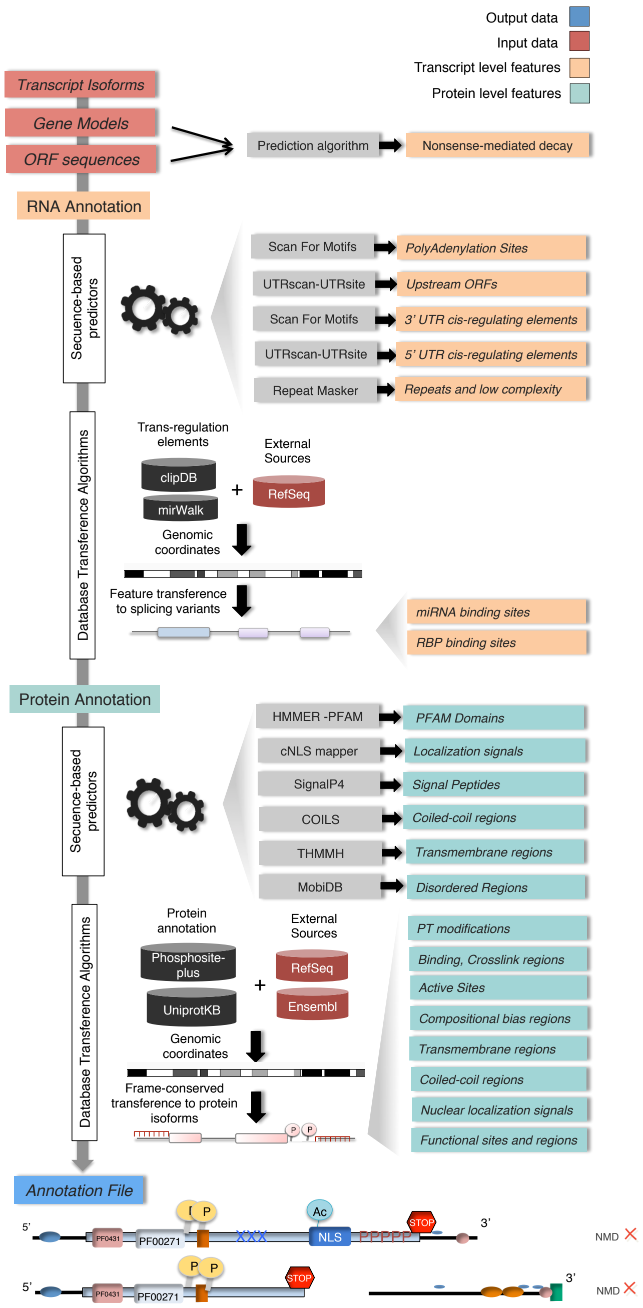


Figure 4.1: Overview of the IsoAnnot pipeline. Isoform-resolved functional annotation is generated by individual interrogation of isoforms using a wide range of methods that generate functional labels at the transcript and protein level.

4.2.2.1 Cis-acting UTR regulatory elements

During last decades, specific functional and regulatory elements at UTR regions have been identified and characterized. Different databases have comprehensively gathered this information and several tools have been developed to annotate them at input RNA sequences. Among them we find RegRNA 2.0 [46], AURA2 [71], AREsite2 [99], UTRsite [252], UTRscan [208] and ScanForMotifs [24], each of them involving a different set of regulatory elements. The ScanForMotifs tool was implemented to identify 3' UTR motifs and UTRsite and UTRscan were used to annotate 5' UTR motifs in the IsoAnnot pipeline. Both are based on the definition of regular expressions and position-frequency matrixes to identify regulatory signals inside query sequences. Positional motifs along isoform-defined UTRs were parsed from output files and assigned independently to each individual splicing variant by IsoAnnot. In the case of ScanForMotifs, a background expectation cut-off (E-value) of 0.175 was used to filter out motifs with a high probability of appearing by chance in a test set of human UTRs [24]. UTR motifs were filtered according to the studied organism. Predicted 3' UTR and 5' UTR elements were labeled by IsoAnnot as two independent annotation layers namely *3UTRmotif* and *5UTRmotif*, respectively.

4.2.2.2 Upstream open reading frame prediction

Upstream Open Reading Frames (uORFs), located in the 5' UTR of mRNAs, are another mayor category of post-transcriptional gene expression regulatory elements. Furthermore, several experimental and bioinformatics studies have revealed that almost half of human transcripts contain uORFs [330][19] and that they are also a common control element in plants [321][342], thus, indicating a conserved functional role. They are usually catalogued as translational repressors because they impact expression of the primary ORF (which encodes the main functional protein) by promoting mRNA decay or decreasing its translation rate.

uORFs are defined as sequence elements with an initiation and a terminationcodon in frame and upstream of the primary ORF. IsoAnnot uses the

UTRscan tool [208] to annotate uORFs by prediction of ORFs (setting ATG as start codon) at 5' UTR regions of query transcripts. Hits are classified into *uORF* annotation category.

4.2.2.3 Repeats and low-complexity elements

Repetitive DNA is a mayor component of eukaryotic genomes. Its regulatory role in transcriptomes has recently been revealed by different studies. As example, Chen et al.reported the gene silencing control effect of a pair of inverted Alus (primate-specific retrotransposed elements) located in the 3' UTR of EGFP transcripts [50].

To account for the functional impact generated by the presence of repeat regions across splicing variants, we implemented RepeatMasker [314], the most commonly used program to search for repeats and low-complexity regions in genomic sequences. It provides a detailed report with information about the nature and the location of each identified repeat within the input sequences. Default cutoff (250) and sensitive mode were specified to guarantee a low proportion of false matches. RMBlast algorithm (http://www.repeatmasker.org/RMBlast.html) was chosen as the search engine to perform the identification of repeats and RepBase database [16], version 20140131) as the repetitive DNA elements reference library. RepBase covers over a hundred model organisms and species of interest including mammalian organisms and plants. Repeat elements were categorized as *repeat* and subdivided into different groups depending on their repeat class or family (low-complexity, LTR, SINE, Simple Repeat,etc).

4.2.2.4 miRNA binding sites

MicroRNAs (miRNAs) are trans-acing elements that post-transcriptionally regulate gene expression by mainly promoting mRNA degradation [14][128]. They are non-coding single-strand RNA molecules (20-25 nt long) which interact with target mRNAs by continuous base-pairing, usually at 3' UTRs [100]. Binding sites are denote as *seed regions* and are located in position 2-7 from 5' end of the miRNAs. Recent experiments have brought new insights into the modulation of pairing affinity based on positional matching and secondary structure and these have been used as rules for the development of predictive algorithms. Thus, currently we find plenty of algorithms that provide miRNA-mRNA interactions based on sequence, physical-chemistry properties or expression levels [310], including DIANAmicroT [241], miRanda [97], miRDB [368], miRMap [334], miRNAMap [146], Pictar2 [174], PITA [167], RNA22 [209], RNAhybrid [176] or Targetscan [2][187]. Moreover, different databases, such as miRWalk [90][307], miRecords [373], TarBase [293], miRTarBase [57] and starBase [189], collect experimentally validated miRNA-mRNA interactions.

However, despite the efforts to collect and predict miRNA binding sites, the miRNA-target annotation remains challenging because of the few number of experimentally validated interactions and the high number of false positives produced by sequence based predictors. Furthermore, there is a lack of consensus among existing predictions meaning that there is very little correlation between predicted interactions. Moreover, the different nature of rules and scoring systems used to measure the probability of binding further complicates their comparison [310].

Because no method has proven preferential performance [310], several unions and intersections between different sources of miRNA binding information has been proposed as a way to improve prediction specificity and sensitivity [293] [177]. A comparison of prediction algorithms by Sethupathy et al. showed that the intersection of results from five different algorithms produced the highest (66.7%) specificity values [293]. Another relevant drawback of several current sources of miRNA binding sites is their lack of comprehensive information about the location of the miRNA seed region in the targeted mRNAs, which is obligaroty for IsoAnnot annotation so that microRNA binding sites can be associated with specific splicing or APA events. Moreover, although recently high-throughput RNA sequencing experiments have led to the definition of a large set of miRNAs, the evidence for some of these flagged miRNAs is dubious [355][134][36], meaning that they must be followed up in subsequent control steps.

To address these problems, we defined an isoform-resolved miRNA binding annotation approach that considers:

1. Several sources of miRNA binding information to intersect results.
2. A microRNA binding site source with information about transcript coordinates.
3. Data about miRNA evidence.

The approach comprises on four steps:

(a) Collecting miRNA binding data. The IsoAnnot miRNA binding approach relies on the mirWalk2.0 database [90], whose predictions are derived from several algorithms: miRWalk, Microt4, miRanda, mirbridge, miRDB, miRMap, miRNAMap, Pictar2, PITA, RNA22, RNAhybrid and Targetscan. Moreover, miRWalk2.0 implements its own predictor, mirWalk, which provides positional information about the seed inside the predicted mRNA target. We downloaded sequence-miRNA interaction information for the total set of annotated genes for each the organism of study. A minimum seed length of 7 bp and a p-value threshold of 0.05 were set as requirements to call miRNA binding sites.

(b) miRNA-binding site filtering: Following previous evaluations [293] [177], we applied a filtering approach for miRNA binding sites based on the number of sources reporting the association. This intersection method allow to decrease the false positive rate and thus, increase the sensitivity. For a mRNA-miRNA interaction to be reported, the association had to be predicted by a minimum of 5 methods which had to include Targetscan, miRanda, and mirWalk, the last one providing transcript coordinate information for miRNA binding sites. For example, in mouse this approach reduced by 92% the number of binding sites reported by 12 prediction methods, going from 33,298,719 to 2,480,531 associations.

(c) Control of miRNA evidence: MirBase [173], a searchable database of published miRNA sequences and annotations, is currently the most complete database of precursorand mature miRNAs. Each miRBae entry represents a hairpin portion of a miRNA transcript, with information on the genome location and sequence of the mature miRNA sequence. All miRBase entries requires an associated publication despite the criteria for calling miRNAs entries can be different. For that reason MirBase annotates the experimental evidence level (non-experimental, cloned, Northern, qPCR, RAKE, miRNA-seq, etc) to each miRNA entry. We parsed the miRBase database in order to define our set of *high confidence* miRNAs by including only those entries with the following experimental evidence: cloned, Northern, PCR, RT-PCR, qRT-PCR, 5'RACE, RTPCR, in-situ, qPCR, miRAP cloned, 3'RACE, insitu, RACE, miRAP, primer-extension, RAKE. Hence, the miRNA information returned from miRWalk2.0 was filtered according to its miRNA confidence level.

(d) Genomic annotation of binding sites: Transference of miRNA binding sites from source to query isoforms is performed by genomic mapping, ensuring seed-region conservation in the 3' UTR region of the query isoforms. As miRNA-mRNA interaction sites are provided by mirWalk2.0 in transcript coordinates, we mapped them to genomic coordinates based on the exon-intron information from the mRNA template. MiRWalk2.0 also uses gene models from RefSeq version 61 as reference templates for the annotation of miRNA binding sites, and so RefSeq 61 sequences and associated exon coordinates were downloaded and used to obtain genome positional information for each reported miRNA binding site, conserving strand and gap alignment information from the seed mapping.

(e) Isoform-specific transference: Finally, the transfer of miRNA binding sites to query splicing variants is performed by using the genome-coordinate annotation generated in the previous step. Only complete, contiguous and strand-specific matches of the seed region in the query isoform are annotated by IsoAnnot. The miRNA binding sites are annotated in the final IsoAnnot output file as *miRNA binding site* elements.

4.2.2.5 RNA binding protein binding sites

RNA binding proteins (RBPs) are *trans* factors that bind to pre-mRNAs, and play essential roles in the fate and metabolism of RNA, controlling processes like transport, AS, RNA editing, polyadenylation, stabilisation, and localisation [198][119]. To study the differential targeting of RPB on transcript variants and their regulatory and functional implications, IsoAnnot incorporates data from crosslinking immunoprecipitation (CLIP) sequencing, a technology widely used for the transcriptome-wide identification of protein binding sites on RNAs [171].

We collected CLIP information data from CLIPdb [378], a public resource that stores more than 426 CLIP datasets from 119 different RBPs in four different organisms (mouse, human, worm and yeast) and several tissues. Data is gathered from public repositories such as the Gene Expression Omnibus (GEO), the National Center of Biotechnology (NCBI) and the European Nucleotide Archive (ENA) from the European Bioinformatics Institute (EBI). Each CLIP dataset is provided already analyzed by two alternative peak calling methods (Piranha, Paralyzer, CIMS, CITS), whose choice depends on their specificity to the CLIP technology used to generate the data (PAR-clip, HITS-CLIP or iCLIP).

Before mapping RBP binding sites to the IsoAnnot query sequences, we assessed the quality of the data stored in CLIPdb. Characterisation of the RBP binding patterns across replicates showed poor binding-site agreement as well as high variation in the number and location of detected binding sites, which was dependent on the peak-identification approach used (data not shown).

Given these limitations, we also considered the use of RBP binding site predictors such as DeepBind, FIMO or Tess. However, evaluation of results revealed a large number of hits for most of the considered RBPs (Data not shown), which is likely due to the short and degenerate nature of RBP motifs [196].

Therefore, our pipeline for annotating RBP binding sites uses CLIP data and comprises two main steps:

(a) Curation of data from the CLIP database: To remove potential false-positive binding sites and improve specificity, we filtered binding sites according to their genomic intersection across alternative peak calling tools. Based on our evaluation of the CLIP data (data not shown), we established a 200 bp window as the cutoff for considering binding sites as originating from the same event. Thus, we minimised non-overlapping events resulting from small genomic shifts during the definition of the binding site across peak-calling strategies and generate a set of confident binding sites that reduce the impact of using alternative programs of analysis in the final outcome.

(b) Transfer of RBP binding-sites to query isoforms: We transferred RBP binding sites to isoforms overlapping genomic positions. As RBPs are known to bind the pre-mRNA in order to modulate processing mRNA steps such as alternative splicing, binding sites falling into intron regions were also considered and annotated. Furthermore, each mRNA-RBP association is further catalogued based on the mRNA region it binds: the 3' UTR, the 5' UTR or the CDS and the intron or the exon. RBP binding sites are annotated within the *RBP binding-site* category.

4.2.2.6 Polyadenylation signals

When using sequencing technologies, the identification of PAS is essential because this allows true APA sites to be discriminated from technical artefacts such as fragmented 3' ends, minimizing the definition of false APA novel transcripts. In mammalian systems, two different motifs are known to provide the signals for the definition of the polyadenylation site: (1) The ubiquitous AAUAAA element, located 20-30 nucleotides upstream of the cleavage site where the poly(A) is added, and (2) a more variable GU-rich sequence located immediately downstream of the previous one [21] [260]. In all cases, the recognition of these sequences by specific proteins leads to mRNA cleavage and subsequent polyadenylation. Methods as UTRscan [208] and ScanForMotifs [24], which make use of sequence patterns and the PatSearch algorithm [252], are available to identify PAS elements in nucleotide sequences. IsoAnnot incorporates both methods in order to identify the position of PASs from each individual query isoform sequence and labels this annotation category as *PAS*.

4.2.2.7 Nonsense-mediated decay

Given the spread and relevant role of NMD couple to AS in the regulation of gene expression programs across eukaryotes and their functional implication in a wide spectrum of biological processes and physiological circumstances (See Section 1.4.0.3), IsoAnnot predicts isoforms containing a premature termination codon (PTC) potencially leading to NMD using the 50-NT rule [389]: a termination codon that falls more than 50-55 nt upstream of an exon-exon junction is a general indication of a PTC while normal termination codons are largely located in the last exon. Isoforms with PTCs will be potential to be detected and degraded via the NMD machinery [47][195].

4.2.3 Functional annotation at protein isoform resolution

To define functional regions inside the coding region of query isoforms, IsoAnnot takes advantage of the wide range of predictive algorithms available and the experimental information stored in UniprotKB [66] and PhosphoSitePlus [143] databases. However, of note, the lack of positional information associated with predicted features considerably limits the spectrum of prediction tools that IsoAnnot can use.

4.2.3.1 Pfam domains

Pfam is one of the largest collections of protein domain families [261]. Each Pfam entry identifies a protein family domain that is represented by multiple sequence alignments and a hidden Markov model (HMMs). Querying a given protein sequence in the Pfam library of HMMs allows its different structural units and domains to be determined.

We used InterProScan5 [263], a package that combines different tools for proteins annotation and which includes a module for predicting functional domains using the HMMER3 algorithm [102] that interrogates query protein sequences with the Pfam HMM database of domains (Pfam version 31). InterProScan5 was run locally for each query protein to retrieve positional information about predicted domains. In-house parser algorithms for InterProScan XML files were then used to collect and adapt the results to the IsoAnnot output annotation file, in which Pfam predicted domains were catalogued as *DOMAIN* entries.

4.2.3.2 Transmembrane domains

Transmembrane (TM) domains are stretches of approximately 25 hydrophobic residues with an occasional polar residue of integral proteins that pass across the membrane[268] and play an important role in signalling, molecule transport, energy and cell adhesion [268]. A wide range of tools for predicting TM regions are currently available (e.g., Phobius, TMHMM, Memsat). We automatically annotate these domains along every query protein isoform by using TMHMM (default parameters) [175] - the most widely used and best-performing tool for the prediction of TM regions [214]. These TM regions are then annotated as *TRANSMEM* in the final IsoAnnot functional annotation file.

4.2.3.3 Signal peptide

Signal pepides are hydrophobic sequences found at the N-terminal of secretory pathway proteins that promote their translocation to the reticulum membrane [343]. Therefore, the presence of these motifs provide information about protein localisation and their destination after synthesis. Many methods have been developed for signal peptide prediction, including SignalP, PrediSi, Phobious, and Signal-BLAST. The IsoAnnot pipeline uses SignalP 4.0 [253] because of its high prediction accuracy [169]. SignalP uses a neural network-based method that allows to discriminate signal peptides from N-terminal transmembrane regions, high hydrophobic regions, and non-contatining signal peptide proteins (cytoplasmatic proteins). The IsoAnnot pipeline captures this signal peptide information by locally running InterProScan5, which implements SignalP 4.0 with its default parameters. We labeled this functional annotation layer as *SignalP*.

4.2.3.4 Coiled regions

The alpha-helical coiled coil is one of the principal structural subunits in proteins. Their main characteristic is to follow a heptad repeat pattern of 3-4 residues whose composition and hydrophobicity is compatible with the structure of alpha helices. Despite their simplicity, coiled-coil motifs have been revealed as a versatile folding motif with important roles in protein refolding processes, signal-transducing events, molecular recognition systems and are involved in the mechanical stability and movement processes inside cells, among others [38].

IsoAnnot pipeline uses COILS [200] to predict protein coiled-coil regions. COILS aligns query sequences to a database of known parallel two-stranded coiled-coils and generates a metric that indicates the probability of a given sequence to form a coil structural motif by computing a similarity score and posterior comparison to the distribution of scores in globular (non-coiled-coil proteins) and coiled-coil proteins. Default parameters and local InterProScan5 were used to ran COILS. Coiled regions were defined as *Coiled-coil* in the IsoAnnot output.

4.2.3.5 Disordered regions

It was recently discovered that alternatively-spliced exons are enriched in intrinsic disordered regions (IDRs) [62] [280] [250] [126]: protein regions that do not adopt a well-defined conformation and have structural plasticity [331]. However, their low evolutionary conservation difficulties their accurately prediction [331]. The IsoAnnot pipeline implements MobiDB Lite [258][224], a novel IDR prediction software that combines 8 different predictors to derive a consensus prediction that discriminates functional IDRs from ambiguous hits and outperforms single methods when annotating long ID regions. InterProScan5, implementing MobiDB Lite, was used to annotated IDRs along protein sequences, namely *DISORDERED* in the IsoAnnot output file.

4.2.3.6 Nuclear localization signals

IsoAnnot also predicts the nuclear localization of proteins by annotating the main elements promoting the subcellular movement to the nucleus: the nuclear localization signal (NLS). There have been many attempts to accurately predict nuclear localisation of proteins. Despite several different signals have been reported, most of prediction methods focused on the inference of the best-characterized nuclear signal, the classical NLS (cNLS), which is recognised by importin-alpa [80]. cNLSs can contain one or two regions of basic amino acids and thus are divided into monopartite and bipartite groups, respectively. However, the exact identification of NLSs in protein sequences is still a task difficult to address due to the fact that NLSs are short and remain poorly defined, hindering the design of accurate predictors [203]. Despite these drawbacks, some NLS prediction tools have been developed, including PredictNLS [61], NucPred [31], WoLF PSORT [144], cNLS mapper [172], NLStradamus [226]. Most of them do not provide NLS positional information in the query protein sequence and hence, cannot be used to associate these motifs to splicing events. Among these tools, both cNLS mapper and NLStradamus are the only programs that provide specific coordinates for predicted NLS motifs in proteins. However, we discarded NLStradamus because it can only be applied to nuclear proteins and IsoAnnot performs functional annotation of coding sequences without previous knowledge.

Therefore, we decided to use cNLS mapper [172] in the IsoAnnot pipeline to predict of cNLS signals. This tool calculates scores for NLS activity instead of using the conventional similarity search or ML strategies. Following authors recommendations, we set a minimum cutoff score of 6 to report sequences as predicted NLSs; scores around 6 are indicators of proteins that are partially localized in the nucleus, while scores higher than 8 identify proteins with stronger NLS activity and exclusively localized in the nucleus. IsoAnnot parses NLS mapper output and extracts the exact localisation of each predicted cNLS signal, recording it as *NLS* in IsoAnnot output.

4.2.3.7 Coordinate-based and in-frame transference of protein functional features

In addition to prediction methods, some protein-centric databases contain a detailed annotation of protein features. Their main drawback is that they are biased towards the annotation of the canonical or the best-documented isoform and hence do not capture the the functional diversity of isoforms. Integration of such a comprehensive set of high-quality manually annotated functional information is essential to improve the prediction of isoforms functions and their contextual modeling. Therefore, we added an extra layer of annotation to the IsoAnnot pipeline to account for this by implementing a module that can assign functional features from public databases to the set of query isoforms.

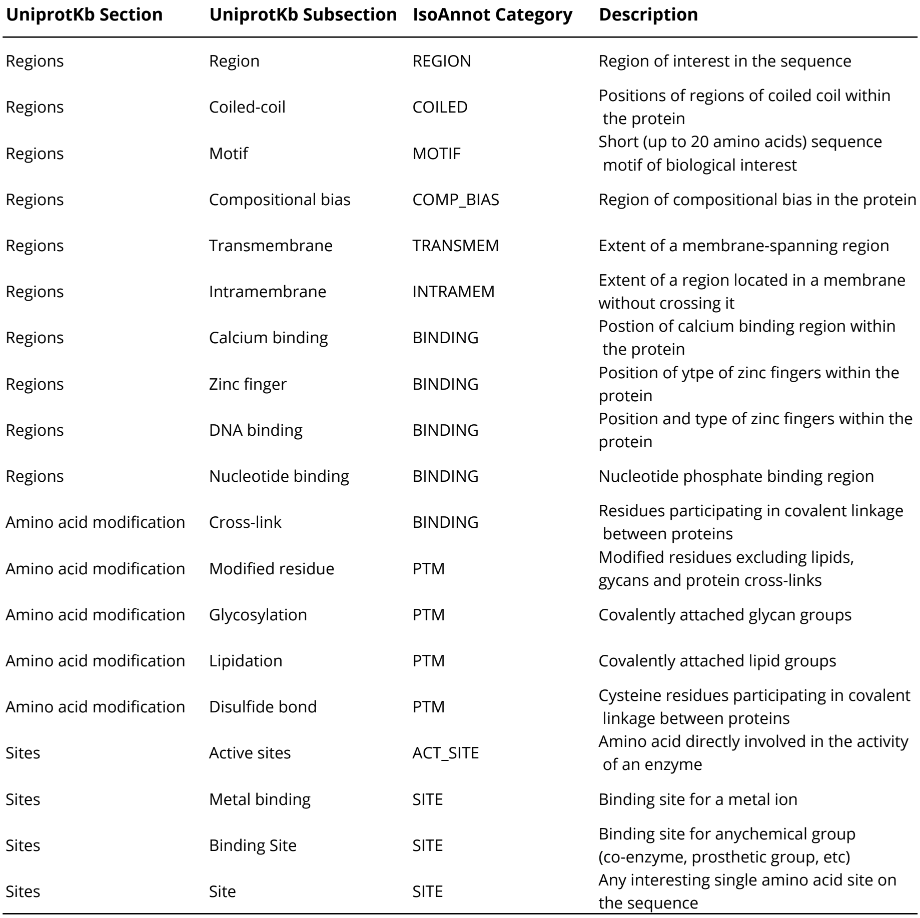
Resources of functional features

Two different sources of functional information were considered:

- **UniProt Knowledgebase:** The Uniprot Knowledgebase (UniprotKB) [8] is the section of Uniprot that contains one of the main collections of functional protein information for more than 10k species. UniprotKB is divided into two sections: Swiss-Prot UniprotKB contains manually-annotated records while Trembl UniprotKB contains computer-generated proteins enriched with automated classification and annotation data. Functional information for the representative canonical sequence is stored in three main sections inside each protein entry: comment lines (CC), feature table (FT) and keyword lines (KW). FT is the section that systematically provides protein coordinates for functional information. UniprotKB integrates the current state-of-the-art protein functional knowledge that can be leveraged to generate a more meaningful functional labels for splicing variants. Table 4.1 shows the different FT categories used for annotation in query isoforms are shown.

- **PhosphoSitePlus:** Post-translational modifications (PTMs) play a fundamental role in the regulation of protein folding, protein targeting to subcellular compartments and signalling. The gain or loss of PTM sites by post-transcriptional mechanisms as APA/AS can increase their molecular versatility by affecting either the contextual control of localisation or signaling of a given gene.

PhosphoSitePlus (PSP) [143] is the main resource dedicated to the annotation of PTMs in mammalians. PSP data is mainly derived from mass spectrometry experiments and includes a wide range of PTM categories including glycosylation, sumoylation, ubiquitination, methylation, phosphorylation and acetylation. All of them were considered by IsoAnnot and catalogued as *PTM* in our output annotation file.



UniprotKb Section	UniprotKb Subsection	IsoAnnot Category	Description
Regions	Region	REGION	Region of interest in the sequence
Regions	Coiled-coil	COILED	Positions of regions of coiled coil within the protein
Regions	Motif	MOTIF	Short (up to 20 amino acids) sequence motif of biological interest
Regions	Compositional bias	COMP_BIAS	Region of compositional bias in the protein
Regions	Transmembrane	TRANSMEM	Extent of a membrane-spanning region
Regions	Intramembrane	INTRAMEM	Extent of a region located in a membrane without crossing it
Regions	Calcium binding	BINDING	Postion of calcium binding region within the protein
Regions	Zinc finger	BINDING	Position of ytpe of zinc fingers within the protein
Regions	DNA binding	BINDING	Position and type of zinc fingers within the protein
Regions	Nucleotide binding	BINDING	Nucleotide phosphate binding region
Amino acid modification	Cross-link	BINDING	Residues participating in covalent linkage between proteins
Amino acid modification	Modified residue	PTM	Modified residues excluding lipids, gycans and protein cross-links
Amino acid modification	Glycosylation	PTM	Covalently attached glycan groups
Amino acid modification	Lipidation	PTM	Covalently attached lipid groups
Amino acid modification	Disulfide bond	PTM	Cysteine residues participating in covalent linkage between proteins
Sites	Active sites	ACT_SITE	Amino acid directly involved in the activity of an enzyme
Sites	Metal binding	SITE	Binding site for a metal ion
Sites	Binding Site	SITE	Binding site for anychemical group (co-enzyme, prosthetic group, etc)
Sites	Site	SITE	Any interesting single amino acid site on the sequence

Table 4.1: Sequence annotations in the parsed UniprotKB database and transferred to query isoforms by IsoAnnot pipeline. UniprotKB functional information describes regions or sites of interest in protein sequences

As PhosphositePlus contains information only for mammalian genomes, it will be considered by IsoAnnot pipeline when annotating mammalian proteomes. Other databases as dbPPT for plants [54], dbPAF for animals and fungi [329] and dbPSP for procaryotes [238] are being in process of implementation. Nevertheless, IsoAnnot currently considers PTM annotation for non-mammalian organisms because UniprotKB provides PTM information for a wide range of available species.

Methodology

We obtained the information that describes protein functional features by parsing UniprotKB and PhosphoSitePlus. In both cases, feature coordinates are referenced to UniProt protein sequences. Therefore, we downloaded the Swissprot, Trembl and VarSplice proteome sequences. Because we are developing approaches to annotate query sequences without gene reference information, the feature transference process must ensure that (1) the query and reference proteins are translated from transcripts belonguing to the same genomic region, and (2) that the feature protein region has a conserved location and ORF.

The first step in the approach is the genomic mapping of reference proteins to genomic coordinates (Figure 4.2). UniprotKB does not contain information about genomic features, but it does provide *cross-reference* information reporting the association between UniProt proteins and Ensembl/RefSeq entries. Thus, we parsed and used gene models and protein sequences from RefSeq and Ensembl repositories together with cross-reference information in order to translate Uniprot proteins containing functional features into genomic coordinates (Figure 4.2). However, because this step involves the integration of isoform information from different databases (Figure 4.2), discrepancies in their isoform representation or ID matching possess significant hurdles and provoke some proteins and features to be discarded from downstream analysis.

These discrepancies include protein differences between cross-linked UniProt and Ensembl/RefSeq entries. We kept UniprotKB-Ensembl/RefSeq associations if the UniprotKB protein matched the protein encoded by the associated Ensembl/RefSeq transcript and allowed a maximum of 3 nt mismatches in order to account for genomic variability between sources. For example, in mouse, we discarded 5% of UniProt entries due to missing cross-reference information and 7% because a lack of protein sequence matching, resulting in a total of 88% of mouse UniProt entries translated into genomic coordinates. Next, the CDS genomic positions of Ensembl/RefSeq transcripts were associated with UniProt proteins in order to obtain the genomic coordinates of the containing UniprotKB or PhosphoSitePlus features (239159 unique functional features mapped to genomic coordinates in mouse, Figure 4.2).

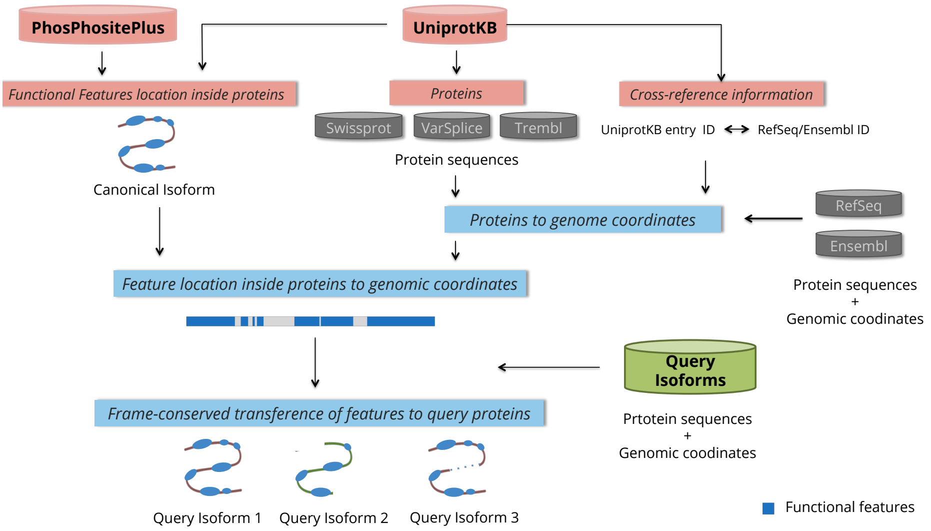

PhosPhositePlus
UniprotKB
Functional Features location inside proteins
Proteins
Cross-reference information
Swissprot
VarSplice
Trembl
UniprotKB entry ID ↔ RefSeq/Ensembl ID
Canonical Isoform
Protein sequences
Proteins to genome coordinates
RefSeq
Ensembl
Feature location inside proteins to genomic coordinates
Protein sequences + Genomic coodinates
Query Isoforms
Frame-conserved transference of features to query proteins
Prtotein sequences + Genomic coodinates
Query Isoform 1
Query Isoform 2
Query Isoform 3
Functional features

Figure 4.2: Pipeline for the isoform-resolved transference of protein functional information contained in UniprotKB and PhosphoSitePlus resources to query isoforms, which must to be defined by their genomic coordinates and ORF sequence.

Finally, query isoforms are interrogated for annotation with UniprotKB and PhosphoSitePlus features by checking that feature coordinates overlap consecutive positions in the query CDS and that the reading frame meets. A splicing event or APA event that either breaks the continuity of the feature or modifies the reading frame provokes the feature to be discarded for annotation and is identified as a lost feature in the query sequence. Genomic coordinates for positive features are recalculated according to the their position in the query sequence, translated into protein coordinates and added to the IsoAnnot output file (Figure 4.2).

The IsoAnnot transfer pipeline is able to populate query sequences with manually-curated features from public databases and resolve annotation at the isoform-resolution level. The approach can be easily adapted to any other source of protein functional features provided that genomic cross-reference information is available.

4.2.4 Annotation of non-positional functional information

We complemented functional information at the isoform-resolution level with labels that describe the processes and pathways in which those genes are involved in, considering data from GO [10], Reactome [98] as well as the manual annotation of protein complexes from CORUM 2.0 [283]. Despite these annotation categories are not isoform specific because isoform resolution is limited in source databases and thus they cannot be used to infer functional AS changes, they contain relevant functional information to categorise genes with AS by their participation in specific pathways or biological processes.

4.2.5 IsoAnnot output

All this functional information is stored in a gff-like file that precisely describes the type, ID, position and source of each collected label, which facilitates computational processing. Additionally, this file also contains information about the structural characterization of isoforms (exons, CDS, UTRs), their mapping to reference genes and proteins and their assignment to unique IDs, all of which are essential to establish relationships between different isoforms regarding the gene they belong to and the protein they are predicted to encode. For example, different isoforms with identical CDS are easily identified in IsoAnnot output since they share common CDS IDs (de novo IDs are automatically generated for CDSs that are not annotated in reference databases and associated to novel isoforms with identical CDSs).

4.3 Results

The IsoAnnot pipeline was used and adapted to annotate different transcriptomes depending on the studies and collaborative projects we were working on. Both long-read defined transcriptomes and reference transcriptomes from RefSeq and Ensembl databases were annotated at the isoform level. Morever, even if IsoAnnot was created for the annotation of mammalian transcriptomes, we are currently expanding it to generate comprenhensive annotation of non-mammalian organisms. Isoform-resolved annotation results for alterantive transcriptomes will be detailed ahead.

4.3.1 Functional annotation of PacBio-defined neural transcriptomes

In this thesis, we have used long-read technologies to define transcriptome complexity in two different murine neural systems:

- **Oligodendrocyte differentiation transcriptome (OLIGd):** Following the strategy defined in Section 3.15, we sequenced NPC and OPC samples using PacBio Iso-Seq and pooled the results to define a transcriptome specific to the oligodendrocyte lineage. A total of 11,970 splicing variants in 7,167 genes were defined; 90.3% of the isoforms were predicted to be coding and 9,546 non-redundant ORFs were obtained.

- **Neural differentiation time course (NEURALtc):** We induced NPCs to differentiate into both OPCs and motor neurons (MNs), harvesting cells from different time points and sequencing their transcriptome using PacBio IsoSeq. This resulted in 34,104 transcript variants; 95.4% were catalogued as coding transcripts and encoded 21,268 non-redundant ORFs.

We ran the IsoAnnot pipeline independently for both the OLIGd and NEURALtc transcriptomes and generated output files which included about 1 M and 3 M elements describing functional and structural mRNA variants, respectively. Among them, 386,114 and 1,171,622 elements represented positional functional features in the OLIGd and NEURALtc transcriptomes, respectively. Tables 4.2 and 4.3 show their respective classification into the different categories considered by IsoAnnot at the transcript and protein levels.

Transcriptome	Source	Category	# Features	# Isoforms
OLGd 11970 isoforms 7167 genes	ScanForMotifs	PAS	8511	5750 (48%)
	ScanForMotifs	3' UTR motifs	11797	5325 (44%)
	UTRscan/UTRsite	5' UTR motifs	325	315 (3%)
	UTRscan/UTRsite	uORF	7444	3045 (25%)
	RepeatMasker	Repeats	19269	7245 (61%)
	miRWalk/miRbase + in-house scripts	3' UTR miRNA binding sites	106392	9474 (79%)
	clipDB + in-house scripts	RNA-binding sites	47821	7279 (61%)
NEURALtc 34104 isoforms 12563 genes	ScanForMotifs	PAS	53815	28629 (84%)
	ScanForMotifs	3' UTR motifs	39831	16873 (49%)
	UTRscan/UTRsite	5' UTR motifs	1069	1021 (3%)
	UTRscan/UTRsite	uORF	23983	11816 (35%)
	RepeatMasker	Repeats	57307	22214 (65%)
	miRWalk/miRbase + in-house scripts	3' UTR miRNA binding sites	288213	28400 (83%)
	clipDB + in-house scripts	RNA-binding sites	157663	22728 (67%)

Table 4.2: IsoAnnot coverage results for features annotated at transcript level.

We found huge differences in the number of annotated elements between the different functional layers (Tables 4.2 and 4.3). While categories such as miRNA binding sites (3' UTR miRNA binding sites) and PTMs include hundreds of thousands of elements, others such as signal peptides (SIGNAL) or 5' UTR motifs included only hundreds of elements in very few isoforms. However, this variability in the annotation coverage of the different functional/regulatory categories certainly agrees with their biological role and their inclusion level in transcriptomes. Signal peptides are only expected to be located in isoforms targeted to the secretory pathway and proteins only need to contain one element to fulfil this role. Consequently, we obtained very few genes with signal peptide features (around 7% of genes for both annotated transcriptomes, Figure 4.3 A-B) and each one had just one annotated signal peptide (Figure 4.4 A-B). In contrast, miRNA binding sites and PTMs are known to be widely expanded layers of post-transcriptional and post-translational regulation, respectively. In aggreement, they are contained in a vast array of genes (60% of genes for miRNA sites and 75% for PTMs, Figure 4.3 A-B) and a median of 10 features per isoform (Figure 4.4 A-B) arising from the known versatility of miRNAs and PTMs in the regulation of single target isoforms.

Other categories as predicted Pfam domains appeared in more than 75% of genes (Figure 4.3 A-B). However, even through Pfam domains populated most of the defined genes, the third quartile for the distribution of the number of domains per isoform was two for both annotated transcriptomes (Figure 4.4 A-B). This aggrees with previous studies that showed an average of 3 domains per protein with a predominance of single-domain proteins [121] [269].

IsoAnnot annotation also revealed relevant differences in the characteristics of our defined transcriptomes such as PAS annotation. Our data revealed a

Transcriptome	Source	Category	# Features	#Isoforms
OLGd 10813 coding isoforms 7167 genes	In house scripts	NMD	329	329 (3%)
	PFAM-HMMER3	DOMAIN	20973	9608 (89%)
	COILS + UniprotKB	COILED	6669	2856 (26%)
	TMHMM+ UniprotKB	TRANSMEM	12543	2061 (19%)
	SignalP	SIGNAL	824	824 (8%)
	MOBIDB	DISORDERED	11256	5626 (52%)
	cNLS mapper + UniprotKB	NLS	7599	4297 (40%)
	PSP + UniprotKB	PTM	100804	8506 (79%)
	UniprotKB	COMBIAS	2260	1480 (14%)
	UniprotKB	MOTIF	6579	2897 (27%)
	UniprotKB	INTRAMEM	159	62 (0.6%)
	UniprotKB	ACTSITE	1770	1168 (11%)
	UniprotKB	BINDING	12790	3339 (31%)
NEURALtc 32119 coding isoforms 12563 genes	In house scripts	NMD	1349	1349 (4%)
	PFAM-HMMER3	DOMAIN	63462	29296 (91%)
	COILS + UniprotKB	COILED	19882	8410 (26%)
	TMHMM + UniprotKB	TRANSMEM	41227	7003 (22%)
	SignalP	SIGNAL	3100	3100 (10%)
	MOBIDB	DISORDERED	39014	18492 (58%)
	cNLS mapper + UniprotKB	NLS	14255	9227 (29%)
	PSP + UniprotKB	PTM	297294	25360 (79%)
	UniprotKB	COMBIAS	8876	5426 (17%)
	UniprotKB	MOTIF	18608	8481 (26%)
	UniprotKB	INTRAMEM	345	151 (0.5%)
	UniprotKB	ACTSITE	5110	3443 (11%)
	UniprotKB	BINDING	37219	9692 (30%)

Table 4.3: IsoAnnot coverage results for features annotated at protein level

lower proportion of isoforms with PAS in the OLIGd transcriptome than in the NEURALtc one (48% vs. 84%, Figure 4.2), likely because the higher proportion of 3' end incomplete isoforms in OLIGd. This correlates with the improvement of the PacBio sequencing technology and the achievement of a higher sequencing depth in the NEURALtc transcriptome sequencing, as will be described in Chapter of this thesis 6.

As previously described, functional labels were assigned to isoforms by following two main different approaches: *a feature-transfer pipeline* based on pub-

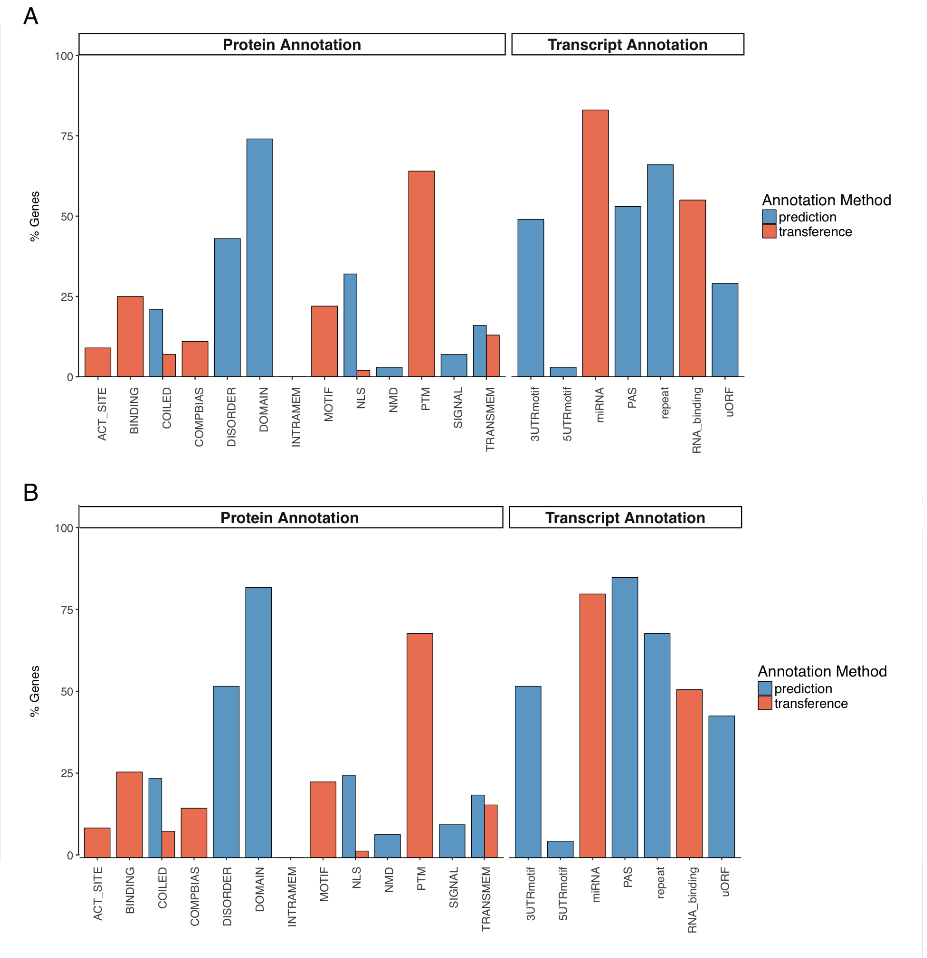


Figure 4.3: Number of genes annotated with each feature type at transcript and protein level in (A) OLGd and (B) NEURALtc transcriptomes. Each different element is further classified in function of the annotation method used for its association.

lic databases and *sequence prediction*. Some functional categories such as coiled-coil regions, transmembrane regions or NLSs were annotated by using both approaches. We found a similar distribution of the number of elements per isoform in these annotation layers (Figure 4.4 A-B) but a higher frequency of annotated genes when using predictive methods (Figure 4.3 A-B) due to the inherent limitation of transferred methods to only consider features and genes contained in the set of databases used. Interestingly, the intersection of genes containing elements annotated by the two different methodologies revealed that around 90% of the elements recovered from databases overlap predicted elements (Figure 4.5), which supports the quality of the prediction algorithms.

Because PacBio-defined transcriptomes includes novel isoforms, we evaluated and compared the annotation coverage across novel and known isoform categories (reference annotated isoforms) to investigate if IsoAnnot transfer methods are biased towards the annotation of known isoforms. We assume that this kind of bias does not exist for prediction methods because they are based on sequences.

First, isoforms from novel genes were compared to isoform from known genes. We found that, at transcript level, novel gene isoforms in both transcriptomes showed a slightly but consistent lower annotation coverage than those from known genes, both for predicted elements such as UTR motifs and for database-transfer elements such as miRNA or RNA binding sites (Figure 4.6.A-B). Conversely, repeats seem to accumulate in isoforms from novel genes (Figures 4.6.A-B), which are highly enriched in non-coding mRNAs, especially in the OLIGd transcriptome where only 19% of novel gene isoforms are coding. This agrees with the reported enrichment of repetitive regions in non-coding transcribed sequences [294]. At protein level, independently of the annotation approach and category, a reduced number of coding isoforms from novel genes contain protein features, indicating that, in general, coding isoforms in novel genes usually have less functional loading (Figures 4.6.A-B). Moreover, this trend is accentuated for categories annotated exclusively by transfer of elements from public databases (such as binding or compositional regions), where

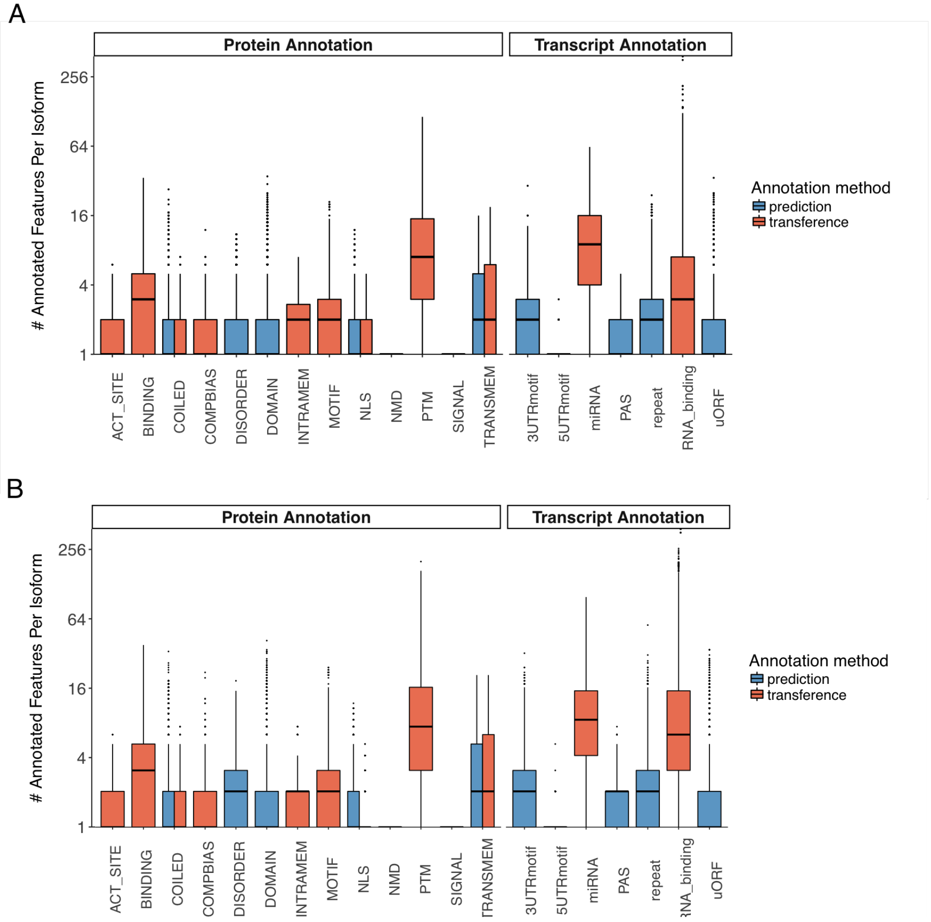


Figure 4.4: Distribution of the number of elements annotated for feature type at transcript and protein level in (A) OLGd and (B) NEURALtc transcriptomes. Each different element is further classified in function of the annotation method used for its association.

null feature coverage is reached in most cases (Figures 4.6.A-B). This verifies the expected inability of transfer methods to populate novel isoforms with functional information.

Secondly, the accurate annotation of features in novel isoforms of known genes represents the major concern of functional annotation because their functional underestimation might lead to notable biases in subsequent analyses designed to assess functional diversity across splicing variants. At protein level, the relationship between the feature coverage of known and novel variants (both from known genes) is maintained across all the categories in our annotation results: a slight decrease in the coverage of protein features for novel variants versus known ones (Figure 4.6.A-B). This pattern is conserved independently of the annotation method, suggesting that the higher feature coverage in known isoforms results from the bias of reference transcriptomes to contain the isoforms with the highest functional load, generally defined as PI. Actually, the CDSs of known isoforms were significantly shorter than CDSs predicted from novel isoforms (Wilcoxon test p = 7.01e-83 and p $<$ 2e-16 for SQANTI curated NEURALtc and OLIGd transcritomes, respectively), what corralates with the resulting pattern. In contrast, at transcript level, we did not find any shared pattern between the different categories. However, it should be note the enrichment of uORFs in novel isoforms, which is consistent across both transcriptomes, in aggreement with the higher 5' UTRs length of novel isoforms categories compared with known isoforms (Wilcoxon test p = 1.22e-95 and p $<$ 2e-16 for the NEURALtc and OLIGd transcritomes, respectively).

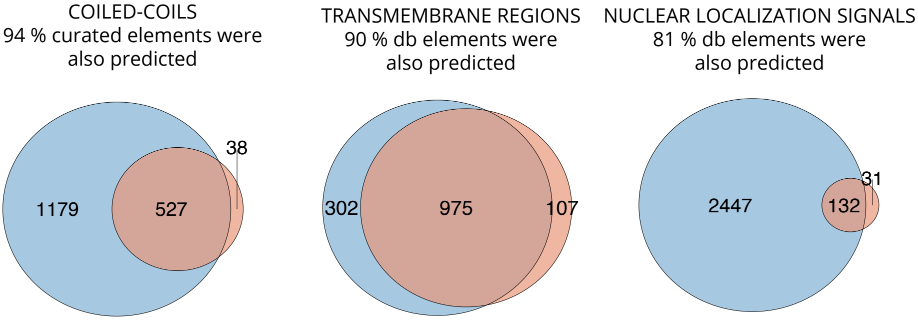


Figure 4.5: Feature-annotated gene intersection between prediction and transference strategies for different functional categories in OLGd transcriptome.

