University of Exeter
Department of Medical Sciences

Genomic characterisation of Alzheimer's disease risk genes using long-read sequencing

Szi Kay Leung

April, 2022

Supervised by Professor Jonathan Mill, Dr Eilis Hannon,
Dr Aaron Jeffries & Professor David Collier

Submitted by Szi Kay Leung to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Medical Sciences in April 2022.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(signature) ....

Abstract

Alzheimer's disease (AD) is a devastating neurodegenerative disorder characterised by progressive intracellular accumulation of hyperphosphorylated tau and extracellular deposition of beta-amyloid. It affects over 50 million people worldwide with numbers expecting to triple by 2050. Despite recent success in identifying genetic risk factors for AD, the mechanisms underpinning disease progression remain unknown. There is increasing evidence for altered transcriptional regulation and RNA splicing in the development of AD pathology. However, current studies exploring isoform diversity in the AD brain are constrained by the inherent limitations of standard short-read RNA-sequencing approaches, which fail to capture full-length transcripts critical for transcriptome assembly.

The primary aim of this thesis was to utilise two long-read sequencing approaches, Pacific Biosciences isoform sequencing and Oxford Nanopore Technologies nanopore cDNA sequencing, to examine isoform diversity and transcript usage in the cortex, and identify alternative splicing events associated with AD pathology in a transgenic model of tau pathology (rTg4510). By generating long reads that span full-length transcripts, our studies revealed widespread RNA isoform diversity with unprecedented detection of novel transcripts not present in existing genome annotations. We further performed ultra-deep targeted long-read sequencing of 20 AD-risk genes, identifying robust expression changes at the transcript level associated with tau accumulation in the cortex. Our analyses provide a systematic evaluation of transcript usage, even in the absence of gene-level expression alterations, and highlight the importance of alternative RNA splicing as a mechanism underpinning gene regulation in the development of tau pathology.

Finally, this thesis presents a laboratory and bioinformatics pipeline for the systematic characterisation of isoform diversity and alternative splicing using long-read sequencing. The data generated as part of this research have implications for our understanding of the mechanisms driving the development of tau pathology, and represent a valuable resource to the wider research community.

To my fiancé, Tim

Acknowledgements

While I had been warned of the difficulties of undertaking a PhD, I never would have imagined just how challenging and mentally taxing it could be. However, I have been blessed with the support of a number of people who have made this journey significantly easier and memorable.

I would like to begin by thanking my supervisors – Professor Jonathan Mill, Dr Eilis Hannon and Dr Aaron Jeffries - for their invaluable guidance, patience and support throughout these four and a half years. Jon, thank you for the countless hours that you have spent patiently advising me throughout this PhD, and for always being so encouraging and enthusiastic about my results (even when I doubted them myself). Eilis, thank you for your insightful feedback and comments, and for being the person I could always turn to for a second, unbiased opinion. Aaron, thank you so much for sharing your invaluable expertise and knowledge be it in the lab or in bioinformatics, and for taking me under your wing; I cannot thank you enough for all those times that you have patiently reassured me in the lab when I thought I had messed up (again). Thank you to the three of you for continuing to inspire and show me what it takes to be a brilliant researcher.

This work would not have been possible in it's current form without the funding from Eli-Lilly, and I would like to express my sincerest gratitude to Professor David Collier, Dr Emma Laing, Dr Zeshan Ahmed and Mr Gopuraja Dharmalingam for their ongoing interest and involvement. Your insightful comments and perspective (i.e. "what question are we trying to answer here?") have shaped the development of this thesis and helped me to think more broadly.

A special thanks to Dr Karen Moore and Paul O'Neill at the Exeter Sequencing Service for sharing their expertise, particularly Karen for all of her guidance in optimising the Oxford Nanopore protocol.

I would like to thank all the members (past and present) of the Complex Disease Epigenetics group for being such a vibrant, enthusiastic and amazing group. Particularly to Thea and Gemma, I am so fortunate to have embarked on this journey with the both of you, to be on the same page and come back home to such a strong support network - thank you for making my day-to-day life at Exeter so much more enjoyable. To Isabel, without your dedication and passion, this PhD, which is partly an extension of your work, might not have existed, so thank you for all the times you have sat me down to discuss and brainstorm ideas. To Stef, thank you for all of your encouraging words both inside and outside of the lab, and for being

exceptionally kind whenever I needed anything.

A special thanks to Alan, who was always one step ahead of me on the PhD journey – thank you for listening to me rant (still) and resisting the urge to say “I told you so”. To all the members of the Belmont PhD group, especially Joshua - thank you for your wisdom, prayers and fellowship. To Lori, Tharini and Tommo - thank you for being the most wonderful friends and just being one phone call away no matter the distance.

My sincere gratitude to my soon to be in-laws, Joe and Sandra, for making me feel so warmly welcomed at their home. Thank you for all the loving care that you have given me, for allowing me to visit almost every weekend, for the lovely meals, walks and spiritual conversations.

To my family – Ma, Han, Dad and Uncle Jabez – for their unconditional love, continuous support and unwavering patience, especially in the last few months of writing. Ma, thank you for always believing in me, for mentoring and guiding me; I would not be where I am without you. Han, thank you for picking up after me and shouldering responsibilities that should have been mine. Dad, thank you for never refusing to drive hours to drop my stuff off only to then pick them back up. Uncle Jabez, thank you for looking after the family without complaint and for taking me out of my research bubble - your resilience, passion and fortitude never ceases to amaze me.

Finally, I would like to dedicate this thesis to my fiancé, Tim. Thank you for essentially being my fifth but uncredited supervisor. Despite having your own PhD, you listened to me go on repeatedly about alternative splicing, reviewed my statistical models and analyses, corrected my programming errors, opened my eyes to LaTeX, checked my grammar (even 5 minutes before I aimed to submit, though I still insist that "Lastly" should be a word), comforted me night and day, the list goes on... I would not have applied for a PhD without your unwavering encouragement, nor would I have finished it in time without your endless support. So thank you for everything, for putting up with me and committing to continually doing so :).

Last but foremost, praise and thanks be to God for keeping me motivated and sane throughout this PhD journey.

Contents

List of Figures

List of Tables

Publications

Chapter 4

(Accepted manuscript presented in **Appendix F**)

Leung SK., Jeffries A.R., Castanho I., Jordan B.T., Moore K., Davies J.P., ... & Mill.J. (2021). Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing.*Cell Reports*,37(7):110022.

Other publications

Marzi S.J., **Leung SK.**, Ribarska T, Hannon E, Smith A.R., Pishva E., ... Mill J. (2018)
A histone acetylome-wide association study of Alzheimer's disease identifies disease-associated H3K27ac differences in the entorhinal cortex. *Nature Neuroscience*,21(11):1618-1627.

Declarations

All mouse samples used in **Chapters 4 - 6** were obtained from Eli Lilly & Company Ltd., Windlesham (United Kingdom).

All laboratory work and analyses were performed by myself, with the following exceptions:

- Animal breeding procedures, sample collection, and subsequent RNA extractions and assessment from mouse samples (**Chapters 4 - 6**) were performed by Dr Isabel Castanho.
- Short-read RNA Sequencing was prepared and performed by Dr Isabel Castanho, Audrey Farbos and Dr Karen Moore at the University of Exeter Sequencing Service. Filtering of raw RNA-Seq reads were further carried out by Dr Isabel Castanho.
- Sample loading and machine operation for Iso-Seq targeted sequencing of the final two Iso-Seq batches (described in **Chapter 6**) were performed by Dr Stefania Policicchio and Dr Aaron Jeffries at the University of Exeter Sequencing Service.
- Nanopore library preparation and sequencing (after target capture, which was performed by myself, as described in **Chapter 6**) were performed with Dr Aaron Jeffries.
- RNA extractions and assessment from AD post-mortem brain tissue (for preliminary results mentioned in **Chapter 7**) were performed by Greg Wheildon. Iso-Seq library preparation and targeted sequencing were subsequently performed by Dr Stefania Policicchio under my guidance.

Gene Nomenclature

Abca1	ATP-binding cassette subfamily A member 1
Abca7	ATP-binding cassette subfamily A member 7
Ank1	Ankyrin 1
Apoe	Apolipoprotein E
App	Amyloid precursor protein
Bin1	Bridging integrator 1
Clu	Clusterin
Fus	Fused in sarcoma
Gfap	Glial fibrillary acidic protein
Mapt	Microtubule-associated protein tau
Picalm	Phosphatidylinositol binding clathrin assembly protein
Psen1	Presenilin 1
Psen2	Presenilin 2
Ptk2b	Protein-tyrosine kinase 2-beta
Rhbdf2	Rhomboid 5 homolog 2
Sorl1	Sortilin related receptor 1
Tardbp	TAR DNA-binding protein
Trem2	Triggering receptor expressed on myeloid cells 2
Trpa1	Transient receptor potential ankyrin 1
Vgf	VGF nerve growth factor inducible

Abbreviations

3'SS	3' splice site
5'SS	5' splice site
A3'	Alternative 3' splice site
A5'	Alternative 5' splice site
AD	Alzheimer's disease
AF	Alternative first exon
AL	Alternative last exon
ALS	Amyotrophic lateral sclerosis
AS	Alternative splicing
ATI	Alternative transcription initiation
ATT	Alternative transcription termination
BACE	β-site APP-cleaving enzyme
CAGE	Cap analysis of gene expression
CCS	Circular consensus sequence
cDNA	Complementary DNA
CDS	Coding sequence
CPM	Counts per million
CSF	Cerebrospinal fluid
DGE	Differential gene expression
DLB	Dementia with Lewy bodies
dNTPs	Deoxynucleotide triphosphates
dsDNA	Double-stranded DNA
DTE	Differential transcript expression
DTU	Differential transcript usage
EOAD	Early-onset Alzheimer's disease
ERCC	External RNA Controls Consortium

ES	Exon skipping
EWAS	Epigenome-wide association study
FAD	Familial Alzheimer’s disease
FDR	False discovery rate
FL	Full-length
FLNC	Full-length non-chimeric
FSM	Full Splice Match
FTD	Frontotemporal dementia
FTDP	Frontotemporal dementia and parkinsonism
GWAS	Genome-wide association study
HQ	High-quality
IR	Intron retention
ISM	Incomplete Splice Match
Iso-Seq	Isoform Sequencing (PacBio)
LncRNAs	Long non-coding RNAs
LOAD	Late-onset Alzheimer’s disease
LOF	Loss-of-function
mRNA	Messenger RNA
MX	Mutually exclusive
NFT	Neurofibrillary tangles
NIC	Novel in Catalogue
NMD	Nonsense-mediated decay
NNC	Novel Not in Catalogue
ONT	Oxford Nanopore Technologies
ORF	Open reading frame
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
PD	Parkinson’s disease
PPT	Polypyrimidine tract
PSI	Percent splicing index
PTC	Premature termination codon
QC	Quality control
RBPs	RNA-binding proteins

RIN	RNA integrity number
RNA-Seq	RNA sequencing
rRNA	Ribosomal RNA
RT	Reverse transcriptase
scRNA-Seq	Single-cell RNA sequencing
SMRT	Single-molecule real-time
SNPs	Single-nucleotide polymorphisms
snRNPs	Small nuclear ribonucleoproteins
ssDNA	Single-stranded DNA
TMM	Trimmed mean of M-values
TPM	Transcripts per million
tRNA	Transfer RNA
TSS	Transcription start site
TTS	Transcription termination site
TWAS	Transcriptome-wide association study
UMI	Unique molecular identifiers
UTRs	Untranslated regions
WGCNA	Weighted gene correlation network analysis
ZMW	Zero-mode waveguide

Chapter 1

Introduction

1.1 Alzheimer's disease (AD)

Alzheimer's disease (AD) is a devastating neurodegenerative disorder, clinically characterised by progressive memory loss, cognitive decline, and behavioural impairment. The most common form of dementia, it is estimated to affect 50 million people worldwide with numbers expecting to triple to 152 million by 2050, ensuing both a heavy global economic and social burden.[1] Despite international efforts to better understand the disorder for drug discovery and development, there is currently no cure and existing medications only act to ameliorate symptoms.

1.1.1 Pathology

The symptoms of AD are underpinned by both morphological and molecular changes in the brain. Neuroimaging and post-mortem brain analyses from patients reveal significant brain atrophy caused by neuronal and synaptic loss[2,3] (**Figure 1.1**). Further microscopic examinations have revealed the accumulation of beta-amyloid (Aβ) as amyloid plaques (**Figure 1.2A**) and the aggregation of tau as neurofibrillary tangles (NFTs) (**Figure 1.2B**), two key hallmarks of AD, which strongly correlate to cognitive decline and are now believed to manifest years before presentation of clinical symptoms.[4] These neuropathological changes are accompanied with heightened neuroinflammation through the abnormal activation and distribution of microglia (the most abundant brain resident immune cells) and astrocytes (glial cells with multiple roles in supporting neuronal function and metabolism)[5] (discussed in **Section 1.1.3**).

The progression of these neuropathological changes have been well mapped in post-mortem tissue, particularly the spread of NFTs which is quantified using Braak staging[6] (**Figure 1.2C**). Pathology is initially apparent in the temporal lobes (hippocampus and entorhinal cortex) with later advancement to the frontal lobes. Conversely, the occipital lobes, motor cortex, and the cerebellum are relatively resistant to neuronal degeneration even in advanced stages of AD.[7]

Of note, it is important to emphasise that aside from Aβ deposition and NFT formation, 15-20% of AD patients also display evidence of Lewy body (LB) pathology,[8,9] the defining pathological hallmark of Parkinson's disease (PD)[10] and dementia with Lewy bodies (DLB)[11]. This is characterised by abnormal aggregation of α-synuclein into intra-neuronal cytoplasmic inclusion bodies. Up to 75% of AD patients further present neuronal cytoplasmic inclusions comprising of aggregates of TDP-43[12–14] (Transactive response DNA binding protein-43), the defining hallmarks of frontotemporal dementia (FTD) and amyotrophic lateral sclerosis[15] (ALS).

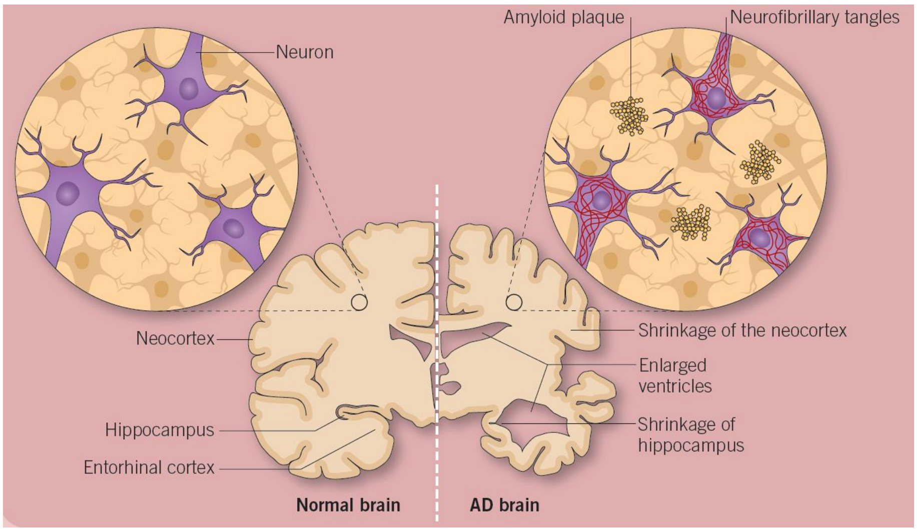


Figure 1.1: Two key hallmarks of AD pathology: amyloid plaques & neurofibrillary tangles. Shown is a schematic figure comparing a normal healthy brain and a diseased brain with advanced AD. AD pathology is well characterised by the presence of extracellular amyloid plaques and intracellular neurofibrillary tangles, accompanied by significant neuronal loss and subsequent shrinkage of the neocortex and hippocampus. Figure is taken from Palmer (2015).[16]

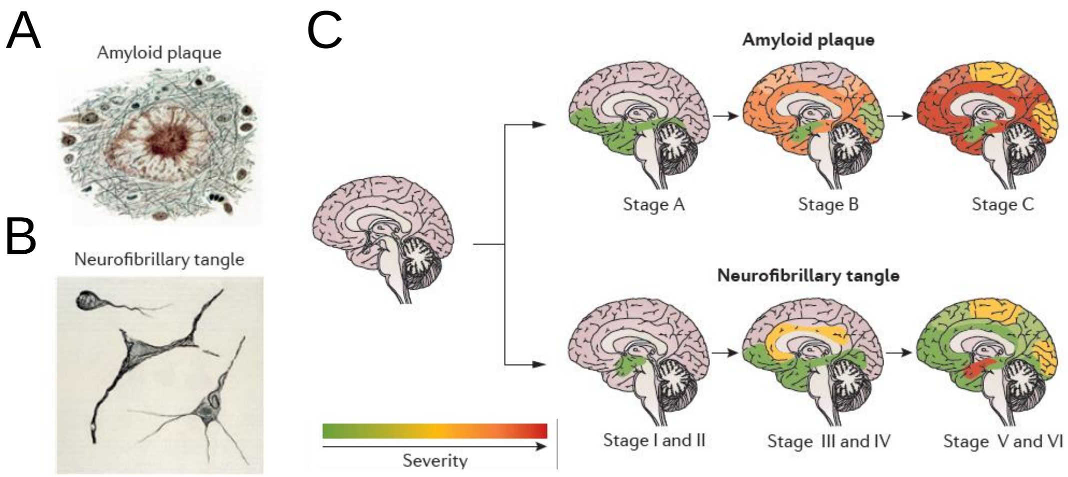


Figure 1.2: Progression of amyloid plaques and neurofibrillary tangles with AD development. Shown is a schematic figure of the progression of **(A, C)** amyloid plaques consisting of Aβ measured according to Thal Phasing,[17] and **(B, C)** neurofibrillary tangles composed of hyper-phosphorylated tau and measured using Braak staging.[6] Figure is taken from Masters et al. (2015).[18]

The deposition of Aβ can be mapped using Thal Phasing from the neocortex (Stage A), to the allocortical regions comprising of the entorhinal cortex and hippocampus, the striatum (Stage B) and finally to the subcortical regions (Stage C).[17] Thal Phasing is based on the detection of immunopositive amyloid.