Taking together all the considered functional categories, except the PAS category, whose presence is expected in all the detected isoforms, more than 90% of mRNA variants and 80% of coding isoforms contain at least one transferred annotation element from public databases and around 84% of trancripts variants and more than 97% of coding variants were populated with predicted functional and regulatory features. Moreover, in both PacBio-defined transcriptomes, we annotated 99% of isoforms with at least one functional or regulatory feature (Table 4.4).

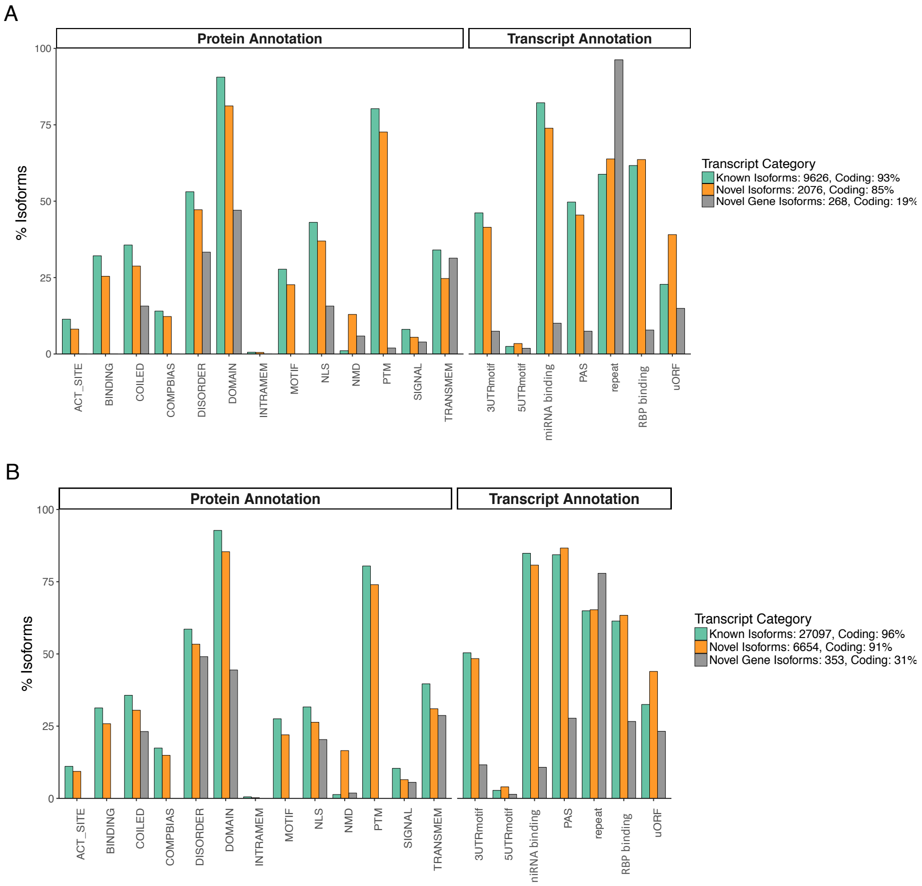


Figure 4.6: Annotation rates for each each functional label and isoform category in (A) OLGd and (B) NEURALtc transcriptomes. The information about the novelty status of isoforms was retrieved from SQANTI classification. Protein annotation rates are computed considering the set of coding isoforms. Known Isoforms: isoforms annotated in reference transcriptomes. Novel Isoforms: isoforms novel transcribed from genes annotated in reference transcriptomes. Novel Gene Isoforms: novel isoforms from unknown genes.

Transcriptome	Level	Method	% Annotated		
OLGd	Transcript	Transference	90.89	98.08	99.73
		Prediction	80.10		
	Protein	Transference	82.33	98.20	
		Prediction	97.16		
NEURALtc	Transcript	Transference	92.19	98.05	99.26
		Prediction	84.55		
	Protein	Transference	81.78	98.80	
		Prediction	98.16		

Table 4.4: Proportion of isoforms annotated at the transcript and protein level for PacBio transcriptomes

4.3.2 Functional annotation of reference transcriptomes and proteomes

As we previously detailed, IsoAnnot provides rich functional and regulatory annotation in isoforms derived from sequencing technologies. However, most transcriptomics studies still consider gene models defined in reference databases such as RefSeq or Ensembl. For that reason, we also ran IsoAnnot pipeline to annotate diverse reference transcriptomes in different organisms.

Mammalian transcriptomes

Homo sapiens and *Mus musculus* are two mammalian organisms widely used in research and so large transcripts and proteins sets are already available for them in public databases. In this sense, two of the most relevant sources are RefSeq, a non-redundant and well-annotated set of mRNA models (mRNAs and protein sequences) maintained by NCBI and Ensembl, a European database that contains gene models from multiple sources but which is less curated than RefSeq. We used IsoAnnot to populate mouse and human reference models (RefSeq78 and Ensembl86 annotation version) with rich regulatory and functional annotation.

IsoAnnot highlighted two clear different patterns in transcript-level annotations which were associated with these two different transcriptome sources, RefSeq and Ensembl. mRNA variants from RefSeq obtained a richer annotation level than those from Ensembl, pattern that maintaned across the two different annotation strategies, transference and prediction (Figure 4.7). Moreover, the annotation of many fewer elements in Ensembl isoforms also affected the PAS category which coincides with the lower proportion of coding mRNAs found in Ensembl versus RefSeq transcriptomes (50% vs. 70%, respectively, Table 4.6) and the low curation level of Ensembl transcripts which led to the presence of incomplete 3’ and 5’ mRNA ends. Hence, the nature of the Ensembl explains why IsoAnnot provides a lower levels of feature coverage at transcript level.

However, the level of feature population at protein level was found similar for both references sources with the exception of NMD category. Moreover, we found important coverage-level differences between mouse and human when transference algorithms are used, in agreement with the higher amount of functional information stored in public databases for human (239,159 vs 438,820 recovered functional features for mouse and human, respectively).

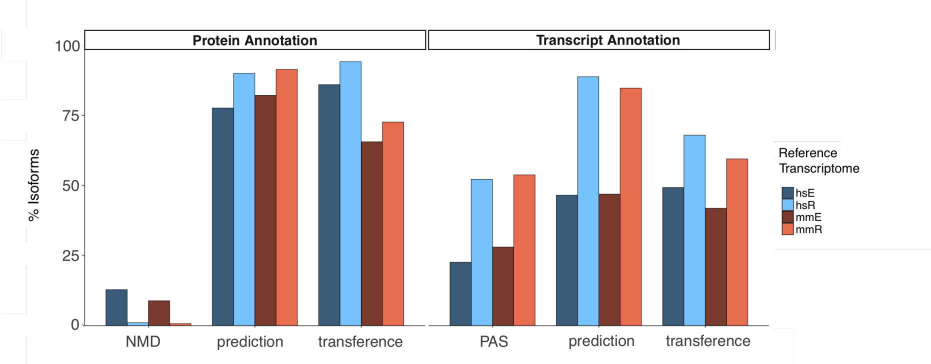


Figure 4.7: Proportion of isoforms annotated for different feature categories and annotation methods in mouse and human reference transcriptomes. hs: *Homo sapiens*; mm: *Mus musculus*; E: Ensembl; R: RefSeq

In summary, IsoAnnot provides rich functional information at the isoform-resolution for well annotated mammmalian organisms. At the protein level, more than 90% of coding isoforms were annotated with at least one functional feature for both databases and organisms (Table 4.5). Despite this, differences associated with the definition, nature and proportion of non-coding isoforms between alternative reference sources (Table 4.6) led to higher differences in the global annotation coverage reached by IsoAnnot (70% vs 90% of isoforms with at least one annotated element for Ensembl and RefSeq, respectively, Table 4.5).

Non-mammalian transcriptomes

Several recent projects studying AS in non-mammalian organisms in which our group has been involved in, have highlighted the lack of isoform-resolved functional annotation. Thus, IsoAnnot analysis was also applied to *Drosophila melanogaster* (Flybase617), *Arabidopsis thaliana* (Ensembl34) and *Zea mays* (Ensembl34) organisms.

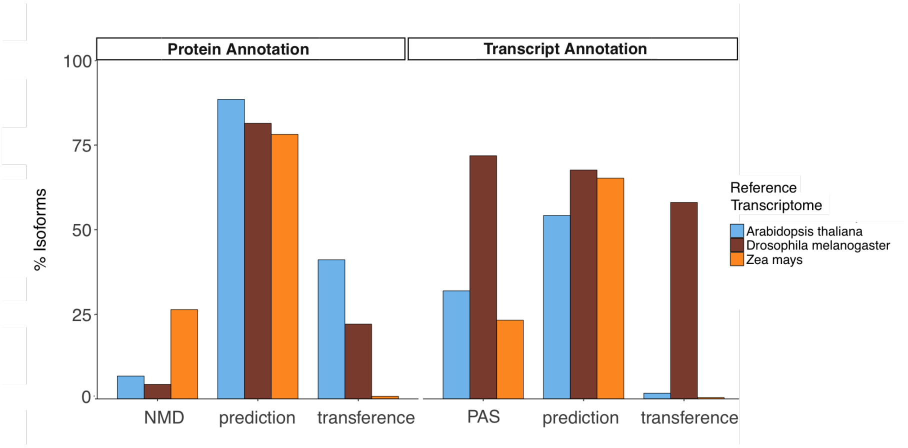


Figure 4.8: Proportion of isoforms annotated for different feature categories and annotation methods in non-mammalian reference transcriptomes

Annotation results (Figure 4.8) indicate a wide range of annotation levels in each organism and with each annotation method (Figure 4.8). Remarkably, at the transcript level, transference algorithms only provided a high coverage for the fly reference transcriptome, mainly because, compared to plants, more miRNAs for Drosophila melanogaster are found in public databases. At the protein level, the main difference was the proportion of isoforms annotated with elements from databases across organisms. We found that almost 50% of reference isoforms in *Arabidopsis thaliana* were populated with experimental features, while this figure was 25% for *Drosophila melanogaster* and was negligible for *Zea mays* (Figure 4.8). These results agree with the their representation levels in the functional databases we considered, such as UniprotKB (764 entries for *Zea mays*, 355 *Drosophila melanogaster* and 14830 entries for *Arabidopsis thaliana*).

Considering annotation results from both transfer and prediction methods, the protein-level annotation depth obtained with our IsoAnnot pipeline for non-mammalian references was much higher than that achieved at the transcript level, reaching an average of 85% of protein isoforms with at least one annotated functional element (Table 4.5).

Organism	**Reference**	**% Isoforms Annotated**		
		Transcript Level	**Protein Level**	**Total**
Mus musculus	RefSeq78	90.24	96.38	94.66
	Ensembl86	59.17	92.70	68.11
Homo sapiens	RefSeq78	94.42	98.19	96.40
	Ensembl86	65.03	95.38	71.88
Drosophila melanogaster	Flybase617	80.85	83.85	85.86
Arabidopsis thaliana	Ensembl34	54.56	90.43	85.35
Zea mays	Ensembl34	65.29	86.24	91.08

Table 4.5: Proportion of isoforms annotated at the transcript and protein levels for the different transcriptomes annotated by IsoAnnot pipeline.

Organism	**Reference**	**# Transcripts**	**# Coding Transcripts**
Mus musculus	RefSeq78	106060	75700 (71.37%)
	Ensembl86	115988	55863 (48.16%)
Homo sapiens	RefSeq78	150558	107211 (71.21%)
	Ensembl86	188667	85103 (45.10%)
Drosophila melanogaster	Flybase617	34740	30353 (87.37%)
Arabidopsis thaliana	Ensembl34	54013	48321 (89.46%)
Zea mays	Ensembl34	138049	133111(96.42%)

Table 4.6: Number of transcripts and the proportion of coding transcripts for the reference transcriptomes annotated by the IsoAnnot pipeline.

4.4 Discussion

Thousands of tissue- and developmental-specific splicing events and alternative polyadenylation sites have been identified during last years through the use of high-thoughput transcriptomics technologies. Among these technologies, single-molecule sequencing of full-length transcripts using long reads has paved the way for the accurate discovery of the huge isoform diversity expressed in both well-annotated species and non-model organisms [319][112][178][84][1] [347][53].

Being able to routinely characterize de novo transcriptome diversity, the main challenge is now to determine how isoform variants impact gene properties to drive context-specific cellular phenotypes. Protein-protein interaction domains [94][37][59], intrinsically disordered regions [126] [37] and nonsense-mediated decay [201] [391] [69] have been described as some of the main properties altered by alterantive processing of pre-mRNAs. Moreover, experimental isoform validations have shown that isoform diversity affect almost all types of transcript and protein functional features including linear motifs, miRNA targets sites, AU-rich elements or sites of post-translational modification [165][304]. However, traditional functional databases are biased to the annotation of the canonical, the most prevalent, the best documented and often the longest isoform so that projection of features onto splicing or alternative polyadenylated variants becomes a manually, time-consuming and tedious task for the scientific community studying the functional divergence of isoforms. Despite the need of databases and methods providing isoform-resolved functional information has been alleviated thanks to the appearance of some resources such as APPRIS [278][277], AS-PicDB [205], VastDB [311] or ASpedia [147], they are static and are restricted to the functional analysis of already annotated variants in well-studied organisms as human.

In this chapter, we developed IsoAnnot, a dynamic pipeline that provides extensive isoform-resolved annotation of novel transcriptomes with both coding and non-coding functional elements. Moreover, each feature is defined by its coordinates, what greatly facilitates the systematic functional comparison of isoforms. Additionally, IsoAnnot output displays the structural characterization of gene models together with automatic relationship information between genes, transcripts and proteins. This is achived by the mapping of sequences to reference protein and transcript IDs and generating unique sequence IDs for novel products, making the study of isoform diversity straightforward for the research community.

Application of IsoAnnot pipeline to the characterization of both long-read and reference transcriptomes demonstrated the comprenhensive and rich annotation achieved. In the case of model organisms as mouse, we found a that 90% of isoforms derived from PacBio sequencing and RefSeq repositories were populated with functional features. In contrast, the annotation of mouse Ensembl gene products hardly reached 70% of transcripts, in aggreement with their higher rate of non-coding transcripts (52% of non-coding mRNAs) and their decreased curation level (27% of coding mRNAs with incomplete ends) compared to RefSeq and Iso-Seq derived models. Thus, IsoAnnot results highlighted some annotation biases consequence of the specific characteristics of each considered transcriptome, demonstrating that the annotation depth is influenced by the nature of the input transcritpome.

IsoAnnot-derived annotations are the result of many sources of data which are divided into two main categories: (i) the sequence prediction methods which provide an efficient way to obtain uniform annotation coverage across novel and known transcripts and independently the amount of functional data in public databases, and (ii) methods projecting functional labels from public databases that complement isoform annotation with experimental and curated data but are highly influenced by the organism under study. This is the case of non-model organisms as *Zea mays*, for which the small scale of curated records collected in databases as UniprotKB (1,832 funtional features) limits the population of isoforms with curated functional information. On the contrary, transcriptomes from model organisms as mouse, human are extensively annotated by transfer methods due to the accumulation of experimental evidence in public databases as

UniprotKB. Even so, we found differences between mouse and human transfer annotation as a result of the difference in the number of elements collected in the considered databases (524,791 features in human vs 281,588 in mouse). Thus, our results highlighted that, as expected, IsoAnnot annotation with curated functional information is strongly influenced by the organism under study.

Moreover, we assessed the possible underestimation of annotated features in novel isoforms from known genes when using transfer methods, which would bias downstream analysis evaluating the functional diversity of splicing variants. Despite our pipeline revealed a richer protein annotation in known isoforms, the pattern was identical for features recovered from transfer and prediction approaches, indicating that the lower annotation of novel isoforms is likely caused by their own nature. Moreover, this agrees with the longer predicted CDS of known isoforms, indicating that reference transcriptomes are enriched in isoforms with the highest functional load. Oppositely, novel genes failed to obtain functional labels from public databases, verifying the inability of transference methods to annotate elements in non-reference locus. Additionally, we found an over-representation of non-coding feature categories in novel genes, in agreement with their higher proportion of non-coding transcripts compared to known genes. All together, we can conclude that prediction methods become the main source of functional information when annotating non-model organisms or novel *isoforms*.

As IsoAnnot was primary developed for the annotation of mammalian transcriptomes, most of the considered databases and prediction tools are mammalian oriented. Currently, IsoAnnot is in process of expansion to generate a more extensive isoform annotation in plants and invertebrates. As example, UTR motif or miRNA annotation in plants would greatly be enhanced if specific databases and tools such as ExactSearch [127], a plant specific 3' UTR motif prediction, were incorporated. Moreover, in next updates, we also plan to maintain and expand our pipeline by adding new prediction algorithms and databases covering additional transcript and proteins properties including protein-protein interaction databases (such as IBIS Shoemaker2012[300] or IntAct [139]), cross-species conservation, structural information (Interactome3D [217]), linear motifs (ELM database [81]) or mitochondrial targeting (TargetP [95]). Finally, currently IsoAnnot is being implemented as a web-based tool to facilitate experimental biologists to dynamically characterize isoform properties from customized sets of isoforms such as long-read derived transcriptomes across a wide range of organisms.

In conclusion, our results indicate that, despite the nature of the transcriptome influences feature coverage, IsoAnnot is able to extensively characterize the functional properties of isoforms without introducing biases that could affect the reliability of results obtained during the study of the functional impact of AS/APA events. Hence, we think that IsoAnnot, a pipeline that dynamically constructs an isoform-level functional database, potentially applicable to any organism and novel isoforms, is intent of become a gold-standard tool for the functional characterization of isoforms. IsoAnnot is able to overcome the limitations of current static databases that either support the annotation of novel events nor several organisms and greatly facilitate biologists and the bioinformatics community to go further in the genome-widstudy of the functional consequences of post-transcriptional mechanisms as APA and AS at genome-wide level.

Chapter 5

Comprehensive framework for the functional analysis of alternative isoform usage

5.1 Introduction

One of the most exciting aspects of transcriptome biology is the contextual adaptability of eukaryotic transcriptomes and proteomes by AS/APA mechanisms. The study of AS and APA mechanisms as a fundamental component of transcriptome biology has traditionally been addressed either by a detailed molecular characterization of context-specific events on single genes [165][304] or by genome-wide studies trying to find global patterns and their phenotypic association [93][348][265][374][311]. The computational approach has been greatly dominated by the analysis of the occurrence of alternative processing events, such as exon spiking, intron retention and polyadenylation sites , and the identification of changes in the usage of these events between conditions [362][33][197][372][149][367]. In parallel, many studies have been conducted to understand the molecular mechanisms behind the dynamic change of AS/APA patterns, which allowed the identification of a large number of RNA binding factors as principal regulators of transcript pre-processing decisions [380][362][366][281][390]. As a consequence, during the last decade, a myriad of tools have been developed to address the analysis of the structural and regulatory aspects of alternative transcript preprocessing and have shaped our current understanding about transcriptome complexity and dynamics.

However, studies on the context specific functional impact of ASA/APA are more limited. First, experimentally, just a bunch of differential splicing events has been experimentally validated and associated with particular properties [165][304] due to the lack of genome-wide approaches able to experimentally determine the functional impact of isoform regulation. Second, at the computational level, even though high-throughput technologies have the ability to accurately reflect the isoform content of individual tissues, developmental stages or environment changes, only a few bioinformatics groups have undertaken the genome-wide study of the alternative isoform usage from a functional perspective. These studies have provided relevant insights into the functional impact of AS, such as how spliced exons are enriched in disordered regions mediating new protein interactions [37], its impact on the dynamic regulation of protein-protein interaction (PPI) networks in a tissue-specific manner [376][94] or the loss of domains leading to the rewiring of PPI in cancer [59]. Similarly, APA has been postulated as a mechanism to escape microRNA regulation by shortening 3' UTR regions [142][109], ATI has been revealed to regulate the inclusion of uORFs that control translational rates [179][356] or NMD has been claimed as an important mechanism of gene expression regulation in cancer and neural systems [150][391]. However, much on the work done to answer transcriptome-wide questions on the functional role of isoform regulation have either involved ad hoc computational pipelines applied to specific biological systems and organisms, address only particular types of events, or interrogate only a limited number of functional properties such as AltAnalyze, which integrates differential splicing (DS) with protein domains, miRNA-binding sites and molecular interactions [96] or SpliceR [337] which predicts protein coding potential and NMD in DS genes. Moreover, recently Tranchavent et al. published the Exon Ontology [322], a resource to study functional enrichment of exon sets based on their annotation with protein functional domains. Using this tool authors were able to show different molecular functionalities directly associated to changes in exon inclusion levels between epithelial and mesenchymal cells. However, the tool does not have isoform resolution, fails to reveal the interconnectivity of functional elements at transcript isoform sequences, does not address the analysis of regulatory signals at alternative UTRs and is just implemented for mouse and human.

Thus, the major shortcoming that has hampered the extended study of the contextual functional impact of isoform regulation is the lack of friendly bioinformatics tools able to take advantage of current sequencing technologies to define, *in silico*, the potential contextual changes in functional properties triggered by post-transcriptional regulation. Most isoform analysis accomplished by groups who are inexperienced in bioinformatics currently rely on typical gene-based functional enrichment analysis of AS-regulated genes, which disregard the majority of the functional potential of the eukaryotic proteome and failing to accurately model the contextual functional consequences of alternative isoform usage in all its complexity.

In this chapter we developed a genome-wide bioinformatics analysis framework to interrogate the potential impact of alternative isoform usage on gene functional properties and the generation of high-confidence functional hypotheses to be validated experimentally. This framework relies on three main pillars: (1) the use of long-read sequencing capable of accurately defining full-length transcripts and their UTR/coding status (Chapter 3), (2) the use of extensive and diverse functional and regulatory annotation at the isoform-resolution level (Chapter 4) and (3) the development of methods to systematically capture functional differences between alternative isoforms as well as statistically evaluate and define the contextual modulation of gene properties triggered by AS, APA or ATI regulation. We have implemented this analysis framework in a user-friendly software called tappAS (www.tappas.org) which is accessible to the broad scientific community, thereby facilitating the genome-wide functional impact of context-specific isoform regulation. We applied our isoform-aware functional profiling approach in a glial differentiation system and experimentally verified some of our findings.

5.2 Methods

In order to study the context-specific functional effect of alternative isoform regulation, we have developed a novel pipeline for the ***Functional Analysis of Alternative Isoform Usage*** (Figure 5.1), which is divided into three main modules (Figure 5.1). Module 1 defines and measures the functional divergence among gene isoforms within the overall transcriptome. Module 2 focuses on the evaluation of expression levels to understand transcriptome dynamics at different levels of resolution. Finally, Module 3 addresses the integration of functional, structural and isoform expression data to unravel the potential effect of differential isoform usage on gene properties (Figure 5.1). The methodology includes already existing methods as well as novel approaches specially designed to tackle the aforementioned functional transcriptomics questions. Moreover, three different experimental designs were considered: pairwise, single-series time course and multiple-series time course designs, thereby providing a broad scope for its application.

5.2.1 Input data

Three pieces of information are required to use our analysis framework (Figure 5.1): (1) An accurate characterization of gene models, including structural definition of transcripts and CDSs and standardized nomenclature; (2) The availability of rich functional annotation at the isoform-resolution. In Chapter 4 we presented IsoAnnot, a pipeline specifically designed to retrieve extensive functional annotation at isoform resolution; (3) An accurate estimation of expression levels at transcript level. Isoform quantification is required in raw count format and two replicates per condition are mandatory for the use of our framework.

5.2.2 Module 1: Isoform functional diversity

To understand functional and regulatory variability between isoforms derived from the same gene, we developed the ***Functional Diversity (FD)*** analysis. FD identifies the nature and measures the magnitude of changes triggered by alternative processing of transcripts by systematically evaluating and comparing isoforms from a functional (positional functional features), regulatory (positional regulatory features) and structural (CDS, UTRs and PolyA sites) perspectives.

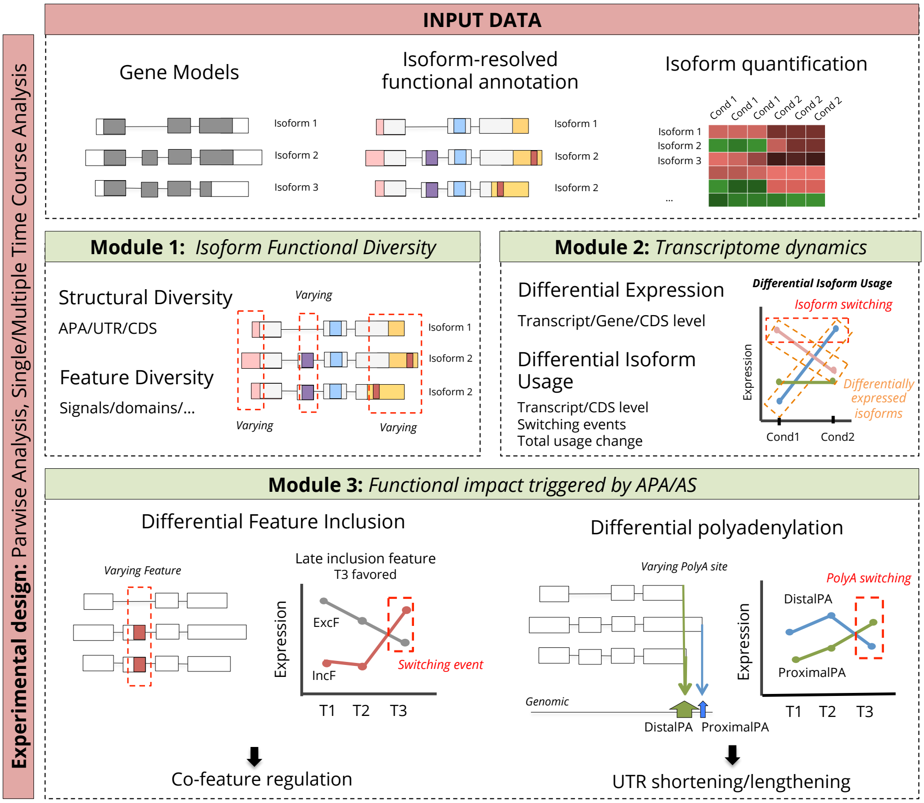


Figure 5.1: Overview of the *Functional Analysis of Alternative Isoform Usage*. Three main pieces of input data are required: gene models, isoform expression and functional annotation at the isoform-resolution level. Methods included in the three analysis modules were adapted to work with both pairwise and time course experimental designs.

5.2.2.1 Structural diversity

The Structural Diversity analysis, part of the FD analysis (Figure 5.2), aims to capture genes with isoforms presenting alternative coding sequences, UTR lengthening/shortening or alternative polyA sites. Structural Diversity catalogues genes as a function of the structural part modulated by alternative pre-mRNA processing.

Alternative polyadenylation

In order to detect APA events, polyA sites are identified as the last genomic position of transcript isoforms and evaluated in a pairwise mode by computing the polyA distance between each pairwise combination of isoforms expressed by a given gene (Figure 5.2). mRNA cleavage is not an exact process and can occur within a small window of positions [247]. To take into account this cleavage

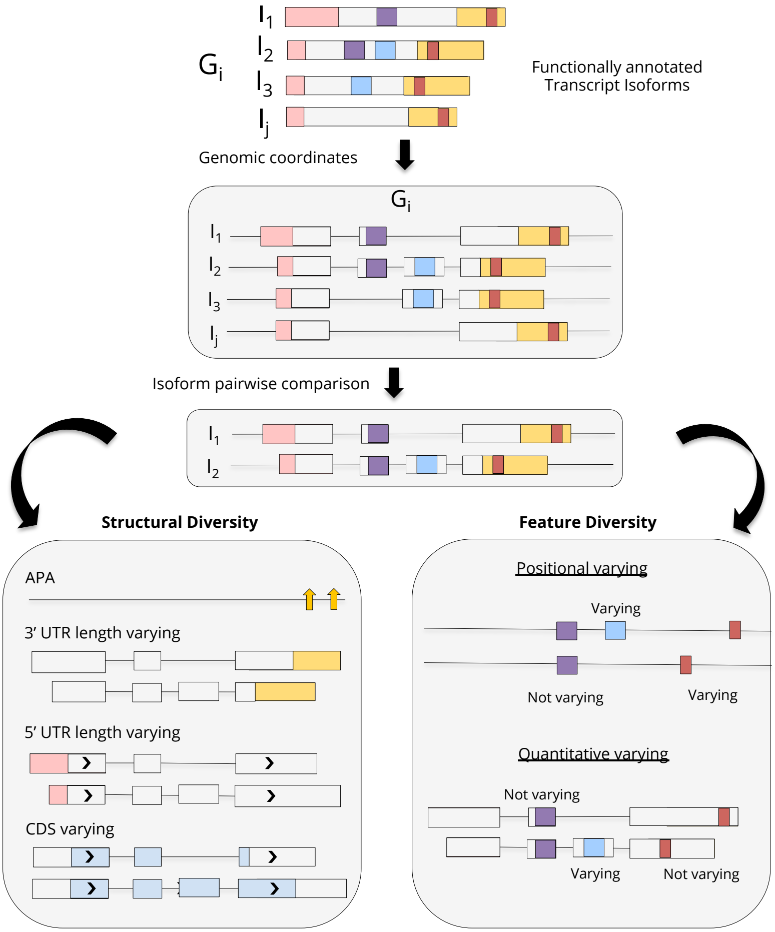


Figure 5.2: Functional Diversity Overview. Gene models and functional annotation is translated into genomic coordinates for subsequent pairwise isoform comparison. While Structural Diversity involves the determination of alternative polyadenlylation, coding sequence (CDS) modification or UTR lengthening/shortening, Feature Diversity focuses on the gain/loss of functional features.

variability, the PacBio Iso-Seq^{TM} Analysis (*Cupcake*) discriminates independent polyA sites by collapsing mRNA variants with a 3' end distance lower than 100 bp. Additionally, Rot et al. considered that auxiliary cleavage RNA motifs appear approximately up to 75 nt upstream and 50 nt downstream of each cleavage site [21][299] and defined independent sites as those located at least 125 nt apart [281]. Therefore, to focus APA analysis in independent polyA sites and remove cleavage variability, our FD analysis labels pair of isoforms undergoing APA by establishing a minimum 100 bp genomic distance (default value) between polyA sites.

UTR length varying

UTR length is obtained from IsoAnnot structural information for subsequent pairwise comparison between coding isoforms derived from the same gene. Pairs of isoforms with 3'/5' UTR differences above a user-specified cutoff (75 bp as default cutoff) are labelled as 3' UTR length varying/5' UTR length varying, respectively (Figure 5.2).

Coding sequence varying

CDSs variability is determined by comparing CDSs both at the sequence and genomic coordinate levels, ensuring that identical CDSs generated from alternative but synonymous events are described as varying. Non-coding isoforms are discarded from CDS diversity analysis.

5.2.2.2 Feature diversity

Feature diversity was developed to identify functional and regulatory elements altered across isoforms. In the same way that structural diversity is evaluated, feature diversity is assessed by comparing each pair of isoforms transcribed from the same gene. However, because isoforms can be annotated with different features, FD individually compares each feature ID present in the transcriptome (Figure 5.2).

Two different approaches are considered in order to classify a feature as varying. The first approach, namely *Positional Varying* (Figure 5.2) compares features by genomic position. In this case, FD maps features to genomic coordinates and classifies them as varying if coordinates are not equivalent between isoforms. The second approach, namely *Quantitative Varying* (Figure 5.2), does not consider genomic coordinates but rather, quantitatively compares the number of elements associated to the feature ID in question within each splicing variant. We considered both approaches because of their complementary goal as well as their different suitability depending on the feature under study. For example, translational repression correlates with the number of uORFs present in the transcript [55] and so analyzing the density of uORFs at transcripts, rather than comparing their genomic position is a more meaningful way to evaluate their putative effect on the translational rate and, consequently, to detect isoforms pairs with varying uORF repressiveness. Similarly, differential stability rates across transcript variants can be assessed by quantitatively comparing the number of predicted AU-elements along the UTR region of isoforms. In contrast, other features such as PTMs are potentially recommended to be analyzed by the positional varying strategy because of the functional independence of most PTMs located at a given protein sequence.

Finally, non-positional features describing transcript characteristics are evaluated by presence/absence of annotation. NMD transcript status is a straightforward case of non-positional feature whose classification as varying is based on a lack of NMD status agreement between the pair of isoforms under study.

5.2.2.3 Overall rate of diversity

FD analysis is applied to all genes described in the studied transcriptome. As previously mentioned, pairwise comparisons are performed for all the combinations of transcript isoforms in a gene. Gene-level overall varying rates are computed per each individual functional category (instead of feature ID) as the percentage of genes with at least one positive (varying) pair-wise comparison. The background is the set of genes annotated with elements associated with the studied category.

Even though genes may code for several isoforms, often only one of them is very different from the others. In these cases, the gene level strategy to assess variation at the transcriptome-wide level might cause overestimation of the functional complexity triggered by AS or APA events: although most of the isoform pairwise comparisons between gene isoforms are not varying, the final gene result is varying because of a single highly different isoform. Moreover, the number of isoforms defined for a gene affects the number of pairwise comparisons, the chance of a gene to acquiring a varying pairwise comparison and consequently, its detection as varying. Thus, we also estimate the overall varying rate of each feature as the percentage of pairwise comparisons catalogued as varying, removing the bias associated to low frequent but rare splicing events as well as discarding the influence of the number of isoforms per gene on the chance to call a gene varying.

5.2.3 Module 2: Transcriptome dynamics

Module II provides tools for studying transcriptome dynamics, including changes in isoform usage and modulation of their absolute levels. To this end, both established methods and novel approaches are used in Module II.

5.2.3.1 Differential expression

Differential expression (DE) analysis in our pipeline comprises three steps:

(a) Data preprocessing: Low-count isoform filtering and normalisation procedures are performed by using NOISeq R package ([312]). We skipped normalization step in those cases where the methods required raw count data.

(b) Aggregation of expression values: we provide expression values at the transcript, gene and CDS-levels with the aim of studying expression dynamics at different resolutions as a function of the purpose of interest. Because only transcript-expression levels are provided as input, gene-level estimations are generated by collapsing the expression levels of their expressed transcripts variants. Similarly, CDS expression levels are estimated by collapsing the expression of transcripts annotated to have identical coding sequences. This collapsing ability stems from using well-structured data as an input, in which the identity relationships between gene transcripts and CDSs are robustly identified.

(c) Algorithms for DE: after revising the literature, we decided to include NOISeq [312] and maSigPro ([63]) methods for the analysis of DE in experimental designs involving two conditions and time course series, respectively. As expression values were computed at the isoform, CDS and gene level, we applied DE methods to the detection of differentially expressed genes (DGE), differentially expressed isoforms (DIE) and differentially Expressed CDSs (DCE), with the latter being informative of the transcriptional regulation associated with specific coding sequences.

5.2.3.2 Differential isoform usage

Our pipeline estimates post-transcriptional regulation by testing the *Differential Usage of Isoforms* (DIU), i.e. the changes in the relative abundance of isoforms derived from the same gene. Note that DIU differs from DIE in that the latter only entails changes in the absolute expression of individual isoforms and does not necessarily involve a change in their relative proportion (Figure 5.3).

Several methods have been proposed to test differences in the usage of isoforms, as detailed in Section 1.3, Chapter 1. However, most of them are focused on detecting the differential inclusion of single events as exons or splice junctions and ignore the exon composition of the full length transcripts. In contrast, our framework has been specially designed to analyze full-length transcript models, and leverage long read sequencing technologies where the exact combination of exons is unambiguously determined. Hence, we considered the analysis of differential usage of full-length isoforms, instead of single events, which provides a more informative approach to understand the combinatorial regulation of exons and polyA sites and dynamics of transcriptome complexity.

Our pipeline to study DIU includes:

(a) Alternative Aggregation Levels: like DE, we evaluate the differential usage of gene products at the CDS and transcript levels. The differential coding sequence usage (DCU) was implemented in our pipeline so that post-transcriptionally

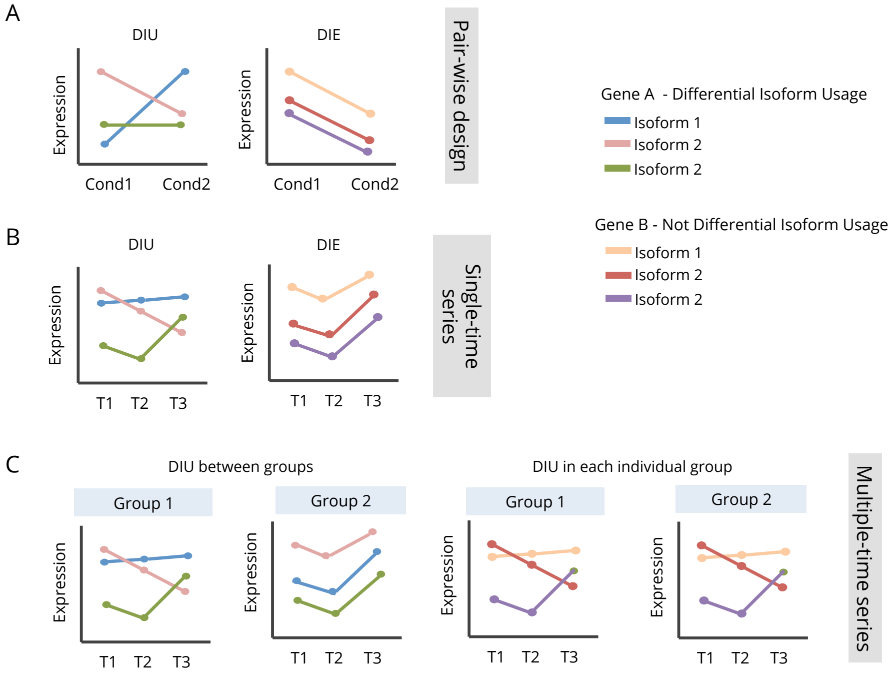


Figure 5.3: Differential Isoform Usage (DIU) defintion in A. Pairwise, B. Single-time course and C. Multiple-time course experimental designs.

regulated genes involving alteration of the coding sequence could be discriminated. Consequently, a novel layer of understanding is added to the differential usage analysis that measures the coding impact. Of note, as transcript expression levels, CDS-computed expression levels are not a direct measurement of protein levels because mRNA stabilisation and translation rate can vary across transcript isoforms, meaning that protein levels do not directly correlated to mRNA levels.

(b) Isoform prefiltering: genes in mammalian transcriptomes usually express a high number of isoforms. However, one of them usually accumulates the major proportion of gene expression [122] and becomes the major isoform. Hence, the remaining isoforms, although detected, have low expression levels. When the minor isoform appears with small expression changes between conditions but they are in the opposite direction to the predominant isoform, significant isoform:condition coefficients may appear during the GLM fitting used by the differential-usage methods considered in this Module (Figure 5.3.A). This scenario is exacerbated when using reference transcriptomes and its influence decreases when using PacBio-defined transcriptomes because of the decreased capacity of long-read technologies to detect low-expression isoforms [313]. To avoid the detection of DIU genes because of the 'flat' behaviour of minor isoforms, an isoform filtering step before gene modelling can be applied (Figure 5.4.A). This filtering follows two alternative approaches. The first approach considers the proportion of the gene expression accounted by each isoform and filters those that do not reach a minimum expression rate. By default, isoforms accumulating less than 10% of the expression of the gene in all the studied conditions are considered as negligible and consequently discarded previous to DIU analysis:

$$\left(\frac{E_{ig}}{\sum_{i=1}^{n} E_{ig}} < p\right) \qquad \gamma_j$$

being *Eig* is the expression value for isoform *i* and gene *g*, *j* the experimental condition, *n* is the number of isoforms in *g*, and *p* is the threshold level.

The second filtering approach, instead of using proportions, considers the fold-change of the minor isoforms in relation to the major one. Minor isoforms are discarded when their expression levels in all the considered are below a specified fold-change (FC) threshold (default FC=2).

(c) Algorithms for DIU: appropriate algorithms for DIU must meet three requirements (1) biological variability between replicates should be taken into account, (2) the significance of DIU must be estimated and (3) the algorithm must accept transcript level expression values. Based on these set of requirements and the results shown from a recent evaluation of DIU methods performed by Merino et al. [207], we chose DEXseq [6] to detect DIU in case-control experimental designs. In addition, we also used the recently released isoMaSigPro approach [233] for DIU testing in time-course experimental designs. Both methods assess DIU by fitting generalised linear models (GLMs) and testing the significance of the isoform-condition interaction coefficient. These methods were adapted for use with transcript and CDS estimations.

(d) Major isoform switching and total usage change: DE methods provide fold-change rates measuring the strength of the change these complement the level of statistical significance. Additionally, differential splicing methods interrogating single splicing events describe the magnitude of change by computing the difference in the relative abundance of a each single splicing event (e.g: ΔPSI, variation in the proportion spliced-in). However, DIU analysis test in a unique model the relative behaviour of multiple isoforms, what hampers the definition of a single measure describing the differential behaviour of them across conditions. Although the coefficient values for fitted GLMs used in this module are a good measure for describing isoform expression patterns, most users find them quite difficult to interpret. Thus, new metrics which catalogue the differential behavior of isoforms are necessary.

We define the *total usage change* as a measure of the magnitude of the redistribution of expression between isoforms across each pair of conditions considered. Because the absolute gene expression levels may be different across conditions, the total-change values are always represented as a function of the gene expression FC. We define *isoform usage* as the relative expression of isoform *i* in gene *g*. The *total usage change* is defined as:

$$\sum_{i=1}^{n} \mid \overbrace{\frac{E_{1ig}}{\sum_{i=1}^{n} E_{1ig}}}^{\text{Isoform Usage C1}} - \overbrace{\frac{E_{2ig}}{\sum_{i=1}^{n} E_{2ig}}}^{\text{Isoform Usage C2}} \mid \times 0.5$$

where *Eig* is the expression value for isoform *i* and gene *g*.

Additionally, *major isoform switching* was defined for genes which switched their most expressed isoform. When analyzing case-control and single series time course experimental designs, the major isoform for each gene is catalogued as the one with the highest mean expression across conditions, while other isoforms of the gene are called minor forms. A major isoform switch is then associated with genes whose major isoform becomes minor at time point or any condition (Figure 5.4.B). In multiple time-course series where two different groups are compared over time, the definition of the major isoform is defined for each experimental group (Figure 5.4). Thus, major isoform switching events are associated with genes with a different predominant isoform between experimental groups.

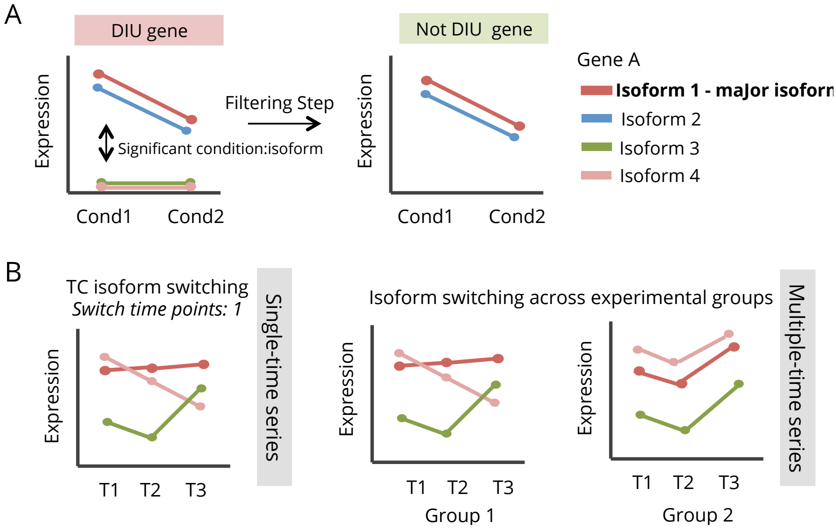


Figure 5.4: Isoform prefiltering and major isoform switching. A. Filtering identifies isoforms with low expression before differential isoform usage analysis and decreases the chance of genes being called DIU without relavant isoform usage regulation. B. Determination of major switching events in time course data.

5.2.4 Module 3: Functional impact triggered by isoform regulation

In Module 3 we included new methods to study the contextual modelling of the functional effect triggered by differential isoform usage based on the evaluation the alteration of functional and regulatory elements such as PTMs or UTRs.

5.2.4.1 Differential feature inclusion

The ***Differential Feature Inclusion (DFI)*** analysis implements an approach to detect positionally-annotated functional features that are significantly altered or disrupted across isoforms in time and which thus, modulate the functional and regulatory outcome of the gene in different experimental conditions.

The process of detecting the differential inclusion of features between conditions consists of several steps. First, like the *FD analysis* detailed in Section 5.2.2.2, each gene is evaluated for the gain or loss of annotated positional features across gene isoforms (Figure 5.5.A). Changes in the feature between isoforms can vary from subtle disruptions to complete skipping. Two scenarios for flagging varying features are considered, to adapt the analysis to different types of variation:

1. Feature disruption: A feature is catalogued as varying when genomic positions across isoforms do not exactly match. Any slight difference between isoforms flags the feature as varying. This would be the most sensitive case of varying since it enables to test the regulation of features with subtle alterations or disruptions across isoforms.

2. Feature exclusion: Feature disruptions across isoforms are not flagged as varying but total feature losses and gains. Thus, only completely excluded/included features are subsequently tested for DFI.

Features catalogued as varying for each gene are subjected to expression aggregation, under the assumption that the level of functionality of the gene in the tested feature can be estimated from the abundance of the isoforms including the feature. Expression for isoforms which partially or completely disrupt a feature (depending on the feature varying option) are collapsed to generate a

Feature-Including Variant; otherwise expression is associated into the Feature-Excluding Variant (Figure 5.5.B). After this transformation, the expression matrix is composed of the set of Feature-Including and Feature-Excluding Variants for all the detected varying features of each gene while observations remain as the different conditions under study. Making use of statistical techniques used by the well-established DEXSeq and maSigpro packages for pairwise and time-course experimental designs, respectively, we test the differential inclusion of features by fitting a generalised linear model:

$$g(\hat{\mu}_{fg}) = \hat{\beta}_0 + \hat{\beta}_1 C_{fg} + \hat{\beta}_1 T_{fg} + \hat{\beta}_2 F_{fg} + \hat{\beta}_3 C_{fg} \cdot F_{fg} + \\ \hat{\beta}_3 C_{fg} \cdot T_{fg} + \hat{\beta}_3 F_{fg} \cdot T_{fg} + \hat{\beta}_3 F_{fg} \cdot T_{fg} \cdot C_{fg} + \hat{\epsilon}_i \quad (5.1)$$

where *g* characterises the GLM, µ*ij=E(yfg)* the expected value of expression *yfg* for observation *g* and feature *f*, *Cfg* is the binary variable that identifies the experimental condition, and *Ffg* is the variable that identifies the variant (Feature-Excluding or Feature-Including Variant).

We individually model each gene-feature (Figure 5.5.C) and evaluate the significance of the interaction condition-variant or condition-variant-time, depending on the experimental designed considered. We control the false positive rate by applying isoform filters (Section 6.3) and FDR corrections. When different functional annotation categories are considered (domains, UTR motifs, disordered region, etc.) we test them independently.

DFI analysis is complemented with additional metrics that make results more meaningful:

- Feature inclusion levels: the magnitude of the feature's inclusion is computed as the relative abundance of the Feature-Including Variant for each condition studied.

  $$FI_{fg} = \frac{EInc_{f_g}}{EInc_{f_g} + EExc_{f_g}}$$

  where EInc is the expression value for the Feature-Including Variant and EExc is the expression value for the Feature-Excluding Variant for gene g and positional feature f.

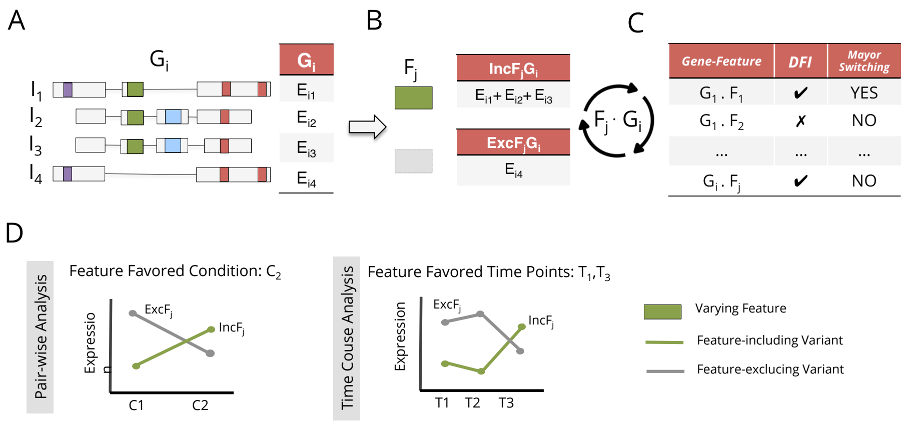


Figure 5.5: Differential feature inclusion. A. The functional diversity of each gene expressing multiple isoforms is evaluated to determine features which vary between isoforms. B. Isoform expression levels are collapsed as a function of the status of the feature being considered and expression is aggregated into two main variants, the Feature-including Variant and the Feature-excluding Variant. C. Varying features are tested for Differential Feature Inclusion. Significance and complementary attributes as major switching events are output. D. Conditions and time points in which the feature's inclusion is enhanced are detected. G: gene; F: feature; I: isoform;

- Major variant switching: major switching events are identified following the rules explain in Section 5.2.3.2. Thus, major switching events reveal features going from a predominant to a minor relative abundance across conditions. major switching features characterize tissue/stage-specific functional outcomes.

- Favored condition: we characterise the direction in which isoform regulation promotes the inclusion of the feature and define the transition points or conditions in which the feature inclusion is enhanced (Figure 5.5.D).

- coDFI: coDFI is assessed by estimating how frequently two different features are differentially co-regulated across genes. We test their association using a Fisher Exact Test and applying multiple testing correction through the Benjamini-Hochberg procedure. CoDFI is computed for each pair of feature IDs and only features detected as DFI in at least 5 genes are considered.

5.2.4.2 Differential polyadenylation and UTR lengthening

We developed the ***Differential Polyadenylation (DPA)*** analysis, a method to test the regulation of polyA sites by leveraging the transcript resolution of PacBio long reads. As seen in Chapter 3, Pacbio sequencing faithfully identifies polyA site variability, which leads to a significant improvement of isoform expression estimations [313] and thereby, provides an accurate measure of polyA site usage to be contextually modelled by our DPA analysis.