In a similar pattern, the progressive spread of NFTs can be classified under the six stages of Braak from the trans-entorhinal regions such as the entorhinal cortex (Stage I and II), to the hippocampus (Stage III), the adjoining neocortex (Stage IV) and finally to other neocortical regions (Stage V and VI).[6] Braak staging is a semi-quantitative measure of the severity of NFTs, which can be visualised using silver stain.

1.1.2 Genetics of AD

Although AD predominantly affects people aged 65 and above (Late-onset Alzheimer's disease, LOAD), 5% of AD cases arise in much younger patients (Early-onset Alzheimer's disease, EOAD). EOAD is typically associated with a clear familial autosomal dominant pattern of inheritance (Familial Alzheimer's disease, FAD).[19] To date, more than 160 highly-penetrant, causative mutations have been identified in EOAD, all located within three genes involved in Aβ formation: *APP* (amyloid precursor protein), *PSEN1* (presenilin 1) and *PSEN2* (presenilin 2).[20,21]

While LOAD does not follow a typical Mendelian inheritance pattern, a relatively high heritability rate has been reported (overall broad-sense heritability of 0.58 to 0.79 if shared environmental influences are removed[22]) in twin studies, indicating that there is still a large genetic predisposition for developing AD in later years. Indeed, genome-wide association studies (GWAS) and subsequent meta-analyses[23–32] have identified over 75 genetic loci associated with an increased risk of developing LOAD (narrow-sense heritability). These GWAS loci are typically changes (or variants) at a single DNA base-pair (single-nucleotide polymorphisms – SNPs) or small insertions and deletions (indels) that are found at a higher frequency in individuals with LOAD than in individuals without the disease.

To date, the strongest genetic risk factor for LOAD is the ϵ4 allele of *APOE*,[27] which encodes the cholesterol transporter apolipoprotein E (ApoE) - an essential protein that is involved in regulating lipid homeostasis and transport critical for synaptic function and maintenance.[33] Notably, harbouring one *APOE*ϵ4 allele increases the risk of developing LOAD by 3-4x, while harbouring two ϵ4 alleles increases the risk by 15x.[34] While the ϵ4 allele is estimated to occur in ~15% of the general population, it has been detected in ~40% of LOAD patients.[34] Conversely, the ϵ2 allele is known to confer a neuroprotective effect against AD.[35,36]

With the exception of *APOE*, all the other GWAS-associated genetics variants are either common but lowly penetrant (i.e. SNPs annotated to *CLU, PICALM*) or highly penetrant but rare (i.e. SNPs annotated to *TREM2*) and collectively only contribute modestly to the risk of developing LOAD, highlighting the polygenic nature of AD (**Figure 1.3**). While the molecular mechanisms through which these variants increase risk remain poorly understood, many are annotated to genes enriched in specific biological pathways (described in **Section 1.1.3**).

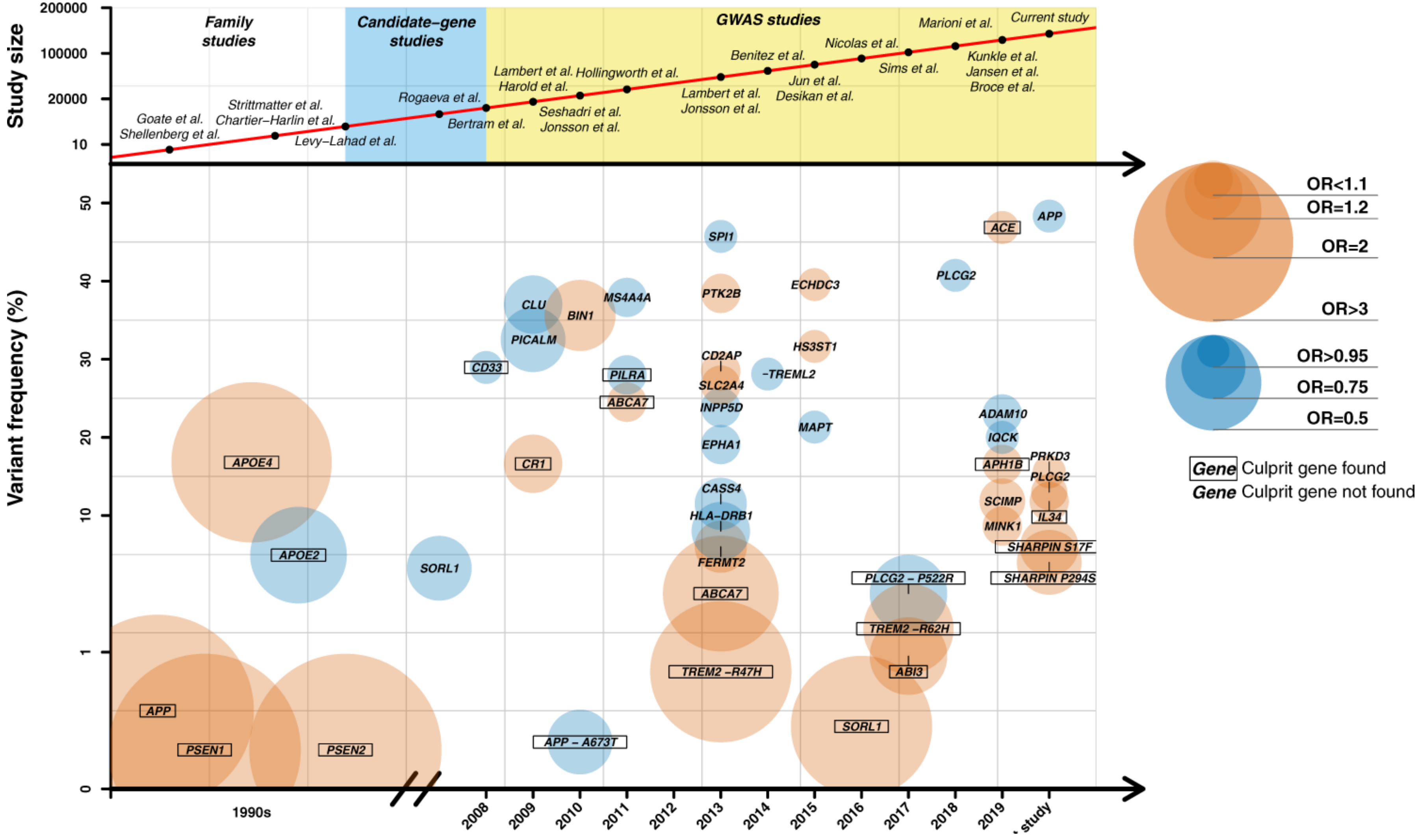


Figure 1.3: The genetic landscape of AD. Shown are the genes implicated in AD from the causative EOAD genes (*APP, PSEN1, PSEN2*) identified from early family studies, to the GWAS-associated genes with either common but lowly penetrant variants or highly penetrant but rare variants. Variant penetrance, or effect size, is measured with the odds ratio (OR) with a larger odds ratio referring to a larger effect size. Variants conferring a protective and negative effect are denoted in orange and blue respectively. Figure was taken from DeRojas et al. (2021).[37] GWAS - Genome-wide association study, OR - odds ratio.

1.1.3 Molecular mechanisms underlying AD pathogenesis

Despite the fact that AD neuropathology has been well described (**Figure 1.1**), the exact biological mechanisms driving AD onset and pathogenesis are still widely unknown. To date, there are two key hypotheses proposed for the progression of AD: i) the amyloid cascade hypothesis, and ii) the tau tangle hypothesis. Results from GWAS, however, implicate other pathways that could be involved including the immune response, lipid metabolism, endocytosis, and cell-adhesion molecule (CAM) pathways for synaptic signalling.

Amyloid cascade hypothesis

The amyloid cascade hypothesis posits that the extracellular accumulation of Aβ is the key driver of AD pathogenesis (**Figure 1.2A**), which initiates a pathological cascade of NFTs, cell loss and vascular damage.[38] Aβ is comprised of short peptides (39-43 amino acids)[39] produced from the amyloidogenic cleavage of APP (a transmembrane protein involved in synapse formation and stability) by β-secretase (BACE, β-site APP-cleaving enzyme 1) and γ-secretase (a complex protein consisting of PSEN1 and PSEN2) (**Figure 1.4**). γ-secretase is further known to cleave APP at various sites, generating multiple Aβ peptides of varying lengths, with 90% secreted as Aβ_{40} and the remaining 10% as Aβ_{42}.[40] In AD, the processing of APP is altered with the vast majority of causative *APP*, *PSEN1* and *PSEN2* mutations favouring the production of the longer and more self-aggregating Aβ_{42},[41–43] thereby promoting the formation of insoluble fibrils and plaques.[43]

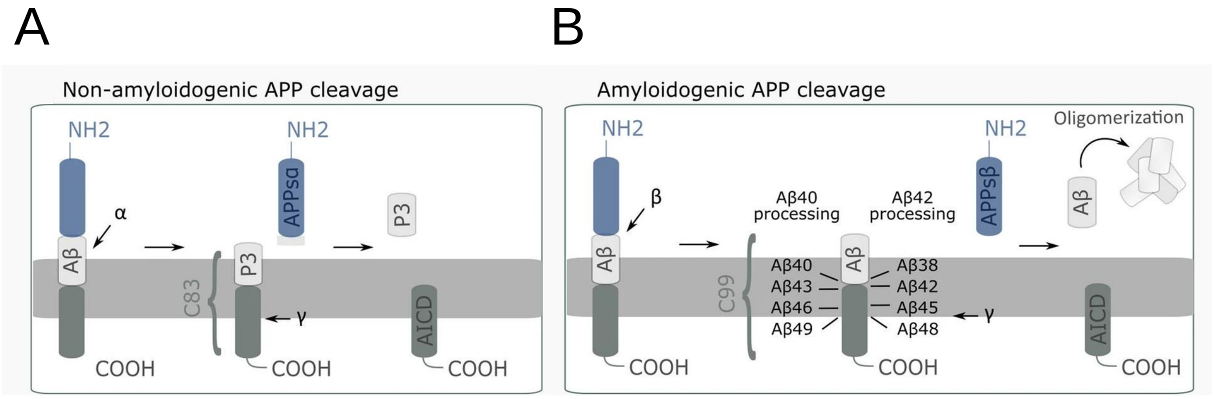


Figure 1.4: Sequential cleavage of APP into Aβ by β-secretase and γ secretase. Shown is a schematic figure depicting sequential cleavage of APP, a transmembrane protein, either through the **(A)** non-amyloidogenic pathway or the **(B)** amyloidogenic pathway.

In the non-amyloidogenic pathway, APP is cleaved by the ADAM protein family (primarily, ADAM10, also known as α-secretases) followed by β-secretase. Conversely in the amyloidogenic pathway, APP is sequentially cleaved by by β-secretase and γ-secretase, which produces Aβ of varying lengths. Monomeric Aβ peptides, particularly Aβ42, have increased propensity to oligomerise and aggregate to form the fibrils and plaques that are characteristic of AD. Figure is taken from Acker et al. (2019).[44]

Tau tangle hypothesis

The tau tangle hypothesis posits that the primary driver of AD is the formation of NFTs from the phosphorylation and aggregation of tau[45] (**Figure 1.2B**), which is also the defining feature of more than 20 other neurodegenerative disorders collectively knowns as tauopathies.[46] Tau, encoded by the *MAPT* gene, is a microtubule-associated protein involved in microtubule maintenance and stability.

Recent studies suggest that tau is hyper-phosphorylated in AD, resulting in dissociation from microtubules and aggregation into filaments[47,48] (components of NFTs) that disrupt axonal transport and signal transmission, ultimately resulting in synpatic degeneration and loss[49] (**Figure 1.5**). Tau mutations associated with frontotemporal dementia and parkinsonism (FTDP) were found to induce conformational changes that promote phosphorylation.[50] While no causative mutations in *MAPT* have been identified in AD, the severity of NFTs has been shown to correlate better with cognitive decline and disease progression than amyloid plaques.[51–53] Notably, regional variations in *MAPT* transcript and protein expression were observed across the brain with a 2-fold increase in the neocortex compared to the cerebellum, potentially explaining the regional vulnerability of different brain regions to tau pathology.[54]

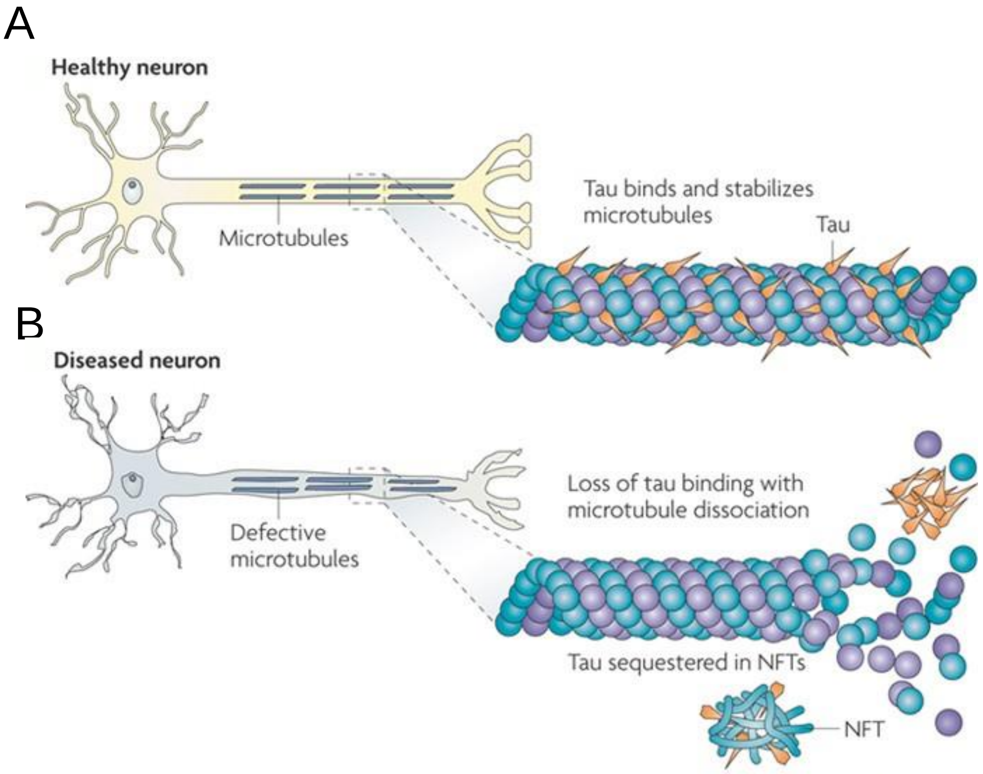


Figure 1.5: Hyperphoshorylated tau dissociation from microtubules and aggregation into NFTs. Shown is a schematic figure depicting **(A)** a healthy neuron with normal tau and **(B)** a diseased neuron with hyper-phosphorylated tau (orange spikes) detaching from microtubules. This results in the formation of NFTs and interruption of neuronal function essential for synaptic transmission. Figure is taken from Brunden et al. (2009).[55]

Endocytosis

Endocytic processing - the internalisation of substrates into the cell - is directly implicated in AD due to the distinct cellular localisation of secretases involved in the amyloidogenic processing of APP[44] (**Figure 1.4B**). Contrary to the non-amyloidogenic pathway that predominantly occurs at the plasma membrane[56] (**Figure 1.4A**), amyloidogenic processing of APP takes place in the endosome and is spatially regulated: BACE1 and PSEN1/γ complex are localised at the plasma membrane and thus must first undergo endocytosis before assemblage with PSEN2/γ secretase at the endosome (**Figure 1.6**). Increasing evidence suggests that regulation of this endocytic pathway is altered in AD, creating an intracellular pool of Aβ peptides[57] that coincide with cognitive deterioration in AD mouse models[58–60] and is more strongly associated to neuronal loss than Aβ plaques.[61] Indeed, several risk genes emerging from recent GWAS are directly involved in the endocytic regulation of APP processing, such as *Bin1*, *Picalm*, and *Sorl1*[62, 63](**Figure 1.6**) (more details of these genes are provided later in **Table 6.2**).

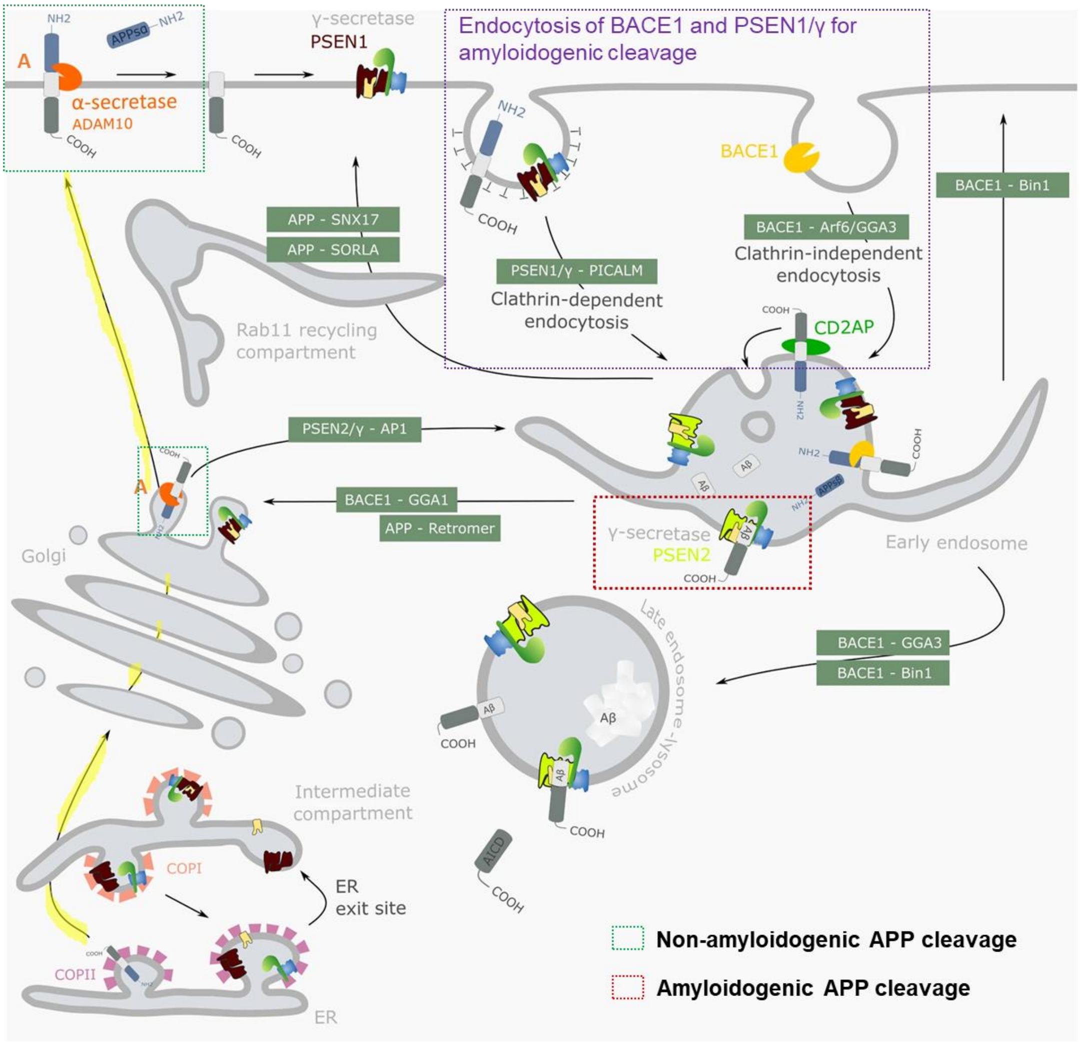


Figure 1.6: Spatial regulation of APP processing. Shown is a schematic figure depicting APP trafficking and processing through the non-amyloidogenic (**Figure 1.4A**), which predominantly occurs at the plasma membrane (boxed green), and the amyloidogenic pathway (**Figure 1.4B**), which preferentially occurs in the endosome (boxed red). APP processing through the amyloidogenic pathway is spatially regulated by the localisation and distinct internalisation of assembled PSEN1/γ complex and BACE1 at the plasma membrane (boxed purple) and PSEN2/γ in the endosome. Figure is adapted from Acker et al. (2019).[44]

Immune response

Profound neuroinflammation - an inflammatory response within the CNS primarily orchestrated by the activation of microglia (microgliosis) and astrocytes (astrogliosis) - is widely implicated in AD development and pathology.[64,65] While the exact role of the immune response is poorly understood in AD, it is widely accepted that an imbalance of the innate immune response is at play. This includes an extensive release of pro-inflammatory neurotoxic cytokines from activated microglia[66] (**Figure 1.7A**), which can trigger neuronal apoptosis[67,68] and β-secretase upregulation[69] (**Figure 1.7B**). Amyloid plaques are further known to be enriched with activated microglia,[70] suggesting that the phagocytic ability of these plaque-associated microglia to remove Aβ is compromised.[71] Notably, this is a complex process that involves the recognition of toxic species by receptors (such as TREM2, CD33 and CR1, **Figure 1.7B**), whose genes have been consistently identified in GWAS as risk genes for AD.

Lipid metabolism

The identification of *APOE* ϵ4 allele as the strongest LOAD-associated genetic variant directly established a link between lipid metabolism and AD. Increasing evidence further postulate that Aβ clearance is regulated in an ApoE isoform-dependent manner (ApoE2, ApoE3 and ApoE4),[73] with ApoE4 having the lowest binding affinity to Aβ and consequently, being the least efficient at Aβ clearance compared to other ApoE isoforms.[74] The lipidation status of ApoE, mediated by ABCA1,[75] is also known to impede Aβ aggregation, with ApoE4 being the least lipidated.[76] Recent studies have further shown that Aβ uptake is reduced in ApoE4-expressing microglia compared to ApoE3-expressing microglia, which is exacerbated upon Trem2 deficiency, revealing an interplay between lipid metabolism and immune response.[77] Notably, carriers of the ϵ4 allele have more pervasive amyloid plaques than non-carriers.[78,79]

More broadly, lipid metabolism is implicated in AD development in that APP, β- and γ-secretase are all transmembrane proteins (**Figure 1.6**). APP trafficking and processing are subsequently influenced by lipid membrane constitution and organisation.[80] Notably, β-secretase activity is indirectly modulated by *ABCA7*, which encodes an ATP-binding cassette transporter essential for regulating lipid membrane composition and was recently identified as an AD risk gene from GWAS.[81,82]

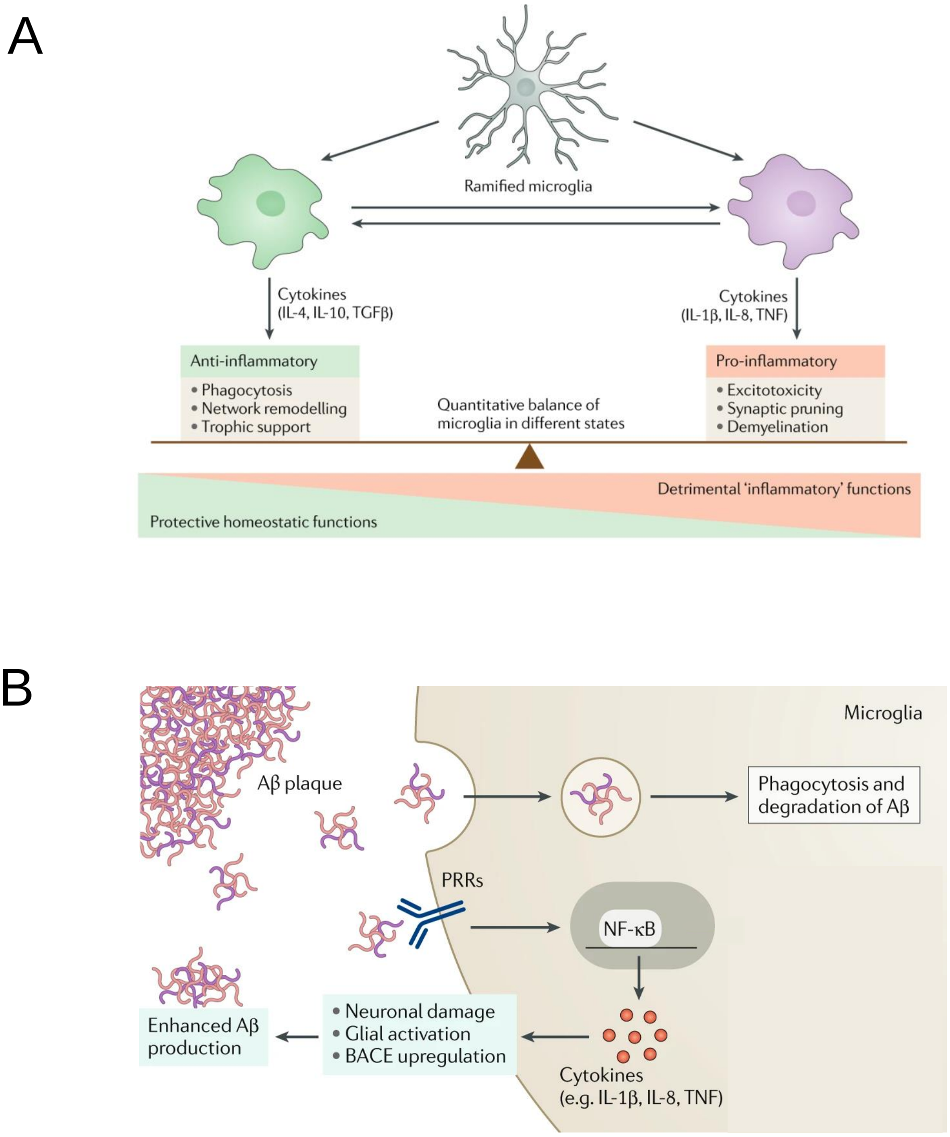


Figure 1.7: Role of microglia in AD development and pathology. Shown is a simplified schematic figure illustrating the **(A)** multifaceted roles of microglia in AD, ranging from a protective to a detrimental role by the respective secretion of anti- and pro-inflammatory cytokines, and **(B)** microglia's dual response to Aβ plaques, either through Aβ clearance or the release of pro-inflammatory cytokines. Under physiological conditions, the microglia is ramified. Pattern recognition receptors (PRRs), such as TREM2 and CD33, are found on the cell surface of microglia and are involved in recognising toxic species for phagocytosis. Both figures are adapted from Leng et al. (2021).[72]

1.1.4 Modelling AD pathology: transgenic mouse models

While the profile of human post-mortem brain tissue is typically considered the gold standard for studying AD pathogenesis, there are many limitations. Various confounding secondary factors (such as environmental exposures including diet, medication, among others) and technical difficulties (such as agonal state and post-mortem interval which impact RNA quality) all need to be considered. It further becomes challenging to resolve age-dependent and disease-associated changes, given post-mortem brain tissue can only be evaluated at the time of death and typically represent the end-stage of the disease. Conversely, mouse models of disease can be tightly controlled (for example, the genotype, living conditions, age and pathological status, among others) to track the progression of pathology in disease-relevant tissue. As such, current mouse models act as a valuable reductionist tool to dissect the processes that drive the onset and progression of AD pathology.[83]

To study the different aspects of pathology, a number of transgenic AD mouse models have been developed with mutations that either result in amyloidopathy (Aβ plaque formation) or tauopathy (NFT formation) (summarised in **Table 1.1**). The development of amyloidopathy is typically achieved through the insertion and overexpression of human *APP*, either alone or in combination with *PSEN1*, whereas tauopathy is recapitulated by overexpressing human *MAPT* with FTD-associated mutations (given no causative *MAPT* mutations have been identified in AD).

In this thesis, we have utilised the rTg4510 mouse model at 4 different ages (2, 4, 6 and 8 months) to profile progressive transcriptomic variation associated with the development of tau pathology, with a focus on the entorhinal cortex - a key region that is implicated early in AD pathogenesis (as depicted in **Figure 1.2**). The range of ages selected reflect the development of age-dependent tauopathy in these mice with the appearance of pretangles from as early as 3 months to synaptic and neuronal loss by 9 months. These mice further develop age-dependent cognitive and behavioural deficits at 6 months, coinciding with the development of mature NFTs.[84] Importantly, the spread of neuropathology in this mouse model closely recapitulates the Braak stages in human AD brains, with progressive accumulation of tau in the entorhinal cortex and hippocampus (**Figure 1.8**).

Table 1.1: Representative AD mouse models. Tabulated is a list of the most widely used AD mouse models developed from overexpression of one or more genes associated with FAD or FTD. Shown is also the age of these mice (in months) at which point they develop the characteristic pathological hallmarks of AD and show signs of cognitive decline. Table is adapted from Hall et al. (2012)[83] and is by no means comprehensive. mo - months.

Mouse models		Mutations	Plaques (mo)	Tangles (mo)	Neuronal loss (mo)	Cognitive deficit (mo)
hAPP	PDAPP	Ind[b]	6	x	x	6
	Tg2576	Swe[a]	11	x	x	>12
	J20	Swe,[a] Ind[b]	6	x	x	4
	APP23	Swe[a]	6	x	14-18	3
hAPP/PS1	PS/APP	Swe,[a] M146L[e]	6	x	22	3
	APP/PS1	Swe,[a] PSEN1dE9	6	x	x	6
	5xFAD	Swe,[a] Lon,[c] Flo,[d] M146L,[e] L28V[e]	2	x	9	4
hTau	hTau.P301S	MAPT P301S	x	4	3	3
	3xTg	Swe,[a] MAPT P301L, M146V[e]	6	12	-	4
	rTg4510	MAPT P301L	x	4	6	3
	htau	Wild-type	x	9	10	6

[a] Swedish APP mutation K670N/M671L

[b] Indiana APP mutation V717F

[c] London APP mutation V717I

[d] Florida APP mutation I716V

[e] Human PSEN1 mutations

rTg4510 mice are produced by crossing the responder line (which carries a human MAPT^{P301L} transgene downstream of a tetracycline operon-responsive element (TRE)) with an activator line (which expresses a tetracycline-controlled transactivator (tTA) under the control of the calcium calmodulin kinase II promoter (CaMK2a). Notably, despite the popularity and wide use of this mouse model, recent studies have reported disruptions of endogenous mouse genes from insertion of the CaMK2a-tTA and MAPT^{P301L} transgenes:[85] i) the CaMK2a-tTA transgene was found inserted on chromosome 12, resulting in a 508 kb deletion that affects *Vipr2*, *Wdr60*, *Esyt2*, *Ncapg2*, and *Ptprn2*, ii) the MAPT^{P301L} transgene was found to integrate within the *Fgf14* gene on chromosome 14, resulting in a 244kb deletion. More details on this mouse model are provided in **Chapter 2**.

It is also important to note that there is currently no AD mouse model that encapsulates all the defining features of AD. While various assays (such as Y Maze Spontaneous Alternation test, Open Field test, among others) can be used to assess cognitive and behavioural changes in AD mouse models associated with pathological progression,[86] one of the major criticisms of current mouse models relates to how representative they are of sporadic LOAD. While there have been recent efforts to generate mouse models that more closely resemble LOAD with the incorporation of AD-associated variants, such as the *APOE* $\epsilon 4$ variant,[87,88] these models are not yet widely used.

Despite these concerns, a recent meta-analysis of differential expression studies in human post-mortem samples revealed that many transgenic mice display gene expression signatures that significantly overlap with human AD-associated co-expression modules, particularly neuronal and microglia-enriched modules.[89] While they concluded that there is a minority of human AD-associated co-expression modules that were poorly recapitulated by current AD mouse models (such as genes involved in proteostasis regulation), they highlighted the utility of mouse transcriptomic data from multiple time points to accelerate discovery of AD progression markers and identify critical time-points for interventions. Notably, they found that memory task impairment and neurodegenerative pathology of transgenic mice from the tau mouse model, rTg4510 (**Table 1.1**), at 4 and 6 months corresponded to activation of the neuronal and microglial expression pattern, respectively.