Although there are some exiting methods designed to infer differential APA sites by using RNA-seq data, each of them has its own limitations: several evaluate polyA usage differences without accounting for biological variability or determining the significance level of the APA modulation [9][129][372][354], some approaches do not deal with multiple polyA sites [372], or just support certain organisms [371] and most of them just consider pairwise experimental designs [9][129][354], which limits meaningful studies across multiple conditions. Our DPA analysis addresses the contextual modelling of APA in different experimental designs and provides statistical significance levels while considering several polyA sites per gene. The approach includes the following steps:

a. Defining a library of polyA sites

Gene models provided by the user are used to define the set of polyA sites that are interrogated for DPA. DPA analysis is coupled to the evaluation UTR lengthening/shortening and so non-coding isoforms as well as predicted nonsense-mediated decay variants are discarded. To build the polyA site database, the genomic coordinate of the last position of the remaining transcript isoforms is extracted. Note that, unlike recently developed tools [129], polyA sites in terminal exons with different 5’ start sites are also considered, what includes the analysis of CDS-APAs and polyA sites in isoforms originated through combination of APA and AS events. Thus, all cleavage sites from stable and coding mRNAs, independently of their splice junction pattern are included in the library of polyA sites so wide APA/UTR regulation is considered.

Next, we perform a series of filtering and collapsing steps in order to define the proximal and distal polyA site for each gene (Figure 5.6.A). First, independent cleavage sites are defined by merging polyA sites located within a 75 bp window (more details in Section 5.2.2.1). Resulting sites are then filtered based on the relative polyA site expression levels. To avoid the definition of a minor polyA site as a distal or proximal site, the proportion-based filtering method (See Section 5.2.3.2) is applied and so, only polyA sites accumulating at least 10% (default threshold) of total gene expression at least in one condition remain for further classification into distal (dPA) or proximal sites (pPA) (Figure 5.6.A). In

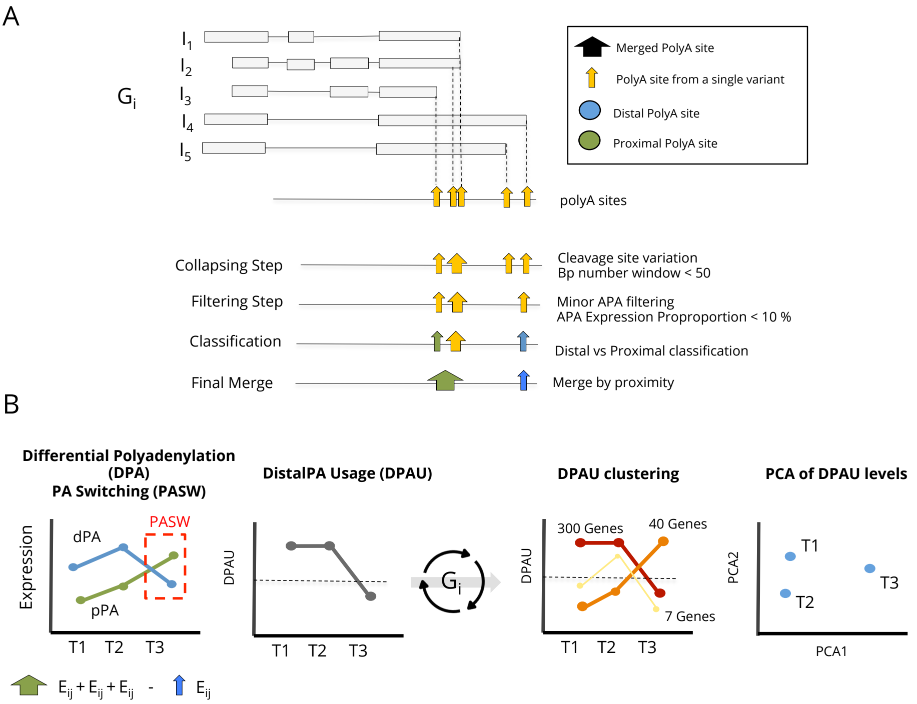


Figure 5.6: Differential Polyadenylation analysis (DPA). A. Definition of the library of polyA sites. Long-read gene models are used to define the expressed polyA variability. DPA analysis performs a series of filtering and collapsing steps in order to define the proximal (pPA) and distal (dPA) polyA sites. B. Expression associated to the dPA and the pPA is estimated by collapsing associated isoforms and subsequently modelled across conditions to detect differentially polyAdenylated genes (DPA genes). Different metrics as the distal polyA site usage (DPAU) and analysis including the polyA switching, the clustering of DPA genes or the principal component analysis on DPAU levels improve the comprenhensive analysis of DPA in different experimental designs.

genes with more than two polyA sites, some in-between sites remain unclassified. In those cases, we perform a final merge of unlabelled sites by assigning them to the most proximal or distal neighbouring site (Figure 5.6.A).

b. Estimation of polyA site associated expression levels and relative usage

Per-gene and per-sample distal and proximal polyA site expression levels are computed by collapsing the expression levels of the associated set of transcript isoforms into polyA site distal and proximal variants. We calculate the relative usage of polyA sites by calculating the relative expression of the distal site over the total polyA site expression level of the gene:

$$DPAU = \frac{E_{dPA}}{E_{dPA} + E_{pPA}}$$

DPAU refers to the distal poly(A) usage and dPA and pPA to the variants defined for distal and proximal polyA sites.

c. Differential polyadenylation and polyA site switching

Like DFI, we also evaluated DPA by fitting regression models to capture significant condition-variant interactions. We based our approach on DEXSeq and Iso-maSigPro for pairwise and time-course series analysis, respectively (Figure 5.6.B). Genes with significant DPA are further assessed for polyA site switching to detect changes in the dominant polyA site across time points or conditions (for further details, see Section 5.2.3.2).

d. APA regulatory dynamics

Characterisation of APA programs requires methods that extract common polyA usage patterns across the genes expressed in the system of study. We use two methods to characterise the genome-wide APA landscape: (1) The Hierarchical Clustering approach identifies groups of genes with similar DPAU profiles over time. Thus, clustering is only applied when time course series are considered. Genes with DPA FDR values under the threshold specified by the user (default=0.05) integrate the set of DPAU profiles used as input for clustering profiling (Figure 5.6.B); (2) multivariate principal component analysis is performed on centered and arscine-transformed DPAU values in order to explore sample relationships related to their APA patterns (Figure 5.6.B).

e. Detecting lengthening and shortening of 3' UTRs

APA regulation is highly coupled with UTR modulation since mRNA polyadenylation cleavage site largely determined the 3' UTR length. Comparing polyA sites from isoforms containing identical CDS end positions (UTR-APA) allows the direct association of distal/proximal polyA site usage and UTR lengthening/shortening events, respectively. However, in cases of CR-APAs, where the polyA site position alters the coding region, it is impossible to directly infer the relationship between distal polyA site and 3' UTR lengthening. As our DPA analysis assesses polyA site regulation independently of the coding sequence, we profile 3' UTR lengthening/shortening by computing the UTR weighted arithmetic mean at each considered condition, using relative isoform usage values as weights:

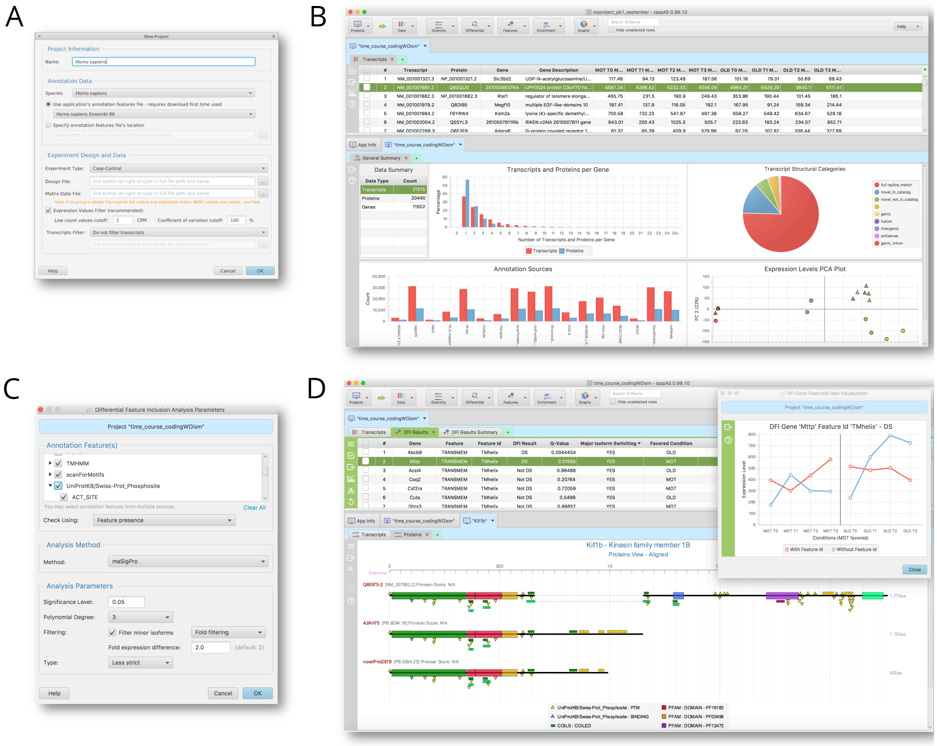
$$\overline{UTR}_w = \frac{\sum_{i=1}^{n} U_{ig} \cdot UTR_{ig}}{\sum_{i=1}^{n} UTR_{ig}}$$

being U the relative usage of isoform i in gene g and UTR its associated UTR length.

Thereby, UTRs from highly expressed isoforms will contribute in a higher proportion to the final UTR mean length. Comparison of weighted UTRs across conditions will be informative about genes which have undergone UTR shortening and lengthening events. Statistical differences are tested by using Wilcoxon rank sum test.

5.2.5 TappAS software implementation

Our framework for the functional analysis of isoforms was implemented in a Java GUI application namely tappAS. Making use of extensive functional annotation at isoform resolution and RNA-seq isoform expression data, tappAS provides the set of methods and approaches described in this chapter together with complementary functionalities (normalisation, PCA analysis, venn diagrams, a visualisation engine...) that enable the easy formulation of many diverse functional hypotheses about the role of isoform regulation in a given system of study.



New Project

Project Information

Name: Homo sapiens

Annotation Data

Species: Homo sapiens

● Use application's annotation features file - requires download first time used

Homo sapiens Ensembl 86

○ Specify annotation features file's location

Use button on right or type in full file path and name

Experiment Design and Data

Experiment Type: Case-Control

Design File: Use button on right or type in full file path and name

Matrix Data File: Use button on right or type in full file path and name

Note: First group in design file must be the control and expression matrix MUST contain raw counts - see Help

☑ Expression Values Filter (recommended):

Low count values cutoff: 1 CPM Coefficient of variation cutoff: 100 %

Transcripts Filter: Do not filter transcripts

myproject_pb1_september - tappAS 0.99.10

Differential Feature Inclusion Analysis Parameters

Project "time_course_codingWOism"

Annotation Feature(s)

☑ TMHMM

☑ scanForMotifs

☑ UniProtKB/Swiss-Prot_Phosphosite

☑ ACT_SITE

You may select annotation features from multiple sources.

Check Using: Feature presence

Analysis Method

Method: maSigPro

Analysis Parameters

Significance Level: 0.05

Polynomial Degree: 3

Filtering: ☑ Filter minor isoforms Fold filtering

Fold expression difference: 2.0 (default: 2)

Type: Less strict

time_course_codingWOism - tappAS 0.99.10

DFI Results
DFI Results Summary
DFI Gene Featured Data Visualization
Project "time_course_codingWOism"
DFI Gene 'Mttp' Feature Id 'TMhelix' - DS
Expression Level
Conditions (MOT favored)
With Feature Id
Without Feature Id
Close
Kif1b - Kinesin family member 1B
Proteins View - Aligned

Figure 5.7: TappAS interface. A. Project definition. B. Structured project data and graphical summary. C. Differential feature inclusion analysis dialogue. D. Differential feature inclusion results, isoform-resolved functional annotation display and multiple-time course expression profile for the Feature-Excluding and Feature-Including Variants.

Analysis in tappAS is organized as projects, which require two pieces of information: an experimental design file and a transcript expression data. It accepts pairwise analysis and single and multiple time-course experimental designs. Users can provide their own isoform-resolved functional annotations or can choose one from among the multiple pre-annotated transcriptomes for different organisms provided by tappAS.

The tappAS interface has two main panels: the upper panel shows structured project data and statistical results in tabular format. Being a modern GUI application, tappAS provides a rich set of features via a JavaFX platform. These features include data tables with customisable columns which can be sorted and filtered; all the application data and images can be exported to files, the windows are resizable and can be zoomed, context-sensitive help and menus are available; data can be drilled down, and displays or individual projects tabs can be customised.

The bottom tappAS panel integrates a comprehensive set of data visualisation tools to help recognise patterns and better understand results and functional isoform variability. It also provides project summary plots, PCA plots, cluster network graphs, Venn diagrams, and exploratory results charts. The tappAS visualisation engine is especially designed to display isoform variability in a user-friendly manner, so that functional and structural differences between isoform variants can easily be visualized and identified. However, the main characteristic of our visualization engine is the mapping of the set of isoform-specific functional and regulatory features. It comprehensively and dynamically displays the whole catalogue of isoform-resolved annotations features at both transcript and protein isoforms, greatly facilitating their study and comparison.

Implementation of the approaches described in this chapter together with the complementary filtering, clustering, normalization, data managing and visualisation features make tappAS an outstanding and unique application for studying isoform complexity and dymamics from a functional point of view.

5.3 Data

We demonstrated the *Functional Analysis of Alternative Isoform Usage* pipeline described in this chapter using a case-control experimental designed. Because this methods leverage long-read technologies to define transcriptome complexity, we used the experimental designed detailed in Chapter 3, Section 3.2.1 involving two different cell types, Neural Precursor Cells (NPC) and Oligodendrocyte Progenitor Cells (OPCs). The expressed transcriptome was defined by using the PacBio sequencing and includes 11,970 transcripts in 7,167 genes (see Chapter 3). Transcripts were functionally annotated by IsoAnnot (see Chapter 4) which generated more than 1M of labels. Isoform-level quantification was computed using RSEM [188] following ENCODE guidelines. Other experimental designs are considered in Chapter 6.

5.4 Results

5.4.1 The impact of neural trancriptome complexity on functional diversity

We ran FD analysis to understand how isoform diversity impacts the coding and regulatory potential of our neural transcriptome. Structural diversity (100 bp for both UTR length cutoff and PolyA site cleavage variability) revealed that around 75% of genes that express more than one isoform code for varying predicted proteins (Figure 5.8.A), representing 24% of the total expressed genes. Moreover, coding sequence modulation was a greater source of transcriptome diversity than UTR variation in our neural system (Proportion test $p < 2.2e16$). Even so, variability at UTR length and polyadenylation sites was present only around 45% of multi-isoform genes, without significant differences between the number of 5' and 3' UTR varying genes (Figure 5.8.A). Statistical testing for gene overlapping across different structural categories (multiple intersection test, [350]) revealed that co-regulation of both transcript ends is far less significant than the simultaneous regulation of a single UTR and the coding sequence (Figure 5.8.B). Moreover, the coding region was altered in 78% of UTR-varying genes, suggesting that UTR length modulation is frequently coupled with coding sequence alteration, likely promoted by CR-APAs. Figure 5.8.C shows an example of a gene which tappAS identifided as containing APA, 3' UTR, 5' UTR and CDS variants.

To determine the impact of these structural differences on functional and regulatory properties we performed the quantitatite FD analysis on this set of positionally-annotated features. Looking at features annotated at the transcript level, we found that the NMD had the highest rate of variation (95% of multi-isoform genes). Moreover, almost 100% of genes containing NMD isoforms code for multiple isoforms, indicating that in our neural system NMD-targeted isoforms are usually co-expressed with functional isoforms. Moreover, UTR motifs, show a high rate of differential isoform inclusion (55% and 90% of multi-isoform genes for 3' and 5' UTR motifs, respectively, Figure 5.8.A). Among 3' UTR motifs, GU-rich (GREs) was the category with the highest enrichment

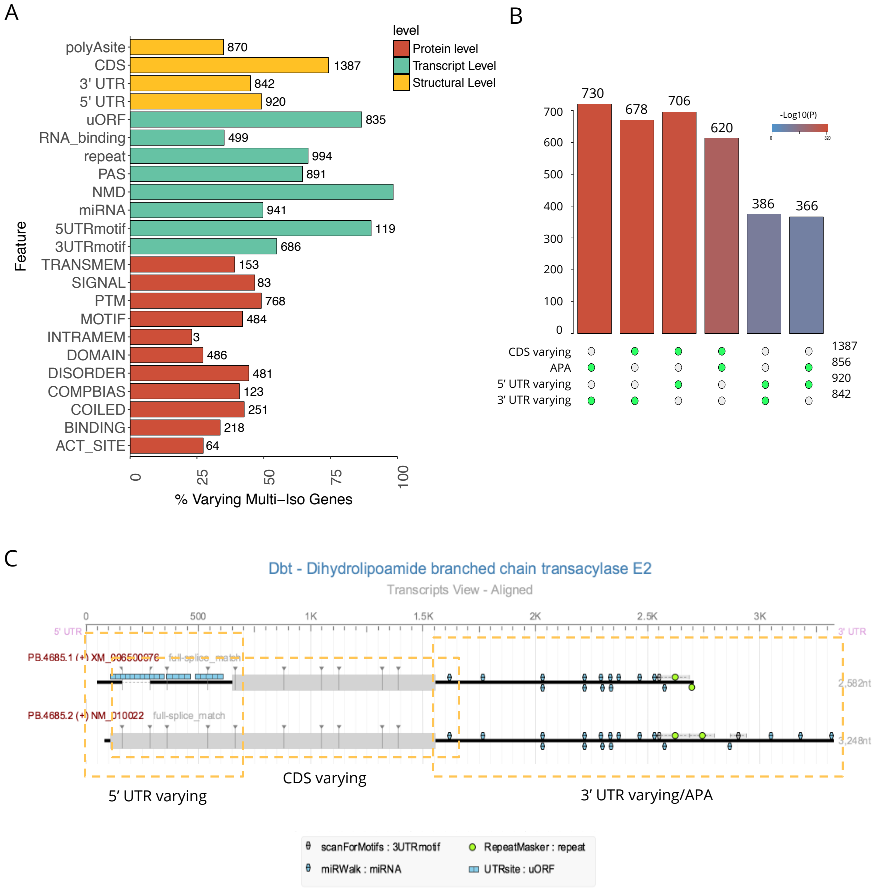


Figure 5.8: Functional diversity in our neural PacBio-defined Transcriptome. A. Based on the set of genes which express more than one isoform, varying rates per category (protein, transcript, or structural-level) and feature level were identified. B. Intersection between genes varying in different structural categories. The significance level was tested by using the SuperExactTest R package [350]. C. Example of a gene detected as CDS, UTR and APA varying.

in varying elements (FET p = 3.10e-3, FDR = 0.08, 63% of varying features in genes expressing multiple isoforms). GREs have recently been associated with mRNA stabilisation [339] and have also been reported as the targets of RNA-binding proteins (RBPs) which are post-transcriptional regulators of polyA, mRNA decay, translation, and pre-mRNA processing [338]. Among the set of 160 genes with differential inclusion of GRE elements in our neural system, we found splicing regulators as *Rbm4* (Figure 5.9.B), which is involved in muscle cell differentiation, and transcription factors such as *Tcf12* (Figure 5.9.A) which plays an important role in controlling proliferating of neural stem cells and progenitor cells during neurogenesis [328].

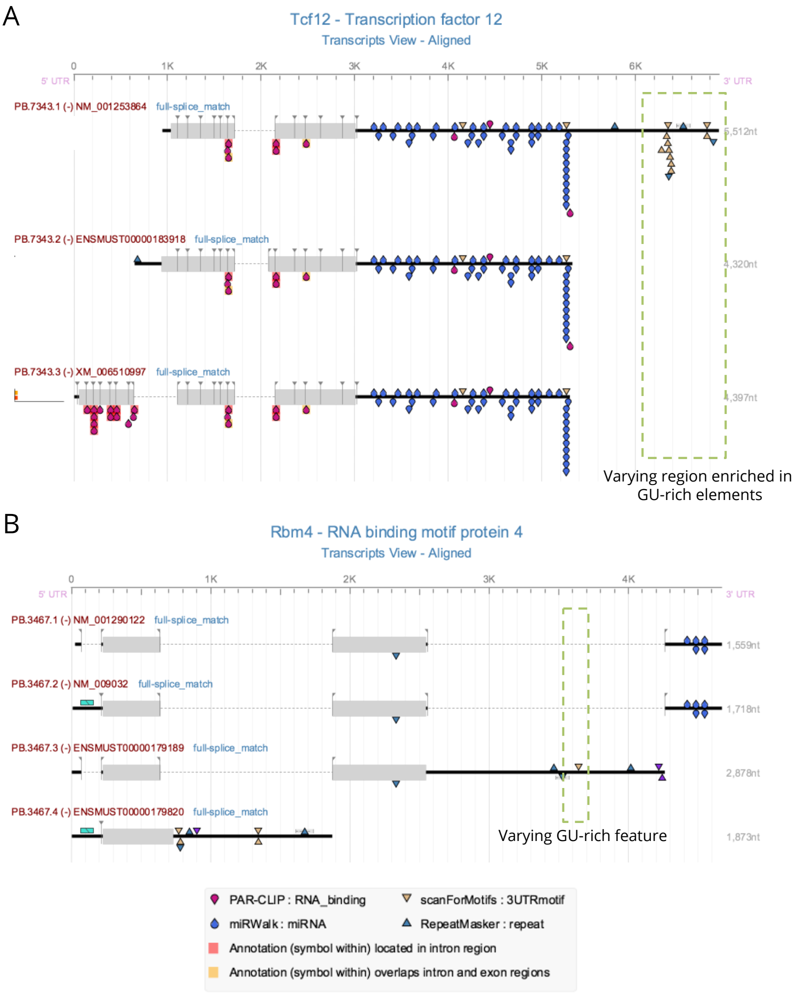

A
Tcf12 - Transcription factor 12
Transcripts View - Aligned
0
1K
2K
3K
4K
5K
6K
5' UTR
3' UTR
PB.7343.1 (-) NM_001253864 full-splice_match
5,512nt
PB.7343.2 (-) ENSMUST00000183918 full-splice_match
4,320nt
PB.7343.3 (-) XM_006510997 full-splice_match
4,397nt
Varying region enriched in GU-rich elements
B
Rbm4 - RNA binding motif protein 4
Transcripts View - Aligned
0
1K
2K
3K
4K
5' UTR
3' UTR
PB.3467.1 (-) NM_001290122 full-splice_match
1,559nt
PB.3467.2 (-) NM_009032 full-splice_match
1,718nt
PB.3467.3 (-) ENSMUST00000179189 full-splice_match
2,878nt
PB.3467.4 (-) ENSMUST00000179820 full-splice_match
1,873nt
Varying GU-rich feature
PAR-CLIP : RNA_binding
scanForMotifs : 3UTRmotif
miRWalk : miRNA
RepeatMasker : repeat
Annotation (symbol within) located in intron region
Annotation (symbol within) overlaps intron and exon regions

Figure 5.9: Examples of genes which express isoforms-specific inclusion of GU-rich elements.

We also ranked miRNAs as a function of their enrichment in alternative isoform targeting. The top five (Figure 5.10.A) included miR-335, which has been associated with oligodendrocyte differentiation [23] and mir-590-3p which presents the higher number of isoform-specific binding sites in our neural system

(Figure 5.10.A). Moreover, mir-590-3p, is strongly associated with proliferation and differentiaion processes [85], and is up-regulated in retinoic acid treated cells [225] - the same treatment our OPCs received. Thus, annotation data reveal a potential isoform-specific layer of expression-regulation involving miR-590-3p binding site gains/losses by alternative processing of immature mRNAs. Interestingly, from the 70 genes in our system with isoform-specific inclusion of mir-590 binding sites, we found examples already described in the literature, such as the transcription factor *Zfp143* (Figure 5.9.C) in which regulation of the 3' UTR impacts on miR-590-3p targeting has been previously characterized [225]. Moreover, Nogando et al. reported the co-exclusion of mir-590 binding sites with AU-Rich elements, which also agrees with our results (Figure 5.9.C) and boost the potential impact of UTR regulation on the *Zfp143* transcript fate.

A

miRNA	p-value	odds ratio	# genes
mmu-miR-335-3p	0.0015	2.01	47
mmu-miR-590-3p	0.0014	1.70	70
mmu-miR-221-3p	0.0064	2.06	30
mmu-miR-511-5p	0.0138	1.79	34
mmu-miR-881-3p	0.0194	1.73	33

B

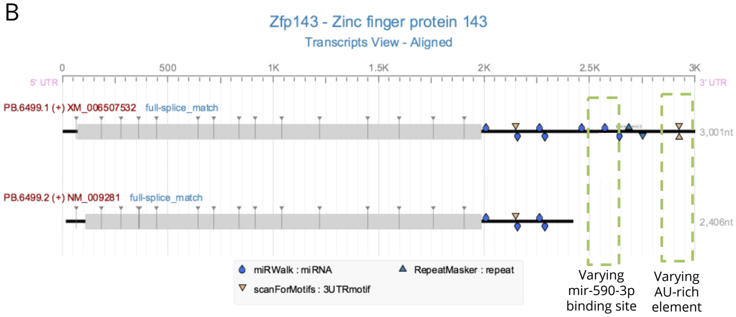


Figure 5.10: Functional impact on miRNA binding site targeting. A. Top 5 miRNAs in differential isoform targeting. Relative over-representation of varying elements by miRNA was evaluated by Fisher Exact Tests (FET). B. The transcription factor *Zfp143* has two alternative polyA sites that generate alternative UTRs containing a differential miR-590 binding site and an AU-rich element.

FD assessment on coding features showed that post-translational modifications (PTMs) varying the most across protein isoforms (49.07%, Figure 5.8.A) and have the highest relative enrichment across categories (FET FDR = 5.7e-

19) in contrast to PFAM domains, the category with the lowest inclusion rates in multiple-isoform genes (27.25%, Figure 5.8.A). However, as mentioned in methods Section 5.2, quantitative FD does not consider partial feature disruptions. Positional versus quantitative FD at the coding level revealed significant changes in the proportion of varying genes for PFAM domains (Proportion Test: p-value = 2.6 e-18) and intrinsically disordered regions (IDRs, Proportion Test: p-value = 6.8 e-18), where positional varying reaches 62.97 % and 50.28 %, respectively (Figure 5.11.A). To understand whether the different PFAM families share this behaviour or, or the contrary, total and partial domain removal depends on the PFAM under study, we interrogate PFAMs at ID level. Among the 15 PFAM families with the highest isoform gain/loss frequency (Figure 5.11.B) categories like zinc finger and KRAB box domains are usually completely contained in alternatively spliced regions because the varying rate only slightly increases when partial disruptions are considered. Figure 5.12A-B shows two examples that illustrate the complete loss of the KRAB box and zinc finger domains in two alternatively spliced transcription factors. In contrast, kinase domains seem to frequently undergo domain disruptions rather than domain skips (Prop.test p.value kinase domains p = 0.02, Figure 5.11.B), likely modulating rather than abolishing its functional role. Genes such as *Cdk10* and *Rbm39* undergo partial kinase-domain deletions (Figure 5.13. A-B).

Moreover positional FD identifies IDRs as having the highest representation of disrupted features in our neural system (FET FDR = 1.06e-20, Figure 5.11.A) and this effect is accentuated when considering the absolute number of disrupted features (FET FDR = 3.186575e-311, Figure 5.11.C). These findings agree with previous studies reporting their enrichment in alternatively spliced regions [280] [37] [62].

Remarkably, we found a highly significant inclusion association between varying IDRs and other varying elements (Figure 5.14), whose presence in IDRs has already been reported such as phosphorilation sites (FET FDR = 2e-67, [148]) or Nuclear Localization Signals (NLS, FET FDR = 3e-45, [82], [375]). *Rbm39* is an example of a completely excluded IDR feature which is associated with several PTMs, an NLS and a small-ligand binding motif (Figure 5.13.B). Thus, FD analysis of our neural transcriptome confirmed the role thay IDRs play in the allocation of PTMs and linear motifs as NLSs as well as their clear association to alternatively-spliced regions.

Finally, at the trancript level, as expected, predicted polyA signal annotation is the most affected category when moving from quantitative to positional FD mode (Figure 5.11.A), most likely because this positional change underlies APA events, an extensive post-transcriptional regulation mechanism in our neural transcriptome. Therefore, our proposed methods for FD interrogation allows analysis to be adapted according to the biological question at hand or the characteristics and regulatory mechanism of the functional category being studied.

Taken together, FD analysis indicated that almost 90% of the multi-transcript genes defined in our PacBio-defined neural transcriptome have coding or non-coding feature labels that vary across their isoforms, suggesting that downstream isoform analysis to interrogate the impact of relative isoform usage in defining cell identity is meaningful.

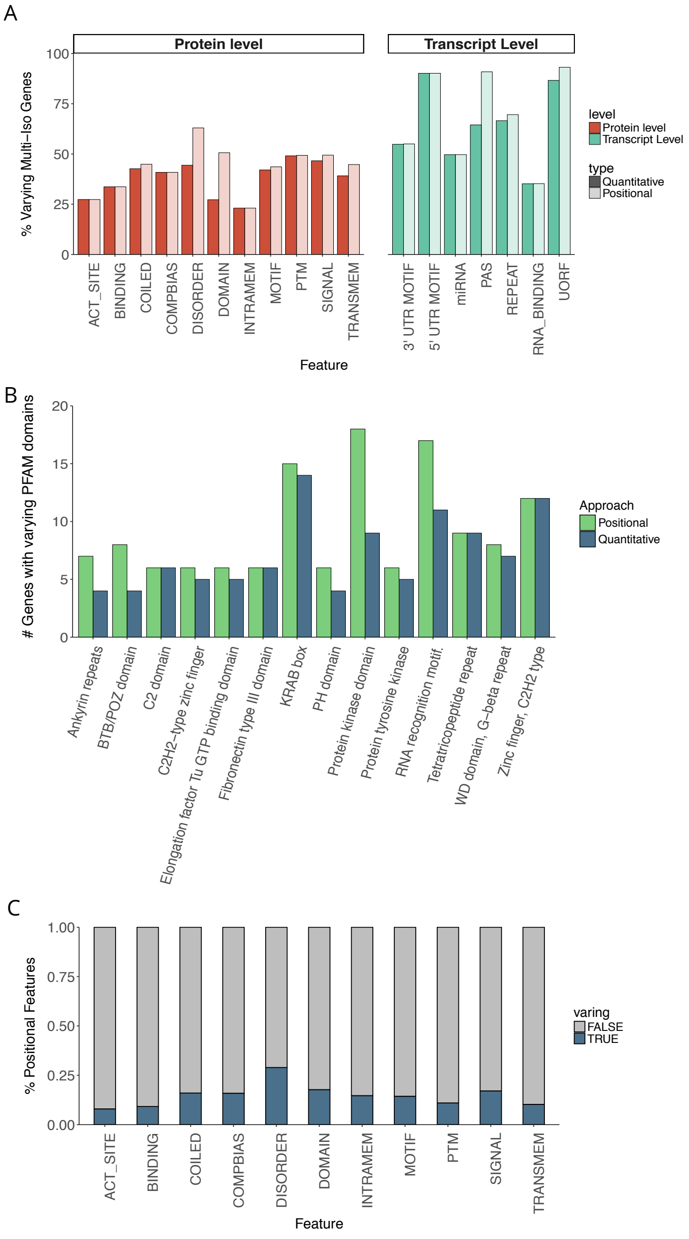


Figure 5.11: Quantitative versus positional feature diversity. A. Proportion of genes with varying features when considering partial disruptions (positional) or complete feature skips (quantitative). B. Number of genes which partially or completely skip domains for the 15 PFAM families with higher positional variation rates. C. Ratio of features in multi-isoform genes detected as varying by the positional diversity analysis. Non-parametric proportion tests were used to test for differences in proportions across methods. (***) p $<$ 0.001; (**) p $<$ 0.01; (*) p $<$ 0.05.

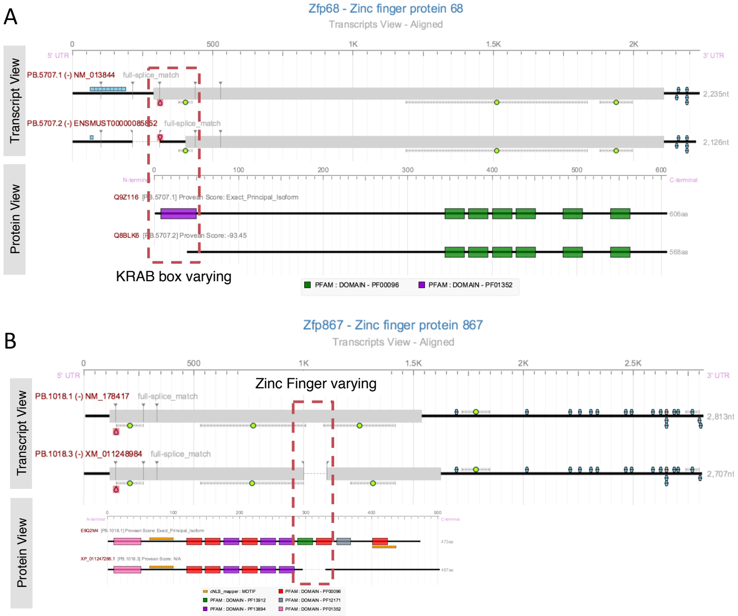

A
Zfp68 - Zinc finger protein 68
Transcripts View - Aligned
Transcript View
0
500
1K
1.5K
2K
5' UTR
3' UTR
PB.5707.1 (-) NM_013844 full-splice_match
2,235nt
PB.5707.2 (-) ENSMUST00000085852 full-splice_match
2,126nt
Protein View
0
100
200
300
400
500
600
N-terminal
C-terminal
Q9Z116 [PB.5707.1] Provean Score: Exact_Principal_Isoform
606aa
Q8BLK6 [PB.5707.2] Provean Score: -93.45
568aa
KRAB box varying
🟩 PFAM : DOMAIN - PF00096
🟪 PFAM : DOMAIN - PF01352
B
Zfp867 - Zinc finger protein 867
Transcripts View - Aligned
Transcript View
0
500
1K
1.5K
2K
2.5K
5' UTR
3' UTR
Zinc Finger varying
PB.1018.1 (-) NM_178417 full-splice_match
2,813nt
PB.1018.3 (-) XM_011248984 full-splice_match
2,707nt
Protein View
0
100
200
300
400
500
N-terminal
C-terminal
E9Q2M4 [PB.1018.1] Provean Score: Exact_Principal_Isoform
473aa
XP_011247286.1 [PB.1018.3] Provean Score: N/A
467aa
🟧 cNLS_mapper : MOTIF
🟩 PFAM : DOMAIN - PF13912
🟪 PFAM : DOMAIN - PF13894
🟥 PFAM : DOMAIN - PF00096
▨ PFAM : DOMAIN - PF12171
🟪 PFAM : DOMAIN - PF01352

Figure 5.12: Complete domain removal by alternative splicing events captured by the quantitative FD analysis. A. Deletion of a KRAB box domain at the N-terminal resulting from an exon-skipping which caused a downstream start coding. B. Complete loss of a Zinc Finger domain caused by an exon-skipping splicing event.

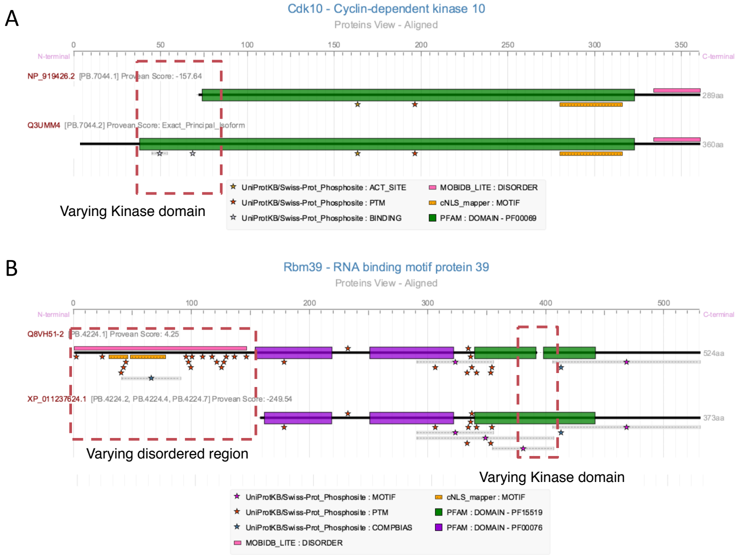

A
Cdk10 - Cyclin-dependent kinase 10
Proteins View - Aligned
0
50
100
150
200
250
300
350
N-terminal
C-terminal
NP_919426.2 [PB.7044.1] Provean Score: -157.64
289aa
Q3UMM4 [PB.7044.2] Provean Score: Exact_Principal_Isoform
360aa
Varying Kinase domain
UniProtKB/Swiss-Prot_Phosphosite : ACT_SITE
UniProtKB/Swiss-Prot_Phosphosite : PTM
UniProtKB/Swiss-Prot_Phosphosite : BINDING
MOBIDB_LITE : DISORDER
cNLS_mapper : MOTIF
PFAM : DOMAIN - PF00069
B
Rbm39 - RNA binding motif protein 39
Proteins View - Aligned
0
100
200
300
400
500
N-terminal
C-terminal
Q8VH51-2 [PB.4224.1] Provean Score: 4.25
524aa
XP_011237624.1 [PB.4224.2, PB.4224.4, PB.4224.7] Provean Score: -249.54
373aa
Varying disordered region
Varying Kinase domain
UniProtKB/Swiss-Prot_Phosphosite : MOTIF
UniProtKB/Swiss-Prot_Phosphosite : PTM
UniProtKB/Swiss-Prot_Phosphosite : COMPBIAS
MOBIDB_LITE : DISORDER
cNLS_mapper : MOTIF
PFAM : DOMAIN - PF15519
PFAM : DOMAIN - PF00076

Figure 5.13: Partial disruption of kinase domains captured by positional functional diversity analysis. A. N-terminal disruption of a kinase domain in the *Cdk10* gene. B. *Rbm39* undergoes an exon-skipping event which caused the truncation of a kinase domain. In addition, N-terminal protein disruption leads to the complete removal of a disordered region enriched in post-translational modifications.

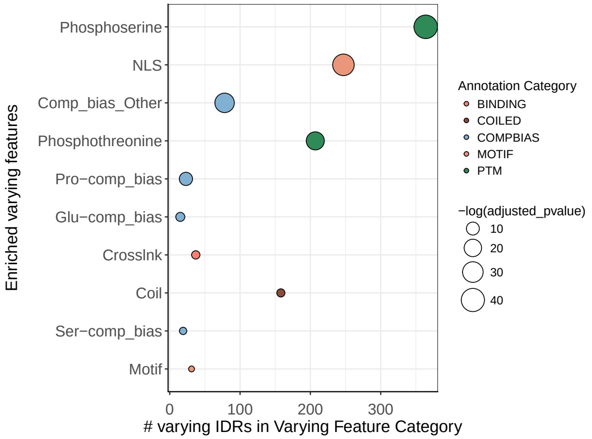


Figure 5.14: Co-inclusion of isoform-specific intrinsically disordered regions and other categories. Fisher exact test analysis was performed for each sub-category to evaluate their overrepresentation in intrinsically disordered region varying genes. P-values were corrected according to the false discovery rate. The dot colour identifies the category of the tested feature, the size of dots is associated with the significance level.

5.4.2 Multi-layer analysis of alternative isoform usage

DE analysis showed a high proportion of differential genes between NPC and OPC cells (3765 genes, FDR$<$0.05 and FC$>$1.5), 32% of them expressing alternative isoforms. Additionally, our analysis identified two sets of genes that, despite not presenting gene expression changes, have differentially expressed isoforms (27% and 19% of multi-isoform genes with DIE and DCE isoforms, respectively), highlighting the significant proportion of genes that might be exclusively subjected to post-transcriptional regulation.

To capture the actual proportion of the transcriptome regulated by DIU between NPCs and OPCs, we ran our set of tools implemented on tappAS (FDR $<$ 0.05). Out of 1,205 differentially expressed multi-isoform genes, 292 were identified as post-transcriptionally regulated (Figure 5.15.A). Additionally, 248 genes were captured in the set of genes without overall changes in gene expression levels, revealing that in our system, approximately 50% of DIU genes are not transcriptionally regulated (Figure 5.15.A). However, 163 of the total DIU genes were not detected when isoforms with low relative abundance were discarded from out analysis (0.1 for the minor isoform proportion threshold), highlighting the relevance of applying filtering strategies based on expression proportions in addition to common absolute low count approaches, in order to capture the most relevant post-transcriptional regulation signal. Two examples are shown in Figure 5.16.

Applying DCU analysis, we discriminated 133 genes whose post-transcriptional regulation does not involve changes in the CDS (Figure 5.15.A). An example that illustrates this behaviour is the myoneurin gene (*Mynn*, Figure 5.17). Three transcript isoforms were detected for *Mynn* via PacBio sequencing, one of them with a longer coding sequenced generated by an exon inclusion that leads to the insertion of a zinc-finger PFAM domain (Figure 5.17.C). Even though post-transcriptional regulation is detected at the transcript level (Figure 5.17.A), predicted CDSs co-express across conditions. Thus, *Mynn* did not involve the differential usage of alternative coding sequences (Figure 5.17 B).

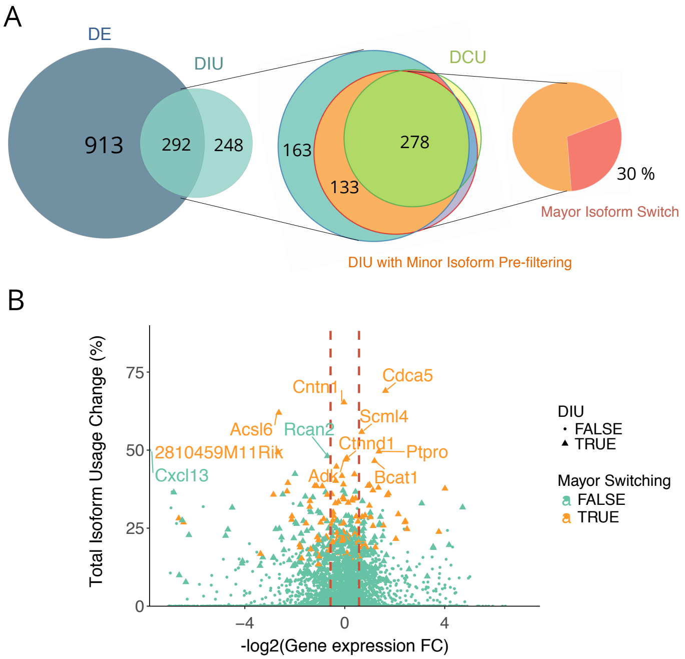


Figure 5.15: Multi-layer analysis of differential isoform usage contributing to the comprenhensive understanding of post-transcriptional regulation by AS and APA mechanisms. A. Venn diagram showing the intersection between different sets of multi-isoform genes tested for differential isoform usage. B. Representation of the proportion of expression redistributed between isoforms (Usage Change measure) in function of gene expression fold-changes. Genes with a switching in the major-expression isoform across conditions are represented in orange.

Finally, out of the 411 DIU genes, a relevant 30% undergo a major isoform switching between NPC and OPC conditions (Figure 5.15.A) despite the mean of the total usage change for DIU genes is around 20% (Figure 5.15.B). *Glyr1*, which switches its most abundant transcript and protein isoforms between conditions (Figure 5.18.A-B) is an example of a gene exclusively regulated by alternative splicing (insertion of a microexon in OPCs) without changes in its gene expression levels.

To try to link post-transcriptonal regulation with function, we ran common approaches as the functional enrichment (FE) and gene-set enrichment (GSE) analyses for all of the above differential analysis results but using, as novelty, all the functional categories available in our IsoAnnot annotation file, including

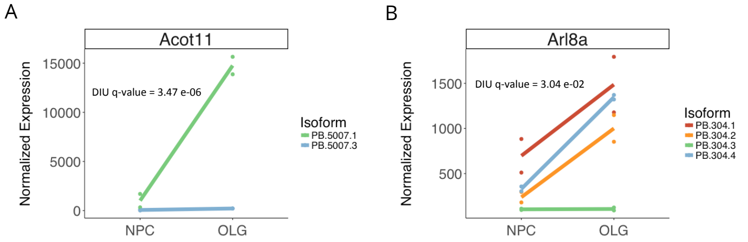


Figure 5.16: Isoform prefiltering of minor isoforms remove false positive DIU calls. Two false positive DIU genes are shown. Their isoforms are differential expressed but not differentially usage.

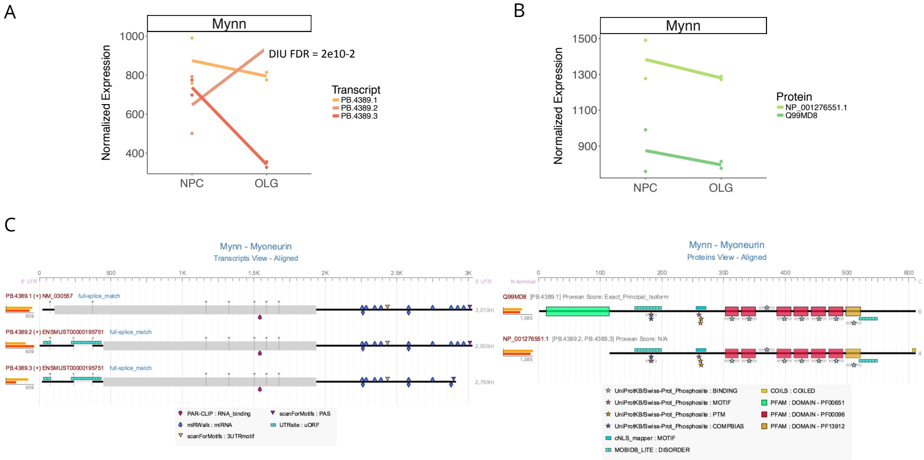


Figure 5.17: The *Mynn* gene undergoes A. differential usage of expressed isoforms but B. predicted coding sequences show the same behaviour across conditions. C. Transcript and coding region view. An exon-skipping event at the 5' end and alternative polyAdenylation events generate three different isoforms and two alternative coding sequences.

positional feature annotations. Thus, the power of the dense transcript-level annotation of functional elements could be used to interrogate differential-splicing functional outcomes.

The set of DE genes are enriched in numerous transcriptionally regulated processes, functions and pathways (FDR $<$ 0.05) that agree with the phenotypic differences expected between the two cell types, such as cell motility, homeostasis, glia cell projection, and lipid metabolic processes. However, to specifically capture the functional contribution of AS/APA regulation to transcriptional reg-

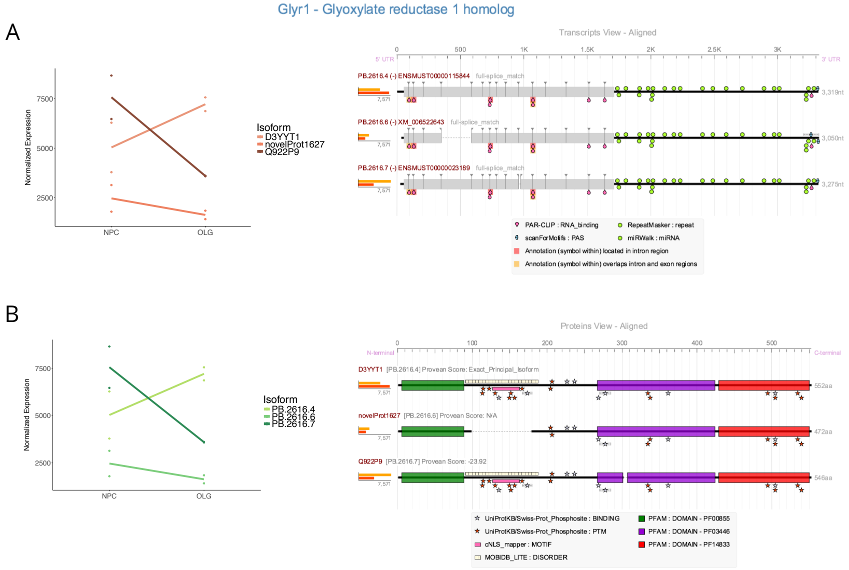


Figure 5.18: Major isoform switching detected in the *Glyr* gene at both A. the transcript-isoform level and B. the coding-sequence level.

ulation, the FE was assessed by testing DIU genes against DE genes. Interestingly, we found processes involved in 3'-end mRNA processing, RNA binding and mRNA splicing (Figure 5.19), indicating that the post-transcriptional machinery accomplishes a high degree of self-regulation between NPCs and OPCs. For example, both *Tardbp* and *Srsf3*, are splicing factors differentially spliced between our two cell types (FDR DIU p = 2.45e-07 and p=4.78e-05 for*Tardbp* and *Srsf3*, respectively) and whose binding sites were actually found enriched in DIU genes (Figure 5.19) and thus, enrichment methods operating over extensive functional annotation link the post-transcriptional regulation of the splicing factor *Tardbp* with its own role as post-transcriptional regulator. Moreover, when we reduced the set of tested genes to DCU genes, structures and cellular components associated with neural development such as synapses (FDR = 0.02), neurite/axon outgrowth (growth cone FDR = 0.01) and cell movement (cell leading edge FDR = 0.02) appear additionally over-represented (Figure 5.20). Remarkably, positional features showed the highest level of enrichment in DCU genes when compared to DE genes including NLSs (FDR = 3.2e-09,NLS), IDRs (FDR

= 1.86e-12) and coiled regions (FDR = 5.36e-07). Genes with NMD isoforms were also significantly enriched (NMD, FDR = 1.61e-11), which correlates with their post-transcriptional origin.

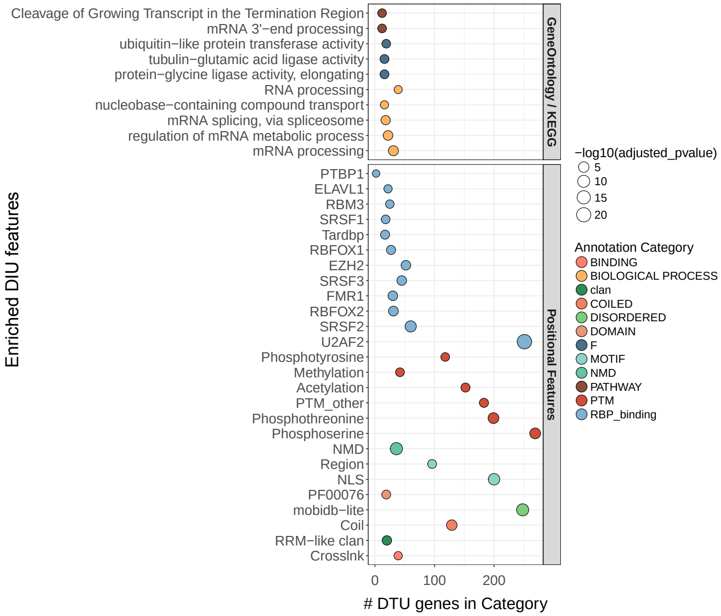


Figure 5.19: Functional enrichment of DIU genes. Functional categories available in our IsoAnnot annotation file were tested for functional enrichment using Fisher exact tests. The significant sub-categories are shown in the graphic plot. The dots are coloured according to their functional category and their size indicates their level of significance (FDR).