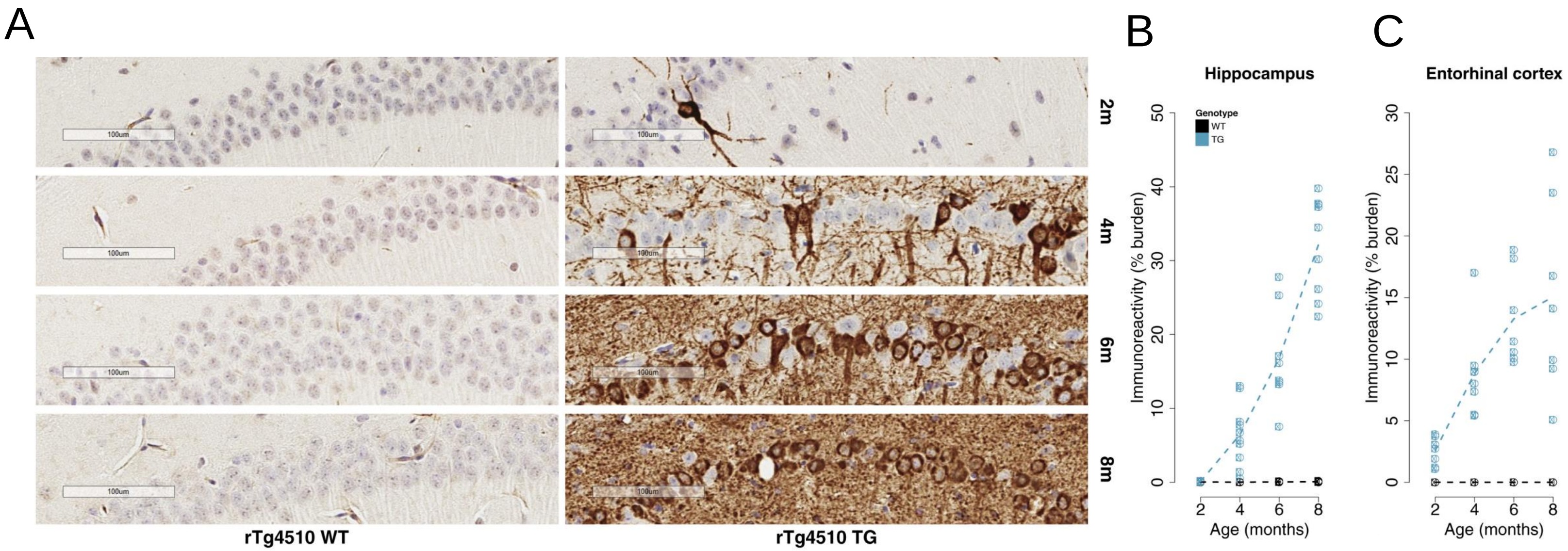


Figure 1.8: Progressive pathological characterisation of rTg4510 mouse model. Shown are **(A)** representative immunohistochemistry images from the hippocampus showing accumulation of tau in rTg4510 transgenic (TG) mice compared with wild-type (WT) mice at 2, 4, 6 and 8 months of age. **(B)** Progressive accumulation of tau is observed in the hippocampus and the **(C)** entorhinal cortex in rTg4510 TG but not WT mice. Figures and legends are adapted from Castanho et al. 2020.[90]

1.2 Transcriptional dysregulation in AD

The vast majority of GWAS variants associated with LOAD are annotated to non-coding regulatory regions of the genome.[25, 27] They are further found to be enriched in open chromatin regions that promote transcription such as enhancers[91] (short DNA sequences containing specific motifs for binding of transcription factors), suggesting that AD-associated risk variants mediate disease associations through gene transcriptional regulation. As such, efforts have been made in profiling the complete set of expressed messenger RNA (mRNA) transcripts (hereby referred to as the "transcriptome") in human post-mortem brain tissue and AD mouse models to better understand the transcriptional variations underlying AD pathogenesis.

A number of theses studies have identified widespread gene differences in human AD post-mortem brain tissue (reviewed in **Table 1.2**) and in transgenic mice harbouring AD-associated mutations (reviewed in **Table 1.3**). Recent works in our group have similarly identified robust differences in gene expression associated with tau pathology development in rTg4510 transgenic mice, with these genes enriched in biological pathways previously implicated in AD pathology.[90]

Despite the power of these studies to identify differences at a gene-level, one major limitation of these studies is that they have broadly ignored identifying differences at the RNA transcript level, given that efforts to perform these analyses up until now have largely been hampered by technology (more details to follow in **Section 1.3.1**). Deeper examination and investigation of variation at the transcript level however will be essential, particularly since differences at the transcript level do not necessarily translate to differences at the gene level. Expression of transcripts in opposite directions may also result in zero net change at the level of gene expression. Transcriptomic profiling of disease-relevant tissue to identify differences at a transcript level associated with AD pathology would therefore be important, especially given that there is an increasing interest in the role of aberrant alternative splicing in Alzheimer's disease.[92]

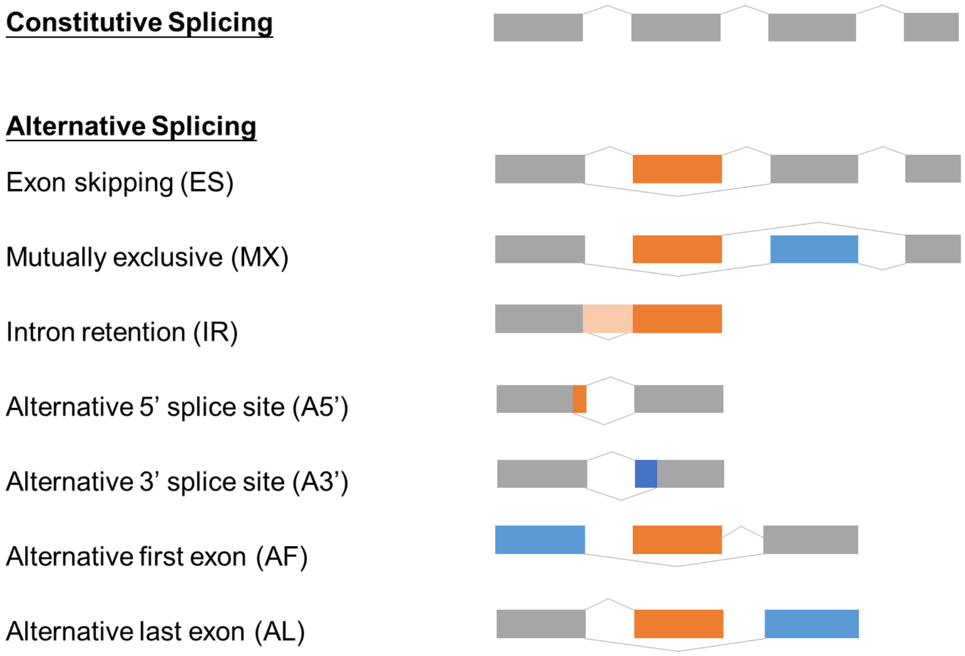
1.2.1 Alternative splicing

Alternative splicing (AS) is a transcriptional regulatory mechanism that produces distinct RNA transcripts (isoforms) from a single gene, which are potentially translated to different protein isoforms with unique, and potentially, antagonistic functions.[93] It is a widespread phenomenon with over 95% of human genes estimated to be influenced,[94] and is most prevalent in the brain,[95] where it impacts upon neuronal development and maintenance.[94,96,97] There is a growing recognition of the key role of aberrant mis-splicing in neurodegenerative and neurodevelopmental disorders,[98,99] including schizophrenia and autism.

Mechanism

Splicing involves the removal of non-coding sequences (introns) from mRNA precursors and the ligation of coding sequences (exons), resulting in isoforms with different exonic structures (**Figure 1.9**). This relies on the concerted and regulated assembly of the spliceosome - a multimegaton, dynamic ribonucleoprotein complex - by its recognition and stepwise-binding to sequence elements within the pre-mRNA (cis-acting elements), and a group of RNA-binding proteins (trans-acting splicing factors) (**Figure 1.10**). This process is highly regulated in a temporal and cell-specific manner, and requires multiple components at work, including i) the sequence of cis-acting elements within the exon or intron, which determines the binding affinity of splicing factors that either enhance or suppress splicing, ii) the availability of such splicing factors, iii) the functional coupling of transcription and splicing, mediated by epigenetics, chromatin and RNA structure, and iv) the polymerase processivity and elongation rate.

Correct splicing first requires recognition of short sequence motifs upstream (5' splice site, 5'SS, donor site) and downstream (3' splice site, 3'SS, acceptor site) of the intron/exon boundary (splice junctions) (**Figure 1.10A**), followed by sequential assembly of the spliceosome components and intron excision[100](**Figure 1.10B**). The 5'SS is typically defined by a conserved 9-nucleotide sequence with a GU(T) dinucleotide, and the 3'SS by a polypyrimidine tract (PPT) followed by a conserved AG dinucleotide.[101] Almost all introns in human and mouse are flanked by the GT-AG splice site dinucleotide,[102] with other dinucleotide variations known to exist in very minute proportions: GC-AG and AT-AC comprises ~0.9% and ~0.09% of human splice sites respectively.[103] An increasing number of diseases are linked to aberrant alternative splicing predominantly by the presence of pathogenic variants disrupting these cis-acting elements and interfering with the functional activity of trans-acting protein splicing factors.[104] Of note, variations in 5'SS and 3'SS can result in exon skipping, exon inclusion, exon extension or exonic splice gain whereas variations in the intron can result in intron gain (intron retention).



Constitutive Splicing
Alternative Splicing
Exon skipping (ES)
Mutually exclusive (MX)
Intron retention (IR)
Alternative 5’ splice site (A5’)
Alternative 3’ splice site (A3’)
Alternative first exon (AF)
Alternative last exon (AL)

Figure 1.9: Types of alternative splicing events. Splicing involves the removal of introns (denoted here by the grey lines) and ligation of exons (represented as boxes) either constitutively (i.e. all the exons are included) or alternatively to generate isoforms with different exonic structure from i) exclusion of an entire exon (orange exon from an exon skipping event, ES), ii) mutual exclusion of two exons (orange and blue exons from a mutually exclusive event, MX) ii) inclusion of an intronic sequence (light orange region from an intron retention event, IR), iii, iv) usage of a different 5’SS (orange region from alternative 5’ splice site, A5’) or 3’SS (blue region from alternative 3’ splice site, A3’) resulting in a novel splice junction (A5’, A3’) and the v, vi) usage of different first (blue exon from alternative first exon, AF) and last exon (blue exon from alternative alternative last exon, AL). Introns and exons are depicted by the box and line respectively.

Nonsense-mediated decay

Up to one-third of alternative splicing events introduce premature termination codons (PTCs) that are typically located between 50 and 55 nucleotides upstream of the splice junction, leading to premature mRNA degradation in a process known as nonsense-mediated decay (NMD)[105] . While NMD was initially considered an RNA surveillance mechanism involved in removing unproductive and detrimental splice variants, it is now widely acknowledged as a key mechanism for gene regulation:[106] up to 25% of transcripts are estimated to be affected.[107] Notably, intron-retaining transcripts often contain a PTC, implicating the coupling of intron retention events and NMD.[108]

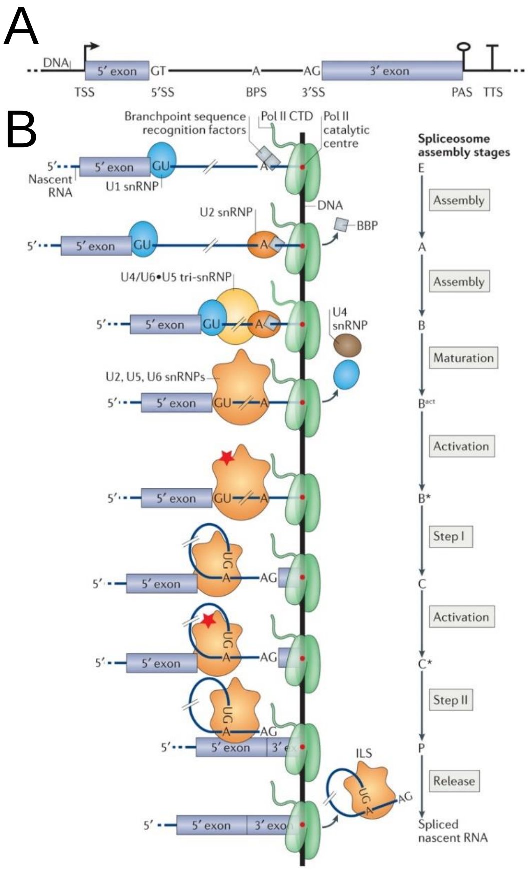


Figure 1.10: Splicing mechanism: spliceosome assembly on nascent RNA. Shown is a schematic figure of the splicing mechanism mediated through the assembly and complex rearrangement of the spliceosome. **(A)** The consensus sequence of the splice junctions that demarcate the intron/exon boundaries essential for recruitment of spliceosomal snRNPs (small nuclear ribonucleoproteins). **(B)** Co-transcriptional assembly of the spliceosome with stepwise interaction of spliceosomal snRNPs, following the formation of:

1. E (Early) commitment complex with the identification and binding of U1 snRNP to the 5’SS and branchpoint binding protein (BPP) to BPS.
2. A (Assembly) catalytically-active complex with association of U2 snRNP to the branch site following the dissociation of BPP.
3. B pre-catalytic spliceosome complex with recruitment of U4, U5 and U6 snRNPs.
4. B pre-catalytic spliceosome complex after major conformational rearrangements within the spliceosome (RNA-protein and RNA-RNA interactions) followed by the release of U1 and U4 snRNP to expose the adenosine from BP to the 5’SS.
5. B* catalytically-active complex with nucleophilic attack of adenosine on 5’SS (1st step of transesterification).
6. C catalytically-active complex with further conformational changes of the U2 snRNP to C* complex, with nucleophilic attack of the 5’SS to 3’SS (2nd step of transesterification).
7. P (post-spliceosome complex) with release of the mRNA transcript from the remaining spliceosome (ILS), now bound to the intron lariat. The snRNPs are then disassociated and recycled for the next cycle of splicing.

BBP - Branchpoint binding protein, BPS - Branch point sequence, CTD - Carboxyl-terminal domain, ILS - Intron lariat spliceosome, PAS - Poly(A) site, SS - Splice site, SnRNP - Small nuclear ribonucleoproteins, TSS - Transcription start site, TTS - Transcription termination site. Figure is taken from Herzel et al. 2017.[100]

1.2.2 Other regulatory mechanisms

Adding to the layer of complexity are other alternative RNA processing regulatory mechanisms that regulate gene expression, such as alternative transcription initiation (ATI) from alternative promoter usage and alternative transcription termination (ATT) from alternative polyadenylation (**Figure 1.11**). More than 70% of mammalian genes are reported to contain multiple polyadenylation sites, and 60% of genes are known to have two or more promoters with alternative transcription start sites.[109] Alternative transcription at these sites generate RNA isoforms with different 5' and 3' untranslated regions (UTRs) that may translate to proteins with varying N- and/or C- terminals. Importantly, the introduction of variable UTRs can further modulate transcription regulation by fine-tuning the mRNA stability, localisation and translation efficiency.[110] This is in contrast to alternative splicing where changes typically lie within the protein sequence, which could potentially influence protein structure and function.