In conclusion, DIU between neural cell types mainly impacts the auto-regulation of mRNA processing while DIU specifically altering the gene coding potential are highly associated with neural-specific compartment-related genes. Moreover, certain functional features as NLSs, IDRs and phosphorilation site appear highly accumulated in genes undergoing alternative transcript processing regulation.

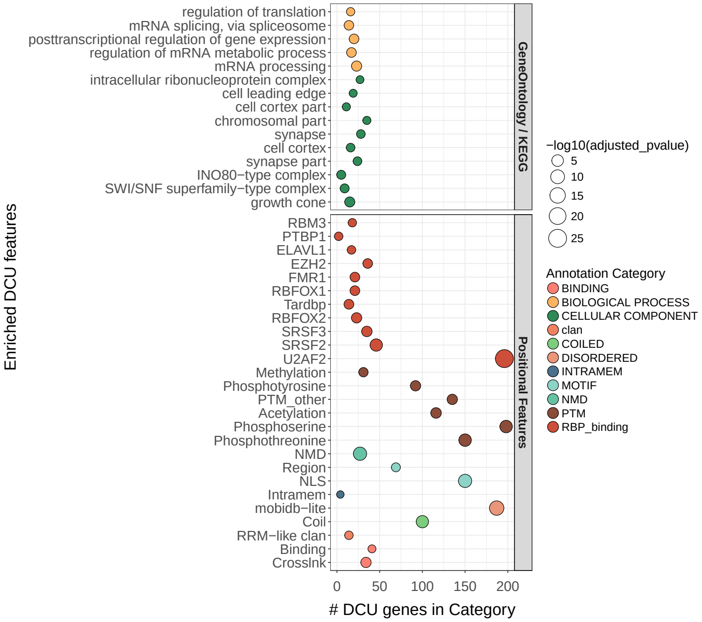


Figure 5.20: Functional enrichment of DCU genes. Functional categories available in our IsoAnnot annotation file were tested for functional enrichment by using Fisther exact test. The significant sub-categories are shown in the graphic plot. The dots are coloured according to their functional category and their size indicates the level of significance (FDR).

5.4.3 Impact of differential isoform usage on functional properties

To understand the functional impact of DIU, we applied the DFI analysis (see in Section 5.2.4.1), which identified significantly regulated functional features for a total of 380 genes, almost 80% of the DIU genes detected in the previous section, highlighting the capacity of our framework to detect the regulation of functional properties in most of post-transcriptionally regulated genes. DFI features were distributed in all the considered categories (Figure 5.21.A), with significant relative enrichment for uORFs (FET p = 5.25e-121), RBP binding sites (FET p = 2.46e-07), regions of compositional bias (FET p = 4.06e-03) and IDRs (FET p = 5.02e-03). To remove the bias of some features to appear repeatedly DFI in a gene because of their multiple occurrences, we collapsed results at gene level, thus finding again IDRs and 5' UTR elements such as uORFs as the categories most overrepresented in differential inclusion (Figure 5.21. B).

Attending to cell-type preferential usage of features, IDR gains and losses were equally distributed in both OPCs and NPCs while signal peptides were found preferentially included in OPCs (Binomial test with probability of succcess 0.5, FDR = 2.10e-2, Figure 5.21.C). At transcript-level annotation miRNA binding sites, uORFs or RBP binding sites are more frequently included in OPCs (Figure 5.21.C, Binomial test with probability of succcess = 0.5, FDR miRNA binding = 3.85e-08, FDR RBP binding = 5.85e-31, FDR uORF = 4.09e-4), suggesting a potential UTR lengthening.

Remarkably, the differences in feature inclusion rates of most categories are around 20%, indicating that alternative isoform processing do not trigger dramatic feature gains and losses between NPCs and OPCs but, in aggreement with differential isoform usage results detailed in Section 5.4.2, slightly modulate their inclusion levels. However, Kruskal test revealed significant differences across categories (Figure 5.21.D, Kruskal test p = 4.30e-18). Coiled regions and IDRs were identified as the feature categories with a significant higher differential-inclusion levels between cell types (Mann-Whitney test disordered FDR = 3.63e-07, coiled FDR= 5.43e-06).

MAP4, a microtubule-associated protein, illustrates the usual complexity of loci and the ability of our DFI analysis to capture features differentially regulated across different conditions or cell types. The *Map4* gene expresses 10 transcript isoforms (6 of them representing novel variants discovered by PacBio sequencing) which encode 10 predicted protein variants. Figure 5.22.A-B shows the 5 top expression isoforms, revealing different splicing events whose combination leads to a high level of transcriptome and predicted proteome complexity. DFI analysis for *Map4* identified several protein elements with differential inclusion rates between NPCs and OPCs: IDRs, a NLS (score=5.2), several phosphorylation sites and a tubulin-binding domain (PF00148) (Figure 5.22.C). The latter was detected preferentially included in OPCs as consequence of the predominant usage of isoforms containing 4 tubulin motifs in OPCs (Figure 5.22.B-C) which may lead to improved MAP4 binding in OPCs. This regulation pattern correlates with the known continuous enrichment of MAP4 protein isoforms containing 4 binding domains during development [48]. Moreover, family-related microtubule-associated genes such as *Mapt* can modulate its microtubule binding properties by skipping a tubulin-binding domain [186] reinforcing the idea that the regulation of *Map4* pre-processing increase its microtubule binding through the inclusion of a fourth domain. Regarding phosphorilation sites, evidence of

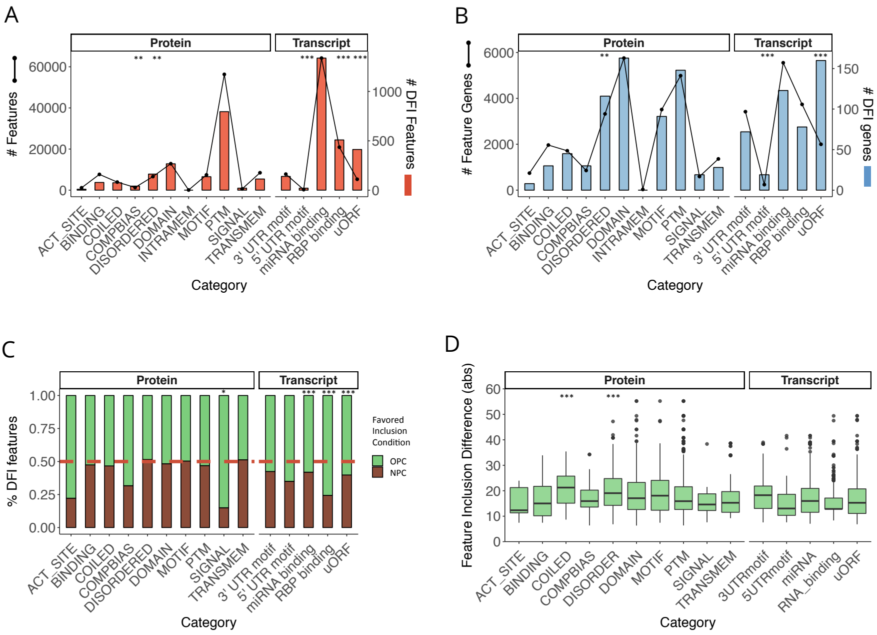


Figure 5.21: Differential Feature Inclusion results between NPCs and OPCs. A. Distribution of the number of features in transcriptome across categories and distribution of differentially included features. The relative over-representation of DFI features in specific categories was statistically evaluated by using Fisher Exact tests to capture enriched categories with Benjamini-Hochberg multiple testing correction. B. Distribution of the number of genes with annotated features in transcriptome across categories regarding the distribution of genes with significantly regulated features for each category. Over-representation test performed as above. C. Distribution of DFI features between favored in OPCs and NPCs. Binomial test with probability = 0.5 and Benjamini-Hochberg multiple testing correction was applied to capture categories enriched in cell-type specific inclusions. D. Absolute levels of differences in inclusion levels across the cell types. Differential distribution across categories was tested with the non-parametric Kruskal test. To identify specific categories, we performed Mann-Whitney tests with Benjamini-Hochberg multiple testing correction. (***) $p < 0.001$; (**) $p < 0.01$; (*) $p < 0.05$.

cooperation between AS and PTMs have been recently reported as a way to provide genomes with signalling plasticity [393]. Indeed, certain MAP4 phosphorylation sites have already been identified as critical for tubulin polymerization activity [**?**]. Hence, our DFI analysis captured the complex pattern of feature inclusion triggered by post-transcriptional regulation in *Map4* gene, and more importantly, allow us to hypothesize about their functional impact on essential properties such as binding and signalling in our system of study,

The effect that AS has on the alteration of binding properties have been largely reported [94][59]. DFI within annotated binding categories (UniprotKB Binding category and Pfam domains) revealed the Ankyrin repeat superfamily, involved in protein-protein interactions, enriched in DFI features with respect to the other binding motif categories (CL0465, 23%, FET p-value = 0.02). Actually, Ankyrin regulation has been previously described for the *Ank2* gene, whose binding affinity to spectrin and tubulin is altered by differential inclusion of an ankyrin repeat [72]. In addition to protein binding effects, numerous genes were detected as differentially regulating the inclusion zinc-finger domains (e.g. *Zfp148, Mbnl2, Zfp961*), RNA-recognition motifs (e.g. *Rbm39, Rbm7, Scml4*) and small-ligand binding motifs (e.g. *H2afy, Gdi1, Nnt, Pycrl, Pcyox1*). Among them, *Nfx1, Mbd1* and *H2afy* have been previously associated with post-transcriptional regulation of their binding properties, leading to altered transcriptional activity [163], recruitment to CpGs [157] and interaction with NAD^{+}-derived small molecules [232], respectively.

Together with the alteration of binding motifs, the change in the intracellular localisation of proteins is the most well-known and studied aspect of post-transcriptional regulation. It involves changes between cytosol and nucleus through the alternation of nuclear localization signals (NLS), the movement from plasma to membrane by the alternative use of transmembrane regions, or even directs proteins across the ER membrane to enter in the secretory pathways by the inclusion of N-terminal signals, among others. Our DFI results showed 74 genes with differential NLS usage, 30 genes altering their transmembrane regions and 19 genes differentially including the secretory-targeting signal. Thus,

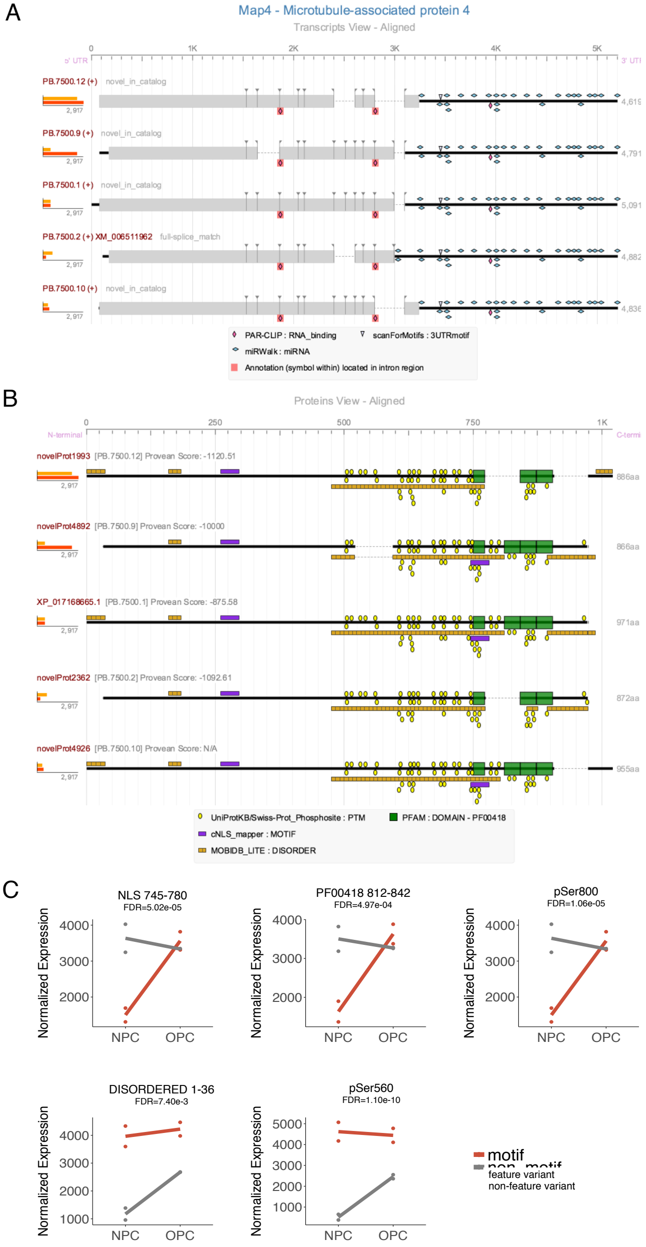


Figure 5.22: Potential functional impact of DIU in Map4. A. The complex combination of splicing events and alternative transcription initiation sites leads to the expression of eight isoforms, five of them novel discoveries of PacBio. B. Comparison of predicted coding regions reveals disruptions and losses of features from different functional categories. C. DFI analysis detected significant context-specific inclusion of NLSs, PFAM domains, Disordered regions and Post-translational modifications.

post-transcriptional regulation promotes the differential inclusion of features associated with intracellular protein localisation for 108 out of 278 DCU genes detected in the previous section, what represents almost 40% of them. Therefore, DFI revealed an spread potential impact of AS/APA regulation between NPCs and OPCs on subcellular re-localization. The functional analysis of genes with regulation of NLSs highlighted transcriptional regulation (*Top2a, Sall1, Kat5, Ctnnd1, Ncapg, Scmh1*, etc.), RNA-binding (*Mbnl2, Wdhd1, Mbnl1, Glod4, Srek1, Rbm39*, etc.), kinase activity (*Adk, Clk4, Clk1, Trpm7, Mapk7, Scyl1*, etc.) and cell differentiation (*Rufy3, Sall1, Ezh1*, etc) as the functional categories (GO terms) with the highest number of NLS regulated genes.

Partial protein translocation to nucleus by exon inclusion or longer N-terminal that code an NLS has recently been reported in some of them such as *Mbln1, Mbnl2* [] or *Adk*[**?**][168]. Two so far non-reported cases discovered in our analysis are the kinases *Clk1* and *Clk4*. In both cases, splicing results in the inclusion of an exon that promotes the use of downstream start codon affecting the inclusion of a NLS in the resulting protein. In both kinases, the NLS-excluding forms increase their relative abundance in OPCs, likely causing the translocation of the kinase activity to the cytoplasm. CLK1 and CLK4 belong to a family of dual specificity proteins which, in the nucleus, phosphorylate serine/arginine-rich proteins that regulate alternative splicing programs [223]. Moreover, CLK1 auto-regulates the splicing of its own pre-mRNA according to its kinase activity, generating the increased expression of an inactive splicing variant that truncates its kinase domain and regulates its activity [89]. From our results, we speculate that, splicing regulation in CLK1 and CLK4 between NPCs and OPCs could, additionally, impact their activity by promoting the expression of kinase isoforms that cannot translocate to nucleus or even acquire alternative functions in the cytoplasm in OPCs, where the NLS is predominantely excluded.

A similar NLS motif inclusion pattern was found for *Ctnnd1*. *Ctnnd1* codes for the p120 protein, a member of the Wnt-B-catenin signalling pathway, that is a key component in neural differentiation [92][219]. In our data, PacBio detected 4 isoforms variants formed by the combination of 2 alternative splicing events, one imprinting an NLS through the exclusion of exon 10 (5.23.A). p120 NLS-containing isoforms were strongly underrepresented in NPCs, while their relative expression levels significantly increase in OPCs, where they become the major isoforms (5.23.B-C). According to the NLS inclusion-pattern data, high cytoplasmic retention of p120 is expected in NPCs, while relatively higher nuclear levels should be found in OPCs. Western blot analysis of nuclear and cytoplasmic fractions in NPCs and OPCs verified this differential localisation pattern leading to the enrichment of the cytoplasmatic *Ctnnd1* form in NPCs (5.23.D) and confirms the functional readout identified by our DFI analysis for the *Ctnnd1* gene.

The co-regulation of functional elements is another intriguing question about transcriptome complexity and function. CoDFS identified pairs of deferentially included features that are actually known to cooperate to generate specific functional readouts such as the NLS and phosphoserine events (FDR = 3e-35): post-translational masking of NLS provides a regulatory mechanism to prevent nuclear import [213] [137] [384] [222] [257] [40]). This pair includes 32 genes (10 % of DFI-associated genes) enriched in mRNA processing (FDR p = 4.97e-05) and suggests complementary PTM masking and differential splicing regulation of NLSs. Regulation by phosphorylation-dependent activation of the NLS, likely as a mechanism for rapid control of protein localization in cells and a more stable regulation by co-skipping of the NLS signal and their regulatory phosphoserine sites. As example, the histone acetyltransferase *Kat5*, which is known to shuttle between the nucleus and the cytoplasm by AS events [270], up-regulates the relative abundance of the short isoforms excluding the NLS and the phosphorilation site in OPCs. Understand if PTM and AS regulatory mechanisms of subcellular localization could complementary regulate Kat acetyltransferanse activity or allow complementary functions in cytoplast is key in our system under study because of the tighly regulated histone acetylase activity during oligodendrocyte lineage progression (ref,ref).

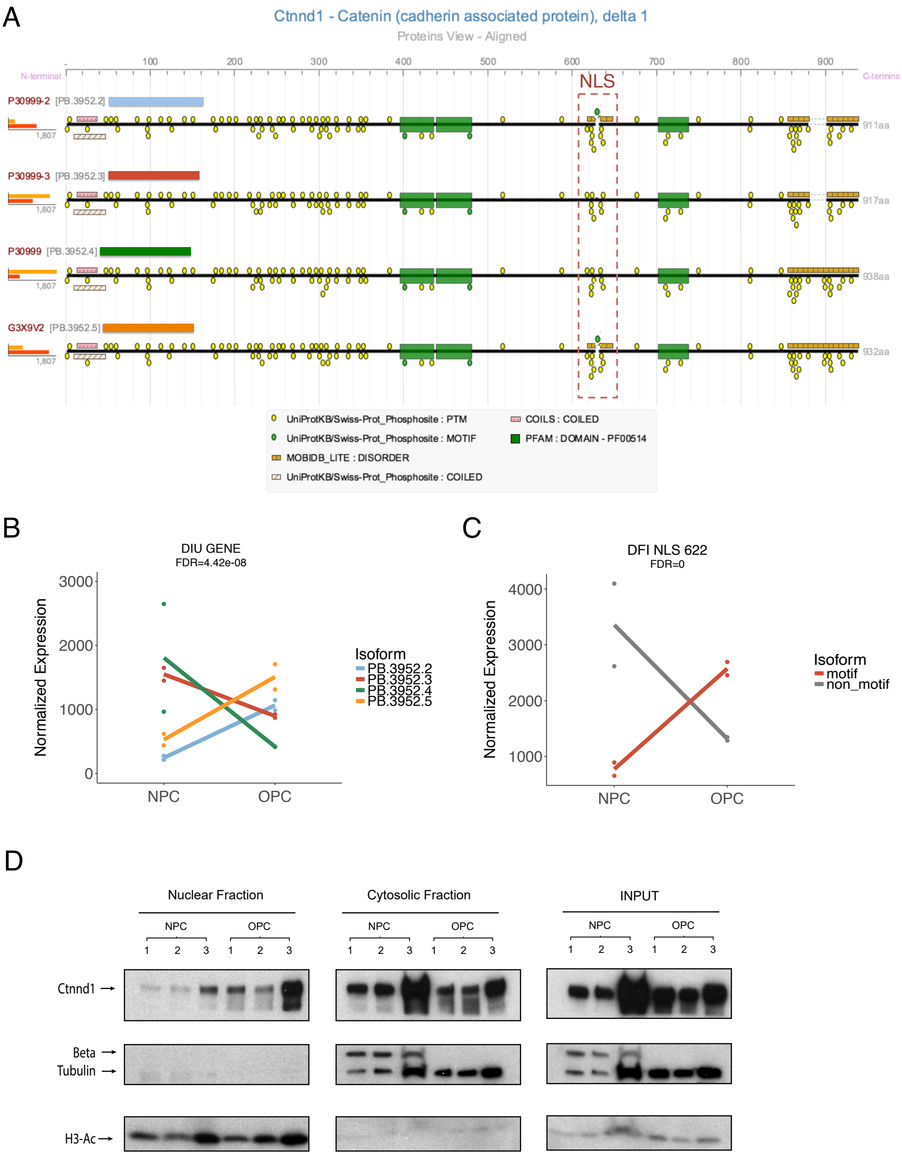


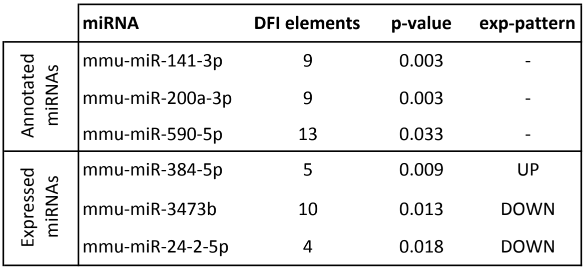
Figure 5.23: Experimentally validated functional impact of *Ctnnd1*. A. The pair-wise combination of two splicing events generates four alternative coding sequences involving the disruption of an IDR and an NLS. B. Although *Ctnnd1* is detected as a DIU with high levels of transcript usage change (47%) and major isoform switching, its differences in gene expression were not significant. C. DFI analysis detected significant inclusion of a NLS in OPCs. D. Western blot analysis of nuclear and cytoplasmic fractions showed the enrichment of the cytoplasmatic form of CTNND1 in NPCs.

5.4.4 Impact of APA events in UTR modulation and containing features

We applied our DPA analysis to study the differential usage of alterantive polyA sites between NPCs and OPCs (polyA distance threshold 75 nt, isoform proportion filtering 0.1, FDR = 0.05). Differential polyadenlyation was detected for 16% of genes expressing alternative polyA sites (128 genes out of 780, FDR $<$ 0.05, Figure 5.24.A). Among them, *Lamp2* displays the highest level of ΔDPAU between conditions (DPA FDR = 9.34e-11; ΔDPAU=38.6). Mediating lysosomal degradation of proteins in response to various stresses, Lamp2 undergoes a polyA site switching favoring the use of the distal polyA site in OPCs and promoting a 3' UTR shortening (Figure 5.24.B), in agreement with previous studies reporting the APA regulation of *Lamp2* [68].

Although there was no general pattern of predominant distal usage in any cell type (Figure 5.24.A), the UTR lengthening/shortening analysis of all the genes expressing varying UTRs (100 bp cutoff) revealed an overall longer median 3' UTR length in OPCs (Figure 5.24.C, WRT test: p = 2.26e-5) in contrast to 5' UTRs, where there was not found any significant UTR regulation pattern.

This 3'-UTR lengthening pattern correlates with the higher proportion of miRNA and RBP binding sites which were detected in DFI analysis as OPC gains (Figure 5.21.C). In total, 123 genes underwent 3'-UTR regulation impacting miRNA binding sites between cell types. Ranking of miRNAs showed that members of the miRNA200 family such as miR-141-3p and miR-200a-3p were the most



	miRNA	DFI elements	p-value	exp-pattern
Annotated miRNAs	mmu-miR-141-3p	9	0.003	-
	mmu-miR-200a-3p	9	0.003	-
	mmu-miR-590-5p	13	0.033	-
Expressed miRNAs	mmu-miR-384-5p	5	0.009	UP
	mmu-miR-3473b	10	0.013	DOWN
	mmu-miR-24-2-5p	4	0.018	DOWN

Table 5.1: Top miRNAs with binding sites over-represented in regulated UTRs between NPC and OPC cell types.

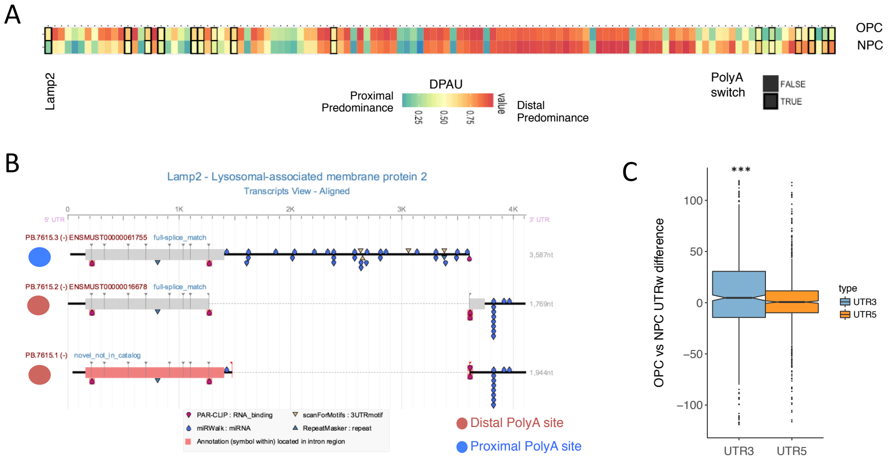


Figure 5.24: Differential polyadenylation (DPA) between NPCs and OPCs. A. Heatmap displays DPAU levels associated to DPA significant genes (FDR $<$ 0.05) for each studied cell type. B. Gene model visualization for *Lamp2* gene, which is subjected to APA modulation between NSC and OPC stages. Red isoforms indicate predicted NMD targeting. C. Boxplots showing the distribution of the weighted 3 and 5' UTR lengths in neural types.

over-represented in differential isoform targeting (Table 5.1). Moreover, in correlation with FD results (Section 5.4.1), miR-590 binding sites were the most frequently regulated between OPCs and NPCs. Surprisingly, in both cases, a high number of targets were related to endocytosis and neural-specific morphological structures such as the somatodendritic compartment, including *Gmp6b*, *Rufy3*, *Lamp2*, *Dync1li2*, *Pcdh2*, and *Vamp4*. For example, 3' UTR regulation in *Rufy3*, involved in neural polarity, promotes the inclusion of the mir590-50 miRNA binding site and is coupled with the deletion of the C-terminal protein region (Figure 5.25.A), responsible of its interaction with RAB5A (GTPases Rab controlling intracellular membrane trafficking [141]) and co-localisation in large vesicle structures [381]. This coupled effect indicates a coordinated and interdependent regulation of isoform-specific miRNA targeting and protein functionality. After filtering out miRNAs by expression (miRNA microarray on NPC and OLG samples), the mir-384-5p, mir-3473b and mir-24-2-5p appeared over-represented in binding sites differentially included between cell types (Table 5.1). Mir-3473b, recently associated to neuroinammation [357] targets specific isoforms of enzymatic genes such as *Ube2l3* (Figure 5.25.B) or *Eif4a2*.

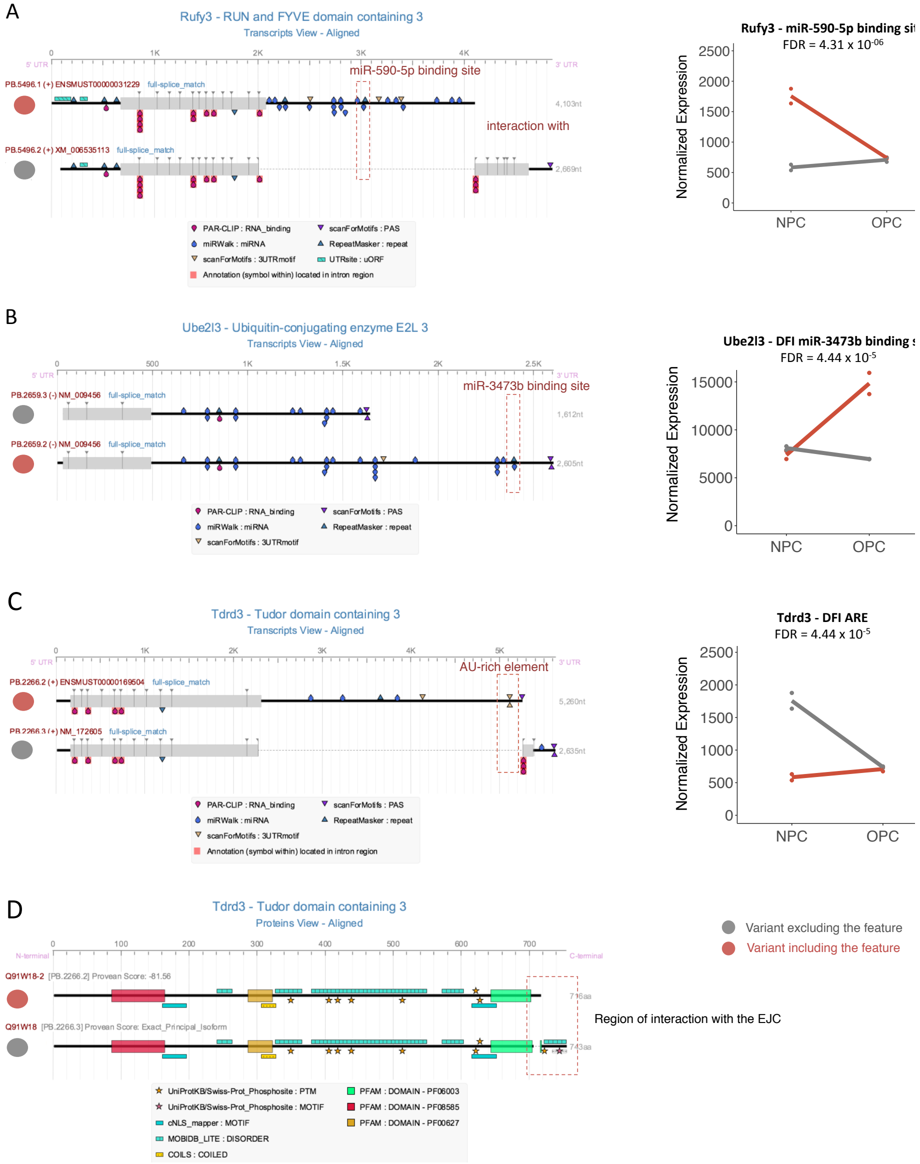


Figure 5.25: Impact of differential polyadenylation (DPA) in the inclusion of regulatory elements at UTR regions between NPCs and OPCs. A-C Gene models for *Rufy3*, *Ube2l3* and *Tdrd3* genes together with functional annotation of alternative UTR regions and the inclusion profile of features across cell types. D. Protein diversity expressed by *Tdrd3*, highlighting the impact of DPA on PTM sites and a region of interaction with the exon junction complex (EJC).

Other elements as AU-rich elements (ARE), well-known to influence target transcript fate [49], were detected with different inclusion rates among transcript variants in 11 genes, including the *Tdrd3* (Figure 5.25.C) which mediates transcriptional activation in nucleus and formation of stress granules and regulation of mRNA translation in the cytoplasm [124]. The use of a CR-APA promotes the inclusion of an AU-rich element along differentiation which is coupled to the lost of two elements associated to the coding part of the protein: a phosphotyrosine site and a region potential to mediate interaction with the exon-junction complex (EJC, Figure 5.25.C-D). This isoform-specific EJC-binding motifs (EBMs) allow the recruitment of proteins with post-transcriptional functions to mRNAs via the EJC [162]. All together indicates that the dual role of *Tdrd3* as disassembler of mRNA stress granules and regulator of translation in cytoplasm and trasnscriptional activator in nucleus is triggered by APA regulation, which leads to regulatory (ARE gain) and functional (EJC motif loss) isoform specificity. Similarly, *Eif4a2*, target of AUF1 (RNA-binding protein mediating the ARE-dependent stabilization process [395]), showed an DPA pattern that leads to the exclusion of an AU-rich element together with the inclusion of an ubiquitination site (K382) at the CDS in OPCs. Previous analysis reported the association of a single residue (K226) of *Eif4a2* with its recruitment to stress granules when re-programming of the translation machinery is needed [156]. It would be interesting to understand if K382 might have a similar role in the recruitment of *Eif4a2* to specific organelles. Thereby, our results suggest the modulation of mRNA stress granules metabolism between NPCs and OPCs by differential polyadenylation and indicate the isoform-specific ARE-mediated destabilization of *Tdrd3* and *Eif4a2* in function of the isoform functional role.

5.5 Discussion

Despite many computational methods have been developed to elucidate the dynamics of isoform regulation, the genome-wide study of the potential functional impact triggered by post-transcriptional mechanims is not a strightforward task because of the lack of tools and methods integrating contextual isoform data and isoform-resolved functional annotation. Here we presented the first comprenhensive computational framework to investigate the effect of differential isoform usage in functional properties which is implemented in an interactive and dynamic tool combining statistical and graphical tools: tappAS (http://tappAS.org).

One fundamental question about isoform function is how post-transcriptional regulation imprints functional complexity to transcriptomes. tappAS implements the FD analysis which systematically evaluates the genome-wide level of structural (i.e. UTR length) and functional (i.e. phosphorylation sites) diversity across isoforms for the wide range of functional annotation categories. Applied to the comparison of murine neural cell types, we found that more than 70% of multi-isoform genes generate alternative CDSs, in contrast to the reduced variability at UTR length and polyadenylation sites (45%). Besides, functional features varying across isoforms were detected for almost 90% of the genes expressing multiple isoforms. Specifically, at transcript level, GU-rich elements and mir590-3p binding sites were over-represented in regulated features in 3' UTR motif and miRNA categories, respectively, illustrating already reported cases as the functional lost of a mir590-3p binding sites in the *Znf123* transcription factor. At protein level, IDRs and PTMs were detected as the most varying features, in agreement with previous studies reporting their enrichment in alternatively-spliced regions [280][393]. Moreover, IDRs were highly associated to linear motifs and PTMs, suggesting overlapping and joint pre-processing. Positional vs. quantitative FD analysis showed zinc fingers and KRAB box domains as usually completely contained in AS exons. In contrast, kinase and RNA binding domains stood out at the positional FD analysis, indicating their frequent partial disruption. In summary, our results illustrate the power of the tappAS FD analysis to explore the isoform functional diversity present in a mammalian transcriptome.

When transcript expression is provided, tappAS assesses transcriptome regulation, including gene, transcript and CDS differential expression, differential isoform usage involving regulation of the coding sequence and/or the untranslated region or isoform switching events. Their flexible combination allows to configure specific questions on transcriptome dynamics and its associated functional enrichment, when combined with functional information. In our murine system, we found 378 genes with isoform usage regulation, half of them without involving gene expression regulation and only a low fraction of them resulting in switches of the major expressed isoform between NPCs and OPCs. Compared to DE genes, we found DIU genes over-represented in processes involved in 3'-end mRNA processing, RNA binding and mRNA splicing as well as targets sites for RNA binding proteins at the enrichment of positional motifs, indicating that the post-transcriptional machinery accomplishes a high degree of self-regulation between neural cell types. Moreover, 75% of DIU genes involve the regulation of the CDS, which, interestingly, are over-represented in neural cellular components. Therefore, combining tappAS differential, filtering and enrichment analysis functions in our neural system, we were able to describe a scenario of post-transcriptional regulation fundamentally decoupled from the regulation of gene expression that mainly impacts the RNA metabolism machinery itself but remarkably, also involves the regulation of genes located in neural-specific structures when the changes affect the coding sequence of the gene.

To link functional diversity with isoform usage dynamics tappAS includes the differential feature inclusion (DFI) analysis that profiles the dynamic change in the feature content of full-length isoforms. Additionally, alternative polyadenilation and UTR shortening/lengthening analysis were coupled to study the impact of isoform regulation on UTR modulation, key for transcript fate regulation. Moreover, joint visualization of expression levels and isoform-resolved functional and structural elements allows the easy mapping of functional differences at transcript models. Applied to our pair-wise neural system in mouse, differentially included features were detected for nearly 80% of the detected DIU genes, indicating that our analysis framework captured regulation of functional properties for most of the post-transcriptionally regulated genes. Moreover, they were found highly distributed across the considered databases, demonstrating that isoform regulation consequences are diverse. An example is *Map4*, whose isoforms, generated by complex combinations of alternative events, are regulated to impact several functional elements such as NLSs, PTMs and IDRs, potentially regulating protein localisation, adding signalling plasticity and modulating protein interactions, respectively. Among all the considered annotation categories, uORFs, RBP binding sites, regions of compositional bias and IDRs were found significantly impacted by post-transcriptional regulation. Moreover, the inclusion of elements at 3'-UTR regions were significantly favored in OPCs, what correlates with the significant 3' UTR lengthening pattern detected in OPCs. MiRNA analysis revealed binding sites for the miRNA 200 family (miR-141-3p and miR-200a-3p) as frequently impacted by UTR regulation and inspection of genes undergoing differential inclusion of AU-rich elements revealed two genes, *Tdrd3* and *Eif4a2*, that are potential to trigger isoform-specific roles in stress granules coupled to isoform-specific ARE-mediated destabilization.

Regarding protein features, ankyring repeat domains, small-ligand binding sites, zinc-finger motifs and RNA-recognition motifs were frequently regulated between NPCs and OPCs, some of them previously validated. Moreover DFI analysis detected a high proportion of genes (40%) that dynamically modulate the inclusion of features associated with the intracellular protein localization (signal peptides, transmembrane regions or NLSs) between our neural cell types. NLSs were differentially included in 74 genes, comprising already reported genes such as the *Mbnl1* and *Mbnl2* RBPs and the *Adk* kinase. Among the set of not reported genes with detected potential impact on nuclear localization, we experimentally validated that the post-transcriptional regulation of the *Ctnnd1* gene, a member of the Β-catenin signaling pathway involved in the differentiation of NPCs into OPCs, leads to its accumulation in cytoplasm in the differentiated stage. Moreover, the complementary coFDS analysis, which explores sets of functional motifs that are processed together, revealed the association between the inclusion of phosphoserines and NLSs signals, suggesting a double layer of NLS regulation comprising the post-translational masking of NLSs as a mechanism for rapid control of protein localization and a more stable regulation by co-skipping of the NLS signal and their regulatory phosphoserine sites in our system. Thus, use of tappAS to the analysis of two neural cell types in mouse demonstrated its ability to recapitulate a great deal of the existing knowledge on isoform function and yet reveal new functional insights.

Thanks to the flexible combination of statistical tests, enrichment methods, filtering and visualization options, tappAS brings the analysis of of isoform function to the reach of experimentalists with little computational skills. Formulate varied functional hypothesis about the role of isoform regulation in a given system of study as analysing if differential splicing across conditions is impacting specific functional elements such as post-translational modifications or nuclear localization signals, or coming up with genes regulating UTR regions and contained motifs are questions that can now be easily answered. We anticipate that tappAS will enable the broad-scientific community to lead advances in the understanding of the functional relevance of the alternative processing of transcripts.

Chapter 6

Functional consequences of differential isoform usage in neural fate determination

6.1 Introduction

During mammalian spinal cord development, motor neurons (MNs) and oligodendrocyte precursos cells (OPCs) are generated in a highly specific manner, both spatially and temporally, from a common pool of neural progenitor cells included in the so-called pMN domain. While MN are responsible for transmitting signals from the spinal cord to muscles and enable muscle contraction, OLGs are glial cells that produce and maintain the myelin sheath that surrounds axons of the central nervous system (CNS), thus forming an electrical insulator that allows rapid signal transmission.

Most research in this area has focused on understanding cell fate decisions during differentiation of MNs and OPCs. Precise modulation of proneuronal and proglial factors [308] [74][161][394], reversible post-translational regulation [188] or differential signaling between neural cells [111][45][243] have been revealed behind the dramatic morphological and functional changes that dictate cell fate specialization. However, the post-transcriptional landscape underlying fate determination of NPCs into different neural subtypes as well as the functional effects of cell-type specific AS and APA events remains poorly understood.

AS is already recognised as an particularly widespread regulatory mechanism in the nervous system (CNS) [265] and involved at every step of neural development, including neuronal migration or establishment of synapses [231] [392] [56]. Genome-wide technologies have revealed the specificity of splicing patterns across neural cell types, brain regions, and developmental stages [362][391][374][345][155] as well as the dynamic and precisely regulation of splicing programs relying on the proper expression and function of splicing regulators [362][388][149][266]. Moreover, APA programs has been shown to be temporally coordinated in an AS-independent manner during neurogenesis and brain development [129][281][7][103]. Dysregulation of splicing regulators or defects in cis-acting splicing elements have been linked to several brain disorders and neurodegerative diseases such as amyotrophic lateral sclerosis (ALS), schizophrenia and autism [75] [345] [265] [149]. However, our understanding of how splicing switches and APA control the functional redout responsible for

NPCs specification into glial and neuronal subtypes during spinal cord development remains limited. Deep understanding on the genome-wide impact of neural-type specific isoform regulation on protein function, mRNA stability, subcellular localization or gene expression regulation via NMD is essential to resolve the functional specificities of alternative lineages as well as decoding the potential basis of several splicing-based neurodegenerative diseases impacting specific neural subtypes.

Thus, in this chapter, we studied the temporal functional impact of differential AS and APA regulation on the cell fate determination of NPCs into MN and OPCs at the genome-wide level. To do this we applied the new paradigm for the functional analysis of differential isoform usage defined in this thesis. Making use of long-read transcriptome sequencing for the definition of transcriptome complexity and functional profiling methods operating at isoform-resolution, we captured the isoform diversity landscape of neural tissues, characterized the potential functional impact of several neural-specific AS and APA events, profiled the dynamic gain and lost of functional features across neural cell types during fate determination.

6.2 Methods

6.2.1 Experimental design

Our experimental design included multiple samples from differentiation systems designed to derive glial and neuronal cells. Neural differentiation experiments were conducted in collaboration with the Neural Regeneration Laboratory at the Príncipe Felipe Centro de Investigación. NPCs were isolated from spinal cords collected from neonatal mice (4 days old) and cultured in suspension for seven days to produce neurospheres. These neurospheres were then put into OPC or MN differentiation conditions for 35 days, as previously described by Moreno et al. [216], using the same NPC pool for both differentiation processes. To account for biological variability, the process was repeated in duplicate. To profile the dynamics of AS and APA regulation, four differentiation stages from each NPC differentiation experimental condition were harvested for total RNA extraction at 0, 9, 28, and 35 days for OPCs and 0, 9, 15, and 35 days for MNs.

6.2.2 RNA-seq by single molecule and short-read sequencing

The total RNA was extracted from samples of NPCs and cells derived from them and was retro-transcribed using the Clontech SMARTer^{TM} cDNA synthesis kit. The samples were randomised in four different batches to account for technical biases during data analysis. Full-length cDNA samples from each time point and lineage (two biological replicates each) were split to prepare Illumina and PacBio sequencing libraries. The PacBio Iso-Seq libraries were sequenced on the PacBio RS II platform using P6-C4 chemistry. To avoid loading bias, which favours the sequencing of shorter transcripts, a BluePippin device was first used to perform multiple size-fractionation (1-2 kb, 2-3 kb, 3-5 kb and 5-10 kb). A total of 135 SMRT cells (a median of 9 SMRT cells per sample) were sequenced following the Iso-Seq PacBio sequencing protocol, providing a total of 10 M of PacBio reads, one of the highest PacBio sequencing depths generated to date. We also conducted 2x75 paired-end Illumina sequencing using the Nextseq platform, which yielding around 50 M of paired-end reads per sample.

Both sequencing approaches were carried out at the Interdisciplinary Center for Biotechnology Research at the University of Florida.

6.2.3 De novo discovery of neural isoforms by Iso-Seq and transcriptome curation

PacBio Iso-Seq data were processed using the command-line version of the PacBio Iso-Seq$^{\mathrm{TM}}$ analysis software (version 3.0). This software version was implemented to allow users to work with data from the Sequel instrument, which provides raw sequences in bam format. Therefore, prior to analysing our PacBio RS II data, we converted reads from bax.h5 to bam format. CCSs were generated by setting a minimum number of 1 passes, a minimum predicted accuracy of 0.8 and a minimum signal-to-noise ratio (SNR) cutoff of 2, yielding a representative consensus sequence ZMWs in which the SMRT adapters were detected. FL CCS classification was ran with default parameters. Both steps were independently performed for each single SMRT cell. Transcripts were then identified using ToFU2 v3.8 inside Iso-Seq Analysis Pipeline. Because of the large amount of data generated (around 10M of PacBio reads), isoform discovery was accomplished by dividing data into five partitions. The ToFU2 pipeline was then independently run for each partition using default parameters (excepting partial hits removal option specified during preCluster step) and taking advantage of extensive parallel computing. We skipped the polishing step from ToFU2 pipeline because the sequencing depth from FL reads was sufficient to generate high quality sequences. Next, GMAP [370] with sense force option was ran to align sequences to the mouse reference genome (mm10 version). Redundant ICE isoforms were collapsed using two different tools: TAMA [178] and PacBio's Cupcake method (`https://github.com/Magdoll/cDNA_Cupcake`).

The SQANTI [313] software developed in this thesis was applied to characterise transcriptomes defined by long-read sequencing (Chapter 3). *Sqanti QC* was run to inspect isoform, providing the junction coverage estimations computed by the STAR aligner (using the parameters specified in Chapter 3) and RefSeq (version 78) genome annotation as input. The filter funtion of SQANTI was subsequently applied to remove potential artefactual isoforms, thus defining a curated transcriptome which from hereon in we refer to as NEURALtc. We set a probability of 65% of adenines at the genomic 3' end on order to flag isoforms as intra-priming and set a probability higher than 0.75 for *Random Forest* to classify positive isoforms.

Finally, we analyzed the contribution of the AS, APA and ATI mechanisms to transcriptome complexity by comparing splice-junction patters across gene isoforms to define the number of unique splicing patterns per gene. The number of APA and ATI variants per gene was then calculated by separate comparison of genomic TTS and TSS coordinates across isoforms derived from the same gene, respectively. A window of 100 bp was established as the threshold for definition of independent TSS or TTS sites.

6.2.4 Isoform quantification and normalisation

The expression quantification of Iso-Seq isoforms for each condition using Illumina short-reads and RSEM software in accordance with ENCODE-recommended guidelines [188]. Prior to statistical analysis, we assessed the effect of sequencing biases on expression quantification so that we could choose the most appropriate normalisation procedure. Based on our exploratory analysis results, we selected TMM normalization method [276]. The NOISeq R package [312] was used for sequencing-depth and TMM normalization. Finally, we also interrogated the data for batch effects and were unable to find any technical confounders influencing the expression estimations.

6.2.5 Principal component analysis for lineage characterisation

To analyse distance and relatedness between differentiation stages and cell types, we performed a principal component analysis (PCA) on gene-expression logarithmic levels and mean centered expression values. The loadings of specific MN and OPC biomarkers were examined for their contribution to the definition of principal components.

6.2.6 Relevance of novel isoforms in defining cell identity

To understand the relevance of novel variants discovered by PacBio sequencing in our neural system, we performed a comparative study between the set of novel and known isoforms based on their contribution to total gene expression. We classified the isoforms as 'principal' (PI) or 'minor' (MI). PIs were catalogued as those with the highest relative proportion in at least one experimental condition. The remaining isoforms were labelled as minor. Only genes expressing both novel and known variants were considered in order to make groups comparable.

We also carried out PCA on the mean-centered logarithmic expression estimations of novel isoforms to analyse the ability of novel calls to define cell identity and maturation timing in multiple differentiation systems. SQANTI categorization was used to classify the transcripts into novel and known using RefSeq (version 78) as the reference gene annotation. Alternatively, CDS novelty-characterisation data was retrieved from IsoAnnot output (Section 4) and RefSeq78, UniProt Trembl and Uniprot Swissprot as reference protein sources.

In addition, the replicability of novel isoforms was characterised by measuring how many of them were detected by Iso-Seq sequencing in two biological replicates. We catalogued isoforms as being found in replicates if the isoform was fully-sequenced in two biological replicates of one studied condition. We used the number of full-length Iso-Seq reads classified by ToFU2 as belonging to the given isoform.

Splicing events in Iso-Seq isoforms were characterised using the Suppa tool with default parametes [3]. Functional enrichment of novel isoforms was performed using GOseq [382], defining the group of genes expressing novel isoforms as the test set and the total set of genes defining our NEURALtc as the background. Finally, the functional network of enriched processes was obtained from tappAS.

6.2.7 Transcriptional and post-transcriptional dynamics in neural differentiation systems

Gene expression and isoform usage dynamics was profiled and statistically interrogated for changes over the course of time in both glial and neural differentiation systems by using tappAS (Chapter 5), running independently time-course analysis for each cell type. The binomial distribution of the data and 3 and 2 regression polynome degrees for DGE and DIU analyses, respectively, were set as the parameters for fitting the model. Degree 2 for DIU was chosen to avoid the loss of statistical power when modeling genes that express several isoforms caused by the increase of explanatory variables. An FDR of 0.05 was set as threshold to call differential genes. DCU and DIU were run after applying filtering steps to discard isoforms with a relative gene expression lower than 0.1. Major isoform switches at different differentiation stages were computed by comparing the predominant gene-isoform at each point in relation to the most-expressed isoform in NPCs. In the case of DGE, developmentally regulated genes were grouped by hierarchical clustering using correlation as a measure of distance, and computing the median expression profile to represent the expression behaviour of genes belonging to each defined cluster. The functional relevance of each layer of gene expression regulation was interrogated by using the multidimensional gene-set analysis implemented in mdgsa package [215], conducting the analysis individually for each cell type differentiation system and using ranking statistics derived from both DIU and DE statistical analysis. We selected GO terms with a FDR less than 0.1 at any of the layers of gene-expression regulation and summarized them using the REVIGO tool [309] for further visualization in Cytoscape [295]. The contribution of each transcriptional layer to modulate a given function was visualized by computing their relative significance in logarithmic p-values.

Because two differentiation systems were considered in this study, next we detected genes whose behaviour was different between MN and OPC development. We interrogated cell-type specific gene regulation by comparing cell types at the three different experimental maturation stages. Lineage-specific transcriptionally and post-transcriptionally regulated genes were detected using DIU and DGE pariwise analysis. Stage-specific differential genes were called at a FDR threshold of 0.05. Similarly, multidimensional analysis was run for each stage of differentiation in single time-course analyses to capture the functional relevance of each regulatory mechanism at each developmental stage. GO terms with bi-dimensional enrichment patterns significantly displaced towards post-transcriptional regulation in at least one differentiation stage were defined as functions predominantly regulated by post-transcriptional mechanisms (genes classified into one of these mdgsa groups: block displaced toward quadrant 4 without interaction (q4f), block displaced toward quadrant 4 with interaction (q4i) and block shifted to positive X values (xh)). These set of GO terms were summarized and further visualized by using REVIGO and Cytoscape tools.

6.2.8 Isoform functional diversity in neural differentiation systems

To unravel the extent to which post-transcriptional mechanisms generate functional complexity in neural systems by altering the functional load of gene products, we functionally characterised our NEURALtc transcriptome using the IsoAnnot pipeline implemented in Chapter 4. The coding regions for defined isoforms were predicted using the GeneMarkS-T tool [30], implemented in SQANTI (Chapter 3). To complementary the RBP binding site annotation provided by IsoAnnot, we incorporated Mbnl1-2 CLIP data from brain samples published by Wang et al. 2012 [349]. We subsequently analysed the FD using the quantitative varying mode for non-coding and signal-peptide categories and the positional varying mode for the other feature categories annotated at coding level. Regarding structural characterisation, 3' and 5'-UTR variability was assessed by length comparison (using 100 bp as the cutoff) while APA and CDS diversity were positionally evaluated.

6.2.9 Differential feature inclusion between glial and neuronal differentiation

DFI analysis was used to obtain the set of functional and regulatory features differentially included by post-transcriptional regulation between the OPC and MN differentiation systems. Isoforms belonging to the ISM category as well as non-coding variants (representing the 3.6% and 4.8% of the NEURALtc transcriptome, respectively) were discarded before the DFI analysis. Moreover, we imposed a minimum isoform relative-expression filtering cutoff of 0.2, thus testing features if relative expression of both feature-included and feature-exclusive variants reached 0.2 of the total expression of the gene in at least one condition. We chose the 'feature disruption' DFI analysis mode to discriminate the varying features, which analyses both partially disrupted or completely excluded features (More details in Section 5.2.4.1). The identification of features with differential inclusion dynamics between cell subtypes across development was performed by using pairwise comparisons at each stage of differentiation. For each stage, a feature was considered differentially regulated between cell types if it two criteria were met: $|FDR < 0.05|$ and ΔFI 0.2. We also defined the cell-type in which the complete inclusion of the feature was promoted and characterised major variant switches. Relative over-representation of DFI features in specific annotation categories was statistically evaluated at the element and gene level using FETs, allowing categories to be ranked by their DFI enrichment levels.

6.2.10 Differential polyadenylation in differentiating glial and neuronal cells

Alternative usage of polyA sites in MN and OLG differentation systems was computed using the DPA method described in Chapter 5. An adjusted p-value threshold of 0.05 and ΔDPAU $>=$ 0.2 were set as paramaters required to call DPA genes. We performed hierarchical clustering of DPA genes on DPAU levels using same parameters described in 6.2.7. 3' and 5'-UTR lengthening was computed as detailed in Chapter 5. PCA analysis on mean-centered 3' UTR lengths was performed in order to explore how far UTR lengthening can define cell maturation and identity. Gene loading values at PC differentiating lineages were used to select genes for further 3’ UTR lengthening analysis.

6.3 Results

6.3.1 Widespread novel post-transcriptional diversity in neural systems captured by long-read sequencing

135 PacBio SMRT cells belonging to 15 samples from our multiple time-course experiment were sequenced by PacBio RS-II, yielding a total of 10M reads, 56% of them catalogued as FL. As shown in Figure 6.1, the proportion of FLs per SMRT cell strongly correlated with the fractionation size range and decreased as the molecules become longer. ICE clustering defined 896,972 consensus sequences (IdentityMean = 0.952; Coverage = 0.971) which yielded around 140,000 unique isoforms after running TAMA and Cupcake redundancy collapsing with comparable parameters (0.95 identity and 5,000 bp threshold at the 5' end). A high proportion of isoforms containing NC SJs, which were strongly associated with retrotranscription and sequencing artefacts, were seen both in TAMA and Cupcake QC-derived transcriptomes (Figure 6.1.B). Moreover, both transcriptomes only presented 25% of known isoforms (FSM and ISM, Figure 6.1.C). Following PacBio Iso-seq guidelines, we took advantage of FL isoform coverage information stored by the Cupcake tool and applied a 3 FL-reads filter, which dramatically increased the number of known isoforms (70% vs. 25%, Figure 6.1.C) and the proportion of isoforms without NC SJs (97% vs 70%, Prop Test p.value=0, Figure 6.1.B). Thus, given the comparable performance of TAMA and Cupcake methods, and the additional FL information provided by Cupcake, we defined our isoform set using Cupcake collapsing with a minimum FL coverage of 3, yielding a total of 45,688 non-redundant PacBio-defined isoforms.

Full-lengthness analysis by SQANTI revealed a bimodal distribution of PacBio-defined TTSs, which were clearly distributed between overlapping reference TTSs or fell more than 200 bp apart and which could represent novel polyA sites (Figure 6.2). Conversely, we found a more spread distribution for TSS completeness with a relevant proportion of isoforms falling short and long from the annotated end by 40 to 100 nt (Figure 6.2), a similar pattern to those found in previously analysed PacBio datasets (Chapter 3), which correlates with reduced 5' end completeness control during RNA retrotranscription.

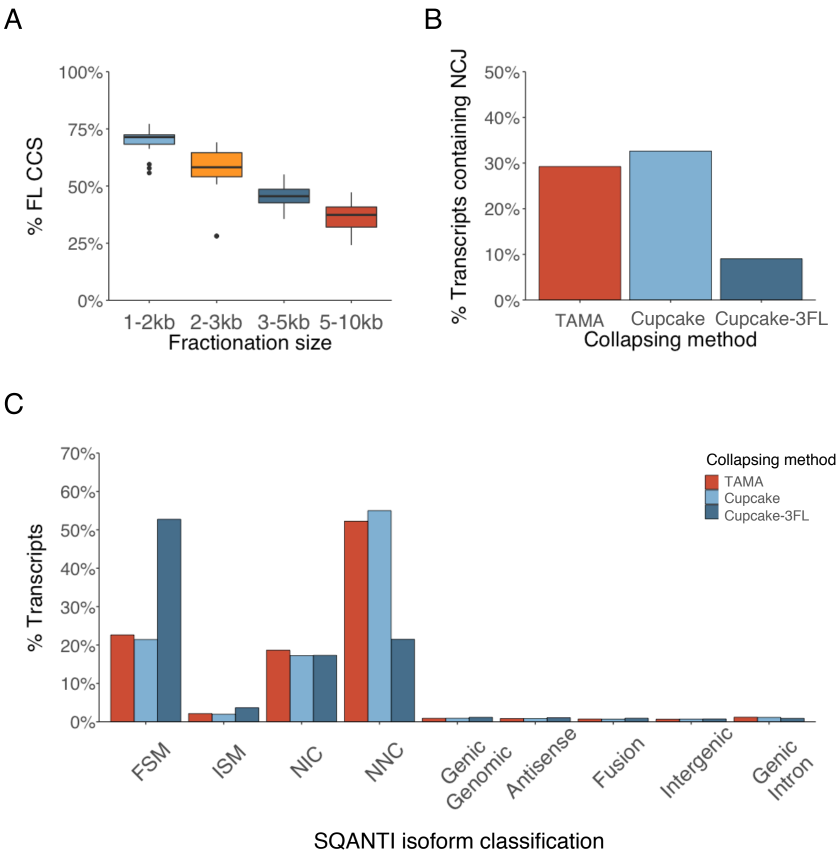


Figure 6.1: Neural transcriptome definition by Iso-Seq sequencing. A. Percentage of FL sequences as a function of the molecule size-fraction. B. Percentage of isoforms containing non-canonical splice junctions (NCJ) for each collapsing strategy applied. C. Isoform distribution into SQANTI categories for each collapsing method.

Furthermore, focusing on the three most populated SQANTI isoform categories, QC revealed the accumulation of low quality attributes in NNC, as expected (Figure 6.2.B). Subsequent SQANTI filtering analysis highlighted bite and the SJ coverage attributes as the most relevant variables for ML classification of artefactual isoforms (Figure 6.3.A). Of note, the FL coverage variable became less important than in our previous PacBio-defined transcriptomes (Chapter 3), likely because its capacity to discriminate artefacts is reduced when all the isoforms have a minimum coverage of 3 FLs. A total number of 12,902 isoforms were discarded (28% of the total), reducing the presence of poor-quality characteristics in the resulting curated transcriptome, which contained 32,787 isoforms (Figure 6.3.B), 80% of them already found in references (FSMs and ISMs) (Figure 6.3.C).

Figure 6.1 revealed that, long molecules are less likely to be fully sequenced. Length comparison between isoforms sequenced by PacBio (32,331 transcripts) and detected by quantification of the RefSeq transcriptome (24,538 transcripts) showed that the transcripts exceeding 6,000 bp were under-represented in the PacBio set. This suggests that an important proportion of long transcripts may be missing in our PacBio-defined transcriptome. To recover long transcripts hidden by PacBio from reference annotations without introducting isoforms mistakenly quantified, we evaluated transcript reliability based on the support level of the isoform containing SJs.

Figure 6.4.B shows that RefSeq-specific transcript recovery was maximal (11,220 transcripts) when we decreased short-read support levels to 0 (mean-

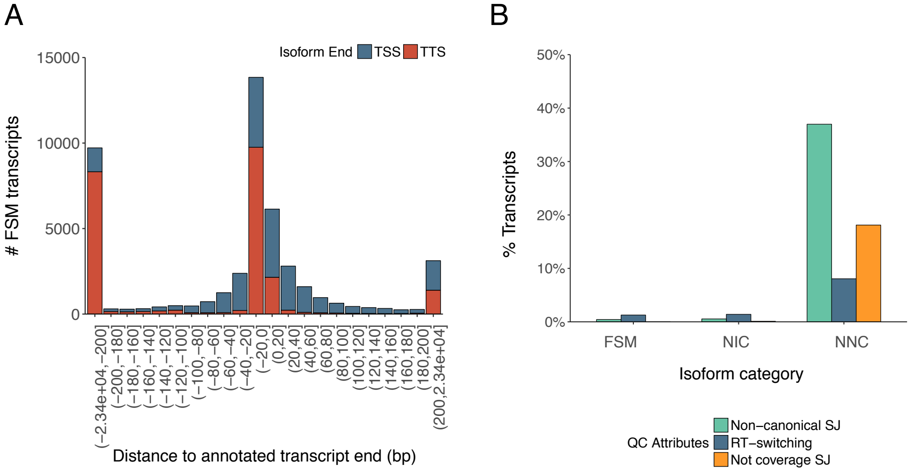


Figure 6.2: Quality control of our NEURALtc transcriptome by SQANTI A. Isoform full-lengthness results provided by SQANTI. Figure modified from QC SQANTI report. Negative values indicate there were insufficient nucleotides to reach the reference transcript end (the sequenced TSS is located downstream of the annotated one or the sequenced TTS is located upstream of the annotated one). B. SQANTI evaluation of quality attributes across the three most populated isoform categories.

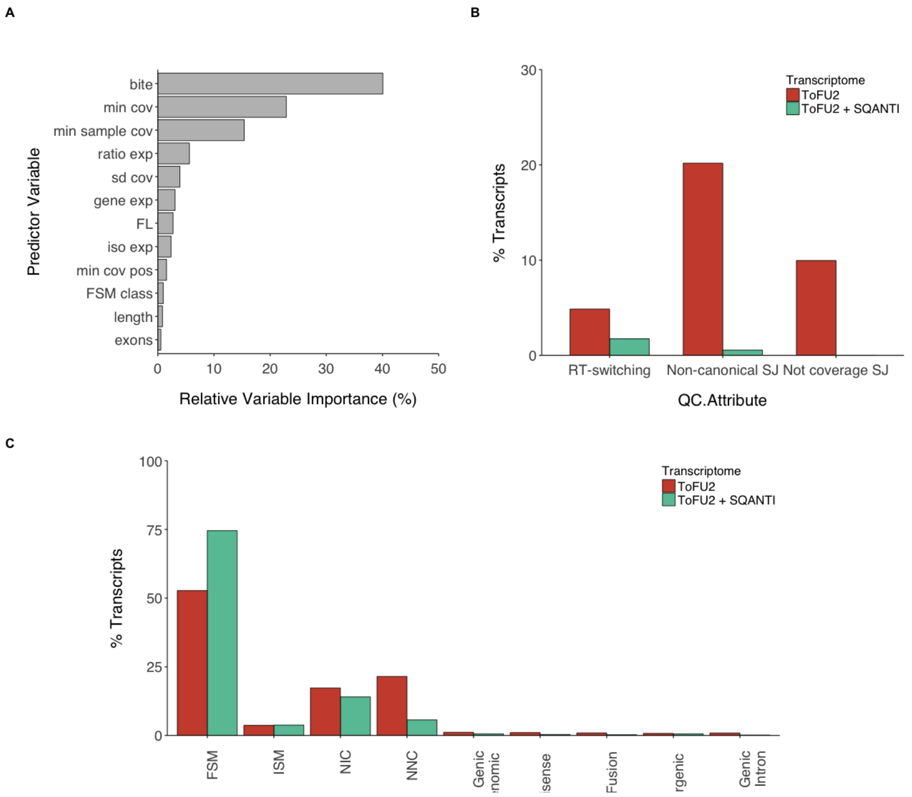


Figure 6.3: Control of false-positive isoforms in our neural transcriptome by the SQANTI filter. A. Ranking of the variables used by SQANTI according to their relative importance for the ML classifier. B. Evaluation of quality attributes by SQANTI in the NNC category before and after applying the SQANTI filter. C. Isoform distribution into SQANTI categories before and after the use of the SQANTI filter.

ing that at least one junction along the transcript had 0 support). To avoid the introduction of this low-expressed or mistakenly detected RefSeq isoforms, the minimum number of short-reads was set to 5 across all the junctions in each isoform, and at least 3 samples had to meet this requirement for the RefSeq isoform to be rescued. Thus, 1,773 transcripts enriched in long transcripts (Figure 6.4.A) and showing SJ coverage and sample-support levels like the isoforms exclusively captured by PacBio were finally added to our long-read defined transcriptome (6.4.C-D).

Hence, our curated NEURALtc comprises on 34,304 isoforms with a percentage of novelty near 20% and 12,563 genes, 99% of them already annotated in reference databases. Regarding isoform complexity, 64% of neural genes expressed alternative variants (Figure 6.5), revealing the high level of transcriptome complexity captured by deep PacBio sequencing. The most common mechanism generating this isoform variability was AS which affected the 49% of total genes, 40% of them expressing more than two splicing variants. In contrast, the ATI site mechanism affected the lowest proportion of genes (24%)

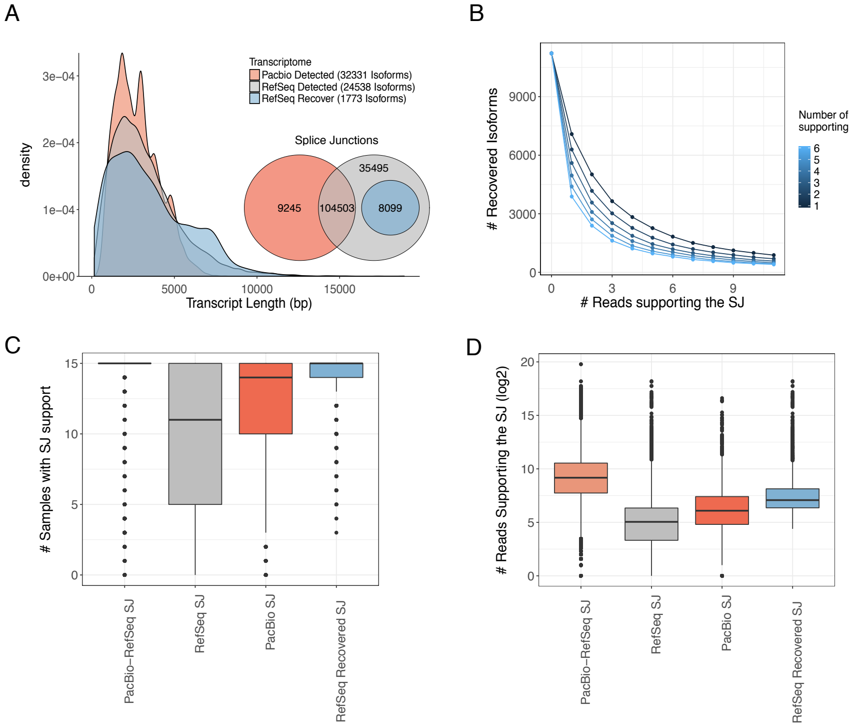


Figure 6.4: Rescue of long isoforms from reference transcriptomes. A. Transcript length distribution and SJ Venn diagram for the Iso-Seq transcriptome, RefSeq detected isoforms, and the set of rescued transcripts. B. Distribution of the number of isoforms recovered from RefSeq as a function of the SJ support level and the number of samples meeting SJ-support requirements. C. Distribution of the samples with SJ-support across the Iso-Seq and the RefSeq SJ subsets. D. Distribution of reads supporting SJs across Iso-Seq and RefSeq splice-junction subsets. Definitions are as follows, SJ: splice-junction, PacBio-RefSeq SJ: set of SJs detected in both long-read sequencing and RefSeq quantification. RefSeq SJ: set of SJ detected only by RefSeq quantification. PacBio SJ: set of splice junctions only detected by long-read sequencing. RefSeq Recovered: set of RefSeq SJs contained in rescued transcripts (rescue thresholds were: coverage of 5 short-reads across all isoform junctions in at least 3 different samples).

and alternative sites per gene (only 1% of ATI genes use more than two alternative sites). Finally, APA was a very common mechanism (40% of genes had APA sites) but very few expressed more than two polyA sites (15%), suggesting dual polyA site use is the norm in neural system APA regulation (Figure 6.5).

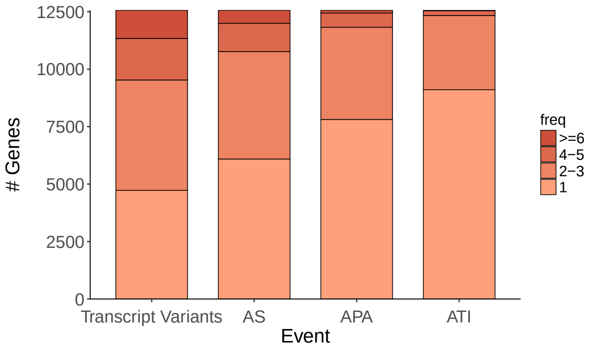


Figure 6.5: Overall distribution of the number of alternative isoforms defined per gene in our neural transcriptome and the level of transcriptome complexity generated by the different mechanisms of transcript processing. AS: alternative splicing. APA: alternative polyadenylation. ATI: alternative transcription initiation.

6.3.2 System characterisation defines oligodendrocyte and motor neuron progenitors as the most mature differentiation stages

Quantififacion of each sample involved in our neural time-course design (Figure 6.6.A) and subsequent PCA showed that PC1, the component accumulating the highest proportion of variability (40 %), was associated with cell maturation levels, while PC2, which explained 26% of variance, discriminated neural identity (Figure 6.6.B). Surprisingly, while MN differentiation pattern suggests that differentiation into the mature state is gradual, NPCs differentiated into OLG appeared to undergo an earlier switching which imprinted the OLG fate in transcriptome characteristics. PCA analysis for OLG samples alone clearly separated early from late states although highlight the similarity between the last two OLG differentiation time points. For example, *Pdgfra*, a cell surface tyrosine kinase receptor gene [364][17][185], as well as other reported OPC markers such as

Cspg4 [256][185] or *Cd9* [315][120] were continuously upregulated throughout differentiation, indicating that OLG differentiation gradually progresses (Figure 6.7.A). However, no expression was found for pre-myelinating and mature OLG markers, indicating that our OLG differentiation protocol produced OPCs.

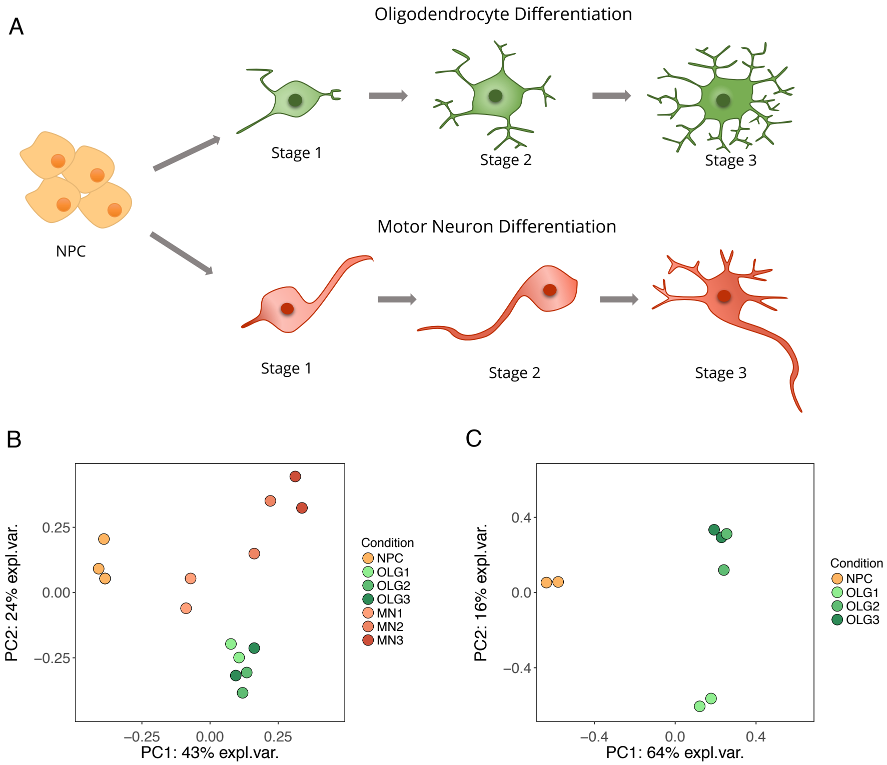


Figure 6.6: Lineage characterization. A. Scheme of our multiple time-course experimental design. Neural precursor cells (NPCs) were put, in parallel, into oligodendrocyte and motor neuron differentiation conditions. B. Principal Component Analysis (PCA) for the NEURALtc expression levels. C. PCA considering only conditions from the oligodendrocyte differentiation time-course series.

In agreement with previously published data [363][34], [73], our data showed that our differentiation of NPCs into MNs activates expression of homeodomain (HD) genes including *Pax6* (Figure 6.7.B). HD proteins are key factors in the repression of MN inhibitors as well as in the activation of *Olig2* gene expression. In accordance with the *Sox21* upregulation we observed in our system, in the presence of retinoic acid (RA) but in the absence of Notch signalling, Olig2-expressing cells express *Sox21* to exit the cell cycle and differentiate into MNs [34] (Figure 6.7.B). Moreover, in agreement with the *Olig2* expression pattern we saw in our system, further MN differentiation is associated with the down-regulation of this gene (Figure 6.7.B). The expression of MN progenitor markers precedes the induction of terminal MN markers such as *Lim3*, *Hb9*, and *Isl1* [73], which were absent in our system, therefore indicating that MN progenitors were the most mature point of our differentiation system.

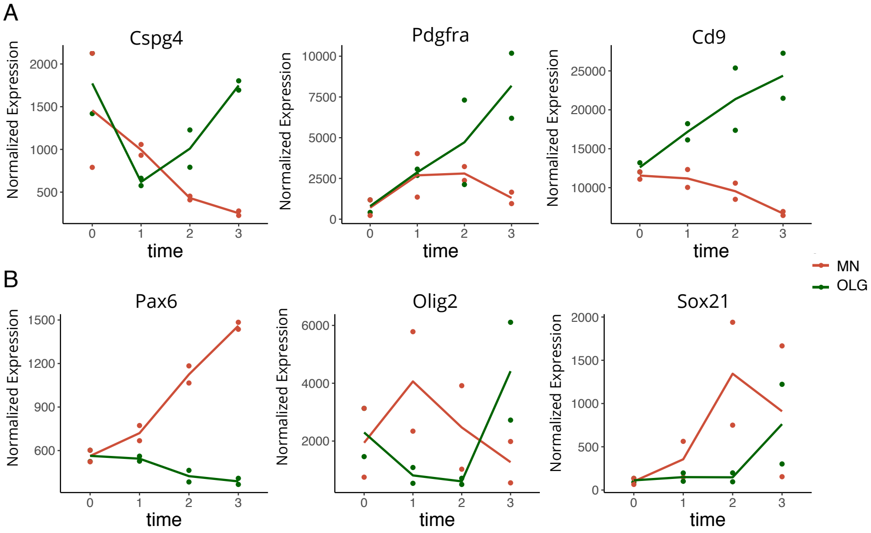


Figure 6.7: Gene expression profiles for OLG (A) and MN (B) differentiation markers.

6.3.3 Novel isoforms are not prevalent but define cell-lineage and differentiation dynamics

The transcriptome generated by PacBio Iso-seq sequencing contained around 20% novel isoforms with about 54% of these coding for novel protein isoforms. Full-length Iso-Seq read-coverage across biological replicates revealed that known isoforms are more frequently captured by PacBio in biological replicates than novel isoforms (68% vs 90% of isoforms detected in both replicates for novel and known isoforms, respectively, prop test p = 1.35e-210, Figure 6.8.A). In agreement, the distribution of novel isoforms was biased towards lower usage levels when compared to already-annotated isoforms, considering both transcript and protein gene products (WRS, p = 0 for both CDS and transcript-level comparisons).

However, categorisation of isoforms into principal and minor gene isoforms revealed that, even though novel isoforms were over-represented among minor isoforms (prop test, $p < 2.2$ 1016, 6.8.B), a third of them were catalogued as PIs. Taken together, the frequently lower expression levels of novel isoforms but the significant proportion of them catalogued as the major isoform in one specific condition indicate the potential stage or cell specific role of novel isoforms in our neural system. Indeed, PCA showed that novel isoforms can separate samples according to their developmental stage and cell type (Figure 6.8.C-D). The loading distribution for PC1 and PC2 verified the high number of isoforms contributing to the definition of PCs. Similar results were obtained for PCA of novel protein isoforms rather than novel transcript isoforms.

To understand the biological role of these novel calls, we analysed the cellular processes they are involved in and found functional enrichment of genes containing novel isoforms in metabolic processes, mitochondria membranes, regulation of neurogenesis and oligodendroglial lineage regulation and mRNA processing and splicing (FET, $p < 0.001$, Figure 6.8. E). Moreover, 18% of total novel isoforms were annotated as undergoing degradation by the NMD surveillance mechanism (representing 75% of the total NMD isoforms, 1313 NMD-predicted isoforms) making novel isoforms highly enriched in NMD targeting (FET p = 3.87e-254). Functional enrichment revealed that the splicing machinery itself, serine-arginine (SR) splicing factors or RBPs were the most over-represented in NMD targets.

Finally, structural characterisation of alternatively-spliced isoforms revealed that, they are evenly distributed across splice junction types: alternative 3' splice site [A3] = 1,358; alternative 5' splice site [A5] = 1,193; alternative first exon [AF] = 1,449; intron retention [IR] = 1,733; exon skipping [ES] = 1,971. However, 80% of all IR events were contained in novel isoforms (Figure 6.8. D) and 65%

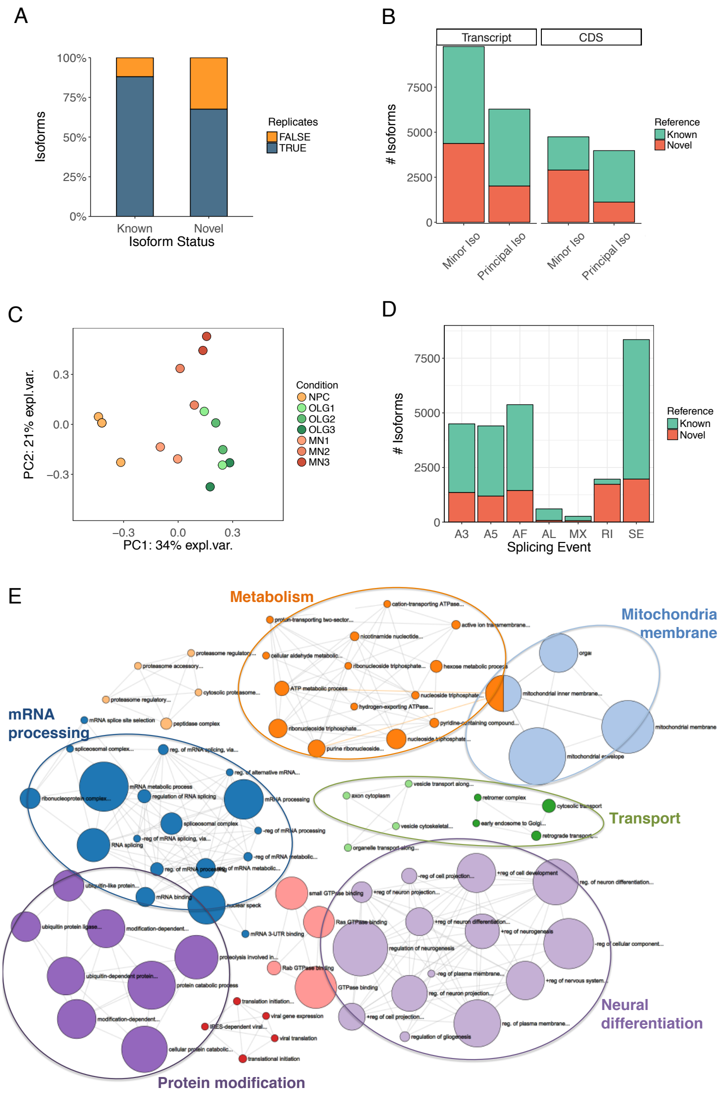


Figure 6.8: Functional Relevance of Iso-Seq defined novel isoforms. A. Novel isoform replicability defined for isoforms found in at least two replicates of a given sample. B. Relative contribution of novel isoforms to gene expression. Isoforms with the highest relative gene expression levels in any condition were defined as 'principal' (PI), otherwise, they are catalogued as 'minor' (MI). C. Principal component analysis on novel isoform expression levels. D. Characterisation of splicing events across novel and known isoforms. E. Network of GO terms enriched in genes containing PacBio-defined novel isoforms ($p < 0.001$) obtained via tappAS software. The size of nodes represents the number of genes while the colour represents the cluster they belong to. A3: Alternative 3' splice site; A5: alternative 5' splice site; AF: alternative first exon; IR: intron retention; ES: exon skipping.

of predicted novel NMD targets contained IR events (FET p = 0), indicating that mRNA degradation by NMD usually results from the introduction of premature termination codons via IR events, both of them highly enriched in the novel isoforms detected by PacBio sequencing,.

6.3.4 Membrane trafficking among the strongest processes specifically regulated by alternative isoform usage in neural determination

Differential gene-expression analysis revealed the presence of an extensive transcriptional regulatory program which affects 25% of genes in MN differentiation and 33% in OLG differentiation (Figure 6.9.A-B). Expression-profile clustering for DE genes highlighted the transcriptional activation of two gene sets in both differentiation courses (red and yellow clusters, Figure 6.9.A-B). This activation was particularly strong in OLG differentiation, where the median expression profile for the most populated cluster 3 reached the highest expression levels at stage 2. Functional analysis of these set of activated genes in OLGs revealed a high metabolic activity (metabolism FDR=6.76e-14) essential to prepare cells for synthesizing, sorting and trafficking high amounts of myelin sheath proteins (myelin sheath FDR = 8.18e-03, integral components of membrane FDR = 2.66e-08). This translates into high mitochondrial respiration (mitochondrion FDR = 4.36e-08; mitochondrial inner membrane FDR = 2.83e-06; respiratory chain FDR = 1.40e-03, oxidation-reduction process FDR = 1.35e-03), endoplasmic reticulum (ER) activity (ER membrane FDR = 3.16e-02), and membrane trafficking (extracellular exosome FDR = 1.993139e-18, transport vesicle membrane FDR = 4.86e-02 and trans-Golgi network FDR = 2.52e-02). Moreover, gene expression in OLGs seems to flatten out during the last stage of differentiation (Figure 6.9.B), in accordance with our previous exploratory results.

Analysis of post-transcriptional regulation indicated that around 14% of genes expressing multiple transcripts show DIU in both differentiation systems, (Figure 6.9.C-D). Interestingly, more that half of them also regulate their CDS usage (56% and 52% for MN and OLG differentiation, respectively), which highlights

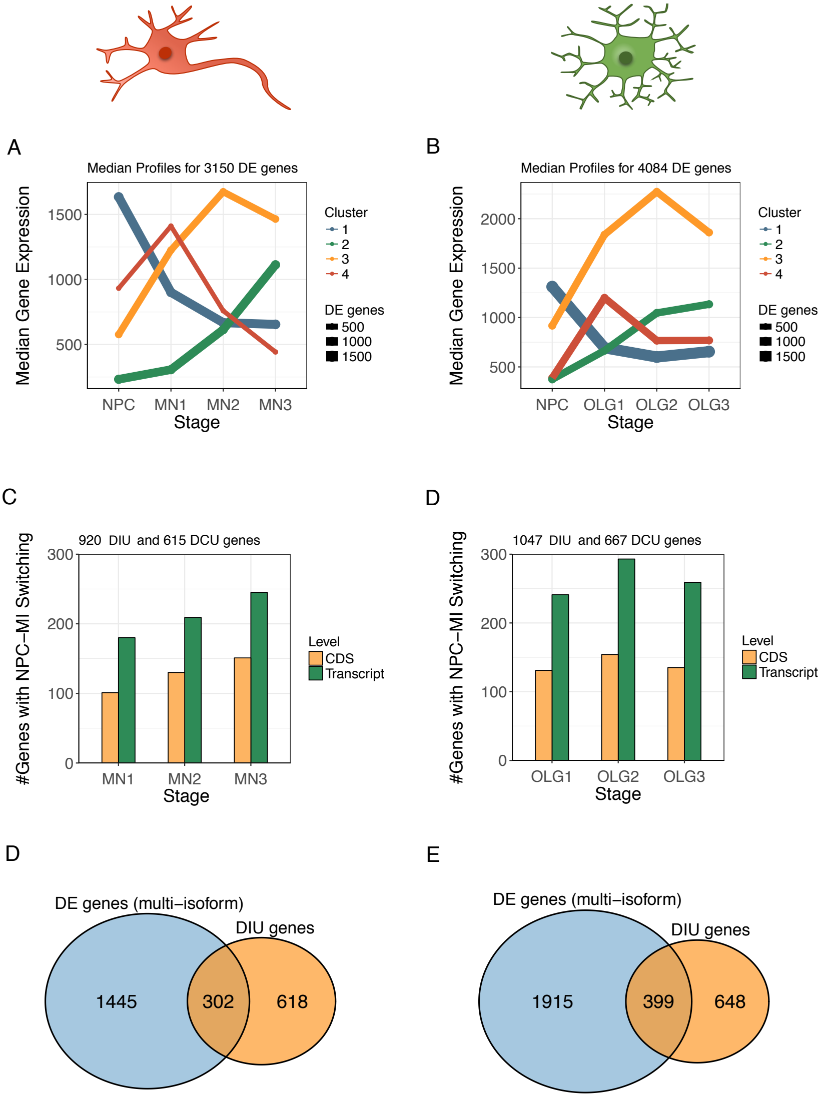


Figure 6.9: Transcriptional dynamics in the glial and neuronal differentiation systems. A-B. Differentially expressed genes during MN and OLG differentiation, clustered, respectively, into groups based on their temporal patterns of expression. Expression profiles are summarised by the median. C-D. Differential isoform usage during MN and OLG differentiation, respectively, computed at the transcript and coding level, and the number of genes undergoing major isoform switches during differentiation stages relative to NPCs. D-E. Venn diagram showing the overlap between differentially expressed genes expressing multiple isoforms and genes with differential isoform usage.

the relevance of post-transcriptional mechanisms in regulating the protein content of our neural system. Furthermore, one third of DIU genes also underwent a switch, with respect to NPC, of their predominant isoform (27% and 35% for MN and OLG differentiation, respectively, Figure 6.9 C-D), with most of present at the earliest differentiation stage (54% and 66% for MN and OLG differentiation, respectively).

Comparison of genes regulated by the two transcriptional programs revealed that around 65% of genes with DIU were not transcriptionally modulated (Figure 6.9.E-F), which indicates that the activity of these two gene-expression regulatory programs could be considered as independent. MN development (see methods section 6.2.7) identified functions including protein transport, protein binding, RNA processing and components of the endosomal system as being generally more associated to post-transcriptional regulation (Figure 6.10.A). For example, components from the endosome compartment, essential for the spatial trafficking of extracellular cues across complex dendritic and extensive axons and the recycling of synaptic vesicles [67] [285] [379], showed a set of regulated genes whose is displaced towards DIU regulation (Figure 6.10.A). Similar functions with over-represented isoform usage regulation were also found during OLG development (Figure 6.10.B) together with additional ones, such as the regulation of mitochondrial dynamics (Figure 6.10.B) which suggests that mechanisms such as AS and APA might be essential for the bioenergetic adaptation of OLGs to their extremely high metabolic rates.

To elucidate the transcriptional program and associated functional changes dictating the fate determination of NPCs into glial cells and neurons, we compared them across the three maturation stages considered. In total, 1,122 genes were detected as undergoing DIU and 6,952 as DE, with a progressive cell-type specialisation reaching 10% of genes with DIU and 35% with DE at the most maturated cell time point (Figure 6.11.A-B). Surprisingly, unlike DGE, post-transcriptional switches were temporary, with almost 80% of them associated with a specific differentiation stage (80% vs. 44% stage-regulated genes, Figure 6.11.C-D), suggesting that isoform regulation is highly dynamic.

The functional analysis of genes regulated by both mechanisms revealed four main functional blocks displaced toward post-transcriptional regulation. The first one is membrane trafficking and includes endosome components such as the SNARE complex and associated proteins, which play a critical role in vesicle docking and coupled exo- and endocytosis [267][352], master regulators of the direction and specificity of endosomal trafficking such as GTPases [193] and protein ubiquitination, which provides sorting signals for plasma membrane internalisation [254]. These results suggest that vesicle exocytosis and endocytosis are tightly regulated between neural cells types by differential isoform

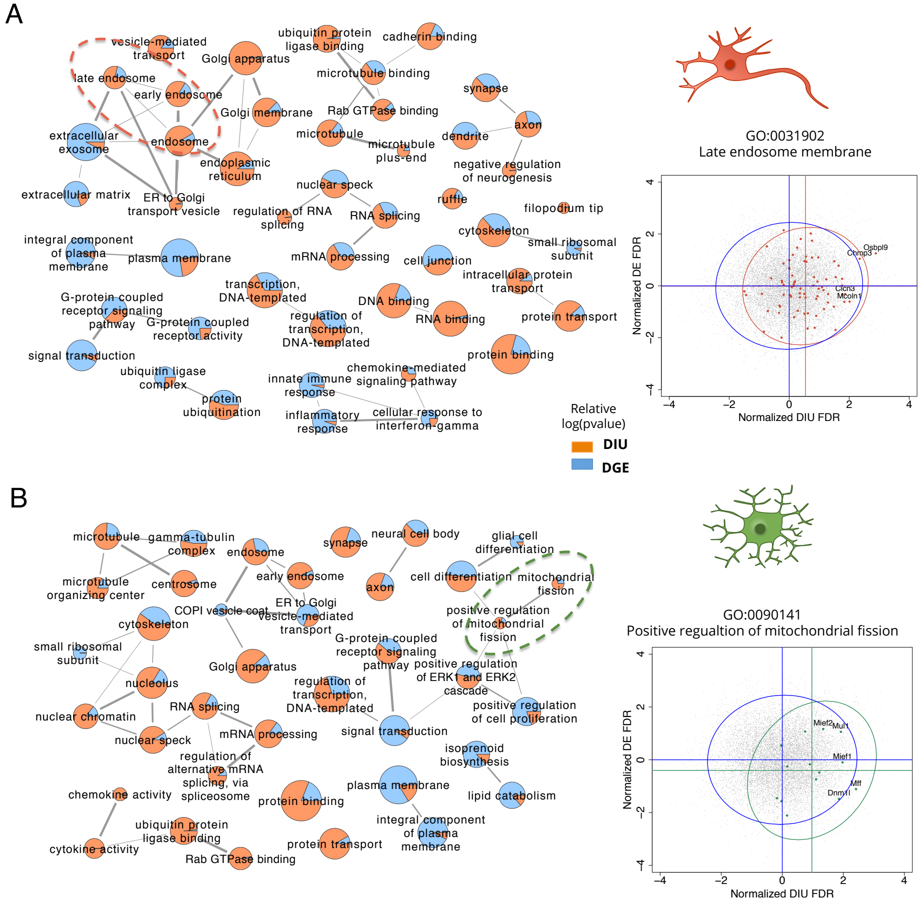


Figure 6.10: Relative functional relevance between differential gene expression and differential isoform usage by multidimensional gene-set analysis, for (A) OLG and (B) MN differentiation. Nodes are represented as pie charts using the relative significance level between both regulatory programs. The ’late endosome membrane’ and ’regulation of mitochondrial fission’ are examples of functional terms more strongly regulated by DIU than by DE.

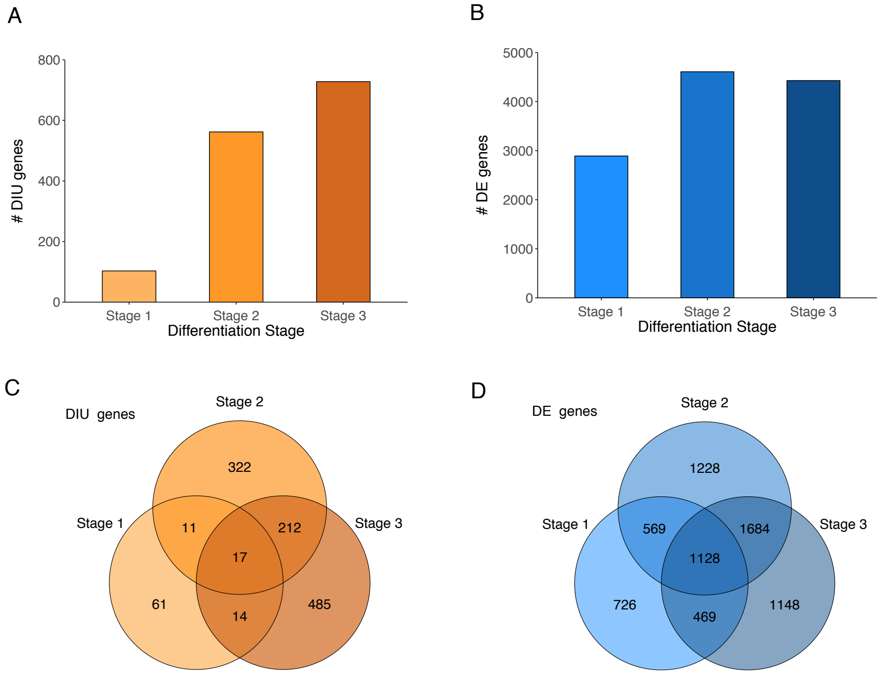


Figure 6.11: Gene expression regulation in glial and neuronal cell types through development in our differentiation system. A. Number of genes with differential isoform usage. B. Number of differentially expressed genes. C. Venn diagram showing the overlap of genes with differential isoform usage genes at different differentiation stages. D. Venn diagram showing the overlap of DE genes at different differentiation stages.

usage to provide trafficking specialisation to neuronal lineages.

Moreover, genes involved in protein transport and components localised at the synapse were also strongly regulated by DIU at differentiation stages 2 and 3 (Figure 6.12.A-B-C). An example is *Cadm1*, a synaptic cell adhesion molecule involved in synapse assembly and axon growth , or *Dtnb*, a poorly-characterised membrane protein component of the dystrophin-associated complex, whose disruption has been associated with various forms of muscular dystrophy (Figure 6.12.B-C). In both cases, several isoforms combining alternative ES events were defined by long-read sequencing with an isoform expression redistribution higher than 40% between cell types (Figure 6.12.D-E) but without altering overall gene expression levels. Thus, cell-type specific transcript variants appeared

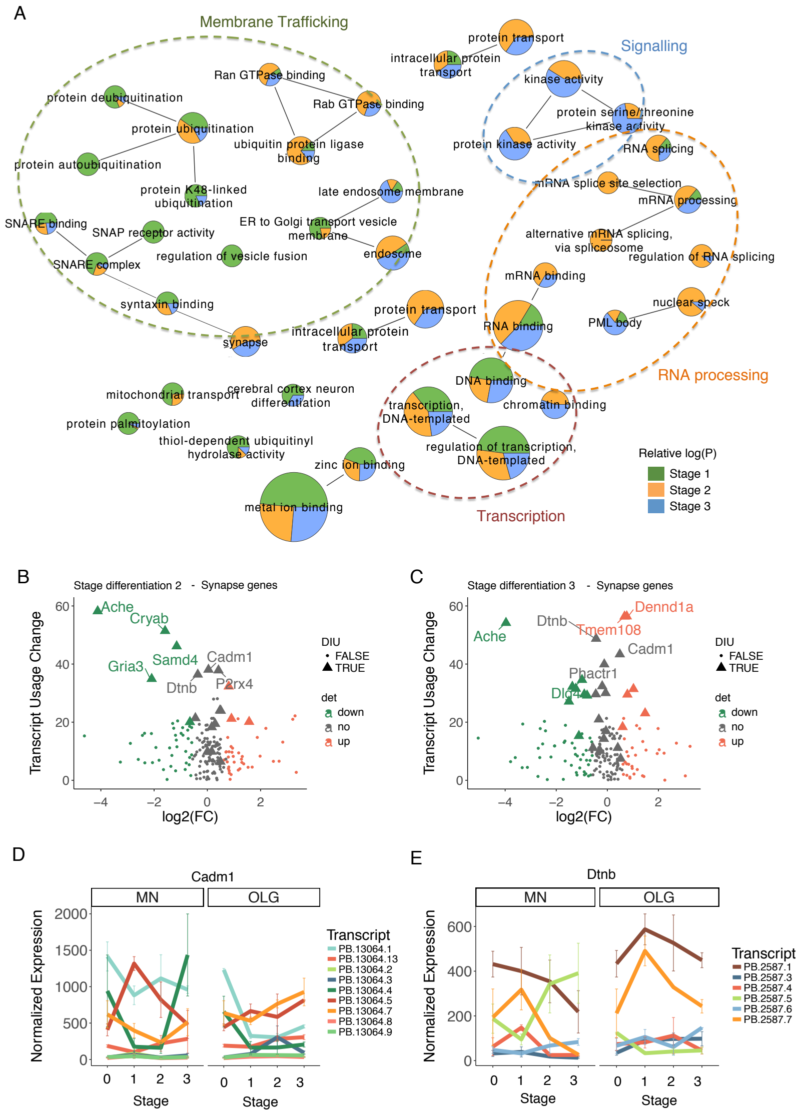


Figure 6.12: Functions significantly over-represented in DIU vs DGE. A. Network of GO terms significantly over-represented in DIU vs DGE. Nodes are represented by pie charts using the logarithmic p-value for DIU enrichment across developmental stages. Node size represents the number of total DIU genes detected with a FDR<0.05. B-C. Proportion of expression redistributed across isoforms (Usage Change) in function of the logarithmic gene expression fold-changes for genes annotated with synapse category at stage 2 and 3. Big triangles represent genes detected as DIU while green and red points represent DE genes. D-E. Transcript level expression dynamics for two synaptic genes, *Cadm1* and *Dtnb*.

such as PB.2587.6 for *Dtnb*, which is only expressed in the late stages of MN differentiation (Figure 6.12.E). In fact, at stage 3, 56% of synapse-associated regulated genes expressed cell-type specific major isoforms.

Apart from protein trafficking and RNA metabolism, DNA binding and transcription processes, there was also significant post-transcriptional regulation of cell signalling at the latest stage, including a total of 65 proteins with kinase activity, some of them strongly associated with brain development such as Pak3 (which plays roles in dendrite spine morphogenesis and synapse formation and plasticity) and Pak1 (found at synapses and involved in vesicle transport and endocytosis). This indicates again potential regulation of signalling affecting endosomal and lysosomal membrane trafficking, likely because they require the coordination of multiple signalling events to control cargo sorting and processing, and endosome maturation.

6.3.5 Functional impact of differential isoform usage on neural fate determination

Structural diversity characterization of our NEURALtc transcriptome revealed that the contribution of UTR and CDS variability to transcriptome complexity was similar (Figure 6.13). Moreover, 66% and 58% of genes with CDS variants also underwent 5' or 3'-UTR length modulation, respectively, indicating the existence of a high level of co-regulation between coding and non-coding regions (co-regulation CDS-3'UTR p = 9.10e-222, CDS-5'UTR p = 0).

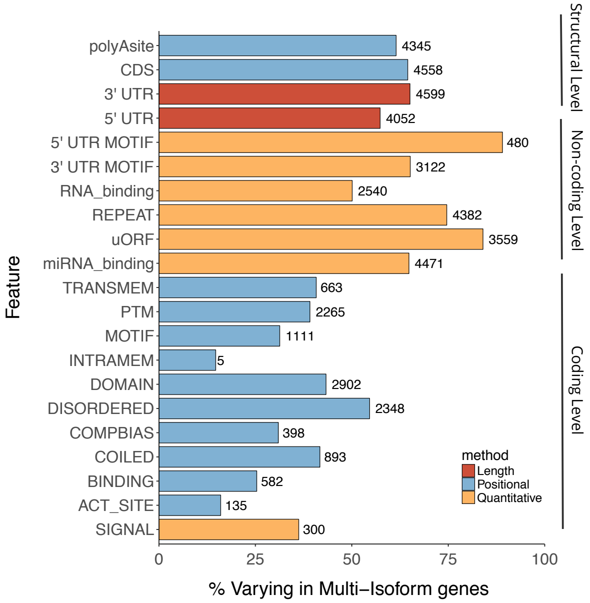


Figure 6.13: Functional diversity in our NEURALtc transcriptome at the structural, coding, and non-coding levels for genes expressing multiple isoforms. Bars are coloured based on the method used to assess isoform diversity.

In terms of features load, 5' UTR elements such as uORFs and 5'-UTR motifs showed the highest variability rates (Figure 6.13). Further analysis of this uORF diversity revealed that 60% of genes with varying uORFs underwent ATI site usage, and 40% underwent AS which affected the definition of the 5' UTR. Even though categories associated with coding features are usually affected less than non-coding ones (Figure 6.13), more than 60% of genes expressing alternative isoforms had at least one coding-feature variation. This rate increases up to

95% when only considering genes expressing alternative predicted CDSs, revealing that the coding diversity in our neural transcriptome generally alters feature content. Among them, IDR elements were the most frequently impacted by post-transcriptional processes (54%, figure 6.13), while domains were disrupted in the highest number of genes (2,902 genes, Figure 6.13).

From the 110,262 isoform-specific functional features, DFI analysis indicated that 4,821 features (5%) in 558 genes have significant differential inclusion levels in the neural subtypes in at least one developmental stage. The number of differentially included features increased through the maturation process (Figure 6.14.A), and a high proportion of them were stage-specific (Figure 6.14.B) which, in agreement with previous DIU results, suggests that post-transcriptional regulation is precisely timed during fate determination. Moreover, IDRs and uORFs were systematically enriched among DFIs during all differentiation stages (Figure 6.14.C). However, other feature categories appeared unevenly impacted through development (Figure 6.14.C). At the earliest stage, PTM site regulation between neural cell subtypes was enriched among DFI (PTM stage 1, FET FDR = 2.03e3; 6.14.C), while differential inclusion of miRNA binding sites and 3'-UTR motifs was over-represented at stage 2 (miRNA binding sites stage 2, FDR = 2.10e39). Finally, RBP binding site regulation increased as the differentiation progresses, and peaks at the final stage of differentiation (Figure 6.14.C). When gene-feature redundancy was removed, leaving only one feature-type per gene, the stage-specific relative over-representation results were similar to the ones found at the feature level (Figure 6.14.D) except for the RBP binding site category, where its lower enrichment levels at gene level are likely due to the accumulation of multiple binding sites in individual genes (similar to the pattern described in Chapter 5).