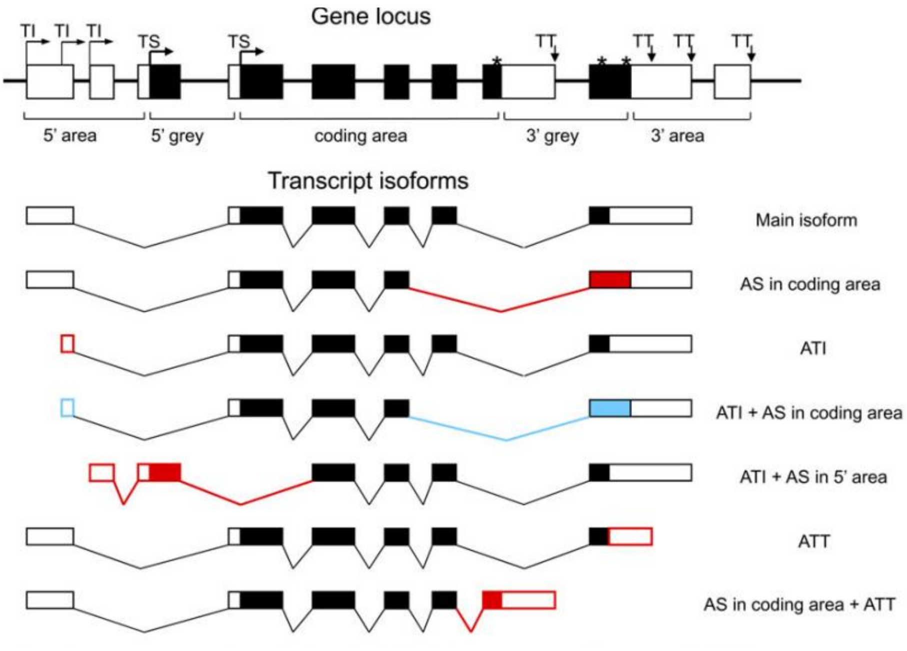


Figure 1.11: Other post-transcriptional regulatory mechanisms. Shown is a schematic figure depicting the gene locus and associated isoforms, generated from alternative splicing (AS) and other alternative transcription mechanisms at the 5' and 3' end: alternative transcription initiation (ATI), alternative transcription termination (ATT). TI - Transcription initiation site, TS - Translation initiation site, TT - Transcription termination site. Protein-coding regions are coloured in black. Frequent alternative exons (AEs) and common combinations of AEs are coloured in red. Rare AEs and avoided combinations of AEs are coloured in blue. Figure and legend are adapted from Shabalina et al. (2010).[111]

1.2.3 Altered splicing in AD

Aside from extending the diversity of the transcriptome and proteome, alternative splicing offers an additional regulatory level of transcript and protein homeostasis. Dysregulation of splicing can have significant functional consequences in driving or contributing to disease progression and susceptibility, by disrupting protein isoform function (loss-of-function or gain-of-function) or generating an imbalanced isoform ratio (i.e. a difference in relative isoform expression). Previous studies that examined the role of alternative splicing in AD have largely focused on identifying mis-splicing variants of FAD genes, such as the detection of the *PSEN1* intron 4 mutation which generates aberrantly-spliced PSEN1 isoforms that were found to increase Aβ_{42} levels *in vitro*.[112] Perhaps more well known is the altered splicing of *MAPT* whereby exclusion or inclusion of exon 10 (E10) generates isoforms with either 3 (3R tau, E10-) or 4 (4R tau, E10+) microtubule-binding repeat domains, with the latter having a greater interaction with microtubules. Of note, over 50 tauopathy-associated intronic mutations have been found clustered around the 5'SS of exon 10, favouring exon 10 inclusion[113,114] and leading to an increased 4R/3R ratio that contributes to tau aggregation.[115] Mimicking this imbalanced ratio in a tauopathy mouse model induced the production of more phosphorylated, self-aggregating tau and seizures.[116]

Notably, multiple spliceosomal components were found to co-aggregate with tau in human AD post-mortem brain tissue,[117] implicating a global change in the core splicing machinery in AD pathogenesis. Recent transcriptomic profiling of various human AD post-mortem brain regions (reviewed in **Table 1.2**) and AD mouse models (reviewed in **Table 1.3**) have revealed aberrant splicing as a hallmark of AD. Hundreds of genes have been reported to be differentially spliced with widespread transcript expression differences and usage of alternative splicing events in genes such as *APOE, BIN1* and *APP*.[92,118] Furthermore, these AD-associated splice variants (splicing quantitive trait loci - sQTL) were enriched in transcriptionally active regions that overlap with other AD-associated SNPs and epigenetic variants, highlighting the genomic complexity underlying AD pathogenesis.

Table 1.2: Transcriptome profiling studies of AD human post-mortem brain tissue. Tabulated is a review of transcriptome profiling studies of human post-mortem brain tissue, revealing mis-splicing as a widespread hallmark of AD.
AS - Alternative splicing, ES - Exon skipping, hnRNPs - Heterogeneous nuclear ribonucleoproteins, IR - Intron retention, LOF - Loss-of-function, NMD - Nonsense-mediated decay, qPCR - Quantitative polymerase chain reaction, RBPs - RNA-binding proteins, RNA-Seq - RNA sequencing, snRNPs - Small nuclear ribonucleoproteins, sQTLs - Splicing quantitive trait loci, TSS - Transcription start site, TWAS - Transcriptome-wide association study.

References	Samples and tissues	Method	Key findings
Twine et al. (2011)[119]	3 AD, 33 Controls Total, frontal & temporal lobe	RNA-Seq	• *APOE* was down-regulated in AD temporal lobe. Identified 3 isoforms with different TSS and isoform expression: ENST00000252486 and ENST00000446996 from TSSA were down-regulated (3.09-fold) whereas ENST00000425718 from TSSB was up-regulated (26.5-fold). • *Ank1* was down-regulated in AD total brain.
Mills et al. (2013)[120]	5 AD, 5 Controls Parietal lobe	RNA-Seq	• Differentially expressed genes were enriched in lipid metabolism (*ACOT1*, *ACOT2* and *DBI/ACBP* were up-regulated, whereas *TECR* was down-regulated). • Differential isoform expression observed in *DBI/ACBP*: non-coding DBI-009 was up-regulated while protein-coding DBI-003 was down-regulated.
Bai et al. (2013)[117]	18 AD, 17 Controls Cortex	Mass Spectrometry, RNA-Seq	• Extranuclear aggregation of 36 proteins, including spliceosomal components (U1 snRNP, U1-70K). • Accumulation of unspliced RNA molecules in AD, with reduced splicing efficiency (increased ratio of pre- and mature RNA) in AD-associated genes (*BACE1, BIN1, CLU, GFAP, PICALM, PSEN1, SORL1*).
Lai et al. (2014)[121]	8 AD, 8 Controls Superior temporal gyrus	Microarray	• 22 genes identified with differential AS events (characterised by differential exon usage). • *GNAL* transcript variant 5 down-regulated in AD whereas transcript variant 1 showed no change. • *MAP4* transcript variant 3 down-regulated in AD whereas transcript variant 1 was up-regulated.
Mills et al. (2014)[122]	14 AD, 16 Controls Superior temporal gyrus	RT-qPCR	• No difference in total *APOE*, *APOE-005*, or *APOE-001* expression between AD superior temporal gyrus and controls, contrary to Twine et al. (2011).[119]

Humphries et al.(2015)[123]	10 AD, 10 Controls, Temporal lobe	RNA-Seq	• 9 genes were differentially expressed (i.e. *ABCA7, CR1*), 5 of which had differential splicing (*ABCA7, TMEM259, EPHA1, MS4A6A, MS4A6E*) defined by differences in overall exon distribution.
Magistri et al.(2015)[124]	4 AD, 4 Controls Hippocampus	RNA-Seq	• Down-regulation of *TAC1* and upregulation of *SERPINE1.* • Pathway analysis indicates dysregulation in neural communication and Aβ clearance.
Alkallas et al. (2017)[125]	6 AD, 5 Controls Dorsolateral cortex	RNA-Seq	• RBFOX1 (RBP) is down-regulated in AD; reduced stability and abundance of RBFOX-regulated transcripts that encode synaptic transmission proteins, contributing to a loss of synaptic function.
Annese et al. (2018)[126]	6 AD, 6 Controls Hippocampus, Temporal and Frontal lobe	RNA-Seq	• 2,122 differentially expressed genes, including upregulation of *TESPA1, CPLX3 SERPINA5, SERPINA1* and dowregulation of *NEUROD6, NEUROD1, LOC400891, CAMK1D.* • Deregulated micro-RNA (miR-132/212) and general decrease in RNA editing in AD.
Raj et al. (2018)[92]	268 AD, 182 Controls Dorsolateral prefrontal cortex	RNA-Seq	• 84 differentially spliced genes defined by differential intron usage, 11 of which were also differentially expressed (*PFKP, NDRG, APP, PICALM, CLU*). • sQTLs enriched in RBPs, including *PTBP1, ELAVL1* and multiple hnRNPs. • TWAS identified 21 genes with differential intron usage associated with AD, including *CR1,PTK2B, CLU, AP2A1, AP2A2, MAP1B.*
Johnson et al. (2018)[127]	20 AD, 13 Controls Dorsolateral prefrontal cortex	Mass-Spectrometry	• More alternative exon-exon junction peptides mapped to *MAPT, BIN1, PTK2B, FERMT2* in AD. • Higher levels of RBPs detected in AD, with enrichment in modules correlated with tau pathology.
Han et al. (2019)[128]	24 AD, 50 Controls Hippocampus	RNA-Seq	• 3 AD-associated ES events in *RELN* & *NOS1*, resulting in truncated proteins with loss of functional domains. A SNP was identified adjacent to *RELN* skipped exon & within splicing regulatory element.

Adusumalli et al. (2019)[129]	4 2AD, 38 Controls[117] Frontal cortex	Mass Spectrometry	• 1,136 differential IR events annotated to 781 genes (including *BIN1, MAPT*), which were enriched in mRNA export and splicing, and had significantly different protein levels between AD and controls. • Differentially retained introns have higher GC content, suggesting that DNA methylation changes may contribute to differential IR.
Fan et al. (2021)[130]	210 AD, 191 Controls[a]	RNA-Seq	• 2 isoform modules with 38 isoforms up-regulated in AD, 33 of which has not been reported as AD-associated (including *ANLN, DOCK5, ERBB3, SEPT8, UGT8*). 67 genes were identified with differentially expressed isoforms in different modules.
Yang et al. (2021)[131]	1074 AD, 608 Controls 9 brain regions	RNA-Seq	• 1,530 differential ES events in 1,103 genes enriched in endocytosis; 2,415 differential ES events in 1,701 genes associated with tau progression and enriched in axon guidance. • *MBP* exon 5 skipping; *ASPH* exon 5 and exon 8 skipping in cerebellum. • 15,556 ES events in 113 RBP (*CHL1* exon 25, *ASPH* exon 5); *RBM3, RBPMS2, AZGP1, RPS16* were differentially expressed in STG; SNP in *ABCA7* donor site was associated with exon 2 skipping. • 70% of genes predicted LOF due to ES, with most significant loss attributed to serine phosphorylation; 86 AD genes with partial function loss were enriched in neuronal development.
Garcia-Escudero et al. (2021)[132]	32 AD (Braak I-VI), 10 Controls	qPCR, Western Blot	• Identified novel human-specific truncated Tau isoform with intron 12 retention, which was down-regulated in AD and less prone to aggregate compared to other tau isoforms
Li et al. (2021)[133]	84 AD, 80 Controls Temporal cortex	RNA-Seq, Mass Spectrometry	• Higher intron-retained levels in AD (including *BACE1, BIN1, PICALM*) but no differential gene expression, suggesting a compensatory mechanism. • HMBOX1, a transcription factor involved in the innate immune response, had the strongest differentially expressed intron level associated with tau pathology. • Increased IR was associated with reduced protein expression, possibly due to NMD.

Table 1.3: Transcriptome profiling studies of AD mouse models. Tabulated is a review of transcriptome profiling studies of AD mouse models, also revealing mis-splicing as a widespread hallmark of AD.
AS - Alternative splicing, hnRNP - Heterogeneous nuclear ribonucleoproteins, FAD - Familial Alzheimer's disease, GWAS - Genome-wide association studies, NFTs - Neurofibrillary tangles, NMD - Nonsense-mediated decay, RBP - RNA-binding protein, RNA-Seq - RNA sequencing, TG - Transgenic, WT - Wild-type.

References	Mouse Models & Tissues	Method	Key Findings
Maziuk et al. (2018)[134]	rTg4510 (8 TG, 8 WT, 2 - 8 months) Frontal cortex	Immuno-histochemistry	• 65% of RBP showed decreased tau association, except EWSR1, TAF15 and hnRNPA0 which co-localised with phosphorylated-tau but not mature NFTs.
Rothman et al. (2018)[135]	TgCRND8 (25 TG, 27 WT, 1.5 - 10 months) Whole cortex	RNA-Seq	• Progressive genotype-associated transcriptomic changes with upregulation of genes overlapping with human inflammation and microglia signatures (i.e. *Trem2, Tyrobp, Cd68. Clec7a, Tspo, Itgfax*). • Vast majority of transcriptional changes were enriched within plaques.
Apicco et al. (2019)[136]	PS19 (3 TG, 3 WT) Cortex	RNA-Seq	• Reduced expression of transcripts encoding spliceosomal protein and RBP with altered splicing of genes involved in synaptic function and glutamatergic synaptic transmission (i.e. *Snap25, Camk2b, Gria2*). • Reducing TIA1 (RBP) partially corrected splicing dysregulation associated with tauopathy.
Salih et al. (2019)[137]	5 mouse models including hTau	RNA-Seq	• *Trem2* significant up-regulated with amyloid deposition in *APP* TG mice. • Identified 4 novel genes (*OAS1, LAPTM5, ITGAM/CD11b and LILRB4*) from mouse amyloid-responsive microglia & 7 GWAS established genes (*TREM2, ABI3, CD33, INPP5D, SP11, PU.1, MS4A6D, GAL3STF*) that overlapped with human AD GWAS studies.
Castanho et al. (2020)[90]	rTg4510 (64 TG, 64 WT, 2 - 8 months) Entorhinal cortex	RNA-Seq	• Genotype-associated differences in rTg4510 with 1,762 genes differentially expressed, including *Gfap, Cd68, Itgax, Clec7a* and others robustly associated with FAD (*App, Trem2, Clu, Picalm, Cd33*). • Differentially-expressed genes were enriched in immune response (upregulation of *C1qa, C1qb*).

1.3 Transcriptome profiling: issues & opportunities

1.3.1 Limitations of short-read RNA sequencing approaches

Transcriptomic profiling of AD pathology has been traditionally performed using exon microarrays and more recently, RNA sequencing (RNA-Seq) (summarised in **Table 1.2**), which involves high throughput parallel sequencing of amplified DNA templates in a "sequence-by-synthesis" fashion. Through larger sample sizes and significant advances in bioinformatic tools, RNA-Seq provides a more comprehensive annotation of the transcriptome and enables deeper interrogation of AS events, particularly exon skipping and intron retention (depicted in **Figure 1.9**).

Despite the power of RNA-Seq to quantify gene expression differences, efforts to characterise isoform diversity and perform transcript-based analyses are constrained by the fact that standard RNA-Seq approaches generate short reads that cannot span full-length (FL) transcripts (**Figure 1.12**). RNA-Seq reads typically have an average length ranging from 50bp to 700bp (depending on the sequencing platform), whereas transcripts are on average 2kb to 3kb; 50% of human transcripts are over 2.5Kb[138] and range from 60bp to 103kb.[138, 139] The longest known human processed transcript to date is Titin with 363 exons spanning over 106kb.[140] Consequently, while short-reads are sufficient for gene-based analyses with accurate exon identification within the associated gene, RNA-Seq fails to capture exon connectivity essential for transcript assembly and transcript-based analyses.[141, 142]

Various bioinformatic tools have been developed to overcome this challenge of transcript reconstruction by probabilistic assignment of short reads to isoforms and exon-exon boundaries.[143–145] However, this is computationally challenging, often resulting in conflicting outcomes and limited success,[146] compounded by the fact that isoforms often have significant overlaps and only a minor proportion of reads span splicing junctions; a survey of current tools revealed that only 40% of known human transcripts were assembled.[146] These tools further rely heavily on reference genome annotations or predefined splicing events, which can be inaccurate and incomplete, resulting in prediction of transcripts that do not exist (false positives) or failure to detect true transcripts (false negatives).[145] Pre-defined transcript models are particularly limiting when comparing splicing profiles between different conditions, as any splicing changes observed are likely to be condition-specific and novel.