Impact on functional protein loading

In our detailed analysis, first we focused on feature types annotated at the coding level (Figure 6.15.A). As previously shown, IDRs were significantly over-represented in differential inclusion in all the developmental stages with a total

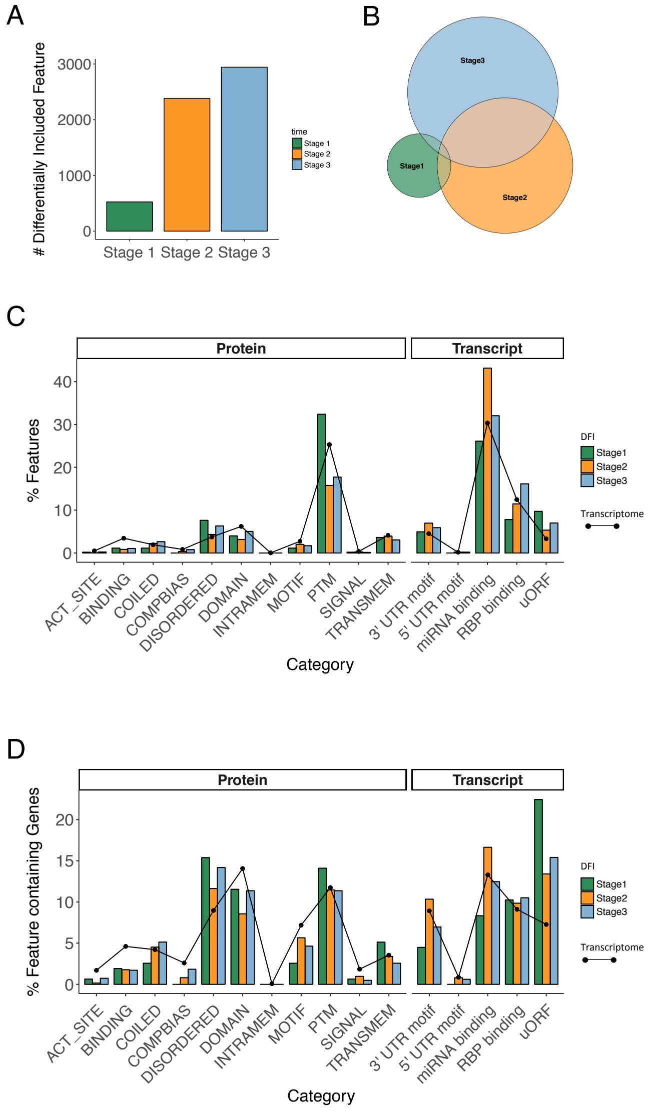


Figure 6.14: Differential feature inclusion between the motor neuron and oligodendrocyte differentiation. A. Number of total features differentially included in each developmental stage. B. Venn Diagram of features regulated in each developmental stage. C. Proportion of total features annotated in each category (black line) relative to the proportion of differentially included features, calculated independently for each considered stage. Statistical analysis of the relative over-representation of DFI features in specific categories relative to the transcriptome distribution was performed using Fisher exact tests with Benjamini-Hochberg multiple-testing correction. D. Feature distribution after removing multiple-feature redundancy from the same category in each gene.

of 172 affected genes (representing around 20% of the total DIU genes, Figure 6.15.A). IDRs were found enriched in protein binding genes with 37% of the genes associated with protein binding properties (FET p = 1.73e-5, FDR = 0.04). Interestingly, the proportion of them undergoing a major inclusion switch between neural cell sub-types increased during differentiation, reaching 62% of total events at the last differentiation stage (Figure 6.15.B). This indicates that IDRs with potential protein binding activity are not only subjected to subtly inclusion modulations between cell types but are predominantly cell-type specific, mainly at late differentiation stages.

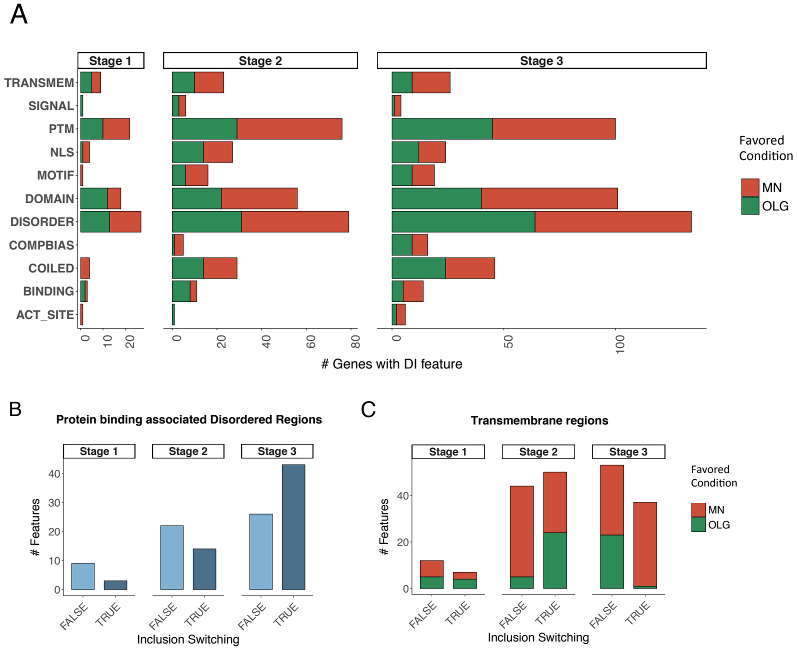


Figure 6.15: Genes with differential inclusion of functional features. A. Distribution of DFI features across protein-feature categories. Bars colours are based on the cell type in which the inclusion of the feature is favored. B. Distribution of differentially included IDRs as a function of their inclusion switching between cell types. C. Distribution of differentially included transmembrane regions as a function of their inclusion switching between cell types. Bars colours are based on the condition in which the inclusion of the transmembrane region is favored.

In addition to protein-binding functions, cell junction (FET p = 3.2e-5, FDR = 0.03) and actin cytoskeleton (FET p = 8.3e-5, FDR = 0.04) were significantly enriched in this set of cell-type specific IDRs. These included presynaptic membrane and postsynaptic regulatory machinery components including *Dennd1a* and *Gphn*, respectively; adherens junction components such as *Plejga7*, and actin-associated proteins such as *Synopo2* and *Phactr1*. Remarkably, a high proportion of these genes are members of the kinase and phosphatase protein families such as *Tjp2*, *Phactr1*, *Gphn*, *Dlg4*, *Wkn4*, or *Pip5k1c*.

For example, *Pip5k1c*, which encodes a kinase participating in cell adhesion, migration and is concentrated at synases in neural tissues [361], was found containing an IDR with a differential inclusion pattern between our neural subtypes (Figure 6.16.A) which results from the regulation of an splicing event modifying the C-terminal region of the coded protein (Figure 6.16.C-D). IDR-containing isoforms were progressively included at late MN differentiation stages in contrast with OLG differentiation, in which splicing variants encoding the disrupted IDP clearly predominate (Figure 6.16.B). We also identified the co-inclusion in the IDR of a region that interacts with the *Tln2* protein (Region636-661; FDR=4.01e-06), an important component of focal adhesion plaques [259] [106] that is also present at synapses and has been reported to concentrate the long isoform of PIP5k1c in focal adhesion contacts [77]. Thus, inclusion of this binding site in the IDR governs *Pip5k1c* expression in MNP and might encourage its recruitment to focal adhesion plaques and synaptic sites to control plasma-membrane pool of phosphoinositides, which are essential for correct synaptic vesicle endocytosis and exocytosis [236][360]. Moreover, several phosphorilation sites were also differentially excluded between cell types, most of them already involved in modulating PIP5K1c activity [183] [77]. Thus, our analysis suggests that differential regulation of *PIP5K1c* isoforms in neural cell types impacts *PIP5K1c* spatial specificity and activity at synapses, specific of MNs, by altering the availability of functional elements such as IDRs and PTMs without involving changes in absolute gene-expresion levels (DGE stage 3 FDR = 0.27).

Notably, IDR enrichment analysis across the different layers of functional annotation also showed their over-representation in genes with domains of the PH superfamily (CL0266, FDR=0.01, p-value=3.20e-05), present in a wide range of cytoskeleton-associated proteins and involved in intracellular signalling. Examples are *Kif1b*, involved in mitochondria transport, synaptic vesicle and myelin transport and *Mapk8ip1*, a scaffold protein mediating signalling and transcrip-

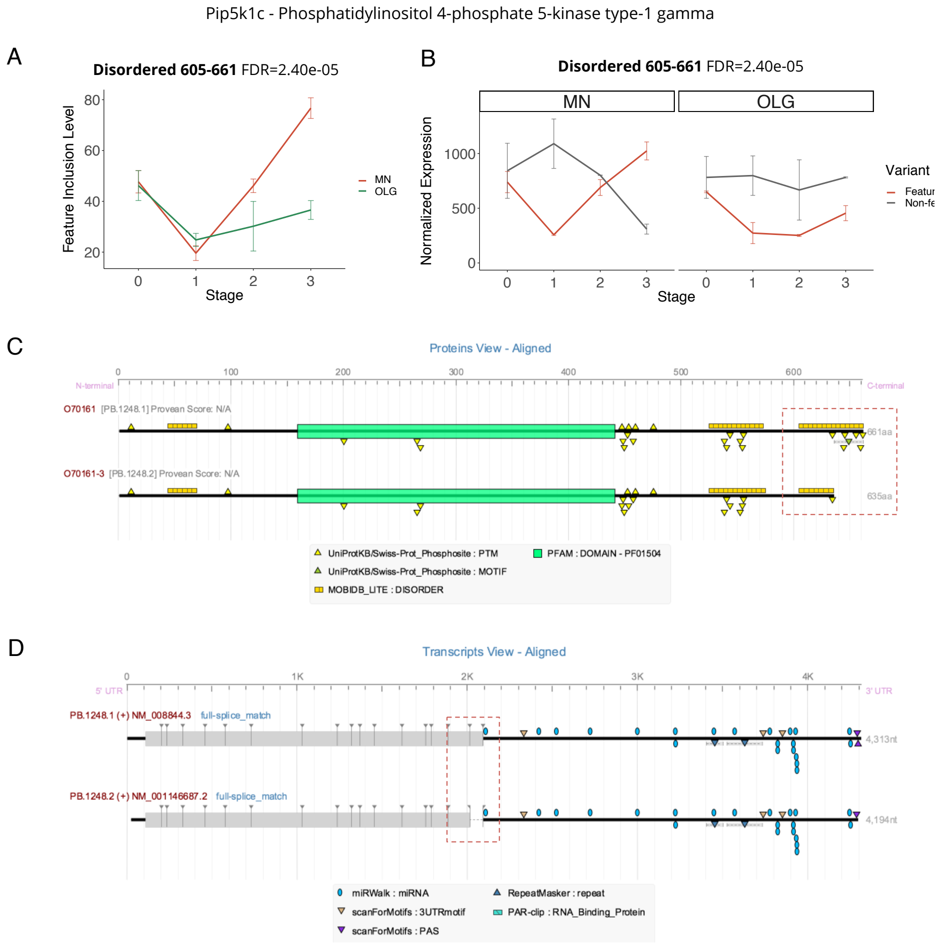


Figure 6.16: Functional impact of isoform regulation in *Pip5k1c*. A. Inclusion profile for the regulated intrinsically disordered region (IDR) in *Pip5k1c* across oligodendrocyte (OLG) and motor neuron (MN) differentiation. B. Cell-type expression profile of the feature-variant associated to the intact and the disrupted IDR. As *Pip5k1c* expresses two isoforms, the expression of these feature-variants correlates with the expression of individual isoforms. C. *Pip5k1c* predicted-proteins visualized by TappAS. The C-terminal variability provokes the disruption of the IDR together with the skipping of several post-translational modification sites and deletion of a *Tln2* binding-motif. D. The disruption of the protein sequence and disruption of funcional features in *Pip5k1c* is caused by an exon-skipping event. Transcript isoforms are also annotated with several functional and regulatory elements within the 3'-UTR region.

tional regulation and also involved in vesicle-mediated transport. *Mapk8ip1* isoforms including the N-terminal IDR dramaticatically accumulate in motor neuron differentiation. In neurons, MAPK8IP1 is known to tightly coordinate motor activity to ensure the fidelity of retrograde autophagosome transport in axons [107], what suggests that the inclusion of IDRs in motor-neurons is providing cell-type specific cargo properties to regulated the specific function of *Mapk8ip1* in neurons.

In contrast to IDRs, Pfam domains were under-represented in differential inclusion through all the differentiation stages. They were regulated in a vast number of genes with significantly more domains becoming distupted or lost in OPC differentiation (Figure 6.15.A, binomial test stage 3 p.value = 1.9e-3). Functional analysis highlighted kinase, ATPase functions, ligand binding motifs and signalling pathways as being enriched in genes in which domains were differentially included at expressed isoforms regulated between cell subtypes. Further functional characterisation of these differential features revealed the Pfam kinase superfamily as the most frequently impacted PFAM clan by post-transcriptional regulation (Figure 6.17.A). Among these are *Clk4* and *Clk1*, dual-specificity protein kinases that were also detected in the previous chapter as accumulating isoforms that loss the two N-terminal NLSs and disrupt a kinase domain through OPC development (Chapter 3). In this more complex differentiation system, *Clk4* expressed 17 alternative isoforms (Figure 6.17.F), 7 of them are PacBio-defined novel variants, each with a different combination of exon skypping events that gave rise to 6 different CDSs which mainly differ at their N-terminal regions (Figure 6.17.E). The inclusion pattern of the kinase domain across cell types show the significant enrichment of isoforms containing the intact domain in OLG differentiation (Figure 6.17.B), a similar pattern to those found for other kinases such as *Dclk1* and *Clk1* (Figure 6.17.C-D). Moreover, the differential isoform usage of *Clk4* between neural subtypes also involves the specific introduction of a NLS during OLG differentiation (Figure 6.17.G). Thus, our results indicate that, during cell fate determination, cell-type specific isoform regulation impacts the activity and nuclear targeting of several kinases, likely as a mechanism to acquire selective functions in nucleus and cytoplasm, as previously seen for other enzymes [303] [88].

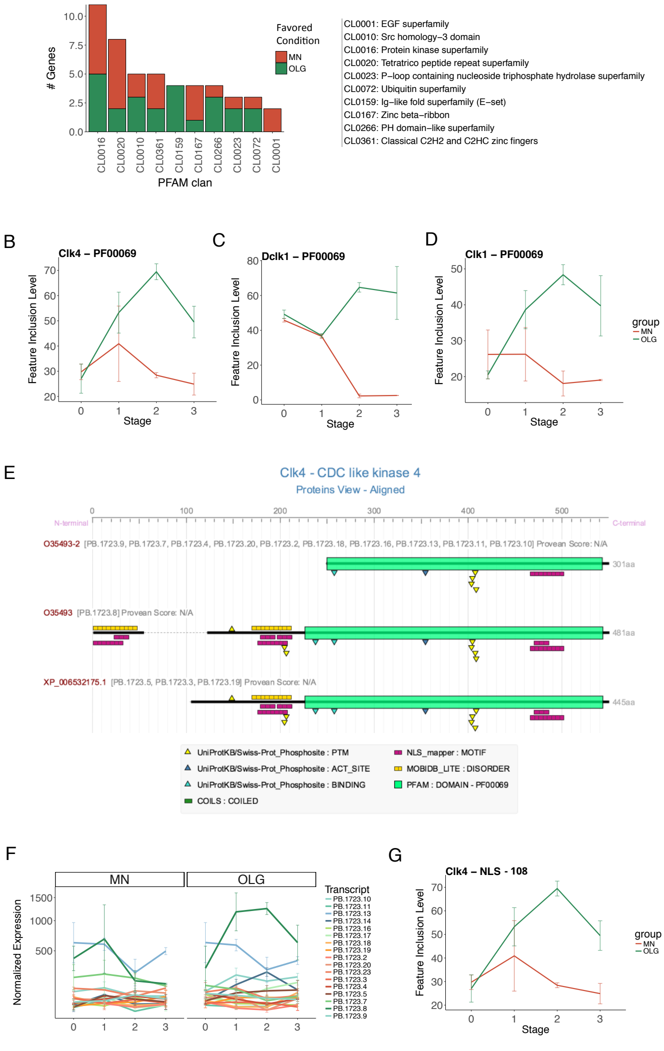


Figure 6.17: Cell type specific inclusion of PFAM domains. A. Distribution of the number of genes regulating the inclusion of PFAM domains between cell types. PFAM family clans are used to collapse domains with related functional role. B-D. Three kinase proteins with domain-inclusion patters favored throughout oligodendrocyte differentiation. E. Clk4 post-transcriptional regulation alters the N-terminal protein region proteins, provoking the disruption of the kinase domain and the skipping of a NLS, two IDRs and some PTMs. F. Expression patterns for the 17 PacBio-defined Clk4 isoforms in both differentiation systems. G. Inclusion pattern for the NLS annotated at position 108.

The transmembrane region (TM) feature-category had the highest bias towards inclusion in a specific cell type (Binomial test, FDR = 1.18e-07), and these were clearly disrupted or lost during OLG differentiation (Figure 6.15.A). This pattern became more accentuated in the last differentiation stage, where almost 100% of the TM switches were preferentially included in MNPs (Figure 6.15.C). Interestingly, we found 2 essential genes in the negative regulation of mitochondrial fusion, *Oma1* [5] and *Mul1* [249]. Both of them underwent a similar pattern of favored transmembrane skipping in OLG differentiation compared to MN development (Figure 6.18). Mitochondrial morphology results from a balance between two processes: fusion and fission. According to cell-context circumstances, the equilibrium between these process may alter which leans the balance towards one of these two processes. Fusion events are controlled by three main proteins: mitofusins 1 and 2 (*Mfn1/Mfn2*) and *Opa1*, mediators of the mitochondrial outer and inner membranes fusion, respectively (Figure 6.18). *Mul1* and *Oma1* are two majors regulators of mitofusins and *Opa1*, respectively, suggesting a potential modulation of mitochondrial dynamics during neural fate determination by post-transcriptional regulation of TM regions (Figure 6.18).

Specifically, the metalloprotease OMA1 controls mitochondrial morphology by proteolytic processing of the dynamin-like GTPase OPA1 protein, localised in the mitochondrial inner membrane [5]. The balance of OPA1 long and short forms governs the delicate balance between fusion and fission (Figure 6.19.B). While OMA1 activation provokes the accumulation of soluble OPA1 forms [15] (Figure 6.19.B), blocking fusion and facilitating mitochondrial fission [302] (Figure 6.19.A), OMA1-deficient cells shift towards fusion, likely because of the decreased proportion of S-OPA1 isoforms [264] (Figure 6.19.A). Our analysis revealed the accumulation of *Oma1* splicing isoforms skipping the transmmebrane region in OPC differentiation compared to the favored transmembrane inclusion observed in MNPs (Figure 6.19.B). This suggests that OMA1 transmembrane anchorage is impaired, likely resulting in OMA1-dependant OPA1 cleavage deficiency. Consequently, the regulation of a transmembrane region by AS in OMA1 might result in a shift in the balance towards fusion during OPC differentiation

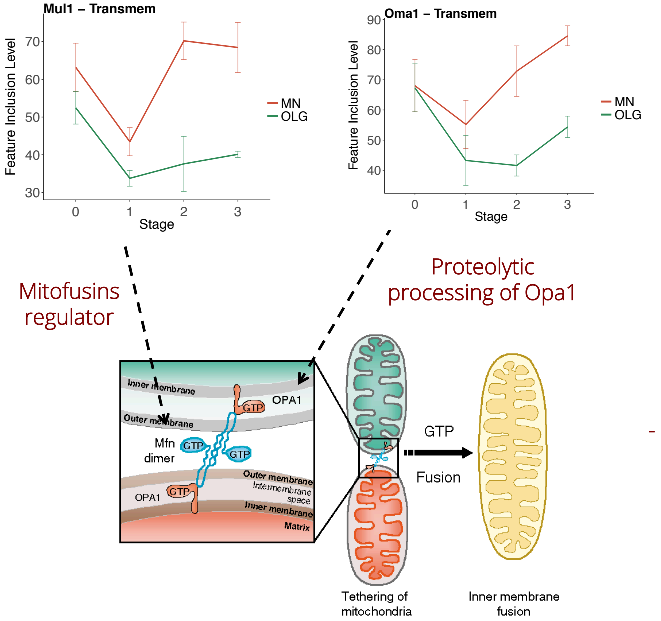


Figure 6.18: Differential inclusion of transmembrane regions in *Mul1* and *Oma1*, the main regulators of mitochondrial fusion proteins. *Mul1* regulates mitofusins, which mediate mitochondrial outer membrane fusion in mammals. *Oma1*, part of the mitochondrial quality control system, is located in the inner membrane and mediates OPA1 cleavage, resulting in mitochondrial inner-membrane fusion.

compared to MN because of the altered proteolytic processing of OPA1. (Figure 6.19.C). Thereby, our functional-impact analysis of DIU across neural cell types suggests a cell-type specific response to energy demands triggered by the differential inclusion of transmembrane regions in key regulators of mitochondrial dynamics such as OMA1 and MUL1.

Impact on UTR length and composition

In the non-coding region, we identified 2,607 features that were differentially included in alternative UTR regions between glial and neuronal subtypes, in a total of 387 genes, with the features most impacted by UTR regulation (in terms of number of genes) being miRNA binding sites and uORFs (Figure 6.20.A). Interestingly, there seemed to be a systematic trend towards 3'-UTR feature gain at stage 2 of MN differentiation (binomial test, miRNA binding sites FDR =

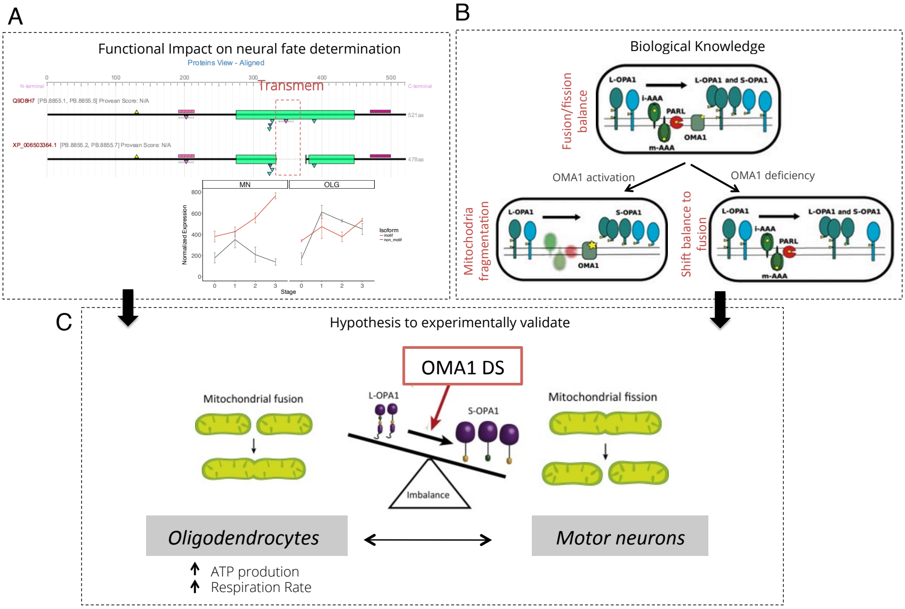


Figure 6.19: Functional effect of alternative splicing in the metalloendopeptidase OMA1. A. Four different transcript isoforms, involving AS events and polyadenylation sites, were defined by PacBio sequencing and predicted to code for two alternative CDSs which differed in the inclusion of a transmembrane region. DFI analysis revealed that the most common AS in MN differentiation was the loss of the TM region, likely hindering its transmembrane anchorage as well as OPA1 cleavage-activity. B. The balance of long and short forms of OPA1 dictates the balance between mitochondrial fusion and fission. OMA1 activation enhances the presence of the OPA1 short form, negatively regulating mitochondrial fusion. C. The differential accumulation of OMA1 splicing isoforms which lack the transmembrane region between differentiation systems suggests that there is cell-type specific regulation of mitochondrial dynamics, promoting fission activity in OPCs in which the TM-included OMA1 variant accumulates.

1.57e-35, 3’ UTR motifs FDR = 1.34e-12, 3’ UTR RBP binding FDR = 4.4e-02).

To elucidate if the increase in elements at 3’ UTR was associated with 3’-UTR lengthening, we ran our lengthening analysis. PCA of the 3’-UTR weighted mean-lengths showed that PC1 explained 28% of the data variance and clearly discriminated MN from OLG lineages (Figure 6.20.B). We defined the set of genes most relevant to the definition of PC1 by selecting genes with PC1 loadings higher to 3.5 (threshold defined based on loadings distribution), revealing a median pattern of gradual 3’-UTR lengthening for MN differentiation but pronounced shortening in the early stages followed by lengthening at the last stage in OLG differentiation (Figure 6.20.C). Thus, the highest 3'-UTR length difference between cell types corresponded to stage 2, in which the 3'-UTRs are preferentially longer for MN differentiating cells, whichs agrees with the significant detected gain of 3' UTR features.

We also applied DPA analysis to investigate the genome-wide landscape of APA across OPC and MNP differentiation, and identified 135 genes with significant DPA between cell subtypes. Clustering analysis for DPAU profiles in significant DPA genes defined four main regulatory patterns (Figure 6.20.D), with two of them characterising genes with increased DPAU in MNs (clusters 1 and 2), and the other two showing favored proximal polyA usage in MNs (clusters 3 and 4). However, the cluster that aggregated the highest number of genes showed a distal polyA usage profile that agrees with the previously-identified primary 3'-UTR lengthening pattern (Figure 6.20.D), and consequently, associates 3'-UTR lengthening with increased DPAU.

We further investigated the identity of features that were differentially included in 3'-UTRs (Figure 6.21.A). Interestingly, 3 of the 5 most differentially included miRNAs have been previously associated to neurite outgrowth [218], some of them in enhancement roles (mir-298) and others preventing neurite growth (mir466 and mir27). Among mir-466d-5p targets, we found a variety of proteins with established roles in neurite development such as *Ncam1*, *Gabbr1* or *Fbox31* and mitochondrial activity such as *Mtch2*, a novel regulator of mitochondrial metabolism, whose deletion has been shown to increase mitochondrial volume in axons and dendrites [284](Figure 6.21.C). Moreover, 81% of mir-466d-5p binding sites are preferentially included in MNs (Figure 6.21.A), suggesting cell-type specific gain of mir-466d-5p binding sites to precisely and autonomously regulate mRNA isoforms involved in neurite growth during motorneuron developement.

There was a similar cell-type inclusion preference for mir-874-3p, a poorly-characterised miRNA with a high proportion of targets associated with extracellular exosomes, including *Prkcz*, *Phb*, *Smim1* and *Cyfip2* (Figure 6.21.C), the

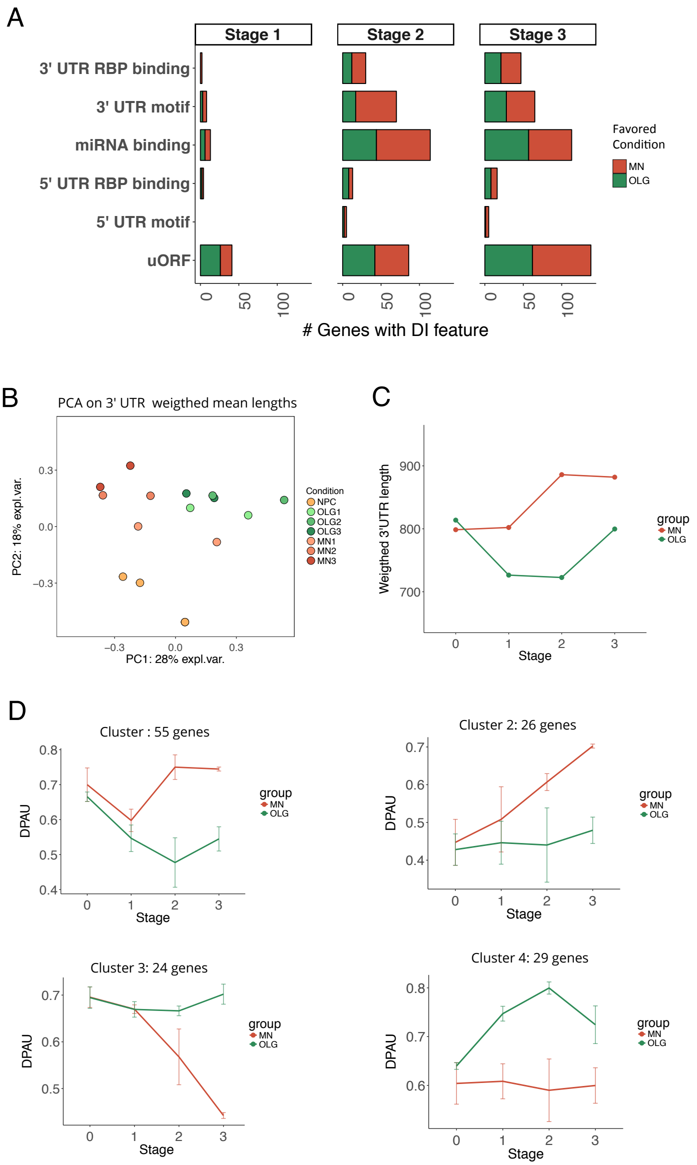


Figure 6.20: Regulation and functional impact of untranslated region (UTR) lengthening or shortening between OPC and MNP differentiation. A. Number of genes with gain or loss of elements between cell types across the different developmental stages and for the different UTR categories. B. PCA on the 3' UTR weighted mean lengths. C. 3' UTR lengthening median profile for the set of genes most contributing to PCA1 definition, the component discriminating OPCs and MNPs. D. Hierarchical clustering on distal polyA site usage levels for genes significantly detected as differentially polyadenylated.

last one found in nucleus and synaptosomes and highly associated to the generation of correct dendritic complexity when located at synapses [244]. Moreover, the long Cyfip2 3'-UTR also include a huge number of binding sites for *Mbnl1-2* (Figure 6.21.D). As MBNL1 binding at 3' UTRs is known to promote localization to membrane compartments [349], we speculate that mir-874-3p could specifically control the stabilization of *Cyfip2* mRNAs targeted to synaptic terminals without affecting *Cyfip2* mRNAs targeted to nucleus, thereby suggesting a cell-type specific spatial, functional and stability regulation of *Cyfip2* mRNA isoforms by differential polyAdenylation between OLD and MN differentiation. Moreover, Mbnl binding sites were also found among the set of most regulated elements (Figure 6.21.A). Besides, these set of genes with isoform-specific Mbnl1 targeting were found enriched in post-synaptic density (FET p.value = 2.8e-4) and significantly co-included with binding sites for CELF4 coDFI (p.value = 4.27e-16, Figure 6.21.D), RBP already reported as associated to the regulation of synaptic plasticity [346]. Among genes undergoing co-inclusion of Mbnl and Celf4 sites we found genes already described in this chapter as regulating the disruption or lost of functional elements between cell types such as Kif1b or Dclk1, correlating differential functional loading with alternative PolyA site choice.

Functional impact of isoform usage regulation in biological pathways

Mapping of genes with context-specific DFI to pathways and networks provides a comprenhensive way to interpret how functional isoform variability modulates the cellular response generated by a set of highly interconnected and coordinated genes under specific environmental or physiological stimuli.

KEGG [158] pathway enrichment analysis using as input the set of genes detected with differentially included features between MNPs and OPCs revealed the significant enrichment of the axon guidance pathway (FDR$<$0.01). Axon guidance is accomplished by signals, called guidance cues, sensed by the growth cones, that activate signalling molecules that eventually affect the cytoskeleton and dictate the axon the direction to grow. Inward signals, combined transcriptional and post-translational modifications of receptors and ligands, results in a

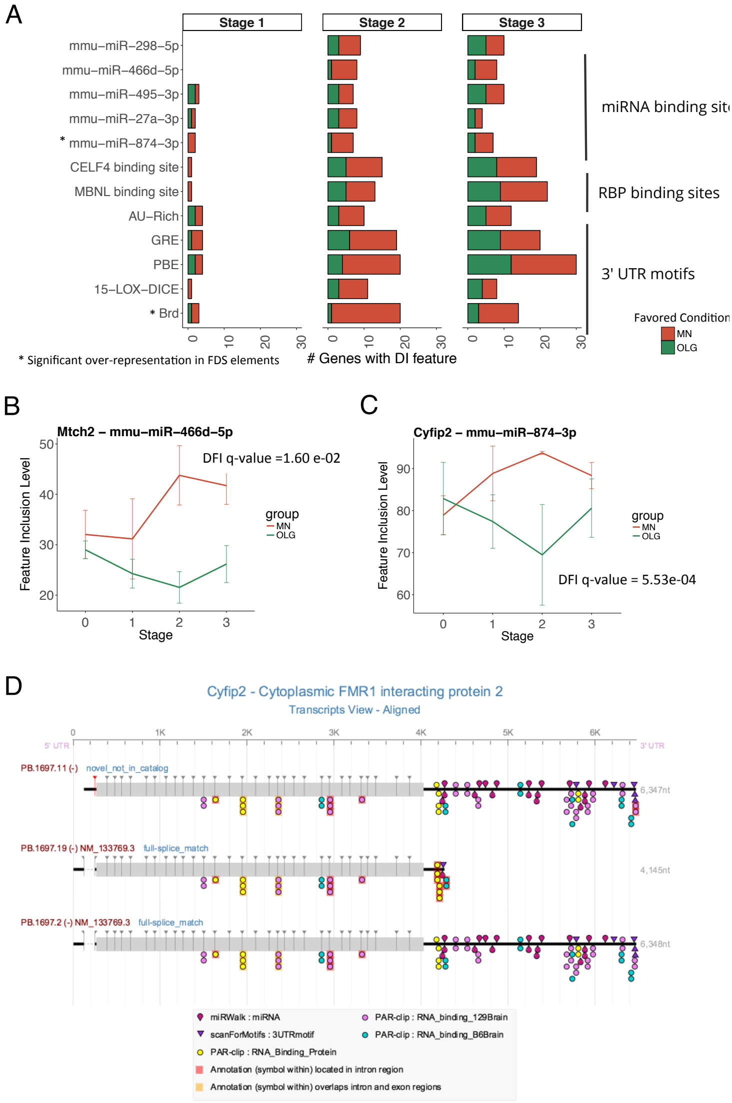
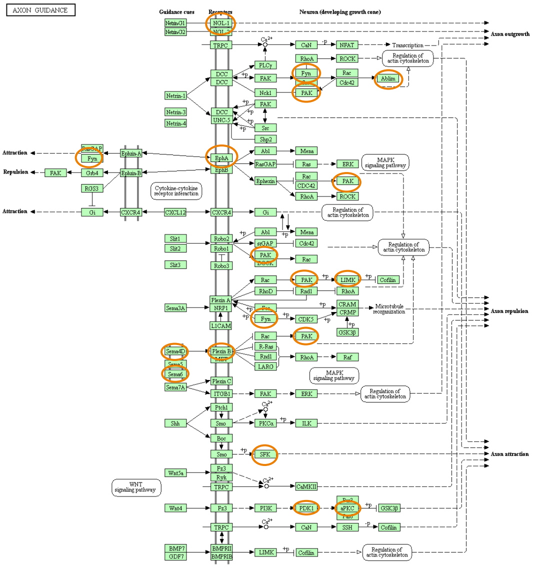
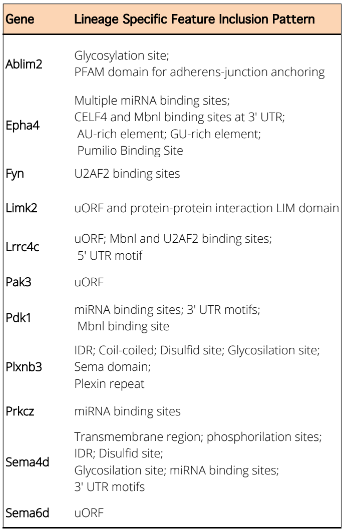


Figure 6.21: Top ranking of differentially included features between neural subtypes. A. Number of genes with gain or loss of 3'-UTR features between cell types across the different developmental stages and for the 3' UTR element categories with the highest number of significant genes. B. Inclusion pattern for the miR-466-5p binding site in *Mtch2*. C. miR-874-5p binding site inclusion pattern in both differentation systems for *Cyfip2*. D. Functional annotation of *Cyfip2* transcript isoforms, revealing two alternative polyA sites that promote the differential inclusion of several miRNA binding sites and multiple MBNL binding sites.

set of complex and dynamic molecular cues that provide either repel or attract axons toward their synaptic targets (Figure 6.22).

As shown in Figure 6.22.A several ligand receptors such as *Epha4*, *Lrrc4c* or *Plxnb3*, signalling proteins such as the tyrosin kinase *Fyn* or semaphorins (*Sema4d* and *Sema6d*), actin binding proteins as *Ablim2* or kinases as Prkcz or *Limk2* underwent the loss or disruption of functional features spread across all the considered experimental and predicted annotation categories, both in the coding and the UTR. For example, *Limk2* was detected expressing a motor-neuron specific isoform variant with an alternative transcription start site that promotes the inclusion of a N-terminal protein-protein interaction LIM domain during motor-neuron differentiation. Based on pathway information, this could alter its phosphorylase activity over cofilin and potentially impacts cytoskeleton dynamics. In contrast, *Pdk1*, which triggers axon attraction, underwent an APA event that promoted the over-representation of UTRs in OPCs containing MBNL1 binding sites, miRNA binding sites and other 3' UTR cis-elements such as Pumilio binding elements.





Gene	Lineage Specific Feature Inclusion Pattern
Ablim2	Glycosylation site; PFAM domain for adherens-junction anchoring

Epha4	Multiple miRNA binding sites;
	CELF4 and Mbnl binding sites at 3' UTR;
	AU-rich element; GU-rich element;
	Pumilio Binding Site

Fyn	U2AF2 binding sites
Limk2	uORF and protein-protein interaction LIM domain
Lrrc4c	uORF; Mbnl and U2AF2 binding sites; 5' UTR motif
Pak3	uORF
Pdk1	miRNA binding sites; 3' UTR motifs; Mbnl binding site

Plxnb3	IDR; Coil-coiled; Disulfid site; Glycosilation site;
	Sema domain;
	Plexin repeat

Prkcz	miRNA binding sites

Sema4d	Transmembrane region; phosphorilation sites;
	IDR; Disulfid site;
	Glycosilation site; miRNA binding sites;
	3' UTR motifs

Sema6d	uORF

Figure 6.22: Enrichment of genes with DFI features on KEGG pathways. A. Axon Guidance pathway. Genes subjected to regulation of functional features are represented with a orange circle. B. Table that shows the elements that are differentially included/disrupted between both differentiation systems.

6.4 Discussion

Neural progenitor cells divide to expand the progenitor population and then differentiate into distinct types of neurons and glial cells. Even though several transcriptional markers and factors have been revealed behind the functional and morphological specification of neural progenitors, the influence of post-transcriptional regulation in neural fate determination and, specially, its functional impact are poorly understood. Deep PacBio sequencing of a neural time-course differentiation system involving MNPs and OPCs generated from common spinal-cord NPCs revealed a transcriptome comprising 34,304 isoforms, 20% of them representing novel calls.

Even though novel isoforms were generally had a significant lower expression distribution than known isoforms, they were able to subtly described cell lineage and differentiaton dynamics. Furthermore, a remarkable 30% of them became the major expressed isoform in specific conditions, what highlighted the stage/cell specific role of novel isoforms and the relevance of capturing these calls to accurately describe transcriptome dynamics. In addition to metabolic and neurogenesis processes, functional profiling showed up the over-representation of novel isoforms in NMD targeting, revealing the strong under-representation of NMD variants in current reference annotations. IR events, were also highly enriched in novel and NMD isoforms, indicating the functional role of IR events in fine tunning gene expression during neural differentiation. In fact, this coupled effect (IR and NMD) has been repeadly claimed as a major driver of several differentiation systems including granulopoiesis [367], or hematopoyesis [114]. Moreover, NMD was found highly enriched in mRNA metabolism, RNA binding and mainly in the SR family of splicing factors, in agreement with the high proportion of SR genes that are known to be affected by alternative splicing couple to NMD [227][181][180]. Thus, NMD was revealed as a potential regulator AS in our neural differentiation system through the regulation of AS drivers.

The extensive program of AS in our neural differentiation was highlighted by the high proportion of genes expressing multiple splicing variants (50 %) with an average of 4 variants per gene. Similarly, 40% of expressed genes also underwent APA. However, the number of alternative polyA sites per gene reduced to 2, in agreement with previous analysis [317], suggesting that APA, in contrast to AS, generally acts like a binary on/off switch. Finally, the expression of alternative TSSs was detected in almost 20% of genes. Taken together, 70% of the genes detected in our neural differentiation system were found expressing alternative variants that are potential to expand the functional and regulatory transcriptome diversity.

Membrane trafficking is highly influenced by differential isoform usage rather than by differential gene expression in neural fate determination.

Transcriptome dynamics profiling revealed that isoform usage regulation occurs as frequently in glia as it does in neurons, and in half of cases, involves the differential usage of alternative CDSs, likely conferring functional specifity to developmental states. Moreover, only around one third of these isoform regulated genes underwent changes in the overall gene expression, what highlighted the independent activity of these two gene-expression regulatory programs during neural differentiation processes.

When we studied the transcriptional program regulating neural fate determination we found an extensive program of gene expression regulation but a highly precise post-transcriptional program, where 80% of isoform switches were associated to specific developmental stages, what highlights the strong temporal regulation to which isoforms are subjected, in agreement with recent studies in neural cell types [362]. Interestingly, a high number of vesicular trafficking, membrane dynamics and synaptic processes were significantly associated to the regulation of specific protein isoforms rather than by global differences in the levels of gene expression, and includes genes such as *Osbpl9*, *Dtnb*, *Camd1*, *Tmem87b*, *Vps39*, *Vps26a*, all of them expressing MN specifc isoforms in the last stage of differentiation or *Exoc3* and *Derl2*, involving modulation of their UTR regions. Moreover, we captured genes already reported to be influenced by AS such as *Stx16*, involved in vesicle-mediated transport and axon regeneration and mis-spliced in autism spectrum disorder (ASD) patiens because of RBFOX dysregulation [340] [28] or *Klc1*, involved in cargo biding and expressing isoforms controlling neurotransmission by specifically regulating the transport of the GABA-B receptor towards dendrites [336][94]. In fact, membrane trafficking has been already reported being developmentally and tissue specifically regulated by alternative splicing [28][33] [79][117][133][149] and the Guidice group has even experimentally studied the functional role of AS in trafficcking functions in heart development and skeletal muscle by validating a bunch of genes among we find some of our cell-type regulated genes such as *Cltc*, a major component of the cytoplasmic face of intracellular organelles [117][118]. Thus, our results indicate that alternative isoform usage also contributes to endosomal and trafficking specification during neural fate determination. During neuron development, an specialzed and sublty orchestrated membrane trafficking machinery is required for the expansión of the plasma membrane and the removal and trafficking of materials and proteins to specific locations. From initial neurite formation to synaptogenesis, vesicle trafficking plays an integral role in neural morphological progession (cell shape and polarization) and function (exocytic fusion of synaptic vesicles or endocytic recyplin of the synaptic machinery). Moreover, the carefully regulated sorting and trafficking of myelin proteins and lipids in oligodendrocytes is key for polarity establishment and maintenance and involves several endosomal compartment and vesicle transport pathways [20][202]. Its physiological importance is highlighted by the number of neurological disorders [110][271][341][353][365] and demyelinating diseases [202] associated with the perturbation of membrane trafficking. Our results highlight post-transcriptional regulation as the mayor contributor of the requiered specialization of membrane trafficking pathways during neural fate determination.

Differential inclusion of TM regions contributes to the regulation of mitochondrial dynamics during neural fate determination

To understand how these isoform transitions alter gene function during neural fate determination we studied their effect on the availability of regulatory features within UTRs and the functional load of coded proteins and found a total of 4,821 functional features differentially included by alternative isoform processing between both neural differentiation, 70% of them representing stage-specific gains and losses, what reveals cell-specific functional modulations highly temporally.

Among the set of functional categories, uORFs, which are predominantly altered by the use of alternative transcription start sites (60%), appeared as the most significantly regulated elements between cell subtypes. Although our results agree with recent studies pointing out ATI as one of the principal drivers of transcript isoform diversity [272], we decided to not characterise in further detail these elements because of the limitation of PacBio sequencing to discriminate 5' incomplete sequences and provide confident transcription start sites. Similar to uORFs, IDRs are sytematically disrupted by alternative transcript preprocessing through all the differentiation stages and affecting to the highest number of genes among considered categories. Moreover, they were enriched in protein binding functions, what correlates with the known role of IDRs on the rewiring of protein interaction networks [37][94]. In addition to the potential spread impact of transcript preprocesssing on protein-protein interactions, genes with regulted IDRs were found also enriched in genes associated to the cell junction and the cytoskeleton and remarkably involving several kinases and phosphatases as *Pip5k1*, which undergoes the inclusion of an IDR during MNP differentiation that is potentially encouraging its recruitment to focal adhesions to control the pool of phosphoinosities implicated in correct synaptic vecicle endocytosis and exocitosis [236][360].

In contast to IDRs and ORFs, other feature categories appeared controlled in a time specific manner due to their enriched regulation in certain differentiation stages: PTM sites were predominantly impacted at early differentiation while a pattern of gain of 3' UTR elements (3' UTR motifs and miRNA binding sites) was found during MNs differentiation, in correlation with the gradual 3' UTR lengthening pattern found during MN differentiation and the pronounced shortening during the first stages of OLG differentiation. Thus, our results verified the already reported widespread elongation of 3' UTRs in neurons [140][153][211][327] and revealed a differential 3'UTR patterning in oligodendrocyte differentiation. Moreover, the role of this differential 3' UTR pattern during neural fate determinantion was determined by the nature of the regulatory elements harboured in these regulated UTRs. Among the most frequently regulated features, we interestely found miRNAS associated to neurite outgrowth, including the mir-466d-5p for which isoform-specific binding sites are gained during MN differentiation. Addditionally, binding sites for the mir-874-3p were significantly enriched in regulated UTRs and essentially contained in genes associated with the extracelullar exosome. Moreover, RBPs as MBNL, controlling mRNA localization to membrane compartments [350], were found frequently targeting regulated UTRs in synaptic genes. Taken all together our results suggest that the diversification of UTRs during neural fate determination regulates the estability and localization of isoforms that are required in neuronal specific compartments as dendrites or synapses and this control is triggered by the favored inclusion of neural specialized cis-acting miRNAs and RBPs in MNs.

Our results also detected the differential inclusion of TM regions in Mul1 and Oma1, which could be potentially impacting their anchored to inner and outer mitochondrial membranes, respectively. These not previously reported post-trascriptional events might be then altering the post-translational proteolysis and/or turnover of their targets, the Mitofusin 1 and 2 (Mfn1 and 2) and OPA1, both of them GTPases coordinating mitochondrial fusion. In brain numerous cellular functions including ATP production, Ca2+ buffering, neurotransmitter synthesis and degradation, ROS production and sequestration, apoptosis and intermediate metabolism are spatially and temporally regulated by mitochondrial localization [104][35][152][194][229][282][298][344] and mitochondrial bioenergetics [282][78], all of which are strongly influenced by mitochondrial dynamics, which entails mitochondrial fission, fusion and transport. Moreover, their physiological relevance has been illustrated by the profound effect that perturbations in mitochondrial dynamics have on neural system development and plasticity. Moreover, numerous neurodegenerative disorders and demyelinating diseases such as the Amyotrophic lateral sclerosis [301], the Huntington's disease [138], the Charcot-Marie–Tooth disease [210] or optic atrophy [383], have been associated with mutations and dysregulation in fission and fusion enzymes, demonstrating that mitochondrial dynamics in both oligodendrocytes and neurons need to be balanced between rates of fusion and fission to properly response to environmental stimuli and pathophysiologic signals. Apart of the known regulation of mitochondrial dynamics accomplished through post-translational modification of mitochondrial fission and fusion enzymes such as OPA1 and Miofusins, here we highlight the potential impact of post-transcriptional regulation of TM regions in fusion regulators for the adaptation of mitochondrial dynamics in neural fate determination. Although we do not know the exact function of Oma1 and Mul1 isoforms, our analysis was able to capture the potential impact of their regulation on gene properties, hypotheisis that the favored exclusion of TM regions during oligodendrocyte differentialtion is altering their enzymatic activity by affecting their membrane anchored and consequently contribute to the regulation of their substrates and influence the fission/fussion balance in neural cells.

Interestingly, the contribution of post-transcriptional preprocessing to the bioenergetic adaptation of cells during neural fate determination is reinforced by the detection of several mitochondrial dynamics associated genes as Mff (mitochondrial fission factor), Mief1 (mitochondrial fission regulator) or Kif1b, expressing neuron-specific isoforms largely reported as responsible of the movement of mitochondria along the axon by modifying regions of cargo binding [65].

All together, our results suggest that post-transcriptional regulation, in addition to alter the post-transcriptional machinery itself, is mainly contributing to modulate the differential spatial localization and movement of gene products in neural cell subtypes and adapt mitochondrial dynamics to specific cellular demands. More importantly, our isoform-resolved functional analysis characterized the functional impact of multiple AS and APA events and profiled the stage-specific pattern of regulation of certain functional features as PTMs and UTR motifs between neural subtypes. Moreover, our study brought out elements such as miR-874-3p binding sites or TM regions as highly impacted by post-transcriptional regulation and playing a potential relevant role in the neural fate determination of NPCs. Thus, our functional isoform analysis framework enables to address the analysis of differential isoform usage in multiple levels:

the structural, by displaying AS and APA patterns, the functional, revealing the processes predominately regulated by post-transcriptional mechanisms, and the feature level, achieving a high degree of specificity by exploring the gain and loss of functional elements, making posible to create mechanistic hypothesis about how alternative splicing and alternative polyadenylation modulate the function of gene products.

Chapter 7

Conclusions

In this thesis we developed a bioinformatics framework to study the potential functional impact of isoform regulation at the genome-wide level and to understand how context-specific alternative splicing and alternative polyadenylation events contribute to phenotype specification by altering the functional and regulatory characteristics of expressed isoforms. This analysis framework is based on the definition of full-length transcriptomes from single-molecule sequencing technologies. The curation of these transcriptomes is essential at a time when long-read sequencing is becoming more popular to define transcriptome complexity. Therefore, we developed SQANTI, a tool which performs the quality control of long-read data.

The second element to consider is the functional characterisation of isoforms. The functional impact of isoforms is only meaningful if we can capture isoform-specific functional features. Thus, we designed IsoAnnot, which annotates functional features at the RNA and protein levels. Its application to several organisms has revealed the relevant variability existing between isoforms.

Finally, we developed the tappAS analysis platform, which, thanks to its flexible combination of expression analysis, feature-inclusion analysis, enrichment methods and visualisation options, enables the formulation of a large range of questions about the functional effect of post-transcriptional regulation. We demonstrated the scope of these tools using a neural cell-fate determination system as proof-of-concept.

We hope these tools are of use to the broad scientific community and will help advance our understanding of the functional relevance of transcript alternative processing.

The conclusions of this thesis are summarised and organised below according to the goals originally defined in 2:

1) Accurately define and quantify long-read defined transcriptomes.

- We developed SQANTI, an analysis tool to boost the quality control of long-read based transcriptome studies by providing the methods to deliver fully characterised and curated long-read transcriptomes.
- We applied our framework to multiple organisms (mouse, human or maize), long-read sequencing platforms (PacBio and Nanopore) and transcriptome reconstruction pipelines (TAPIS, IDP and Iso-Seq Analysis^{TM}) and illustrated how SQANTI can effectively characterise and evaluate the composition and peculiarities of each full-length transcriptome.
- Evaluation of a PacBio-defined murine neural transcriptome highlighted the enrichment of low-quality SQANTI attributes in novel transcripts associated with sequencing errors and RT-switching and intra-priming events originated during the library preparation.
- Our machine-learning filtering strategy efficiently discriminated artefactual transcripts from long-read transcriptomes and outperformed previous alternative approaches.
- We detected that non-annotated variability at 3' transcript ends has a strong negative impact in the accuracy of current isoform expression quantification algorithms and this effect is alleviated when an expressed, full-length reference transcriptome is used.
- Long-read technologies tend to accumulate false novel calls if proper quality standards are not established. However, if the quality control is adequate, long-read technologies can effectively characterise transcriptome complexity and accurately estimate isoform expression levels.

2) **Implement a pipeline for the extensive functional annotation of gene products at the isoform-resolution.**

- We designed the IsoAnnot pipeline, which dynamically constructs an isoform-resolved database of functional and regulatory annotations from a set of input sequences by integrating information disseminated across several databases and prediction methods, thus facilitating the study of the functional divergence of isoforms.
- We implemented a functional transfer module that populates query sequences with experimental and manually-curated features stored in gene-centric public databases and resolves annotations at the isoform-level.
- IsoAnnot was designed to be potentially applicable to any organism, independently of its novelty rate, overcoming the limitations of current static databases that do not support the annotation of novel isoforms or multiple organisms.
- Application of IsoAnnot to long-read defined and reference transcriptomes in different organisms such as mouse, human, maize or fruit flies revealed that the nature of the transcriptome influences feature coverage without introducing biases that could affect the reliability of results obtained while studying the functional isoform variability.

3) **Develop an analysis framework to address the functional impact of context-specific isoform regulation.**

- We developed the *Functional Diversity* analysis, which processes gene-models and positional functional labels at isoform resolution to describe the level of structural and functional diversity between isoforms at a genome-wide scale.
- For contextual modelling, we developed the *Differential Feature Inclusion* analysis, which, provided transcript-level expression data, defines the set of functional features that are included or excluded within transcripts because of differential isoform usage. We also added complementary functionalities such as the co-inclusion analysis to explore pairs of functional elements that are processed together.

- We developed the *Differential Polyadenylation* analysis to model the context-specific polyA site regulation by leveraging the resolution of PacBio to identify polyA site variability and the UTR lengthening analysis to study the impact of post-transcriptional regulation on UTR length modulation.

- Our analyses were designed to tackle three different experimental designs (pairwise, single-series time course and multiple-series time course) and to comprehensively display results according to the selected design, thus providing a wide scope of application.

- This isoform-resolved functional analysis framework was implemented in a highly interactive, graphical and user-friendly tool called tappAS.

4) Understand the functional consequences of isoform-usage regulation on neural cell-fate determination.

- High-throughput Iso-Seq PacBio sequencing of differentiating murine neural cells revealed an extensive PTR program, in which 70% of the detected genes express multiple transcript variants as a combination of multiple AS variants, dual polyA site choice and, to a lesser extent, alternative transcription start sites.

- Around 20% of Iso-Seq full-length isoforms were characterised as novel and were enriched in IR and NMD targeting. Despite their reduced expression levels, novel variants subtly defined differentiation timing, were able to discriminated cell types and a significant proportion of them became the major expressed variant under specific conditions, potentially indicating their developmental-stage specific role.

- Isoform usage regulation is highly dynamic and occurs as frequently in OPC differentiation as it does in MNP differentiation.

- Membrane trafficking is highly regulated by differential isoform usage rather than by differential gene expression in OPC and MNP developmental systems, suggesting that the endosomal and trafficking specification during neural fate determination is strongly influenced by PTR mechanisms.

- Isoform-resolved functional analysis across our differentiation system profiled the pattern of gains and losses between cell types for multiple functional features and characterised the potential functional impact of multiple AS and APA events.
- IDRs were systematically enriched in feature-skipping events between differentiation systems and enriched in protein-binding functions and cell junction and cytoskeleton components, while the regulation of elements as PTM sites and miRNA binding sites was clearly over-represented during specific developmental stages.
- MNPs included a higher number of regulatory elements at their expressed 3' UTRs compared to OPCs during development, including miRNAS binding sites associated with neurite outgrowth such as the mir-466d-5p. This pattern correlates with the MNP-specific gradual lengthening of 3' UTRs.
- The differential inclusion of TM regions in mitochondrial fusion regulators between neural subtypes indicated the potential influence of post-transcriptional regulation on the modulation of GTPase activity and its contribution to the adaptation of the fission/fussion balance to meet specific neural cell subtype demands.

Appendix 1:
SQANTI attributes at transcript level

Column	Feature Name	Description
1	isoform	Unique ID conferred by Pacbio to ToFU transcripts
2	chrom	Chromosome where the transcript aligns
3	strand	Strand to which the transcript belongs
4	length	Length of the transcript in nucleotides
5	exons	Number of exons
6	structural_category	Splice junction based classification of the transcript against a annotated reference transcriptome
7	associated_gene	Gene to which the transcript maps
8	associated_transcript	Transcript in the reference annotation which has the same splice pattern as the isoform. Only applicable to FSM, ISM and UTR3 Fragment transcripts.
9	ref_length	Length of the associated reference transcript in nts. Only applicable to FSM, ISM and UTR3 Fragment transcripts.
10	ref_exons	Number of exons of the associated reference transcript. Only applicable to FSM, ISM and UTR3 Fragment transcripts.
11	diff_to_TSS	Difference in nucleotides between the Transcription Start Site (TSS) of the associated reference transcript and the sequenced transcript. Only applicable to FSM, ISM and UTR3 Fragment transcripts.
12	diff_to_TTS	Difference in nucleotides between the Transcription Termination Site (TTS) of the associated reference transcript and the sequenced transcript. Only applicable to FSM, ISM and UTR3 Fragment transcripts.
13	subcategory	Type of subevent associated with the transcript. Applicable for ISM and novel transcripts.
14	RTS_stage	Detection of hallmarks of RT switching within at least one of the splice junctions of the sequenced transcript
15	all_canonical	Labelling of the type of splice junction in the transcript. If all the splice junctions of the transcript are canonical this field will be "canonical". If there is at least one non canonical junction in the transcript will be labelled as non-canonical.
16	min_sample_cov	Lowest number of biological samples showing the presence of a splice junction within the transcript. Each time a splice junction within a transcript is covered by short reads of a different sample the value is augmented by 1. After mapping all available samples, the number of samples supporting each junction is computed and the minumun value of all transcript junctions is taken as min_sample_cov value
17	min_cov	Lowest number of short reads supporting the presence of a splice junction within the transcript. The number of short reads mapping to each splice junction are obtained for each available sample. After mapping all the available samples, the total number of short reads mapped per junction is computed and the lowest value is taken as minCov value
18	min_cov_pos	Position in nts relative to the TSS of the splice junction with the minCov value. If two splice junctions show the same minCov, the minCovPos selected is the one closest to the TSS.
19	sd_cov	Standard Deviation of the splice junction short read coverage per transcript
20	FL	Number of Full Length reads per transcript
21	n_indels	Number of indels in the ToFU transcript before genome-based correction
22	n_indels_junc	Number of junctions with indels around the splice junction of the ToFU transcript before correction
23	bite	The transcript has at least one junction labelled as bite
24	iso_exp	Expression of the transcript in TPMs calculated by short reads
25	gene_exp	Sum of all isoExp values of transcripts belonging to the same gene.
26	ratio_exp	Ratio between the isoExp and the geneExp
27	FSM_class	This feature classifies the transcript according to the expression of other isoforms in the gene to which the transcript belongs. Transcripts belonging to genes that only express one isoform are classified as A. Transcripts belonging to genes that express more than one isoform but none is a FSM are classified as B. Transcripts belonging to genes which express more than one isoform and other isoforms and at least one is a FSM are classified as C
28	coding	Logical indicating if the transcript is predicted to have an ORF by GMST
29	ORF_length	Length in aminoacids of the Open RF predicted by GMST
30	CDS_length	Length in nts of the CDS predicted by GMST
31	CDS_start	Position in nucleotides relative to the TSS of the start codon
32	CDS_end	Position in nucleotides relative to the TSS of the stop codon
33	perc_A_downstream_TTS	Percentage of adenines in a window of genomic DNA inmmediately downstream the TTS. The deffault parameter for window size is 20 nucleotides

Appendix 2:
SQANTI attributes at splice junction level

Column	Feature Name	Description
1	isoform	Unique ID conferred by Pacbio to ToFU transcripts.
2	chrom	Chromosome where the transcript aligns
3	strand	Strand to which the transcript belongs
4	junction_number	Position order, starting from the 5' end, of the junction in the transcript
5	genomic_start_coord	Start coordinate position of the junction in the genome
6	genomic_end_coord	End coordinate position of the junction in the genome
7	transcript_coord	Coordinate of the junction inside the transcript
8	junction_category	Type of junction. "Known" if both the acceptor and the donor sites are annotated in the provided reference annotation file and "Novel" otherwise.
9	start_site_category	"Known" or "novel" according to reference annotated splice sites
10	end_site_category	"Known" or "novel" according to reference annotated splice sites
11	diff_to_Ref_start_site	Nearest annotated splice site in the genome
12	diff_to_Ref_end_site	Nearest annotated splice site in the genome
13	bite_junction	Applies only to novel splice junctions. If the novel intron partially overlaps annotated exons the bite value is TRUE, otherwise it is FALSE.
14	splice_site	Splice site sequence
15	canonical	Indicates whether the junction is canonical or not. The set of splice sites that are considered by SQANTI as canonical are GTAG,GCAG,ATAC. Canonical junction sequences can be modified by --sites option.
16	RTS_junction	Logical, indicating the detection of hallmarks of RT switching
17	indel_near_junc	Logical, indicating the existence of indels around the junction
18	samples_with_cov	Number of samples that support the splice junction
19	total_coverage	Short-read coverage sum across input samples
20	coverage_per_sample	Average short-read coverage across input sample

References

[1] Abdel-Ghany, S.E., Hamilton, M., Jacobi, J.L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A. & Reddy, A.S.N. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications*, **7**, 11706. 9, 13, 23, 35, 41, 86, 92, 125

[2] Agarwal, V., Bell, G.W., Nam, J.W. & Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, **4**. 98

[3] Alamancos, G.P., Pages, A., Trincado, J.L., Bellora, N. & Eyras, E. (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA (New York, N.Y.)*, **21**, 1521–1531. 192

[4] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–410. 51

[5] Anand, R., Wai, T., Baker, M.J., Kladt, N., Schauss, A.C., Rugarli, E. & Langer, T. (2014). The i-AAA protease YME1L and OMA1 cleave OPA1 to balance mitochondrial fusion and fission. *The Journal of cell biology*, **204**, 919–929. 222

[6] Anders, S., Reyes, A. & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, **22**, 2008–2017. 15, 142

[7] Andreassi, C. & Riccio, A. (2009). To localize or not to localize: mRNA fate is in 3'UTR ends. *Trends in cell biology*, **19**, 465–474. 92, 94, 187

[8] Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. & Yeh, L.S.L. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic acids research*, **32**, D115–9. 107

[9] Arefeen, A., Liu, J., Xiao, X. & Jiang, T. (2018). TAPAS: tool for alternative polyadenylation site analysis. *Bioinformatics (Oxford, England)*, **34**, 2521–2529. 148

[10] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, **25**, 25–29. 91, 111

[11] Au, K.F., Jiang, H., Lin, L., Xing, Y. & Wong, W.H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic acids research*, **38**, 4570–4578. 12

[12] Au, K.F., Underwood, J.G., Lee, L. & Wong, W.H. (2012). Improving PacBio long read accuracy by short read alignment. *PloS one*, **7**, e46679. 11, 12, 42, 67

[13] 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. & Wong, W.H. (2013). Characterization of the human ESC transcriptome by hybrid sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, E4821–30. 12, 23, 35, 41, 42, 86

[14] Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P. & Bartel, D.P. (2008). The impact of microRNAs on protein output. *Nature*, **455**, 64. 97

[15] Baker, M.J., Lampe, P.A., Stojanovski, D., Korwitz, A., Anand, R., Tatsuta, T. & Langer, T. (2014). Stress-induced OMA1 activation and autocatalytic turnover regulate OPA1-dependent mitochondrial dynamics. *The EMBO journal*, **33**, 578–593. 222

[16] Bao, W., Kojima, K.K. & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 11. 97

[17] Baracskay, K.L., Kidd, G.J., Miller, R.H. & Trapp, B.D. (2007). NG2-positive cells generate A2B5-positive oligodendrocyte precursor cells. *Glia*, **55**, 1001–1010. 202

[18] Baralle, F.E. & Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, **18**, 437. 16, 23, 91

[19] Barbosa, C., Peixeiro, I. & Romao, L. (2013). Gene expression regulation by upstream open reading frames and human disease. *PLoS genetics*, **9**, e1003529. 96

[20] Baron, W. & Hoekstra, D. (2010). On the biogenesis of myelin membranes: Sorting, trafficking and cell polarity. *FEBS Letters*, **584**, 1760–1770. 233

[21] Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M. & Gautheret, D. (2000). Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Research*, **10**, 1001–1010. 5, 102, 137

[22] Belfiore, A., Frasca, F., Pandini, G., Sciacca, L. & Vigneri, R. (2009). Insulin receptor isoforms and insulin receptor/insulin-like growth factor receptor hybrids in physiology and disease. *Endocrine reviews*, **30**, 586–623. 18

[23] Birch, D., Britt, B.C., Dukes, S.C., Kessler, J.A. & Dizon, M.L.V. (2014). MicroRNAs participate in the murine oligodendroglial response to perinatal hypoxia–ischemia. *Pediatric Research*, **76**, 334. 157

[24] Biswas, A. & Brown, C.M. (2014). Scan for Motifs: a webserver for the analysis of post-transcriptional regulatory elements in the 3' untranslated regions (3' UTRs) of mRNAs. *BMC bioinformatics*, **15**, 174. 96, 102

[25] Blair, J.D., Hockemeyer, D., Doudna, J.A., Bateup, H.S. & Floor, S.N. (2017). Widespread Translational Remodeling during Human Neuronal Differentiation. *Cell Reports*, **21**, 2005–2016. 19, 23

[26] Blazie, S.M., Geissel, H.C., Wilky, H., Joshi, R., Newbern, J. & Mangone, M. (2017). Alternative Polyadenylation Directs Tissue-Specific miRNA Targeting in Caenorhabditis elegans Somatic Tissues. *Genetics*, **206**, 757–774. 23

[27] Blencowe, B.J. (2017). The Relationship between Alternative Splicing and Proteomic Complexity. *Trends in Biochemical Sciences*, **42**, 407–408. 18

[28] Blue, R.E., Curry, E.G., Engels, N.M., Lee, E.Y. & Giudice, J. (2018). How alternative splicing affects membrane-trafficking dynamics. *Journal of cell science*, **131**. 232, 233

[29] Bolisetty, M.T., Rajadinakaran, G. & Graveley, B.R. (2015). Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biology*, **16**, 204. 9

[30] Borodovsky, M. & Lomsadze, A. (2011). Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. *Current protocols in bioinformatics*, **Chapter 4**, Unit 4.5.1–17. 49, 81, 194

[31] Brameier, M., Krings, A. & MacCallum, R.M. (2007). NucPred—Predicting nuclear localization of proteins. *Bioinformatics*, **23**, 1159–1160. 106

[32] Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**, 525. 14

[33] Brinegar, A.E., Xia, Z., Loehr, J.A., Li, W., Rodney, G.G. & Cooper, T.A. (2017). Extensive alternative splicing transitions during postnatal skeletal muscle development are required for calcium handling functions. *eLife*, **6**, e27192. 131, 233

[34] Briscoe, J. & Novitch, B.G. (2008). Regulatory pathways linking progenitor patterning, cell fates and neurogenesis in the ventral neural tube. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **363**, 57–70. 203, 204

[35] Brodin, L., Bakeeva, L. & Shupliakov, O. (1999). *Presynaptic mitochondria and the temporal pattern of neurotransmitter release*, vol. 354. 235

[36] Brown, M., Suryawanshi, H., Hafner, M., Farazi, T.A. & Tuschl, T. (2013). Mammalian miRNA curation through next-generation sequencing. *Frontiers in genetics*, **4**, 145. 98

[37] BULJAN, M., CHALANCON, G., EUSTERMANN, S., WAGNER, G.P., FUXREITER, M., BATEMAN, A. & BABU, M.M. (2012). Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Molecular Cell*, **46**, 871–883. 23, 125, 132, 159, 234

[38] BURKHARD, P., STETEFELD, J. & STRELKOV, S.V. (2001). Coiled coils: a highly versatile protein folding motif. *Trends in Cell Biology*, **11**, 82–88. 105

[39] BYRNE, A., BEAUDIN, A.E., OLSEN, H.E., JAIN, M., COLE, C., PALMER, T., DUBOIS, R.M., FORSBERG, E.C., AKESON, M. & VOLLMERS, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature communications*, **8**, 16027. 9, 40

[40] CAI, X. & LIU, X. (2008). Inhibition of Thr-55 phosphorylation restores p53 nuclear localization and sensitizes cancer cells to DNA damage. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 16958–16963. 175

[41] CARBONNEAU, M.A., CHEPLYGINA, V., GRANGER, E. & GAGNON, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, **77**, 329–353. 91

[42] CARNEIRO, M.O., RUSS, C., ROSS, M.G., GABRIEL, S.B., NUSBAUM, C. & DEPRISTO, M.A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC genomics*, **13**, 375. 35

[43] CARTOLANO, M., HUETTEL, B., HARTWIG, B., REINHARDT, R. & SCHNEEBERGER, K. (2016). cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing. *PLOS ONE*, **11**, e0157779. 87

[44] CHAISSON, M.J.P., WILSON, R.K. & EICHLER, E.E. (2015). Genetic variation and the de novo assembly of human genomes. *Nature reviews. Genetics*, **16**, 627–640. 9

[45] CHAMBERS, C.B., PENG, Y., NGUYEN, H., GAIANO, N., FISHELL, G. & NYE, J.S. (2001). Spatiotemporal selectivity of response to Notch1 signals in mammalian forebrain precursors. *Development*, **128**, 689–702. 187

[46] CHANG, T.H., HUANG, H.Y., HSU, J.B.K., WENG, S.L., HORNG, J.T. & HUANG, H.D. (2013). An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. *BMC bioinformatics*, **14 Suppl 2**, S4–S4. 96

[47] CHANG, Y.F., IMAM, J.S. & WILKINSON, M.F. (2007). The nonsense-mediated decay RNA surveillance pathway. *Annual review of biochemistry*, **76**, 51–74. 103

[48] CHAPIN, S.J., LUE, C.M., YU, M.T. & BULINSKI, J.C. (1995). Differential expression of alternatively spliced forms of MAP4: a repertoire of structurally different microtubule-binding domains. *Biochemistry*, **34**, 2289–2301. 171

[49] CHEN, C.Y. & SHYU, A.B. (1995). AU-rich elements: characterization and importance in mRNA degradation. *Trends in biochemical sciences*, **20**, 465–470. 180

[50] CHEN, L.L., DECERBO, J.N. & CARMICHAEL, G.G. (2008). Alu element-mediated gene silencing. *The EMBO Journal*, **27**, 1694–1705. 97

[51] CHEN, S.Y., DENG, F., JIA, X., LI, C. & LAI, S.J. (2017). A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. *Scientific Reports*, **7**, 7648. 9

[52] CHEN, W., JIA, Q., SONG, Y., FU, H., WEI, G. & NI, T. (2017). Alternative Polyadenylation: Methods, Findings, and Impacts. *Genomics, Proteomics & Bioinformatics*, **15**, 287–300. 5, 23, 91, 92

[53] CHENG, B., FURTADO, A. & HENRY, R.J. (2017). Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. *GigaScience*, **6**, gix086–gix086. 9, 125

[54] CHENG, H., DENG, W., WANG, Y., REN, J., LIU, Z. & XUE, Y. (2014). dbPPT: a comprehensive database of protein phosphorylation in plants. *Database : the journal of biological databases and curation*, **2014**, bau121–bau121. 108

[55] CHEW, G.L., PAULI, A. & SCHIER, A.F. (2016). Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nature Communications*, **7**, 11663. 19, 138

[56] CHIH, B., GOLLAN, L. & SCHEIFFELE, P. (2006). Alternative splicing controls selective trans-synaptic interactions of the neuroligin-neurexin complex. *Neuron*, **51**, 171–178. 187

[57] Chou, C.H., Shrestha, S., Yang, C.D., Chang, N.W., Lin, Y.L., Liao, K.W., Huang, W.C., Sun, T.H., Tu, S.J., Lee, W.H., Chiew, M.Y., Tai, C.S., Wei, T.Y., Tsai, T.R., Huang, H.T., Wang, C.Y., Wu, H.Y., Ho, S.Y., Chen, P.R., Chuang, C.H., Hsieh, P.J., Wu, Y.S., Chen, W.L., Li, M.J., Wu, Y.C., Huang, X.Y., Ng, F.L., Buddhakosai, W., Huang, P.C., Lan, K.C., Huang, C.Y., Weng, S.L., Cheng, Y.N., Liang, C., Hsu, W.L. & Huang, H.D. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic acids research*, **46**, D296–D302. 98

[58] Cieply, B. & Carstens, R.P. (2015). Functional roles of alternative splicing factors in human disease. *Wiley interdisciplinary reviews. RNA*, **6**, 311–326. 16

[59] Climente-Gonzalez, H., Porta-Pardo, E., Godzik, A. & Eyras, E. (2017). The Functional Impact of Alternative Splicing in Cancer. *Cell reports*, **20**, 2215–2226. 125, 132, 172

[60] Cocquet, J., Chong, A., Zhang, G. & Veitia, R.A. (2006). Reverse transcriptase template switching and false alternative transcripts. *Genomics*, **88**, 127–131. 46

[61] Cokol, M., Nair, R. & Rost, B. (2000). Finding nuclear localization signals. *EMBO reports*, **1**, 411–415. 106

[62] Colak, R., Kim, T., Michaut, M., Sun, M., Irimia, M., Bellay, J., Myers, C.L., Blencowe, B.J. & Kim, P.M. (2013). Distinct Types of Disorder in the Human Proteome: Functional Implications for Alternative Splicing. *PLoS Computational Biology*, **9**, e1003030. 105, 159

[63] Conesa, A., Nueda, M.J., Ferrer, A. & Talon, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics (Oxford, England)*, **22**, 1096–1102. 140

[64] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, **17**, 13. 11, 15, 91

[65] Conforti, L., Anne Buckmaster, E., Tarlton, A., C. Brown, M., F. Lyon, M., Perry, V. & Coleman, M. (1999). *The major brain isoform of Kif1b lacks the putative mitochondria-binding domain*, vol. 10. 236

[66] Consortium, U. (2008). The universal protein resource (UniProt). *Nucleic acids research*, **36**, D190–D195. 92, 103

[67] Cosker, K.E. & Segal, R.A. (????). Neuronal signaling through endocytosis. *Cold Spring Harbor perspectives in biology*, **6**, a020669. 209

[68] Creemers, E.E., Bawazeer, A., Ugalde, A.P., van Deutekom, H.W.M., van der Made, I., de Groot, N.E., Adriaens, M.E., Cook, S.A., Bezzina, C.R., Hubner, N., van der Velden, J., Elkon, R., Agami, R. & Pinto, Y.M. (2016). Genome-Wide Polyadenylation Maps Reveal Dynamic mRNA 3'-End Formation in the Failing Human Heart. *Circulation research*, **118**, 433–438. 177

[69] da Costa, P.J., Menezes, J. & Romao, L. (2017). The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease. *The international journal of biochemistry & cell biology*, **91**, 168–175. 125

[70] Daguenet, E., Dujardin, G. & Valcárcel, J. (2015). The pathogenicity of splicing defects: mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. *EMBO reports*, **16**, 1640–1655. 16

[71] Dassi, E., Re, A., Leo, S., Tebaldi, T., Pasini, L., Peroni, D. & Quattrone, A. (2014). AURA 2: Empowering discovery of post-transcriptional networks. *Translation (Austin, Tex.)*, **2**, e27738. 96

[72] Davis, L.H., Davis, J.Q. & Bennett, V. (1992). Ankyrin regulation: an alternatively spliced segment of the regulatory domain functions as an intramolecular modulator. *The Journal of biological chemistry*, **267**, 18966–18972. 172

[73] Davis-Dusenbery, B.N., Williams, L.A., Klim, J.R. & Eggan, K. (2014). How to make spinal motor neurons. *Development (Cambridge, England)*, **141**, 491–501. 203, 204

[74] Deneen, B., Ho, R., Lukaszewicz, A., Hochstim, C.J., Gronostajski, R.M. & Anderson, D.J. (2006). The transcription factor NFIA controls the onset of gliogenesis in the developing spinal cord. *Neuron*, **52**, 953–968. 187

[75] Deshaies, J.E., Shkreta, L., Moszczynski, A.J., Sidibe, H., Semmler, S., Fouillen, A., Bennett, E.R., Bekenstein, U., Destroismaisons, L., Toutant, J., Delmotte, Q., Volkening, K., Stabile, S., Aulas, A., Khalfallah, Y., Soreq, H., Nanci, A., Strong, M.J., Chabot, B. & Vande Velde, C. (2018). TDP-43 regulates the alternative splicing of hnRNP A1 to yield an aggregation-prone variant in amyotrophic lateral sclerosis. *Brain : a journal of neurology*, **141**, 1320–1333. 187

[76] Di Giammartino, D.C., Nishida, K. & Manley, J.L. (2011). Mechanisms and consequences of alternative polyadenylation. *Molecular cell*, **43**, 853–866. 5

[77] Di Paolo, G., Pellegrini, L., Letinic, K., Cestra, G., Zoncu, R., Voronov, S., Chang, S., Guo, J., Wenk, M.R. & De Camilli, P. (2002). Recruitment and regulation of phosphatidylinositol phosphate kinase type 1 gamma by the FERM domain of talin. *Nature*, **420**, 85–89. 218

[78] Dickey, A.S. & Strack, S. (2011). PKA/AKAP1 and PP2A/Bβ2 regulate neuronal morphogenesis via Drp1 phosphorylation and mitochondrial bioenergetics. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **31**, 15716–15726. 235

[79] Dillman, A.A., Hauser, D.N., Gibbs, J.R., Nalls, M.A., McCoy, M.K., Rudenko, I.N., Galter, D. & Cookson, M.R. (2013). mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex. *Nature Neuroscience*, **16**, 499. 233

[80] Dingwall, C. & Laskey, R.A. (1991). Nuclear targeting sequences — a consensus? *Trends in Biochemical Sciences*, **16**, 478–481. 106

[81] Dinkel, H., Van Roey, K., Michael, S., Kumar, M., Uyar, B., Altenberg, B., Milchevskaya, V., Schneider, M., Kühn, H., Behrendt, A., Dahl, S.L., Damerell, V., Diebel, S., Kalman, S., Klein, S., Knudsen, A.C., Mäder, C., Merrill, S., Staudt, A., Thiel, V., Welti, L., Davey, N.E., Diella, F. & Gibson, T.J. (2016). ELM 2016–data update and new functionality of the eukaryotic linear motif resource. *Nucleic acids research*, **44**, D294–D300. 128

[82] Dixon, S.E., Bhatti, M.M., Uversky, V.N., Dunker, A.K. & Sullivan, W.J. (2011). Regions of intrinsic disorder help identify a novel nuclear localization signal in Toxoplasma gondii histone acetyltransferase TgGCN5-B. *Molecular and biochemical parasitology*, **175**, 192–195. 159

[83] Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, **29**, 15–21. 43, 46

[84] Dong, L., Liu, H., Zhang, J., Yang, S., Kong, G., Chu, J.S.C., Chen, N. & Wang, D. (2015). Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. *BMC Genomics*, **16**, 1039. 9, 125

[85] Dong, Y. & Qiu, G.B. (2017). Biological functions of miR-590 and its role in carcinogenesis. *Frontiers in Laboratory Medicine*, **1**, 173–176. 158

[86] Dredge, B.K. & Darnell, R.B. (2003). Nova regulates GABA(A) receptor gamma2 alternative splicing via a distal downstream UCAU-rich intronic splicing enhancer. *Molecular and cellular biology*, **23**, 4687–4700. 5

[87] Drewe, P., Stegle, O., Hartmann, L., Kahles, A., Bohnert, R., Wachter, A., Borgwardt, K. & Rätsch, G. (2013). Accurate detection of differential RNA processing. *Nucleic Acids Research*, **41**, 5189–5198. 15

[88] Duncan, P.I., Howell, B.W., Marius, R.M., Drmanic, S., Douville, E.M. & Bell, J.C. (1995). Alternative splicing of STY, a nuclear dual specificity kinase. *The Journal of biological chemistry*, **270**, 21524–21531. 221

[89] Duncan, P.I., Stojdl, D.F., Marius, R.M. & Bell, J.C. (1997). In vivo regulation of alternative pre-mRNA splicing by the Clk1 protein kinase. *Molecular and Cellular Biology*, **17**, 5996–6001. 174

[90] Dweep, H. & Gretz, N. (2015). miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nature Methods*, **12**, 697. 98, 99

[91] Eksi, R., Li, H.D., Menon, R., Wen, Y., Omenn, G.S., Kretzler, M. & Guan, Y. (2013). Systematically Differentiating Functions for Alternatively Spliced Isoforms through Integrating RNA-seq Data. *PLOS Computational Biology*, **9**, e1003314. 91

[92] Elia, L.P., Yamamoto, M., Zang, K. & Reichardt, L.F. (2006). p120 Catenin Regulates Dendritic Spine and Synapse Development through Rho-Family GTPases and Cadherins. *Neuron*, **51**, 43–56. 174

[93] Elkon, R., Ugalde, A.P. & Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nature reviews. Genetics*, **14**, 496–506. 18, 131

[94] 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. & Blencowe, B.J. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular cell*, **46**, 884–892. 4, 23, 91, 125, 132, 172, 233, 234

[95] Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of molecular biology*, **300**, 1005–1016. 128

[96] Emig, D., Salomonis, N., Baumbach, J., Lengauer, T., Conklin, B.R. & Albrecht, M. (2010). AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic acids research*, **38**, W755–W762. 132

[97] Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C. & Marks, D.S. (2003). MicroRNA targets in Drosophila. *Genome biology*, **5**, R1. 98

[98] Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C.D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H. & D'Eustachio, P. (2018). The Reactome Pathway Knowledgebase. *Nucleic acids research*, **46**, D649–D655. 91, 111

[99] Fallmann, J., Sedlyarov, V., Tanzer, A., Kovarik, P. & Hofacker, I. (2016). AREsite2: an enhanced database for the comprehensive investigation of AU/GU/U-rich elements. *Nucleic Acids Research*, **44**, D90–D95. 96

[100] Fang, Z. & Rajewsky, N. (2011). The impact of miRNA target sites in coding sequences and in 3'UTRs. *PloS one*, **6**, e18067. 97

[101] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–874. 49

[102] Finn, R.D., Clements, J. & Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, **39**, W29–W37. 103

[103] Flavell, S.W., Kim, T.K., Gray, J.M., Harmin, D.A., Hemberg, M., Hong, E.J., Markenscoff-Papadimitriou, E., Bear, D.M. & Greenberg, M.E. (2008). Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron*, **60**, 1022–1038. 6, 187

[104] Flippo, K.H. & Strack, S. (2017). Mitochondrial dynamics in neuronal injury, development and plasticity. *Journal of cell science*, **130**, 671–681. 235

[105] Floor, S.N. & Doudna, J.A. (2016). Tunable protein synthesis by transcript isoforms in human cells. *eLife*, **5**, e10921. 18

[106] Franco, S.J., Rodgers, M.A., Perrin, B.J., Han, J., Bennin, D.A., Critchley, D.R. & Huttenlocher, A. (2004). Calpain-mediated proteolysis of talin regulates adhesion dynamics. *Nature cell biology*, **6**, 977–983. 218

[107] Fu, M.m. & Holzbaur, E.L.F. (2014). MAPK8IP1/JIP1 regulates the trafficking of autophagosomes in neurons. *Autophagy*, **10**, 2079–2081. 220

[108] Fu, X.D. & Ares Jr, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, **15**, 689. 5

[109] Fu, Y., Chen, L., Chen, C., Ge, Y., Kang, M., Song, Z., Li, J., Feng, Y., Huo, Z., He, G., Hou, M., Chen, S. & Xu, A. (2018). Crosstalk between alternative polyadenylation and miRNAs in the regulation of protein translational efficiency. *Genome research*, **28**, 1656–1663. 132

[110] Fuchs-Telem, D., Stewart, H., Rapaport, D., Nousbeck, J., Gat, A., Gini, M., Lugassy, Y., Emmert, S., Eckl, K., Hennies, H.C., Sarig, O., Goldsher, D., Meilik, B., Ishida-Yamamoto, A., Horowitz, M. & Sprecher, E. (2011). CEDNIK syndrome results from loss-of-function mutations in SNAP29. *The British journal of dermatology*, **164**, 610–616. 233

[111] Gaiano, N., Nye, J.S. & Fishell, G. (2000). Radial glial identity is promoted by Notch1 signaling in the murine forebrain. *Neuron*, **26**, 395–404. 187

[112] Gao, S., Ren, Y., Sun, Y., Wu, Z., Ruan, J., He, B., Zhang, T., Yu, X., Tian, X. & Bu, W. (2016). PacBio full-length transcriptome profiling of insect mitochondrial gene expression. *RNA Biology*, **13**, 820–825. 9, 125

[113] Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., McNeill, L., Wallace, E.J., Jayasinghe, L., Wright, C., Blasco, J., Young, S., Brocklebank, D., Juul, S., Clarke, J., Heron, A.J. & Turner, D.J. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, **15**, 201. 9

[114] Ge, Y. & Porse, B.T. (2014). The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression. *BioEssays*, **36**, 236–243. 231

[115] Gehman, L.T., Meera, P., Stoilov, P., Shiue, L., O'Brien, J.E., Meisler, M.H., Ares, M.J., Otis, T.S. & Black, D.L. (2012). The splicing regulator Rbfox2 is required for both cerebellar development and mature motor function. *Genes & development*, **26**, 445–460. 5

[116] Gibilisco, L., Zhou, Q., Mahajan, S. & Bachtrog, D. (2016). Alternative Splicing within and between Drosophila Species, Sexes, Tissues, and Developmental Stages. *PLOS Genetics*, **12**, e1006464. 3

[117] Giudice, J., Xia, Z., Wang, E.T., Scavuzzo, M.A., Ward, A.J., Kalsotra, A., Wang, W., Wehrens, X.H.T., Burge, C.B., Li, W. & Cooper, T.A. (2014). Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. *Nature communications*, **5**, 3603. 233

[118] Giudice, J., Loehr, J.A., Rodney, G.G. & Cooper, T.A. (2016). Alternative Splicing of Four Trafficking Genes Regulates Myofiber Structure and Skeletal Muscle Physiology. *Cell reports*, **17**, 1923–1933. 233

[119] Glisovic, T., Bachorik, J.L., Yong, J. & Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, **582**, 1977–1986. 101

[120] Goldman, S.A. & Kuypers, N.J. (2015). How to make an oligodendrocyte. *Development (Cambridge, England)*, **142**, 3983–3995. 203

[121] Gomez, S., H Lo, S. & Rzhetsky, A. (2001). *Probabilistic Prediction of Unknown Metabolic and Signal-Transduction Networks*, vol. 159. 114

[122] Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology*, **14**, R70. 141

[123] Gordon, S.P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., Kang, D., Underwood, J., Grigoriev, I.V., Figueroa, M., Schilling, J.S., Chen, F. & Wang, Z. (2015). Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. *PloS one*, **10**, e0132628. 12, 41

[124] Goulet, I., Boisvenue, S., Mokas, S., Mazroui, R. & Cote, J. (2008). TDRD3, a novel Tudor domain-containing protein, localizes to cytoplasmic stress granules. *Human molecular genetics*, **17**, 3055–3074. 180

[125] Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. & Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, **29**, 644–652. 7

[126] Gueroussov, S., Weatheritt, R.J., O'Hanlon, D., Lin, Z.Y., Narula, A., Gingras, A.C. & Blencowe, B.J. (2017). Regulatory Expansion in Mammals of Multivalent hnRNP Assemblies that Globally Control Alternative Splicing. *Cell*, **170**, 324–339.e23. 105, 125

[127] Gunasekara, C., Subramanian, A., Avvari, J.V.R.K., Li, B., Chen, S. & Wei, H. (2016). ExactSearch: a web-based plant motif search tool. *Plant methods*, **12**, 26. 127

[128] Guo, H., Ingolia, N.T., Weissman, J.S. & Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**, 835–840. 97

[129] Ha, K.C.H., Blencowe, B.J. & Morris, Q. (2018). QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biology*, **19**, 45. 148, 187

[130] Hackl, T., Hedrich, R., Schultz, J. & Forster, F. (2014). proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics (Oxford, England)*, **30**, 3004–3011. 11, 41, 61

[131] Hall, M.N., Gabay, J., Débarbouillé, M. & Schwartz, M. (1982). A role for mRNA secondary structure in the control of translation initiation. *Nature*, **295**, 616–618. 19

[132] Han, H., Irimia, M., Ross, P.J., Sung, H.K., Alipanahi, B., David, L., Golipour, A., Gabut, M., Michael, I.P., Nachman, E.N., Wang, E., Trcka, D., Thompson, T., O'Hanlon, D., Slobodeniuc, V., Barbosa-Morais, N.L., Burge, C.B., Moffat, J., Frey, B.J., Nagy, A., Ellis, J., Wrana, J.L. & Blencowe, B.J. (2013). MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature*, **498**, 241–245. 5

[133] HANNIGAN, M.M., ZAGORE, L.L. & LICATALOSI, D.D. (2017). Ptbp2 Controls an Alternative Splicing Network Required for Cell Communication during Spermatogenesis. *Cell reports*, **19**, 2598–2612. 233

[134] HANSEN, T.B., KJEMS, J. & BRAMSEN, J.B. (2011). Enhancing miRNA annotation confidence in miRBase by continuous cross dataset analysis. *RNA biology*, **8**, 378–383. 98

[135] HARDY, J. & NORBURY, C. (2016). Cleavage factor Im (CFIm) as a regulator of alternative polyadenylation. *Biochemical Society Transactions*, **44**, 1051–1057. 6

[136] HARGREAVES, A.D. & MULLEY, J.F. (2015). Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. *PeerJ*, **3**, e1441. 9

[137] HARREMAN, M.T., KLINE, T.M., MILFORD, H.G., HARBEN, M.B., HODEL, A.E. & CORBETT, A.H. (2004). Regulation of nuclear import by phosphorylation adjacent to nuclear localization signals. *The Journal of biological chemistry*, **279**, 20613–20621. 175

[138] HAUN, F., NAKAMURA, T., SHIU, A.D., CHO, D.H., TSUNEMI, T., HOLLAND, E.A., LA SPADA, A.R. & LIPTON, S.A. (2013). S-nitrosylation of dynamin-related protein 1 mediates mutant huntingtin-induced mitochondrial fragmentation and neuronal injury in Huntington's disease. *Antioxidants & redox signaling*, **19**, 1173–1184. 235

[139] HERMJAKOB, H., MONTECCHI-PALAZZI, L., LEWINGTON, C., MUDALI, S., KERRIEN, S., ORCHARD, S., VINGRON, M., ROECHERT, B., ROEPSTORFF, P., VALENCIA, A., MARGALIT, H., ARMSTRONG, J., BAIROCH, A., CESARENI, G., SHERMAN, D. & APWEILER, R. (2004). IntAct: an open source molecular interaction database. *Nucleic acids research*, **32**, D452–D455. 128

[140] HILGERS, V., PERRY, M.W., HENDRIX, D., STARK, A., LEVINE, M. & HALEY, B. (2011). Neural-specific elongation of 3' UTRs during Drosophila development. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 15864–15869. 234

[141] HOFFENBERG, S., LIU, X., NIKOLOVA, L., HALL, H.S., DAI, W., BAUGHN, R.E., DICKEY, B.F., BARBIERI, M.A., ABALLAY, A., STAHL, P.D. & KNOLL, B.J. (2000). A novel membrane-anchored Rab5 interacting protein required for homotypic endosome fusion. *The Journal of biological chemistry*, **275**, 24661–24669. 178

[142] HOFFMAN, Y., BUBLIK, D.R., UGALDE, A.P., ELKON, R., BINIASHVILI, T., AGAMI, R., OREN, M. & PILPEL, Y. (2016). 3'UTR Shortening Potentiates MicroRNA-Based Repression of Pro-differentiation Genes in Proliferating Human Cells. *PLoS genetics*, **12**, e1005879–e1005879. 23, 132

[143] HORNBECK, P.V., ZHANG, B., MURRAY, B., KORNHAUSER, J.M., LATHAM, V. & SKRZYPEK, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic acids research*, **43**, D512–20. 103, 107

[144] HORTON, P., PARK, K.J., OBAYASHI, T., FUJITA, N., HARADA, H., ADAMS-COLLIER, C.J. & NAKAI, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic acids research*, **35**, W585–W587. 106

[145] HOUSELEY, J. & TOLLERVEY, D. (2010). Apparent Non-Canonical Trans-Splicing Is Generated by Reverse Transcriptase In Vitro. *PLOS ONE*, **5**, e12271. 46

[146] HSU, P.W.C., HUANG, H.D., HSU, S.D., LIN, L.Z., TSOU, A.P., TSENG, C.P., STADLER, P.F., WASHIETL, S. & HOFACKER, I.L. (2006). miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes. *Nucleic acids research*, **34**, D135–9. 98

[147] HYUNG, D., KIM, J., CHO, S.Y. & PARK, C. (2018). ASpedia: a comprehensive encyclopedia of human alternative splicing. *Nucleic Acids Research*, **46**, D58–D63. 92, 125

[148] IAKOUCHEVA, L.M., RADIVOJAC, P., BROWN, C.J., O'CONNOR, T.R., SIKES, J.G., OBRADOVIC, Z. & DUNKER, A.K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic acids research*, **32**, 1037–1049. 159

[149] 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. & BLENCOWE, B.J. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*, **159**, 1511–1523. 49, 131, 187, 233

[150] JAFFREY, S.R. & WILKINSON, M.F. (2018). Nonsense-mediated RNA decay in the brain: emerging modulator of neural development and disease. *Nature Reviews Neuroscience*, **19**, 715–728. 132

[151] JAMBHEKAR, A. & DERISI, J.L. (2007). Cis-acting determinants of asymmetric, cytoplasmic RNA transport. *RNA (New York, N.Y.)*, **13**, 625–642. 92, 94

[152] Jayashankar, V. & Rafelski, S.M. (2014). Integrating mitochondrial organization and dynamics with cellular architecture. *Current opinion in cell biology*, **26**, 34–40. 235

[153] Ji, Z., Lee, J.Y., Pan, Z., Jiang, B. & Tian, B. (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 7028–7033. 6, 234

[154] Jin, L., Li, G., Yu, D., Huang, W., Cheng, C., Liao, S., Wu, Q. & Zhang, Y. (2017). Transcriptome analysis reveals the complexity of alternative splicing regulation in the fungus Verticillium dahliae. *BMC genomics*, **18**, 130. 3

[155] Johnson, M.B., Kawasawa, Y.I., Mason, C.E., Krsnik, Z., Coppola, G., Bogdanovic, D., Geschwind, D.H., Mane, S.M., State, M.W. & Sestan, N. (2009). Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron*, **62**, 494–509. 187

[156] Jongjitwimol, J., Baldock, R.A., Morley, S.J. & Watts, F.Z. (2016). Sumoylation of eIF4A2 affects stress granule formation. *Journal of cell science*, **129**, 2407–2415. 180

[157] Jorgensen, H.F., Ben-Porath, I. & Bird, A.P. (2004). Mbd1 is recruited to both methylated and nonmethylated CpGs via distinct DNA binding domains. *Molecular and cellular biology*, **24**, 3387–3395. 172

[158] Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**, 27–30. 91, 227

[159] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic acids research*, **44**, D457–62. 91

[160] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, **45**, D353–D361. 91

[161] Kang, P., Lee, H.K., Glasgow, S.M., Finley, M., Donti, T., Gaber, Z.B., Graham, B.H., Foster, A.E., Novitch, B.G., Gronostajski, R.M. & Deneen, B. (2012). Sox9 and NFIA coordinate a transcriptional regulatory cascade during the initiation of gliogenesis. *Neuron*, **74**, 79–94. 187

[162] Kashima, I., Jonas, S., Jayachandran, U., Buchwald, G., Conti, E., Lupas, A.N. & Izaurralde, E. (2010). SMG6 interacts with the exon junction complex via two conserved EJC-binding motifs (EBMs) required for nonsense-mediated mRNA decay. *Genes & development*, **24**, 2440–2450. 180

[163] Katzenellenbogen, R.A., Vliet-Gregg, P., Xu, M. & Galloway, D.A. (2009). NFX1-123 Increases hTERT Expression and Telomerase Activity Posttranscriptionally in Human Papillomavirus Type 16 E6 Keratinocytes . *Journal of Virology*, **83**, 6446–6456. 172

[164] Keirstead, H.S., Nistor, G., Bernal, G., Totoiu, M., Cloutier, F., Sharp, K. & Steward, O. (2005). Human embryonic stem cell-derived oligodendrocyte progenitor cell transplants remyelinate and restore locomotion after spinal cord injury. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **25**, 4694–4705. 38

[165] Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M. & Stamm, S. (2013). Function of alternative splicing. *Gene*, **514**, 1–30. 16, 17, 18, 23, 92, 125, 131

[166] Kent, W.J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, **12**, 656–664. 40

[167] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nature genetics*, **39**, 1278–1284. 98

[168] Kino, Y., Washizu, C., Kurosawa, M., Oma, Y., Hattori, N., Ishiura, S. & Nukina, N. (2015). Nuclear localization of MBNL1: splicing-mediated autoregulation and repression of repeat-derived aberrant proteins. *Human molecular genetics*, **24**, 740–756. 174

[169] Klee, E.W. & Ellis, L.B.M. (2005). Evaluating eukaryotic secreted protein prediction. *BMC Bioinformatics*, **6**, 256. 104

[170] Klepikova, A.V., Kasianov, A.S., Gerasimov, E.S., Logacheva, M.D. & Penin, A.A. (2016). A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *The Plant journal : for cell and molecular biology*, **88**, 1058–1070. 3

[171] König, J., Zarnack, K., Luscombe, N.M. & Ule, J. (2012). Protein–RNA interactions: new genomic technologies and perspectives. *Nature Reviews Genetics*, **13**, 77. 101

[172] Kosugi, S., Hasebe, M., Tomita, M. & Yanagawa, H. (2009). Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 10171–10176. 106

[173] Kozomara, A. & Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, **42**, D68–D73. 99

[174] Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M. & Rajewsky, N. (2005). Combinatorial microRNA target predictions. *Nature genetics*, **37**, 495–500. 98

[175] Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*, **305**, 567–580. 104

[176] Kruger, J. & Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic acids research*, **34**, W451–4. 98

[177] Kuhn, D.E., Martin, M.M., Feldman, D.S., Terry, A.V.J., Nuovo, G.J. & Elton, T.S. (2008). Experimental validation of miRNA targets. *Methods (San Diego, Calif.)*, **44**, 47–54. 98, 99

[178] Kuo, R.I., Tseng, E., Eory, L., Paton, I.R., Archibald, A.L. & Burt, D.W. (2017). Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics*, **18**, 323. 9, 125, 190

[179] Kurihara, Y., Makita, Y., Kawashima, M., Fujita, T., Iwasaki, S. & Matsui, M. (2018). Transcripts from downstream alternative transcription start sites evade uORF-mediated inhibition of gene expression in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 7831–7836. 132

[180] Lareau, L.F. & Brenner, S.E. (2015). Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Molecular biology and evolution*, **32**, 1072–1079. 231

[181] Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C. & Brenner, S.E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, **446**, 926–929. 231

[182] Lau, A.G., Irier, H.A., Gu, J., Tian, D., Ku, L., Liu, G., Xia, M., Fritsch, B., Zheng, J.Q., Dingledine, R., Xu, B., Lu, B. & Feng, Y. (2010). Distinct 3'UTRs differentially regulate activity-dependent translation of brain-derived neurotrophic factor (BDNF). *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 15945–15950. 92, 94

[183] Le, O., Cho, O.Y., Tran, M., Ah Kim, J., Chang, S., Jou, I. & Yoon Lee, S. (2015). *Phosphorylation of phosphatidylinositol 4-phosphate 5-kinase γ by Akt regulates its interaction with talin and focal adhesion dynamics*, vol. 1853. 218

[184] Lee, C. & Roy, M. (2004). Analysis of alternative splicing with microarrays: successes and challenges. *Genome biology*, **5**, 231. 7

[185] Levine, J.M., Reynolds, R. & Fawcett, J.W. (2001). The oligodendrocyte precursor cell in health and disease. *Trends in neurosciences*, **24**, 39–47. 202, 203

[186] Levy, S.F., Leboeuf, A.C., Massie, M.R., Jordan, M.A., Wilson, L. & Feinstein, S.C. (2005). Three- and four-repeat tau regulate the dynamic instability of two distinct microtubule subpopulations in qualitatively different manners. Implications for neurodegeneration. *The Journal of biological chemistry*, **280**, 13520–13528. 171

[187] Lewis, B.P., Burge, C.B. & Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. 98

[188] Li, B. & Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323. 14, 43, 154, 187, 191

[189] Li, H.D., Menon, R., Omenn, G.S. & Guan, Y. (2014). Revisiting the identification of canonical splice isoforms through integration of functional genomics and proteomics evidence. *Proteomics*, **14**, 2709–2718. 98