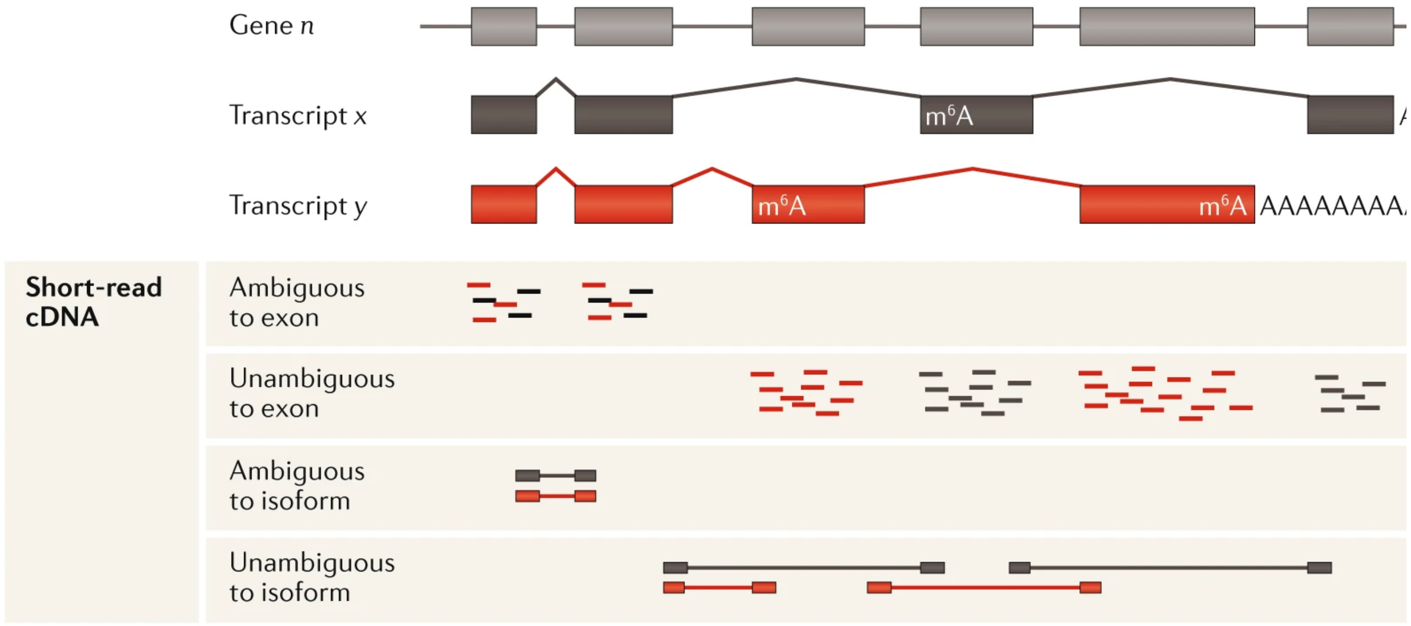


Figure 1.12: Challenges of using short-read RNA-Seq data for transcript assembly due to the generation of short-reads. Over 90% of human genes (Gene n) are alternatively spliced to generate multiple distinct isoforms (Transcript x and y). The ability to recapitulate the structure of these transcripts is limited in short-read RNA sequencing, due to the generation of short reads that cannot span the full-length transcript. Consequently, a significant number of short reads either map ambiguously to shared exons or to common junctions between isoforms. Reads that span unique exon–exon junctions can be used, however these are confounded by other limitations such as misalignment and the usage of an incomplete reference genome annotation. Figure is adapted from Stark et al. (2019).[147]

1.3.2 Leveraging long-reads for isoform annotation

The major challenges of using reads from short-read RNA-Seq for transcript assembly have been addressed with the recent emergence of long-read sequencing technologies. Instead of sequencing cDNA templates in a "wash-and-scan" fashion that result in de-phasing and the subsequent generation of shorter reads (**Figure 1.13A**), long-read sequencing technologies capitalise on real-time sequencing of templates in an uninterrupted and processive manner. This provides an unprecedented ability to generate longer reads that span the entire lengths of transcripts from the 5' end to the poly-A tail, thereby resolving splicing junctions and relinquishing the need for transcriptome assembly (**Figure 1.13D**). Pacific Bioscience (PacBio) single-molecule real-time (SMRT) and Oxford Nanopore Technologies (ONT) nanopore sequencing currently dominate this space (**Figure 1.13B,C**), with both platforms generating reads over 10kb (~15kb for PacBio and >30kb for ONT) when sequencing the whole genome or transcriptome.

With incremental improvements in chemistry, subsequent increases in throughput and diminishing sequencing costs, an increasing number of studies have leveraged both long-read platforms to characterise isoform diversity and splicing with notable success (summarised in **Table 1.4** and **Table 1.5** for PacBio Isoform Sequencing (Iso-Seq) and ONT nanopore sequencing respectively). A common finding across all these studies is the identification of widespread isoform diversity in the human and mouse transcriptome,[138, 145, 148, 149] revealing a significant number of genes with >10 isoforms, many of which were novel and not currently annotated in the reference genome; a comparative analysis with 100 million RNA-Seq reads showed that the number of isoforms and exons was saturated at four even with increased sequencing depth.[149] Validation of these novel isoforms with proteomic approaches further suggest that some of these isoforms with novel splice junctions and splicing events are functionally relevant with biological implications.[150]

However, the majority of existing long-read transcriptome studies on the human or mouse transcriptome have either been performed on cell lines or involved profiling a relatively small number of tissue samples. Targeted sequencing of mouse or human post-mortem brain tissue were further constrained to a few selective genes.

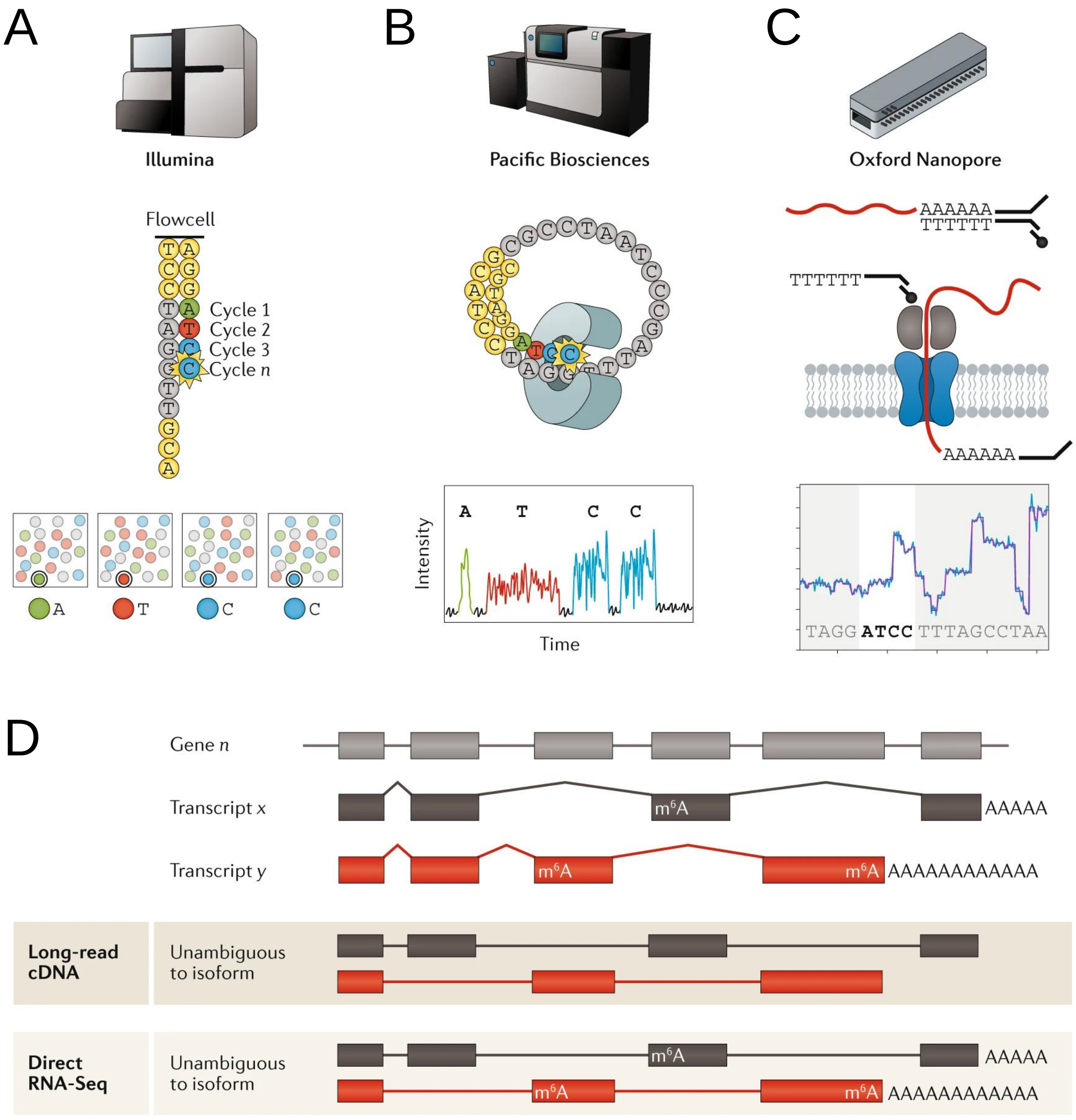


Figure 1.13: Long-read sequencing approaches capitalise on real-time sequencing of templates, generating long-reads that span the entire transcript. Shown is an overview of the three main sequencing approaches for transcriptome profiling:

- **(A)** Short-read RNA-Seq on the Illumina platform: cDNA is sequenced in a "sequence-by-synthesis" fashion with complementary binding, scanning and washing of fluorescently-labelled nucleotides at each sequencing cycle.
- **(B)** Long-read RNA-Seq on the PacBio platform: cDNA is sequenced in real-time by the incorporation of fluorescently-labelled nucleotides with immobilised polymerases.
- **(C)** Long-read RNA-Seq on the ONT platform: cDNA or RNA is translocated through a nanopore, causing a base-dependent change electric current.

(D) Long-read sequencing approaches generate long reads that span the full-length transcript, relinquishing the need for transcript assembly (see comparison with **Figure 1.12**). Figures and legends are adapted from Stark et al. (2019).[147]

Table 1.4: Long-read sequencing studies using PacBio Iso-Seq. Tabulated is a review of recent studies that leveraged the power of PacBio Iso-Seq for transcriptome profiling.
AF - Alternative first exon, FL - Full-length, GC - Gastric cancer, hiPSC - Human induced pluripotent stem cell, HCC - Hepatocellular carcinoma, HCM - Hypertrophic cardiomyopathy, hESCs - Human embryonic stem cells, Iso-Seq - Isoform Sequencing, ISS - Intronic splice site, NSC - Neural stem cell, SVA - SINE-VNTR-Alu, SE - Skipped exon, TSS - Transcription start site, TTS - Transcription termination site, XDP - X-linked Dystonia-Parkinsonism.

References	Samples and Tissues	Key Findings
Sharon et al. (2013)[138]	20 human tissue	• RNA transcripts up to 1.5kb were sequenced without fragmentation or amplification. • Identified 14,000 genes with >10% of reads mapping to novel transcripts.
Au et al. (2013)[145]	hESCs	• Error-correction of long-reads with short-reads enabled detection of 8,084 known and 1,800 novel isoforms.
Tilgner et al. (2014)[151]	Human lymphoblastoid	• Identified FL reads for genes < 3kb long and novel isoforms, assigned transcripts to original transcribed allele.
Treutlein et al. (2014)[152]	Mouse prefrontal cortex (n = 1)	• Targeted sequencing of *Nrxn* identified novel, abundantly-used alternatively spliced exons and splice sites potentially resulting in partial or complete deletion of domains. • Canonical AS events appear to be independent of each other, suggesting greater isoform diversity.
Schreiner et al. (2014)[153]	Mouse cortex (n = 1)	• Complementary to Treutlein et al.(2014) with deeper sequencing coverage of *Nrxn*, detecting 1,364 isoforms
Tseng et al. (2017)[154]	Human cerebellum (3 Carriers, 3 Controls)	• Targeted sequencing of *FMR1* in pre-mutation carriers at risk of fragile X syndrome identified 49 isoforms, with increased expression of novel truncated isoforms in pre-mutation group.
Aneichyk et al. (2018)[155]	NSC cell lines (n = 112)	• Targeted sequencing of *TAF1* in NSCs from XDP patients identified a novel isoform with cryptic exon inclusion from aberrant splice junctions in intron 32, coinciding with SVA insertion. • Significant down-regulation of canonical isoform coupled with upregulation of aberrant isoform in XDP NSCs.
Nattestad et al. (2018)[156]	Breast cancer cell line	• Identify novel full-gene fusion isoforms with 2-3 structural variants captured within a single read (such as *KLHDC2-SNTB1* through fusion of chromosome 8,14,17).
Dainis et al. (2019)[157]	Human heart tissue (4 HCM, 6 Controls)	• Sequencing of *MYBPC3* in HCM patients with ISS variant (E19-E20) detected abundant isoforms missing E20. • Novel isoforms identified from mutant allele (retained introns, extended & cryptic exon, & premature stop codons).

Flaherty et al. (2019)[158]	hiPSCs (4 Psychosis, 4 Controls)	• Patient-derived hiPSC-neurons (psychosis-diagnosed individuals with rare *NRXN1* heterozygous deletions) displayed aberrant *NRXN1* isoform expression, with down-regulation of *NRXN1*α owing to reduced abundance of normal isoforms and expression of 31 novel, mutant isoforms with reduced neuronal activity.
Chen et al. (2019)[159]	HCC cell cultures (n = 8)	• Identified candidate tumour-specific novel isoforms mostly from intron retention and early termination codon.
Tseng et al. (2019)[148]	Human frontal cortex (4 PD, 4 DLB, 4 Controls)	• Targeted sequencing of *SNCA* revealed usage of alternative 5'start sites, variable 3'UTR lengths and known exon skipping events (Exon3 skipping - SNCA126 & Exon5 skipping - SNCA112). • Canonical *SNCA* isoform was most abundant with isoforms containing all 6 exons accounting for 95% of abundance.
Mays et al. (2019)[149]	Human brain marrow cells (n = 2)	• Mass-spectrometry validation of a novel *EEF1A1* isoform by detection of its unique tryptic peptide fragment. • Iso-Seq identified 10-fold more isoforms than RNA-Seq, which plateaued at 4 isoforms irrespective of exon number.
Lian et al. (2019)[160]	Breast cancer cell line	• Down-regulation of novel isoform in *BAK1* in paclitaxel-resistant cells as target of chemotherapy resistance.
Huang et al. (2021)[150]	Gastric cancer cell lines (n = 10)	• Cell-line cancer specific novel isoforms with functional implications (i.e. *CD44* with novel domain). • Widespread use of alternative promoters (represented by AF) validated by mass-spectrometry data, which is up-regulated in GC of known oncogenes; novel promoters predicted to disrupt signal peptide sequence essential for cell localisation.

Table 1.5: Long-read sequencing studies using ONT nanopore sequencing. Tabulated is a review of recent studies that leveraged the power of ONT nanopore sequencing for transcriptome profiling. All studies were performed with ONT MinION unless otherwise stated.
ES - Exon skipping, IR - Intron retention, lncRNA - Long non-coding RNA, NIID - Neuronal intranuclear inclusion disease, NMD - Nonsense-mediated decay, NSCLC - Non-small cell lung cancer, ONT - Oxford Nanopore Technologies, PTC - Premature termination codon, PFC - Prefrontal cortex, SZ - Schizophrenia, SupCol- Superior colliculus, VCx - Primary visual cortex, TSS - Transcription start site.

References	Samples and Tissue	Key Findings
Bolisetty et al. (2015)[161]	Drosophila	• First paper to use ONT MinION to characterise exon connectivity; identifying 7,900 full-length isoforms from targeted sequencing of *Dscam, MRP, Mhc* and *Rdl.*
DeRoeck et al. (2017)[162]	Human cortex, blood, lymphocytes (7 EOAD)	• Targeted sequencing of *ABCA7* validated 7 known PTC mutations, identifying deleterious out-of-frame IR and in-frame skipping of respective PTC-bearing exon from usage of cryptic splice site (potential rescue mechanism).
De Jong et al. (2017)[163]	Human lymphoblastoid cell line	• Targeted sequencing of *BRCA1* identified 32 isoforms; 18 novel isoforms with multiple concurrent known ES events generating out-of-frame coding sequences with missing functional domains (majority predicted for NMD). • Enrichment of *BRCA1* performed with long-range RT-PCR resulted in biased amplification of short transcripts (< 4kB), and many of the novel isoforms had < 3 MinION reads (> 10% error rate).
Hardwick et al. (2019)[164]	Human PFC, VCx, caudate, SupCol (n = 3)	• Targeted sequencing of GWAS neuropsychiatric-associated haplotype blocks containing non-coding SNPs identified 107 novel inter-genic transcripts (novel genes) classed as putative lncRNAs. • Detected novel splicing events of known neuropsychiatric-associated genes (i.e. novel TSS of *NRGN* 20kb upstream of annotated TSS resulting in novel introns overlapping SZ-associated SNP).
Clark et al. (2019)[165]	Human brain tissue (7 regions) (n = 3)	• Long-range RT-PCR (target capture) and ONT cDNA-Seq of psychiatric risk *CACNA1C* revealed 251 isoforms, majority novel (96%); detected 5 transcripts with in-frame deletions with potential functional implications. • Brain-regional isoform expression differences with notable isoform switch between cerebellum and cortex.

Tang et al. (2020)[166]	Human CLL PBMCs (3 *SF3B1*$^{\mathrm{WT}}$, 3 *SF3B1*$^{\mathrm{k700E}}$, 3 Controls)	• Novel bioinformatics tool (*FLAIR*) to correct ONT reads with short reads, identifying aberrant 3'SS & retained intron usage in CLL samples with *SF3B1* mutation; down-regulation of intron-retained isoforms containing conserved motif upstream of 3'SS (motif only found using short-read RNA-Seq due to nanopore length bias), enriched in NMD.
Tian et al. (2020)[167]	Human cell lines & PMBCs (1 CLL) Mouse muscle stem cells	• Novel sequencing method and tool (FLT-Seq, *FLAIR*) combining scRNA-Seq, ONT cDNA-Seq & Illumina RNA-Seq. • Sequenced 2,800 single cells, identifying thousands of cell-specific novel isoforms and differential isoform usage between cell lines for genes enriched in mRNA splicing and cell-surface receptors. • Novel alternative promoters from novel isoforms overlapped with open promoter regions from scATAC-Seq data.
Robinson et al. (2021)[168]	Human and mouse macrophages	• 50% splicing changes classified as AF events following inflammation; no expression change in genes with AF usage. • Identified inflammatory-regulated *AIM2* novel isoform with novel promoter upstream (as supported by Chip-Seq), regulated by transcription factor binding, and reduced translational efficiency due to binding of iron to IRE motif.
Oka et al. (2021)[169]	Human NSCLC lines	• Identified 2021 novel isoforms (validated with mass-spectrometry), a significant proportion (30%) of which were predicted for NMD.

1.3.3 Advances in single-cell and direct RNA sequencing

During my research for this thesis, significant technological advances have been made in the realm of long-read sequencing of isoforms, predominantly in the unprecedented ability to perform sequencing at a single-cell level (scRNA-Seq) using micro-fluidic or droplet-based technology, and sequencing of native RNA molecules (rather than cDNA) using nanopore sequencing (Direct RNA-Seq) (depicted in **Figure 1.14** and reviewed in **Table 1.6**). Both sequencing approaches address limitations of current transcriptome profiling: expression analysis at the resolution of individual "single" cells allows the identification of cell-specific changes that would have otherwise been masked and averaged across in "bulk" tissue studies, whereas sequencing of direct RNA molecule reduces the risks of generating artefacts from library preparation and allows elucidation of RNA epigenetic modifications.[170]

Single cell transcriptomic analyses of AD

Motivated by the promising potential to study DNA and RNA at the single-cell level, several studies have examined the single-cell transcriptional landscape of human AD post-mortem brain tissue[171–176] and AD mouse models.[177,178] These studies have identified cell-specific transcriptional signatures associated with disease progression, and detected subpopulations of microglia and astrocytes that have an altered molecular expression profile. Proliferation of these distinct AD-associated microglial cells were accompanied with release of pro-inflammatory cytokines[178] with altered expressions of *Trem2* and *Cd33*,[171,179] attesting to the role of immune response. Importantly, these studies identified enrichment of AD-associated non-coding SNPs in microglia enhancers with cell-type specific regulation of gene expression,[172,176,180,181] corroborating the role for transcriptional dysregulation in AD pathogenesis.

While scRNA-Seq has revolutionised our understanding of the transcriptome's cellular heterogeneity, it is not without its challenges; namely, significantly low starting materials coupled with low capture efficiency renders high "dropout" events where one transcript may be highly expressed in one cell while missing in another.[182,183] This inflation of observed zeros, sparsity, can impede downstream analyses with statistical and interpretative challenges.[183] Furthermore, any increase in resolution entails an increase in dimension and stochasticity, calling for the development of novel computational tools and a scalable data framework.[182]

Direct RNA sequencing

To date, there have been no studies of direct RNA sequencing of AD-relevant tissues, primarily due to the novelty of this approach. The first study to utilise direct RNA sequencing for transcriptome profiling was published in 2019, and observed high sequencing error rate with prevalent read truncation at the 5' end.[184] Furthermore, a large amount of poly(A) RNA (500ng) is required as input, which is unfeasible to obtain from frozen tissues, rendering this approach only currently applicable to the transcriptome profiling of cell lines.

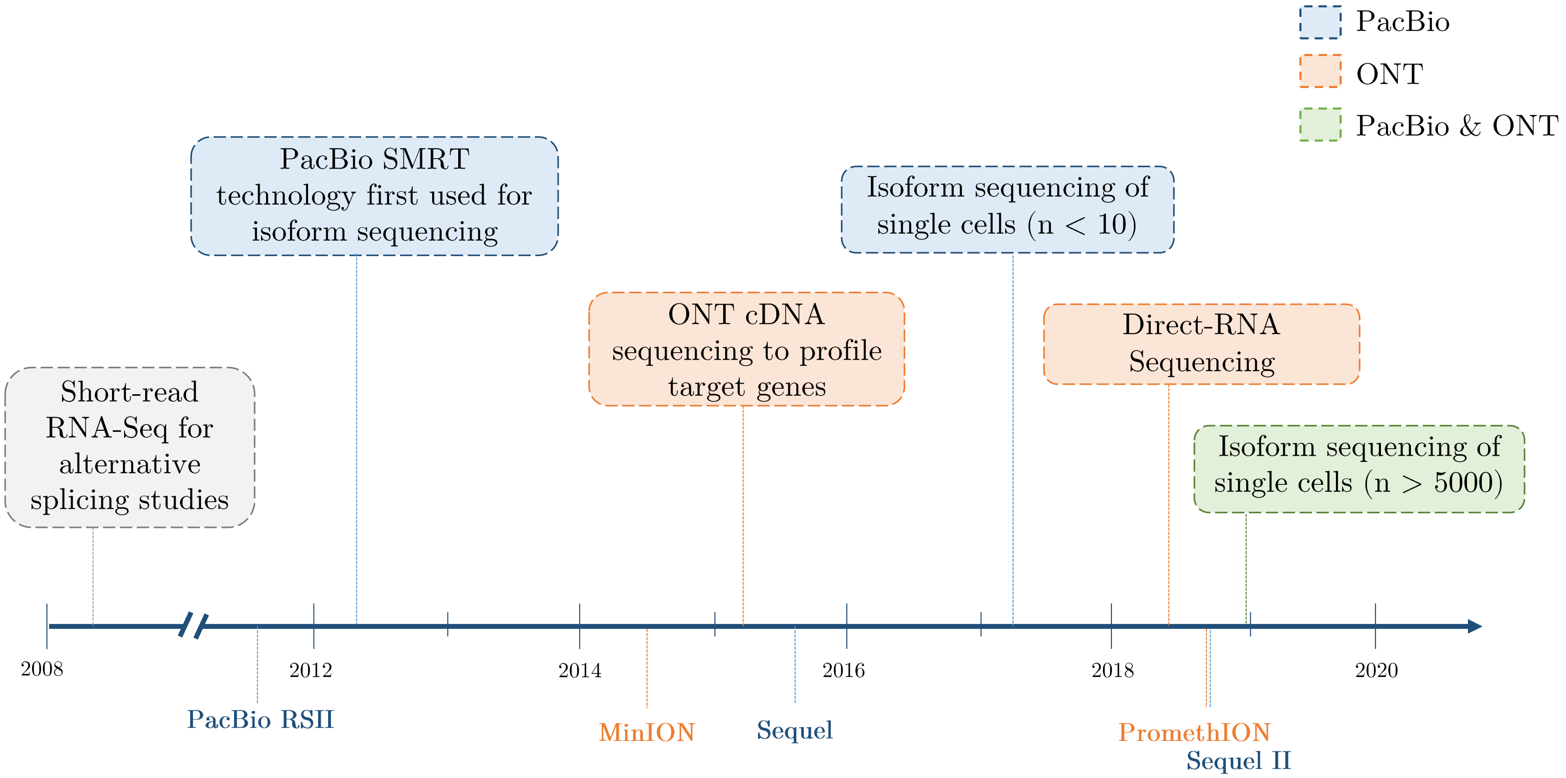

PacBio
ONT
PacBio & ONT
PacBio SMRT technology first used for isoform sequencing
Isoform sequencing of single cells (n < 10)
ONT cDNA sequencing to profile target genes
Direct-RNA Sequencing
Short-read RNA-Seq for alternative splicing studies
Isoform sequencing of single cells (n > 5000)
2008
2012
2014
2016
2018
2020
PacBio RSII
MinION
Sequel
PromethION
Sequel II

Figure 1.14: Significant advances in long-read sequencing technology & approaches. Shown is a timeline highlighting the major breakthroughs in long-read sequencing approaches using Pacific Biosciences (PacBio) single-molecule real-time sequencing (SMRT) (boxed blue) and Oxford Nanopore Technologies nanopore sequencing (ONT, boxed orange). The commercial release of respective sequencing platforms are also marked underneath the timeline.

Table 1.6: Advances in Long-read sequencing technologies: single-cell RNA-Seq and Direct RNA-Seq. Tabulated is a review of recent studies that leveraged the power of single-cell and direct RNA sequencing.
CCS - Circular Consensus Sequence, DTE - Differential transcript expression, ONT - Oxford Nanopore Technologies, ES - Exon skipping, TSS - Transcription start site, TTS - Transcriptional termination site.

References	Samples and Tissue	Key Findings
Karlsson & Linnarsson (2017)[185]	Mouse single cels (n = 6)	• High isoform diversity observed within single-cell oligodendrocytes with ~1000 distinct isoforms mapped to 700 genes with low overlap between cells, predominantly driven by alternative TSS and TTS.
Gupta et al. (2018)[186]	Mouse cerebellum (n = 1)	• Long-read sequencing of >5000 single cells (microglia, astrocytes, neurons) after isolation & barcoding. • Identified cell-specific *Bin1* isoforms with skipping of A1 and A2-A6 alternative exons (separated by constitutive exons) in all microglia, some astrocytes but not in neuronal cell-types, indicating cell-specific ES coordination.
Byrne et al. (2017)[187]	Mouse single B1a cells (n = 7)	• Identified thousands of novel TSS and TTS (within 20bp bins due to lower error rate) & hundreds of splicing events. • 160 genes with complex isoforms, 55 of which showed differential isoform usage (including B cell receptors).
Volden et al. (2018)[188]	Human single B cells (n = 96 from 1 donor)	• Circularising input cDNA and generating a CCS read (R2C2) significantly improved raw (316,000 cDNA reads at 94% accuracy) and splice site accuracy (92% vs ONT 1D raw reads at 80%, Iso-Seq CCS reads at 97%, based on SIRVs). • Ability to accurately demultiplex reads based on 7-8nt barcodes enabling mass sequencing of single cells with accurate gene quantification (strongly correlated with RNA-Seq, r = 0.79) and identification of cell-specific isoforms.
Garalde et al. (2018)[189]	Yeast ERCC RNA-Spike in mix	• Direct RNA Sequencing of yeast poly(A) RNA achieved good coverage (2.8M reads vs 5.7M reads using ONT cDNA) with negligible effect on transcript length and GC content. • Accurately identification of splice variants with no missing or novel exons from spike-in, and able to rudimentally discriminate RNA modifications (m^6A, 5-mC) using trained datasets.

Workman et al. (2019)[184]	Human B lymphocyte cell line (n = 30)	• Direct RNA sequencing of human cell line documented a high proportion (52.6%) of novel isoforms. • High error rate (14%) & significant 5'truncation due to technical issues (rapid translocation through pore, signal artefacts from enzyme stalling or strand breaks), making it difficult to ascertain TSS. • Differences in poly(A) length distribution between mitochondrial and nuclear genes, and between different isoforms of the same gene (increase in polyA tail-lengths of intron-retaining isoforms).
Sessegolo et al. (2019)[190]	Mouse brain & liver (n = 3)	• Benchmark study of Illumina RNA-Seq, ONT cDNA-Seq with/without 5'cap, & ONT Direct RNA-Seq. • Biased ONT cDNA-Seq of truncated transcripts with internal runs of poly(T) (15nt) due to cDNA synthesis; ONT RNA-Seq most accurately quantified gene expression using spike-ins, followed by RNA-Seq, cDNA-Seq.
Singh et al. (2019)[191]	Human T- & B-cell lines Tumour & paired lymph node (n = 1)	• Novel sequencing method (RAGE-Seq) combining droplet-based scRNA-Seq with target capture ONT cDNA-Seq. • Able to differentiate naive and mature B cells, and subpopulation by accurate identification of antigen receptor; track clonally related cells across tissues revealing cell-specific expression changes between tumour & lymph node.
Joglekar et al. (2021)[192]	Mouse hippocampus & prefrontal cortex (n = 2)	• Identified 400 differentially expressed genes between brain regions using gene-wise test (nx2 table with isoform counts per gene), which was governed predominantly by splice variant changes in one single cell type. • Spatial transcriptomics with Iso-Seq (Sl-ISO-Seq) confirmed localisation of brain-region specific DTE (exon-based).

Table 1.7: AD single-cell sequencing studies. Tabulated is a review of recent single cell sequencing studies of human AD post-mortem brain tissue and AD mouse models.
ARM - Activated response microglia, DEG - Differentially expressed gene, IRM - Interferon response microglia, scRNA-Seq - Single-cell RNA sequencing

References	Samples and Tissue	Key Findings
Keren-Shaul et al. (2017)[177]	5xFAD AD mouse model (1 - 8 months)	• Few isoforms shared between cells (7%) highlighting the importance of scRNA-Seq. • Identified subsets of protective microglia (disease-associated microglia - DAM), with a characteristic transcriptional activation profile: Trem2-independent manner that involves down-regulation of microglia checkpoints, followed by activation of a Trem2-dependent program for upregulation of phagocytic-related genes, essential for Aβ clearance.
Frigerio et al. (2019)[179]	App$^{\text{NL-G-F}}$ AD mouse model	• Two activated states (reactive) of microglia: i) activated response microglia (ARM) characterised by up-regulated expression of immune cells, and ii) interferon response microglia (IRM) characterised by up-regulated expression of innate immune response and interferon response pathway. • ARM was enriched with GWAS AD risk genes: *Trem2* upregulation, *Bin1, Cd33, Picalm* down-regulation. • ARM promoted *Apoe* expression in microglia, whereby deletion of *Apoe* ablated ARM expression and density of microglia around amyloid deposits.
Mathys et al. (2019)[171]	Human prefrontal cortex (24 AD, 24 Controls)	• Identified cell-specific response with upregulation of microglial-expressed genes (i.e. *TREM2,PICALM*). • 95% of DEGs were observed in one cell type, indicating perturbations are strongly cell specific. However, top DEGs enriched were in myelination across multiple cell types (i.e. *ERBIN, CNTNAP2, NEGR1, BEX1, NTNG1*).
Grubman et al. (2019)[193]	Human entorhinal cortex (6 AD, 6 Controls)	• *APOE* is specifically repressed in oligodendrocyte progenitor cells and astrocyte subpopulations, while up-regulated in microglial subpopulation.
Leng et al. (2021)[175]	AD entorhinal cortex (6 Control, 6 AD)	• Subset of AD-associated astrocytes, likely to represent reactive astrocytes, characterised with up-regulated expression of *GFAP* and *CD44*, and down-regulation of genes associated with homeostasis.

1.4 Aims and objectives

An increasing number of studies implicate a role of transcriptional dysregulation and aberrant splicing in AD disease development and pathogenesis (reviewed in **Table 1.2**). However, investigation of splicing and transcript expression variation is typically performed by profiling the transcriptome using standard short-read RNA sequencing approaches, which are inherently-constrained at transcript assembly and subsequent isoform characterisation essential for splicing analyses (described in **Section 1.3.1**). My PhD thus aims to overcome these challenges by leveraging the use of long-read sequencing to accurately characterise isoform diversity and splicing patterns associated with Alzheimer's disease (**Figure 1.15**).

Hypothesis: Transcriptomic dysregulation plays a fundamental role in development of AD pathology. This includes alterations in gene splicing, which results in differential and novel expression of transcripts that are translated to generate isoforms with functional biological implications.

Main Objectives:

1. Optimise long-read sequencing approaches, PacBio Iso-Seq and ONT complementary DNA (cDNA) nanopore sequencing, for profiling of full-length transcripts (**Chapter 3**).
2. Characterise global isoform diversity and splicing events in the mouse cortex using optimised long-read sequencing approaches (**Chapter 4**).
3. Identify global transcriptional and splicing variations associated with progression of tau pathology using a well-characterised AD mouse model, rTg4510 (**Chapter 5**).
4. Comprehensively characterise isoform diversity and splicing events of 20 AD-associated genes, and differences in transcript expression associated with tau pathology in rTg4510 mice (**Chapter 6**).

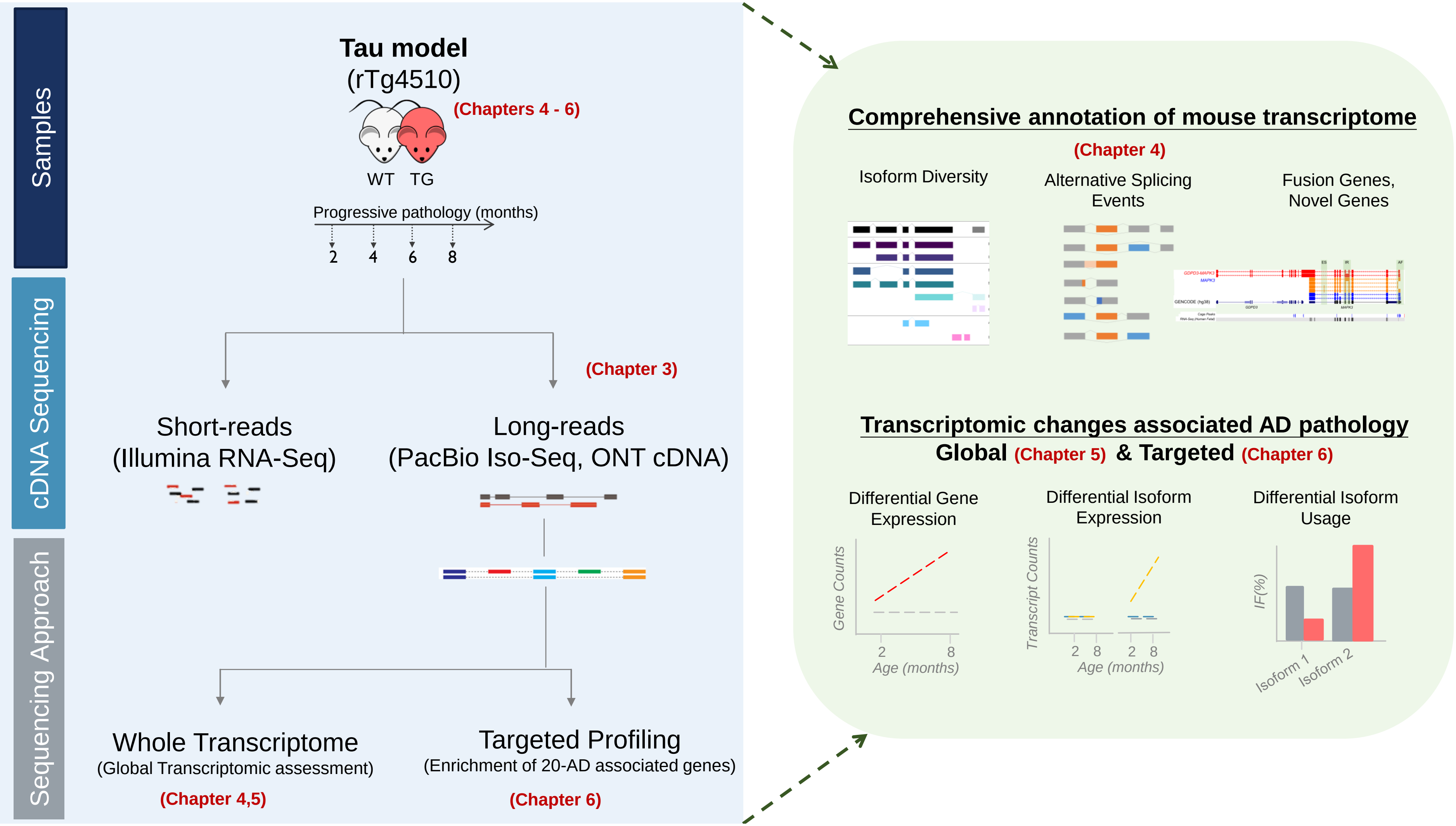

Samples
cDNA Sequencing
Sequencing Approach
Tau model
(rTg4510)
(Chapters 4 - 6)
WT
TG
Progressive pathology (months)
2
4
6
8
(Chapter 3)
Short-reads
(Illumina RNA-Seq)
Long-reads
(PacBio Iso-Seq, ONT cDNA)
Whole Transcriptome
(Global Transcriptomic assessment)
(Chapter 4,5)
Targeted Profiling
(Enrichment of 20-AD associated genes)
(Chapter 6)
Comprehensive annotation of mouse transcriptome
(Chapter 4)
Isoform Diversity
Alternative Splicing Events
Fusion Genes, Novel Genes
Transcriptomic changes associated AD pathology
Global (Chapter 5) & Targeted (Chapter 6)
Differential Gene Expression
Differential Isoform Expression
Differential Isoform Usage
Gene Counts
Transcript Counts
IF(%)
2
8
Age (months)
Isoform 1
Isoform 2

Figure 1.15: Study designs and analyses overview. Shown is an overview of the research presented in this thesis. To identify transcriptomic and splicing variations associated with AD pathology, this thesis aims to characterise isoform diversity and splicing events at a global and targeted level from the rTg4510 mouse model. ONT - Oxford Nanopore Technologies, PacBio Iso-Seq - Pacific Bioscience Isoform Sequencing

Chapter 2

General Methodology

This chapter describes the general methods that were applied in the long-read sequencing experiments in **Chapters 4 and 6**. Experimental methods specific to individual results chapters can be found in the Methods section of the relevant chapter. Methods pertaining to the library preparation for PacBio Isoform Sequencing (Iso-Seq) and ONT nanopore cDNA sequencing can be found in **Section 3.1.2** and **Section 3.2.2**, respectively. Standard manufacturer's protocols used in this thesis can be found in **Appendix A and B**. All reagents mentioned in this Chapter were provided with the respective kits unless otherwise stated.

2.1 Mouse tissue samples and RNA isolation

2.1.1 Mouse model of AD tauopathy: rTg4510

rTg4510 mice recapitulate AD tauopathy through the overexpression of the human tau transgene, MAPT^{P301L}, which harbours the FTD-associated P301L mutation. It contains four microtubule-binding domains while lacking the N-terminal segment (0N4R), and exons 2-3 of the mouse prion protein gene *Prnp*. The transgene expression is controlled under the CaMK2a promoter and is largely restricted to the forebrain, with rapid age-dependent spread of neuropathology starting from as early as 2 months in the neocortex to the hippocampus by 5 months (**Figure 1.8**). Neuronal and synaptic loss are also observed from 9 months, with these mice exhibiting cognitive and behavioural impairments. Sex differences in pathology have been reported with female mice exhibiting earlier and more severe cognitive and behavioural impairments than transgenic male mice.[194]

The rTg4510 mouse model is particularly informative as tau expression can be induced through the tetracycline operon-responsive element and suppressed upon doxycycline treatment.[195] However, a recent study reported disruption of several endogenous mouse genes due to the random integration of MAPT^{P301L} (as previously described in **Section 1.1.4**), which has additional off-target effects that may potentially contribute to the neurodegenerative phenotype associated with rTg4510 transgenic mice.[85]

2.1.2 Animal breeding and sample preparation

All animal procedures were carried out at Eli Lilly and Company (Windlesham, UK), in accordance with the UK Animals (Scientific Procedures) Act 1986 and with approval of the local Animal Welfare and Ethical Review Board. rTg4510 mice, licensed from the Mayo Clinic (Jacksonville, US), were bred on a mixed 129S6/SvEvTac + FVB/NCrl background (heterozygous tau responder x heterozygous tTA effector).[90] Six breeding colonies were required to produce the rTg4510 transgenic mice (**Table 2.1**). The mice were housed under standard conditions (constant temperature and humidity with a 12-hour light/dark cycle in individually ventilated cages) before terminal anaesthesia with pentobarbital and transcardial perfusion with phosphate-buffered saline (PBS).[90]

The entorhinal cortex was dissected from the left-brain hemisphere on wet ice using common neuroanatomical landmarks, as described by Heffner et al.(1980),[196] to ensure reproducibility. Total RNA was then extracted[90] using the AllPrep DNA/RNA Mini Kit (Qiagen) according to the manufacturer's protocol, and converted to cDNA for library preparation (described later in **Section 2.2.1**). Of note, more than 80% of total RNA is comprised of ribosomal RNA (rRNA), with the remaining 15% representing transfer RNA (tRNA) and 5% representing mRNA.

2.1.3 Assessment of nucleic acid quality and quantity

Acquiring high-quality RNA, generating full-length cDNA and successfully performing the multi-step library preparation are all crucial for optimal sequencing experiments, particularly long-read sequencing. The assessment of the purity and integrity of extracted RNA, followed by cDNA quality and quantity, was therefore required throughout library preparation and quality control (QC) stages of my sequencing experiments. This was undertaken using the RNA/DNA ScreenTape and Bioanalyzer assays for qualitative assessment, and the Qubit for

Table 2.1: rTg4510 breeding procedures. Tabulated are the strain backgrounds and breeding schemes required to generate the rTg4510 transgenic mice. Table is taken from Castanho (2019).[197]

	Line	Strain background	Breeding scheme	Offspring
1	Background Line 1	129S6/SvEvTac (Taconic)	129S6/SvEvTac x 129S6/SvEvTac	129S6/SvEvTac
2	Background Line 2	FVB/NCrl (Charles River)	FVB/NCrl x FVB/NCrl	FVB/NCrl
3	Tau responder Parental Line 1	FVB/NCr	Tau P301L x FVB/NCrl	Heterozygous Tau Responder
4	tTA Effector Parental Line 2	29S6/SvEvTac	Camk2a-tTa x 129S6/SvEvTac	Heterozygous tTA Effector
5	rTg4510	Mixed	HET Tau Responder x HET tTA Effector	rTg4510 wildtype (tTA WT) rTg4510 transgenic (tTA with Tau transgene)

DNA quantification.

2.1.4 ScreenTape and Bioanalyzer assays

ScreenTape and Bioanalyzer assays are commonly used to provide an accurate and automated assessment of nucleic acid quality and size by electrophoresis. It works on the principle that upon applying an electric field, negatively-charged DNA migrates through a gel matrix towards the positive anode at a rate that is dependent on DNA size; smaller DNA fragments migrate faster, and thus move further through the gel within a specific time frame. The separated DNA can be then visualised using a fluorescent dye that intercalates into the double-stranded DNA (dsDNA) structure and fluoresces under ultraviolet light.

Both RNA ScreenTape and Bioanalyzer assays further provide a numeric evaluation of the quality of an RNA sample using a score between 1 and 10, known as a RNA integrity number (RIN); a RIN score of 1 is indicative of high degradation and poor quality RNA, whereas a RIN score of 10 indicates minimal degradation (**Figure 2.1**). The purity and quantity of extracted RNA was assessed using the RNA Bioanalyzer assay with Agilent RNA 6000 Nano Kit (Agilent Technologies) and Agilent 2100 Bioanalyzer instrument (Agilent Technologies).

Assessment of cDNA quality during various QC stages of long-read library preparation was mostly performed using the DNA Bioanalyzer assay with Agilent D1200 Kit (Agilent Technologies), particularly where accurate determination of library molarity was critical (given that the Bioanalyzer assay is more sensitive than the ScreenTape assay). However in QC stages where assessment is optional, the D5000 ScreenTape (Agilent Technologies) and 4200 TapeStation (Agilent Technologies) were used instead as the ScreenTape assay was more cost-effective and less time-consuming to run. Both assays were performed following the standard manufacturer's protocol. Briefly, this involved mixing the sample with the ladder and buffer (if using the ScreenTape), or with the marker and gel-dye mix (if using the Bioanalyzer assay), before loading the sample into the machine to be assayed. Detailed lab instructions for Bioanalyzer and ScreenTape assays are detailed in **Appendix A.2.0.3**.

2.1.5 Qubit

Qubit assays (Invitrogen) allow accurate nucleic acid quantification by the selective binding of fluorescent Qubit dyes to dsDNA or RNA, rendering it more sensitive and specific than the NanoDrop 8000 spectrophotometer (Thermo Fisher Scientific) which uses UV absorbance. Following RNA extraction, the RNA concentration was determined using Qubit assays to ensure that the same amount of total RNA for each sample was used for library preparation. Many of the QC steps post DNA purification (later discussed in **Section 2.2.4**) throughout library preparation also required Qubit assays to determine the cDNA concentration prior to proceeding with downstream experiments. Briefly, this involved running two "standard samples" (prepared with Qubit reagent in a 10:200 ratio) and the "test samples" (prepared with the same Qubit reagent in a 1:200 ratio) on the Qubit 3.0 Fluorometer (Thermo Fisher Scientific) following the standard manufacturer's protocol (detailed in **Appendix A.2.0.2**). Of note, all quantity assessments in this thesis were performed using the Qubit dsDNA High Sensitivity assay.

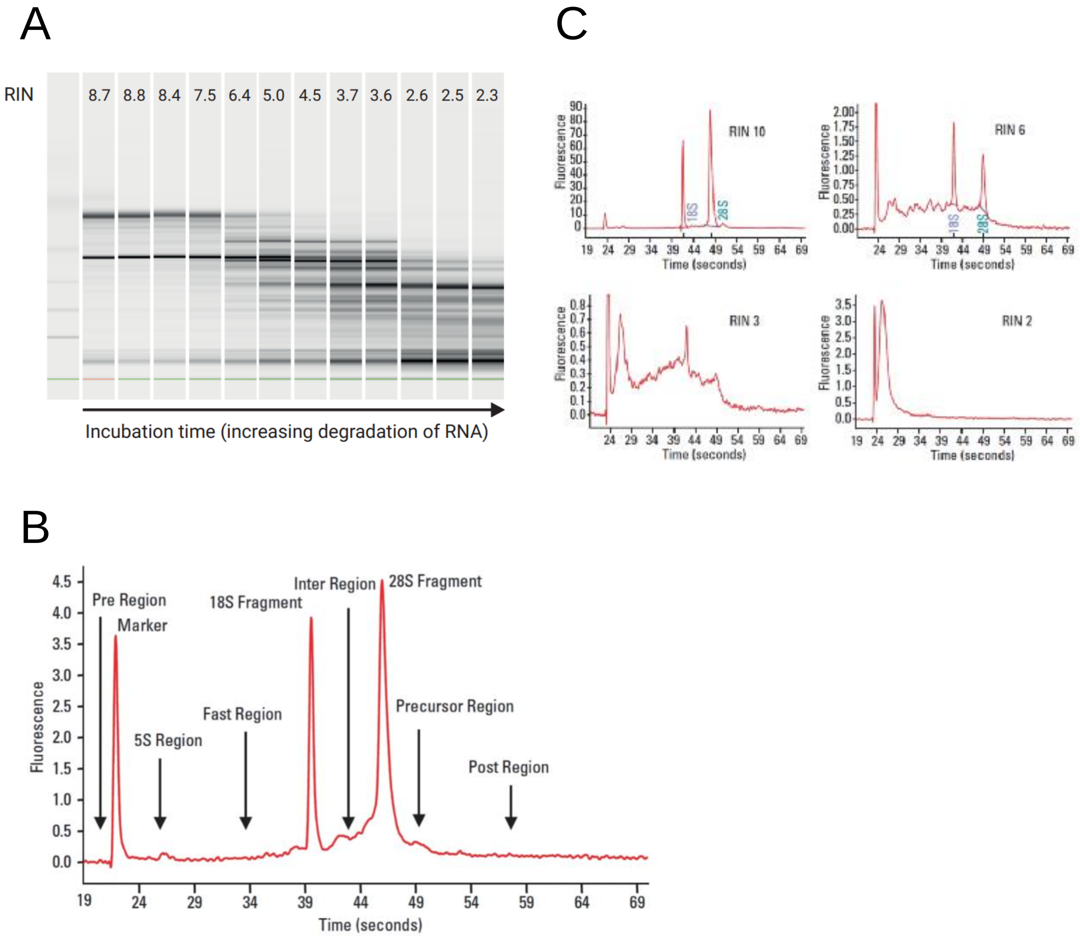


Figure 2.1: Evaluation of RNA integrity using ScreenTape and Bioanalyzer assays. Shown is a **(A)** gel image from a Bioanalyzer assay demonstrating progressive total RNA degradation over a prolonged period of incubation. Degradation is indicated by a general shift to the right with more bands representing shorter fragments and a decrease in RIN. **(B)** An alternative assessment of total RNA quality and integrity is represented with the Bioanalyzer electropherogram with two distinctive peaks corresponding to 18S and 28S fragment of rRNA, and a marker peak. The RIN is calculated using the relative ratio of the Fast Region and 18S, 28S fragment. **(C)** Bioanalyzer electropherograms depicting total RNA with varying degrees of degradation: i) minimal (RIN = 10) as indicated by the two distinct peaks corresponding to 18S and 28S fragment, ii) small degree (RIN = 6) with the two peaks still visible but also detection of other smaller peaks in the Fast Region, corresponding to fragmented RNA, iii) large degree of degradation (RIN = 3) with inability to detect the two peaks, and iv) significant degradation (RIN = 2), indicated by the absence of the 18S and 28S fragment. Figures and legends are adapted from Mueller et al. (2004).[198]

2.2 cDNA synthesis, amplification and purification

After RNA isolation, integrity assessment and quantification, total RNA was converted to cDNA. Given the low frequency of mRNA (< 5% of total RNA) and low sensitivity of current sequencing platforms, the converted cDNA was subsequently amplified using polymerase chain reaction (PCR) and assessed by using agarose gel electrophoresis.

2.2.1 Complementary DNA synthesis

Recommended as part of the Iso-Seq protocol, the SMARTer PCR cDNA Synthesis Kit (Clontech) was used to convert 200ng total RNA to cDNA. Unlike other cDNA synthesis methods, the SMARTer PCR cDNA synthesis relies on a modified oligo(dT) primer and a reverse transcriptase (RT) that has an inherent terminal