[190] Li, H.D., Omenn, G.S. & Guan, Y. (2016). A proteogenomic approach to understand splice isoform functions through sequence and expression-based computational modeling. *Briefings in bioinformatics*, **17**, 1024–1031. 91

[191] Li, W., Kang, S., Liu, C.C., Zhang, S., Shi, Y., Liu, Y. & Zhou, X.J. (2014). High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. *Nucleic Acids Research*, **42**, e39–e39. 91

[192] Li, W., Liu, C.C., Kang, S., Li, J.R., Tseng, Y.T. & Zhou, X.J. (2016). Pushing the annotation of cellular activities to a higher resolution: Predicting functions at the isoform level. *Methods*, **93**, 110–118. 91

[193] Li, X., Garrity, A.G. & Xu, H. (2013). Regulation of membrane trafficking by signalling on endosomal and lysosomal membranes. *The Journal of physiology*, **591**, 4389–4401. 210

[194] Li, Z., Okamoto, K.I., Hayashi, Y. & Sheng, M. (2004). The importance of dendritic mitochondria in the morphogenesis and plasticity of spines and synapses. *Cell*, **119**, 873–887. 235

[195] Lindeboom, R.G.H., Supek, F. & Lehner, B. (2016). The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nature genetics*, **48**, 1112–1118. 103

[196] Liu, Y., Gonzàlez-Porta, M., Santos, S., Brazma, A., Marioni, J.C., Aebersold, R., Venkitaraman, A.R. & Wickramasinghe, V.O. (2017). Impact of Alternative Splicing on the Human Proteome. *Cell Reports*, **20**, 1229–1241. 18, 101

[197] Llorian, M., Gooding, C., Bellora, N., Hallegger, M., Buckroyd, A., Wang, X., Rajgor, D., Kayikci, M., Feltham, J., Ule, J., Eyras, E. & Smith, C.W.J. (2016). The alternative splicing program of differentiated smooth muscle cells involves concerted non-productive splicing of post-transcriptional regulators. *Nucleic acids research*, **44**, 8933–8950. 131

[198] Lunde, B.M., Moore, C. & Varani, G. (2007). RNA-binding proteins: modular design for efficient function. *Nature Reviews Molecular Cell Biology*, **8**, 479. 101

[199] Luo, T., Zhang, W., Qiu, S., Yang, Y., Yi, D., Wang, G., Ye, J. & Wang, J. (2017). *Functional Annotation of Human Protein Coding Isoforms via Non-convex Multi-Instance Learning*. 91

[200] Lupas, A., Van Dyke, M. & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science (New York, N.Y.)*, **252**, 1162–1164. 105

[201] Lykke-Andersen, S. & Jensen, T.H. (2015). Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nature Reviews Molecular Cell Biology*, **16**, 665. 19, 125

[202] Maier, O., Hoekstra, D. & Baron, W. (2008). Polarity development in oligodendrocytes: sorting and trafficking of myelin components. *Journal of molecular neuroscience : MN*, **35**, 35–53. 233

[203] Marfori, M., Mynott, A., Ellis, J.J., Mehdi, A.M., Saunders, N.F.W., Curmi, P.M., Forwood, J.K., Bodén, M. & Kobe, B. (2011). Molecular basis for specificity of nuclear import and prediction of nuclear localization. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, **1813**, 1562–1577. 106

[204] Marquez, Y., Brown, J.W.S., Simpson, C., Barta, A. & Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome research*, **22**, 1184–1195. 3

[205] Martelli, P.L., D'Antonio, M., Bonizzoni, P., Castrignanò, T., D'Erchia, A.M., D'Onorio De Meo, P., Fariselli, P., Finelli, M., Licciulli, F., Mangiulli, M., Mignone, F., Pavesi, G., Picardi, E., Rizzi, R., Rossi, I., Valletti, A., Zauli, A., Zambelli, F., Casadio, R. & Pesole, G. (2011). ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. *Nucleic Acids Research*, **39**, D80–D85. 92, 125

[206] Mayr, C. & Bartel, D.P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684. 23, 92, 94

[207] Merino, G.A., Conesa, A. & Fernandez, E.A. (2017). A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies. *Briefings in bioinformatics*. 142

[208] Mignone, F., Grillo, G., Licciulli, F., Iacono, M., Liuni, S., Kersey, P.J., Duarte, J., Saccone, C. & Pesole, G. (2005). UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Research*, **33**, D141–D146. 96, 97, 102

[209] Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B. & Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217. 98

[210] MISKO, A.L., SASAKI, Y., TUCK, E., MILBRANDT, J. & BALOH, R.H. (2012). Mitofusin2 mutations disrupt axonal mitochondrial positioning and promote axon degeneration. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **32**, 4145–4155. 235

[211] MIURA, P., SHENKER, S., ANDREU-AGULLO, C., WESTHOLM, J.O. & LAI, E.C. (2013). Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome research*, **23**, 812–825. 234

[212] MODREK, B. & LEE, C. (2002). A genomic view of alternative splicing. *Nature Genetics*, **30**, 13. 7

[213] MOLL, T., TEBB, G., SURANA, U., ROBITSCH, H. & NASMYTH, K. (1991). The role of phosphorylation and the CDC28 protein kinase in cell cycle-regulated nuclear import of the S. cerevisiae transcription factor SWI5. *Cell*, **66**, 743–758. 175

[214] MOLLER, S., CRONING, M.D. & APWEILER, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics (Oxford, England)*, **17**, 646–653. 104

[215] MONTANER, D. & DOPAZO, J. (2010). Multidimensional Gene Set Analysis of Genomic Data. *PLOS ONE*, **5**, e10348. 193

[216] MORENO-MANZANO, V., RODRIGUEZ-JIMENEZ, F.J., GARCIA-ROSELLO, M., LAINEZ, S., ERCEG, S., CALVO, M.T., RONAGHI, M., LLORET, M., PLANELLS-CASES, R., SANCHEZ-PUELLES, J.M. & STOJKOVIC, M. (2009). Activated spinal cord ependymal stem cells rescue neurological function. *Stem cells (Dayton, Ohio)*, **27**, 733–743. 189

[217] MOSCA, R., CÉOL, A. & ALOY, P. (2012). Interactome3D: adding structural details to protein networks. *Nature Methods*, **10**, 47. 128

[218] MOTTI, D., LERCH, J.K., DANZI, M.C., GANS, J.H., KUO, F., SLEPAK, T.I., BIXBY, J.L. & LEMMON, V.P. (2017). Identification of miRNAs involved in DRG neurite outgrowth and their putative targets. 225

[219] MUNJI, R.N., CHOE, Y., LI, G., SIEGENTHALER, J.A. & PLEASURE, S.J. (2011). Wnt signaling regulates neuronal differentiation of cortical intermediate progenitors. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **31**, 1676–1687. 174

[220] NAGALAKSHMI, U., WAERN, K. & SNYDER, M. (2010). RNA-Seq: A Method for Comprehensive Transcriptome Analysis. *Current Protocols in Molecular Biology*, **89**, 4.11.1–4.11.13. 7

[221] NAM, D.K., LEE, S., ZHOU, G., CAO, X., WANG, C., CLARK, T., CHEN, J., ROWLEY, J.D. & WANG, S.M. (2002). Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6152–6156. 47

[222] NARDOZZI, J.D., LOTT, K. & CINGOLANI, G. (2010). Phosphorylation meets nuclear import: a review. *Cell communication and signaling : CCS*, **8**, 32. 175

[223] NAYLER, O., STAMM, S. & ULLRICH, A. (1997). Characterization and comparison of four serine- and arginine-rich (SR) protein kinases. *The Biochemical journal*, **326 (Pt 3**, 693–700. 174

[224] NECCI, M., PIOVESAN, D., DOSZTANYI, Z. & TOSATTO, S.C.E. (2017). MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics (Oxford, England)*, **33**, 1402–1404. 105

[225] NGONDO, R.P. & CARBON, P. (2014). ZNF143 is regulated through alternative 3'UTR isoforms. *Biochimie*, **104**, 137–146. 158

[226] NGUYEN BA, A.N., POGOUTSE, A., PROVART, N. & MOSES, A.M. (2009). NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC bioinformatics*, **10**, 202. 106

[227] NI, J.Z., GRATE, L., DONOHUE, J.P., PRESTON, C., NOBIDA, N., O'BRIEN, G., SHIUE, L., CLARK, T.A., BLUME, J.E. & ARES, M.J. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes & development*, **21**, 708–718. 231

[228] NI, T., YANG, Y., HAFEZ, D., YANG, W., KIESEWETTER, K., WAKABAYASHI, Y., OHLER, U., PENG, W. & ZHU, J. (2013). Distinct polyadenylation landscapes of diverse human tissues revealed by a modified PA-seq strategy. *BMC Genomics*, **14**, 615. 6

[229] NIESCIER, R.F., CHANG, K.T. & MIN, K.T. (2013). Miro, MCU, and calcium: bridging our understanding of mitochondrial movement in axons. *Frontiers in cellular neuroscience*, **7**, 148. 235

[230] NIU, L., HUANG, W., UMBACH, D.M. & LI, L. (2014). IUTA: a tool for effectively detecting differential isoform usage from RNA-Seq data. *BMC Genomics*, **15**, 862. 14

[231] NORRIS, A.D. & CALARCO, J.A. (2012). Emerging Roles of Alternative Pre-mRNA Splicing Regulation in Neuronal Development and Function. *Frontiers in neuroscience*, **6**, 122. 187

[232] NOVIKOV, L., PARK, J.W., CHEN, H., KLERMAN, H., JALLOH, A.S. & GAMBLE, M.J. (2011). QKI-mediated alternative splicing of the histone variant MacroH2A1 regulates cancer cell proliferation. *Molecular and cellular biology*, **31**, 4244–4255. 172

[233] NUEDA, M.J., MARTORELL-MARUGAN, J., MARTI, C., TARAZONA, S. & CONESA, A. (2018). Identification and visualization of differential isoform expression in RNA-seq time series. *Bioinformatics (Oxford, England)*, **34**, 524–526. 143

[234] OIKONOMOPOULOS, S., WANG, Y.C., DJAMBAZIAN, H., BADESCU, D. & RAGOUSSIS, J. (2016). Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Scientific Reports*, **6**, 31602. 9, 35

[235] O'LEARY, N.A., WRIGHT, M.W., BRISTER, J.R., CIUFO, S., HADDAD, D., MCVEIGH, R., RAJPUT, B., ROBBERTSE, B., SMITH-WHITE, B., AKO-ADJEI, D., ASTASHYN, A., BADRETDIN, A., BAO, Y., BLINKOVA, O., BROVER, V., CHETVERNIN, V., CHOI, J., COX, E., ERMOLAEVA, O., FARRELL, C.M., GOLDFARB, T., GUPTA, T., HAFT, D., HATCHER, E., HLAVINA, W., JOARDAR, V.S., KODALI, V.K., LI, W., MAGLOTT, D., MASTERSON, P., MCGARVEY, K.M., MURPHY, M.R., O'NEILL, K., PUJAR, S., RANGWALA, S.H., RAUSCH, D., RIDDICK, L.D., SCHOCH, C., SHKEDA, A., STORZ, S.S., SUN, H., THIBAUD-NISSEN, F., TOLSTOY, I., TULLY, R.E., VATSAN, A.R., WALLIN, C., WEBB, D., WU, W., LANDRUM, M.J., KIMCHI, A., TATUSOVA, T., DICUCCIO, M., KITTS, P., MURPHY, T.D. & PRUITT, K.D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, **44**, D733–45. 42

[236] OSBORNE, S.L., MEUNIER, F.A. & SCHIAVO, G. (2001). Phosphoinositides as Key Regulators of Synaptic Function. *Neuron*, **32**, 9–12. 218, 234

[237] PAN, Q., SHAI, O., LEE, L.J., FREY, B.J. & BLENCOWE, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, **40**, 1413. 3, 7

[238] PAN, Z., WANG, B., ZHANG, Y., WANG, Y., ULLAH, S., JIAN, R., LIU, Z. & XUE, Y. (2015). dbPSP: a curated database for protein phosphorylation sites in prokaryotes. *Database : the journal of biological databases and curation*, **2015**, bav031. 108

[239] PANWAR, B., MENON, R., EKSI, R., LI, H.D., OMENN, G.S. & GUAN, Y. (2016). Genome-Wide Functional Annotation of Human Protein-Coding Splice Variants Using Multiple Instance Learning. *Journal of Proteome Research*, **15**, 1747–1753. 91

[240] PARADA, G.E., MUNITA, R., CERDA, C.A. & GYSLING, K. (2014). A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic acids research*, **42**, 10564–10578. 45

[241] PARASKEVOPOULOU, M.D., GEORGAKILAS, G., KOSTOULAS, N., VLACHOS, I.S., VERGOULIS, T., RECZKO, M., FILIPPIDIS, C., DALAMAGAS, T. & HATZIGEORGIOU, A.G. (2013). DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic acids research*, **41**, W169–W173. 98

[242] PARK, E., PAN, Z., ZHANG, Z., LIN, L. & XING, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *American journal of human genetics*, **102**, 11–26. 3, 4, 5

[243] PARK, H.C. & APPEL, B. (2003). Delta-Notch signaling regulates oligodendrocyte specification. *Development*, **130**, 3747–3755. 187

[244] PATHANIA, M., DAVENPORT, E.C., MUIR, J., SHEEHAN, D.F., LOPEZ-DOMENECH, G. & KITTLER, J.T. (2014). The autism and schizophrenia associated gene CYFIP1 is critical for the maintenance of dendritic complexity and the stabilization of mature spines. *Translational psychiatry*, **4**, e374. 227

[245] PATRO, R., MOUNT, S.M. & KINGSFORD, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, **32**, 462. 14

[246] PATRO, R., DUGGAL, G., LOVE, M.I., IRIZARRY, R.A. & KINGSFORD, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, **14**, 417. 14

[247] PAUWS, E., VAN KAMPEN, A.H., VAN DE GRAAF, S.A., DE VIJLDER, J.J. & RIS-STALPERS, C. (2001). Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. *Nucleic acids research*, **29**, 1690–1694. 136

[248] PECCARELLI, M. & KEBAARA, B.W. (2014). Regulation of Natural mRNAs by the Nonsense-Mediated mRNA Decay Pathway. *Eukaryotic Cell*, **13**, 1126–1135. 19

[249] PENG, J., REN, K.D., YANG, J. & LUO, X.J. (2016). Mitochondrial E3 ubiquitin ligase 1: A key enzyme in regulation of mitochondrial dynamics and functions. *Mitochondrion*, **28**, 49–53. 222

[250] PENTONY, M.M. & JONES, D.T. (2010). Modularity of intrinsic disorder in the human proteome. *Proteins*, **78**, 212–221. 105

[251] PERCIFIELD, R., MURPHY, D. & STOILOV, P. (2014). Medium throughput analysis of alternative splicing by fluorescently labeled RT-PCR. *Methods in molecular biology (Clifton, N.J.)*, **1126**, 299–313. 7

[252] PESOLE, G., LIUNI, S. & D'SOUZA, M. (2000). *PatSearch: A pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance*, vol. 16. 96, 102

[253] PETERSEN, T.N., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. 104

[254] PIPER, R.C., DIKIC, I. & LUKACS, G.L. (????). Ubiquitin-dependent sorting in endocytosis. *Cold Spring Harbor perspectives in biology*, **6**, a016808. 210

[255] PLEISS, J., WHITWORTH, G., BERGKESSEL, M. & GUTHRIE, C. (2007). *Rapid, Transcript-Specific Changes in Splicing in Response to Environmental Stress*, vol. 27. 16

[256] POLITO, A. & REYNOLDS, R. (2005). NG2-expressing cells as oligodendrocyte progenitors in the normal and demyelinated adult central nervous system. *Journal of anatomy*, **207**, 707–716. 203

[257] POON, I.K.H. & JANS, D.A. (2005). Regulation of Nuclear Transport: Central Role in Development and Transformation? *Traffic*, **6**, 173–186. 175

[258] POTENZA, E., DOMENICO, T.D., WALSH, I. & TOSATTO, S.C.E. (2015). MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Research*, **43**, D315–D320. 105

[259] PRIDDLE, H., HEMMINGS, L., MONKLEY, S., WOODS, A., PATEL, B., SUTTON, D., DUNN, G.A., ZICHA, D. & CRITCHLEY, D.R. (1998). Disruption of the talin gene compromises focal adhesion assembly in undifferentiated but not differentiated embryonic stem cells. *The Journal of cell biology*, **142**, 1121–1133. 218

[260] PROUDFOOT, N. & O'SULLIVAN, J. (2018). Polyadenylation: A tail of two complexes. *Current Biology*, **12**, R855–R857. 5, 102

[261] PUNTA, M., COGGILL, P.C., EBERHARDT, R.Y., MISTRY, J., TATE, J., BOURSNELL, C., PANG, N., FORSLUND, K., CERIC, G., CLEMENTS, J., HEGER, A., HOLM, L., SONNHAMMER, E.L.L., EDDY, S.R., BATEMAN, A. & FINN, R.D. (2012). The Pfam protein families database. *Nucleic acids research*, **40**, D290–301. 103

[262] PYL, P.T., ANDERS, S. & HUBER, W. (2014). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169. 15

[263] QUEVILLON, E., SILVENTOINEN, V., PILLAI, S., HARTE, N., MULDER, N., APWEILER, R. & LOPEZ, R. (2005). InterProScan: protein domains identifier. *Nucleic acids research*, **33**, W116–20. 103

[264] QUIROS, P.M., RAMSAY, A.J., SALA, D., FERNANDEZ-VIZARRA, E., RODRIGUEZ, F., PEINADO, J.R., FERNANDEZ-GARCIA, M.S., VEGA, J.A., ENRIQUEZ, J.A., ZORZANO, A. & LOPEZ-OTIN, C. (2012). Loss of mitochondrial protease OMA1 alters processing of the GTPase OPA1 and causes obesity and defective thermogenesis in mice. *The EMBO journal*, **31**, 2117–2133. 222

[265] RAJ, B. & BLENCOWE, B.J. (2015). Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron*, **87**, 14–27. 131, 187

[266] RAJ, B., O'HANLON, D., VESSEY, J.P., PAN, Q., RAY, D., BUCKLEY, N.J., MILLER, F.D. & BLENCOWE, B.J. (2011). Cross-regulation between an alternative splicing activator and a transcription repressor controls neurogenesis. *Molecular cell*, **43**, 843–850. 187

[267] RAMAKRISHNAN, N.A., DRESCHER, M.J. & DRESCHER, D.G. (2012). The SNARE complex in neuronal and sensory cells. *Molecular and cellular neurosciences*, **50**, 58–69. 210

[268] RAMASARMA, T., JOSHI, N.V., SEKAR, K., UTHAYAKUMAR, M. & SHERLIN, D. (2012). Transmembrane Domains. 104

[269] RAMÍREZ-SÁNCHEZ, O., PÉREZ-RODRÍGUEZ, P., DELAYE, L. & TIESSEN, A. (2016). Plant Proteins Are Smaller Because They Are Encoded by Fewer Exons than Animal Proteins. *Genomics, Proteomics & Bioinformatics*, **14**, 357–370. 114

[270] RAN, Q. & PEREIRA-SMITH, O.M. (2000). Identification of an alternatively spliced form of the Tat interactive protein (Tip60), Tip60(beta). *Gene*, **258**, 141–146. 175

[271] RAPAPORT, D., LUGASSY, Y., SPRECHER, E. & HOROWITZ, M. (2010). Loss of SNAP29 impairs endocytic recycling and cell motility. *PloS one*, **5**, e9759. 233

[272] REYES, A. & HUBER, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic acids research*, **46**, 582–592. 234

[273] RHOADS, A. & AU, K.F. (2015). PacBio Sequencing and Its Applications. *Genomics, proteomics & bioinformatics*, **13**, 278–289. 9

[274] RITCHIE, M.E., PHIPSON, B., WU, D., HU, Y., LAW, C.W., SHI, W. & SMYTH, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, **43**, e47. 15

[275] ROBERTS, A. & PACHTER, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature methods*, **10**, 71–73. 14

[276] ROBINSON, M.D. & OSHLACK, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**, R25. 73, 191

[277] RODRIGUEZ, J.M., MAIETTA, P., EZKURDIA, I., PIETRELLI, A., WESSELINK, J.J., LOPEZ, G., VALENCIA, A. & TRESS, M.L. (2013). APPRIS: annotation of principal and alternative splice isoforms. *Nucleic acids research*, **41**, D110–7. 50, 74, 92, 125

[278] RODRIGUEZ, J.M., RODRIGUEZ-RIVAS, J., DI DOMENICO, T., VÁZQUEZ, J., VALENCIA, A. & TRESS, M.L. (2018). APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Research*, **46**, D213–D217. 92, 125

[279] ROGERS, M.F., THOMAS, J., REDDY, A.S.N. & BEN-HUR, A. (2012). SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biology*, **13**, R4–R4. 13, 35, 42, 67

[280] ROMERO, P.R., ZAIDI, S., FANG, Y.Y., UVERSKY, V.N., RADIVOJAC, P., OLDFIELD, C.J., CORTESE, M.S., SICKMEIER, M., LEGALL, T., OBRADOVIC, Z. & DUNKER, A.K. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 8390–8395. 92, 105, 159, 181

[281] ROT, G., WANG, Z., HUPPERTZ, I., MODIC, M., LENCE, T., HALLEGGER, M., HABERMAN, N., CURK, T., VON MERING, C. & ULE, J. (2017). High-Resolution RNA Maps Suggest Common Principles of Splicing and Polyadenylation Regulation by TDP-43. *Cell reports*, **19**, 1056–1067. 131, 137, 187

[282] RUEDA, C.B., LLORENTE-FOLCH, I., AMIGO, I., CONTRERAS, L., GONZALEZ-SANCHEZ, P., MARTINEZ-VALERO, P., JUARISTI, I., PARDO, B., DEL ARCO, A. & SATRUSTEGUI, J. (2014). Ca(2+) regulation of mitochondrial function in neurons. *Biochimica et biophysica acta*, **1837**, 1617–1624. 235

[283] RUEPP, A., WAEGELE, B., LECHNER, M., BRAUNER, B., DUNGER-KALTENBACH, I., FOBO, G., FRISHMAN, G., MONTRONE, C. & MEWES, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic acids research*, **38**, D497–501. 111

[284] RUGGIERO, A., ALONI, E., KORKOTIAN, E., ZALTSMAN, Y., ONI-BITON, E., KUPERMAN, Y., TSOORY, M., SHACHNAI, L., LEVIN-ZAIDMAN, S., BRENNER, O., SEGAL, M. & GROSS, A. (2017). Loss of forebrain MTCH2 decreases mitochondria motility and calcium handling and impairs hippocampal-dependent cognitive functions. *Scientific reports*, **7**, 44401. 225

[285] SAITO, A. & CAVALLI, V. (2016). Signaling Over Distances. *Molecular & cellular proteomics : MCP*, **15**, 382–393. 209

[286] SALMELA, L. & RIVALS, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics (Oxford, England)*, **30**, 3506–3514. 11

[287] SANDBERG, R., NEILSON, J.R., SARMA, A., SHARP, P.A. & BURGE, C.B. (2008). Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science (New York, N.Y.)*, **320**, 1643–1647. 92, 94

[288] SCHAEFKE, B., SUN, W., LI, Y.S., FANG, L. & CHEN, W. (2018). The evolution of posttranscriptional regulation. *Wiley interdisciplinary reviews. RNA*, e1485. 3

[289] SCHMUCKER, D., CLEMENS, J.C., SHU, H., WORBY, C.A., XIAO, J., MUDA, M., DIXON, J.E. & ZIPURSKY, S. (2000). Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. *Cell*, **101**, 671–684. 3

[290] SCHUELER, M., MUNSCHAUER, M., GREGERSEN, L.H., FINZEL, A., LOEWER, A., CHEN, W., LANDTHALER, M. & DIETERICH, C. (2014). Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biology*, **15**, R15. 40

[291] SCOTTI, M. & SWANSON, M. (2015). *RNA mis-splicing in disease*, vol. 17. 16

[292] SEDLAZECK, F.J., LEE, H., DARBY, C.A. & SCHATZ, M.C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, **19**, 329–346. 13

[293] SETHUPATHY, P., MEGRAW, M. & HATZIGEORGIOU, A.G. (2006). A guide through present computational approaches for the identification of mammalian microRNA targets. *Nature methods*, **3**, 881–886. 98, 99

[294] SHABALINA, S.A. & SPIRIDONOV, N.A. (2004). The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biology*, **5**, 105. 116

[295] SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N.S., WANG, J.T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. & IDEKER, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13**, 2498–2504. 193

[296] SHARON, D., TILGNER, H., GRUBERT, F. & SNYDER, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nature Biotechnology*, **31**, 1009. 23, 35, 85, 86

[297] SHARP, P.A. (1994). Split Genes and RNA Splicing (Nobel Lecture). *Angewandte Chemie International Edition in English*, **33**, 1229–1240. 3

[298] SHENG, Z.H. (2014). Mitochondrial trafficking and anchoring in neurons: New insight and implications. *The Journal of Cell Biology*, **204**, 1087–1098. 235

[299] SHI, Y. (2012). Alternative polyadenylation: New insights from global analyses. *RNA*, **18**, 2105–2117. 137

[300] SHOEMAKER, B.A., ZHANG, D., THANGUDU, R.R., TYAGI, M., FONG, J.H., MARCHLER-BAUER, A., BRYANT, S.H., MADEJ, T. & PANCHENKO, A.R. (2010). Inferred Biomolecular Interaction Server–a web server to analyze and predict protein interacting partners and binding sites. *Nucleic acids research*, **38**, D518–24. 127

[301] SONG, W., SONG, Y., KINCAID, B., BOSSY, B. & BOSSY-WETZEL, E. (2013). Mutant SOD1G93A triggers mitochondrial fragmentation in spinal cord motor neurons: Neuroprotection by SIRT3 and PGC-1α. *Neurobiology of Disease*, **51**, 72–81. 235

[302] SONG, Z., CHEN, H., FIKET, M., ALEXANDER, C. & CHAN, D.C. (2007). OPA1 processing controls mitochondrial fusion and is regulated by mRNA splicing, membrane potential, and Yme1L. *The Journal of cell biology*, **178**, 749–755. 222

[303] SRINIVASAN, M., EDMAN, C.F. & SCHULMAN, H. (1994). Alternative splicing introduces a nuclear localization signal that targets multifunctional CaM kinase to the nucleus. *The Journal of cell biology*, **126**, 839–852. 221

[304] STAMM, S., BEN-ARI, S., RAFALSKA, I., TANG, Y., ZHANG, Z., TOIBER, D., THANARAJ, T.A. & SOREQ, H. (2005). Function of alternative splicing. *Gene*, **344**, 1–20. 16, 18, 23, 92, 125, 131

[305] STEIJGER, T., ABRIL, J.F., ENGSTROM, P.G., KOKOCINSKI, F., HUBBARD, T.J., GUIGO, R., HARROW, J. & BERTONE, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature methods*, **10**, 1177–1184. 7

[306] STERNE-WEILER, T., MARTINEZ-NUNEZ, R.T., HOWARD, J.M., CVITOVIK, I., KATZMAN, S., TARIQ, M.A., POURMAND, N. & SANFORD, J.R. (2013). Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome research*, **23**, 1615–1623. 18

[307] STICHT, C., DE LA TORRE, C., PARVEEN, A. & GRETZ, N. (2018). miRWalk: An online resource for prediction of microRNA binding sites. *PLOS ONE*, **13**, e0206239. 98

[308] STOLT, C.C., LOMMES, P., SOCK, E., CHABOISSIER, M.C., SCHEDL, A. & WEGNER, M. (2003). The Sox9 transcription factor determines glial fate choice in the developing spinal cord. *Genes & development*, **17**, 1677–1689. 187

[309] SUPEK, F., BOŠNJAK, M., ŠKUNCA, N. & ŠMUC, T. (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE*, **6**, e21800. 193

[310] TABAS-MADRID, D., MUNIATEGUI, A., SÁNCHEZ-CABALLERO, I., MARTÍNEZ-HERRERA, D.J., SORZANO, C.O.S., RUBIO, A. & PASCUAL-MONTANO, A. (2014). Improving miRNA-mRNA interaction predictions. *BMC genomics*, **15 Suppl 1**, S2–S2. 98

[311] TAPIAL, J., HA, K.C.H., STERNE-WEILER, T., GOHR, A., BRAUNSCHWEIG, U., HERMOSO-PULIDO, A., QUESNEL-VALLIERES, M., PERMANYER, J., SODAEI, R., MARQUEZ, Y., COZZUTO, L., WANG, X., GOMEZ-VELAZQUEZ, M., RAYON, T., MANZANARES, M., PONOMARENKO, J., BLENCOWE, B.J. & IRIMIA, M. (2017). An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome research*, **27**, 1759–1768. 92, 125, 131

[312] TARAZONA, S., FURIÓ-TARÍ, P., TURRÀ, D., PIETRO, A.D., NUEDA, M.J., FERRER, A. & CONESA, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, **43**, e140–e140. 139, 140, 191

[313] 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. & CONESA, A. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome research*. 142, 148, 190

[314] TEMPEL, S. (2012). Using and Understanding RepeatMasker BT - Mobile Genetic Elements: Protocols and Genomic Applications. 29–51, Humana Press, Totowa, NJ. 97

[315] TERADA, N., BARACSKAY, K., KINTER, M., MELROSE, S., BROPHY, P.J., BOUCHEIX, C., BJARTMAR, C., KIDD, G. & TRAPP, B.D. (2002). The tetraspanin protein, CD9, is expressed by progenitor cells committed to oligodendrogenesis and is linked to beta1 integrin, CD81, and Tspan-2. *Glia*, **40**, 350–359. 203

[316] TIAN, B. & MANLEY, J.L. (2017). Alternative polyadenylation of mRNA precursors. *Nature reviews. Molecular cell biology*, **18**, 18–30. 18

[317] TIAN, B., HU, J., ZHANG, H. & LUTZ, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic acids research*, **33**, 201–212. 5, 232

[318] TILGNER, H., RAHA, D., HABEGGER, L., MOHIUDDIN, M., GERSTEIN, M. & SNYDER, M. (2013). Accurate Identification and Analysis of Human mRNA Isoforms Using Deep Long Read Sequencing. *G3: Genes, Genomes, Genetics*, **3**, 387–397. 67

[319] TILGNER, H., GRUBERT, F., SHARON, D. & SNYDER, M.P. (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proceedings of the National Academy of Sciences*, **111**, 9869–9874. 8, 9, 23, 35, 85, 125

[320] TILGNER, H., JAHANBANI, F., BLAUWKAMP, T., MOSHREFI, A., JAEGER, E., CHEN, F., HAREL, I., BUSTAMANTE, C.D., RASMUSSEN, M. & SNYDER, M.P. (2015). Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nature Biotechnology*, **33**, 736. 8, 35, 85

[321] TRAN, M.K., SCHULTZ, C.J. & BAUMANN, U. (2008). Conserved upstream open reading frames in higher plants. *BMC genomics*, **9**, 361. 96

[322] TRANCHEVENT, L.C., AUBE, F., DULAURIER, L., BENOIT-PILVEN, C., REY, A., PORET, A., CHAUTARD, E., MORTADA, H., DESMET, F.O., CHAKRAMA, F.Z., MORENO-GARCIA, M.A., GOILLOT, E., JANCZARSKI, S., MORTREUX, F., BOURGEOIS, C.F. & AUBOEUF, D. (2017). Identification of protein features encoded by alternative exons using Exon Ontology. *Genome research*, **27**, 1087–1097. 132

[323] TRAPNELL, C., WILLIAMS, B.A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M.J., SALZBERG, S.L., WOLD, B.J. & PACHTER, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**, 511. 7, 14

[324] TRESS, M.L., ABASCAL, F. & VALENCIA, A. (2017). Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends in Biochemical Sciences*, **42**, 98–110. 18

[325] TRINCADO, J.L., ENTIZNE, J.C., HYSENAJ, G., SINGH, B., SKALIC, M., ELLIOTT, D.J. & EYRAS, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome biology*, **19**, 40. 15

[326] TSENG, E., UNDERWOOD, J.G. & TSENG, E. (2013). Full Length cDNA Sequencing on the PacBio® RS. *Journal of Biomolecular Techniques : JBT*, **24**, S45–S45. 85

[327] TUSHEV, G., GLOCK, C., HEUMULLER, M., BIEVER, A., JOVANOVIC, M. & SCHUMAN, E.M. (2018). Alternative 3' UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. *Neuron*, **98**, 495–511.e6. 234

[328] UITTENBOGAARD, M. & CHIARAMELLO, A. (2002). Expression of the bHLH transcription factor Tcf12 (ME1) gene is linked to the expansion of precursor cell populations during neurogenesis. *Brain research. Gene expression patterns*, **1**, 115–121. 157

[329] ULLAH, S., LIN, S., XU, Y., DENG, W., MA, L., ZHANG, Y., LIU, Z. & XUE, Y. (2016). dbPAF: an integrative database of protein phosphorylation in animals and fungi. *Scientific reports*, **6**, 23534. 108

[330] V KOCHETOV, A., AHMAD, S., IVANISENKO, V., VOLKOVA, O., A KOLCHANOV, N. & SARAI, A. (2008). *uORFs, reinitiation and alternative translation start sites in human mRNAs*, vol. 582. 96

[331] VAN DER LEE, R., BULJAN, M., LANG, B., WEATHERITT, R.J., DAUGHDRILL, G.W., DUNKER, A.K., FUXREITER, M., GOUGH, J., GSPONER, J., JONES, D.T., KIM, P.M., KRIWACKI, R.W., OLDFIELD, C.J., PAPPU, R.V., TOMPA, P., UVERSKY, V.N., WRIGHT, P.E. & BABU, M.M. (2014). Classification of intrinsically disordered regions and proteins. *Chemical reviews*, **114**, 6589–6631. 105

[332] VAQUERO-GARCIA, J., BARRERA, A., GAZZARA, M.R., GONZALEZ-VALLINAS, J., LAHENS, N.F., HOGENESCH, J.B., LYNCH, K.W. & BARASH, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, **5**, e11752. 3

[333] VEGRAN, F., BOIDOT, R., OUDIN, C., RIEDINGER, J.M. & LIZARD-NACOL, S. (2005). Distinct expression of Survivin splice variants in breast carcinomas. *International journal of oncology*, **27**, 1151–1157. 16

[334] VEJNAR, C.E., BLUM, M. & ZDOBNOV, E.M. (2013). miRmap web: Comprehensive microRNA target prediction online. *Nucleic acids research*, **41**, W165–W168. 98

[335] VENABLES, J.P., TAZI, J. & JUGE, F. (2012). Regulated functional alternative splicing in Drosophila. *Nucleic Acids Research*, **40**, 1–10. 3, 15

[336] VIDAL, R.L., RAMIREZ, O.A., SANDOVAL, L., KOENIG-ROBERT, R., HARTEL, S. & COUVE, A. (2007). Marlin-1 and conventional kinesin link GABAB receptors to the cytoskeleton and regulate receptor transport. *Molecular and cellular neurosciences*, **35**, 501–512. 233

[337] VITTING-SEERUP, K., PORSE, B.T., SANDELIN, A. & WAAGE, J. (2014). spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC Bioinformatics*, **15**, 81. 132

[338] VLASOVA, I.A. & BOHJANEN, P.R. (2008). Posttranscriptional regulation of gene networks by GU-rich elements and CELF proteins. *RNA biology*, **5**, 201–207. 156

[339] VLASOVA, I.A., TAHOE, N.M., FAN, D., LARSSON, O., RATTENBACHER, B., STERNJOHN, J.R., VASDEWANI, J., KARYPIS, G., REILLY, C.S., BITTERMAN, P.B. & BOHJANEN, P.R. (2008). Conserved GU-rich elements mediate mRNA decay by binding to CUG-binding protein 1. *Molecular cell*, **29**, 263–270. 156

[340] VOINEAGU, I., WANG, X., JOHNSTON, P., LOWE, J.K., TIAN, Y., HORVATH, S., MILL, J., CANTOR, R.M., BLENCOWE, B.J. & GESCHWIND, D.H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, **474**, 380–384. 232

[341] VOLDERS, K., NUYTENS, K. & CREEMERS, J.W.M. (2011). The autism candidate gene Neurobeachin encodes a scaffolding protein implicated in membrane trafficking and signaling. *Current molecular medicine*, **11**, 204–217. 233

[342] VON ARNIM, A.G., JIA, Q. & VAUGHN, J.N. (2014). Regulation of plant translation by upstream open reading frames. *Plant science : an international journal of experimental plant biology*, **214**, 1–12. 96

[343] VON HEIJNE, G. (1990). The signal peptide. *The Journal of membrane biology*, **115**, 195–201. 104

[344] VOS, M., LAUWERS, E. & VERSTREKEN, P. (2010). Synaptic mitochondria in synaptic transmission and organization of vesicle pools in health and disease. *Frontiers in synaptic neuroscience*, **2**, 139. 235

[345] VUONG, C.K., BLACK, D.L. & ZHENG, S. (2016). The neurogenetics of alternative splicing. *Nature reviews. Neuroscience*, **17**, 265–281. 187

[346] WAGNON, J.L., BRIESE, M., SUN, W., MAHAFFEY, C.L., CURK, T., ROT, G., ULE, J. & FRANKEL, W.N. (2012). CELF4 regulates translation and local abundance of a vast set of mRNAs, including genes associated with regulation of synaptic function. *PLoS genetics*, **8**, e1003067. 227

[347] WANG, B., TSENG, E., REGULSKI, M., CLARK, T.A., HON, T., JIAO, Y., LU, Z., OLSON, A., STEIN, J.C. & WARE, D. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature Communications*, **7**, 11708. 9, 35, 39, 92, 125

[348] WANG, E.T., SANDBERG, R., LUO, S., KHREBTUKOVA, I., ZHANG, L., MAYR, C., KINGSMORE, S.F., SCHROTH, G.P. & BURGE, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476. 3, 4, 5, 7, 15, 131

[349] WANG, E.T., CODY, N.A.L., JOG, S., BIANCOLELLA, M., WANG, T.T., TREACY, D.J., LUO, S., SCHROTH, G.P., HOUSMAN, D.E., REDDY, S., LÉCUYER, E. & BURGE, C.B. (2012). Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell*, **150**, 710–724. 194, 227

[350] WANG, M., ZHAO, Y. & ZHANG, B. (2015). Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports*, **5**, 16923. 155, 156, 235

[351] WANG, P., YAN, B., GUO, J.T., HICKS, C. & XU, Y. (2005). Structural genomics analysis of alternative splicing and application to isoform structure modeling. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 18920–18925. 92

[352] WANG, T., LI, L. & HONG, W. (2017). SNARE proteins in membrane trafficking. *Traffic*, **18**, 767–775. 210

[353] WANG, W., QIN, Z., FENG, Z., WANG, X. & ZHANG, X. (2013). Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene*, **518**, 164–170. 15, 233

[354] WANG, W., WEI, Z. & LI, H. (2014). A change-point model for identifying 3'UTR switching by next-generation RNA sequencing. *Bioinformatics (Oxford, England)*, **30**, 2162–2170. 148

[355] WANG, X. & LIU, X.S. (2011). Systematic Curation of miRBase Annotation Using Integrated Small RNA High-Throughput Sequencing Data for C. elegans and Drosophila. *Frontiers in genetics*, **2**, 25. 98

[356] WANG, X., HOU, J., QUEDENAU, C. & CHEN, W. (2016). Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. *Molecular systems biology*, **12**, 875. 132

[357] WANG, X., CHEN, S., NI, J., CHENG, J., JIA, J. & ZHEN, X. (2018). miRNA-3473b contributes to neuroinflammation following cerebral ischemia. *Cell Death & Disease*, **9**, 11. 178

[358] WEATHERITT, R.J., STERNE-WEILER, T. & BLENCOWE, B.J. (2016). The ribosome-engaged landscape of alternative splicing. *Nature structural & molecular biology*, **23**, 1117–1123. 18

[359] WEIRATHER, J.L., AFSHAR, P.T., CLARK, T.A., TSENG, E., POWERS, L.S., UNDERWOOD, J.G., ZABNER, J., KORLACH, J., WONG, W.H. & AU, K.F. (2015). Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. *Nucleic acids research*, **43**, e116. 40

[360] WENK, M.R. & DE CAMILLI, P. (2004). Protein-lipid interactions and phosphoinositide metabolism in membrane traffic: insights from vesicle recycling in nerve terminals. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 8262–8269. 218, 234

[361] WENK, M.R., PELLEGRINI, L., KLENCHIN, V.A., DI PAOLO, G., CHANG, S., DANIELL, L., ARIOKA, M., MARTIN, T.F. & DE CAMILLI, P. (2001). PIP kinase Igamma is the major PI(4,5)P(2) synthesizing enzyme at the synapse. *Neuron*, **32**, 79–88. 218

[362] WEYN-VANHENTENRYCK, S.M., FENG, H., USTIANENKO, D., DUFFIÉ, R., YAN, Q., JACKO, M., MARTINEZ, J.C., GOODWIN, M., ZHANG, X., HENGST, U., LOMVARDAS, S., SWANSON, M.S. & ZHANG, C. (2018). Precise temporal regulation of alternative splicing during neural development. *Nature Communications*, **9**, 2189. 131, 187, 232

[363] WICHTERLE, H., LIEBERAM, I., PORTER, J.A. & JESSELL, T.M. (2002). Directed differentiation of embryonic stem cells into motor neurons. *Cell*, **110**, 385–397. 203

[364] WILSON, H.C., SCOLDING, N.J. & RAINE, C.S. (2006). Co-expression of PDGF alpha receptor and NG2 by oligodendrocyte precursors in human CNS and multiple sclerosis lesions. *Journal of neuroimmunology*, **176**, 162–173. 202

[365] WINKLE, C.C. & GUPTON, S.L. (2016). Membrane Trafficking in Neuronal Development: Ins and Outs of Neural Connectivity. *International review of cell and molecular biology*, **322**, 247–280. 233

[366] WITTEN, J.T. & ULE, J. (2011). Understanding splicing regulation through RNA splicing maps. *Trends in genetics : TIG*, **27**, 89–97. 131

[367] WONG, J.J.L., RITCHIE, W., EBNER, O.A., SELBACH, M., WONG, J.W.H., HUANG, Y., GAO, D., PINELLO, N., GONZALEZ, M., BAIDYA, K., THOENG, A., KHOO, T.L., BAILEY, C.G., HOLST, J. & RASKO, J.E.J. (2013). Orchestrated intron retention regulates normal granulocyte differentiation. *Cell*, **154**, 583–595. 131, 231

[368] WONG, N. & WANG, X. (2015). miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic acids research*, **43**, D146–D152. 98

[369] WONG, Q.W.L., VAZ, C., LEE, Q.Y., ZHAO, T.Y., LUO, R., ARCHER, S.K., PREISS, T., TANAVDE, V. & VARDY, L.A. (2016). Embryonic Stem Cells Exhibit mRNA Isoform Specific Translational Regulation. *PloS one*, **11**, e0143235. 23

[370] WU, T.D. & WATANABE, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875. 13, 41, 80, 190

[371] WU, X., ZHANG, Y. & LI, Q.Q. (2016). PlantAPA: A Portal for Visualization and Analysis of Alternative Polyadenylation in Plants. *Frontiers in plant science*, **7**, 889. 148

[372] XIA, Z., DONEHOWER, L.A., COOPER, T.A., NEILSON, J.R., WHEELER, D.A., WAGNER, E.J. & LI, W. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3-UTR landscape across seven tumour types. *Nature Communications*, **5**, 5274. 131, 148

[373] XIAO, F., ZUO, Z., CAI, G., KANG, S., GAO, X. & LI, T. (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic acids research*, **37**, D105–10. 98

[374] XU, Q., MODREK, B. & LEE, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic acids research*, **30**, 3754–3766. 4, 91, 131, 187

[375] YAMAGISHI, R., OKUYAMA, T., OBA, S., SHIMADA, J., CHAEN, S. & KANEKO, H. (2015). Comprehensive analysis of the dynamic structure of nuclear localization signals. *Biochemistry and Biophysics Reports*, **4**, 392–396. 159

[376] YANG, I.S., SON, H., KIM, S. & KIM, S. (2016). ISOexpresso: a web-based platform for isoform-level expression analysis in human cancer. *BMC Genomics*, **17**, 631. 132

[377] 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., TRIGG, 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. & VIDAL, M. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, **164**, 805–817. 23

[378] YANG, Y.C.T., DI, C., HU, B., ZHOU, M., LIU, Y., SONG, N., LI, Y., UMETSU, J. & LU, Z.J. (2015). CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC genomics*, **16**, 51. 101

[379] YAP, C.C. & WINCKLER, B. (2012). Harnessing the power of the endosome to regulate neural development. *Neuron*, **74**, 440–451. 209

[380] YEE, B.A., PRATT, G.A., GRAVELEY, B.R., VAN NOSTRAND, E.L. & YEO, G.W. (2019). RBP-Maps enables robust generation of splicing regulatory maps. *RNA (New York, N.Y.)*, **25**, 193–204. 131

[381] YOSHIDA, H., OKUMURA, N., KITAGISHI, Y., SHIRAFUJI, N. & MATSUDA, S. (2010). Rab5(Q79L) interacts with the carboxyl terminus of RUFY3. 178

[382] YOUNG, M.D., WAKEFIELD, M.J., SMYTH, G.K. & OSHLACK, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, **11**, R14. 192

[383] ZANNA, C., GHELLI, A., PORCELLI, A.M., KARBOWSKI, M., YOULE, R.J., SCHIMPF, S., WISSINGER, B., PINTI, M., COSSARIZZA, A., VIDONI, S., VALENTINO, M.L., RUGOLO, M. & CARELLI, V. (2008). OPA1 mutations associated with dominant optic atrophy impair oxidative phosphorylation and mitochondrial fusion. *Brain : a journal of neurology*, **131**, 352–367. 235

[384] ZHANG, F., WHITE, R.L. & NEUFELD, K.L. (2000). Phosphorylation near nuclear localization signal regulates nuclear import of adenomatous polyposis coli protein. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 12577–12582. 175

[385] ZHANG, H., HU, J., RECCE, M. & TIAN, B. (2005). PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic acids research*, **33**, D116–20. 5

[386] ZHANG, H., LEE, J.Y. & TIAN, B. (2005). Biased alternative polyadenylation in human tissues. *Genome biology*, **6**, R100. 5

[387] ZHANG, S.J., WANG, C., YAN, S., FU, A., LUAN, X., LI, Y., SUNNY SHEN, Q., ZHONG, X., CHEN, J.Y., WANG, X., CHIN-MING TAN, B., HE, A. & LI, C.Y. (2017). Isoform Evolution in Primates through Independent Combination of Alternative RNA Processing Events. *Molecular biology and evolution*, **34**, 2453–2468. 9, 14, 15

[388] ZHANG, X., CHEN, M.H., WU, X., KODANI, A., FAN, J., DOAN, R., OZAWA, M., MA, J., YOSHIDA, N., REITER, J.F., BLACK, D.L., KHARCHENKO, P.V., SHARP, P.A. & WALSH, C.A. (2016). Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. *Cell*, **166**, 1147–1162.e15. 187

[389] ZHANG, Z., XIN, D., WANG, P., ZHOU, L., HU, L., KONG, X. & D HURST, L. (2009). *Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay*, vol. 7. 103

[390] ZHENG, D., WANG, R., DING, Q., WANG, T., XIE, B., WEI, L., ZHONG, Z. & TIAN, B. (2018). Cellular stress alters 3UTR landscape through alternative polyadenylation and isoform-specific degradation. *Nature Communications*, **9**, 2268. 131

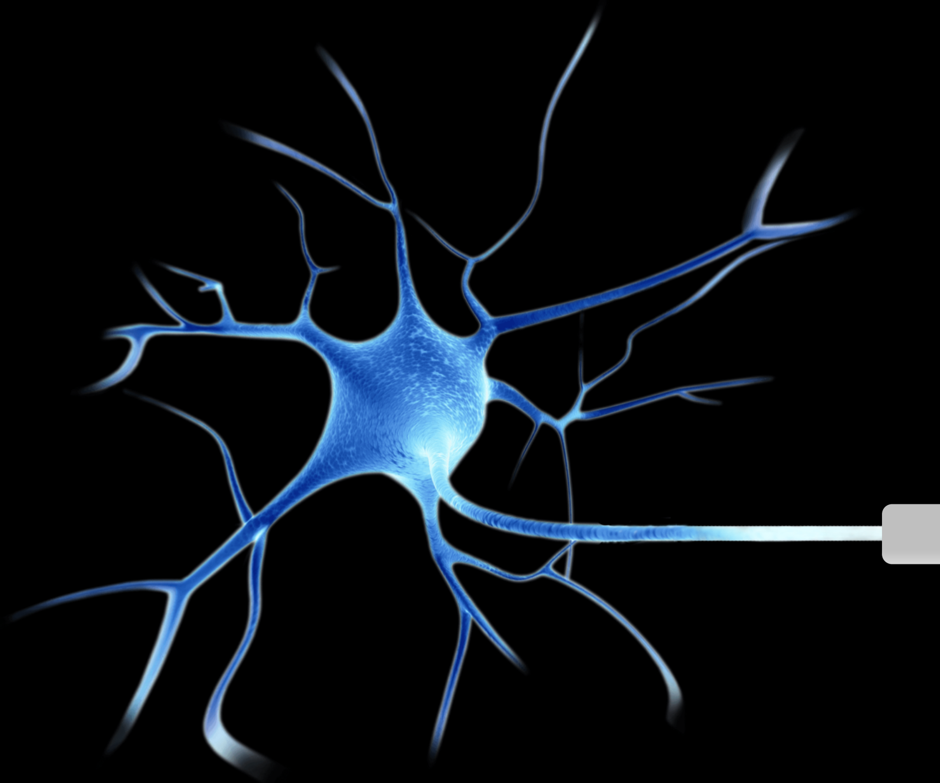
[391] ZHENG, S. (2016). Alternative splicing and nonsense-mediated mRNA decay enforce neural specific gene expression. *International journal of developmental neuroscience : the official journal of the International Society for Developmental Neuroscience*, **55**, 102–108. 125, 132, 187

[392] ZHENG, S. & BLACK, D.L. (2013). Alternative pre-mRNA splicing in neurons: growing up and extending its reach. *Trends in genetics : TIG*, **29**, 442–448. 187

[393] ZHOU, J., ZHAO, S. & DUNKER, A.K. (2018). Intrinsically Disordered Proteins Link Alternative Splicing and Post-translational Modifications to Complex Cell Signaling and Regulation. *Journal of molecular biology*, **430**, 2342–2359. 172, 181

[394] ZHOU, Q., CHOI, G. & ANDERSON, D.J. (2001). The bHLH transcription factor Olig2 promotes oligodendrocyte differentiation in collaboration with Nkx2.2. *Neuron*, **31**, 791–807. 187

[395] ZUCCONI, B.E. & WILSON, G.M. (2011). Modulation of neoplastic gene regulatory pathways by the RNA-binding factor AUF1. *Frontiers in bioscience (Landmark edition)*, **16**, 2307–2325. 180



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA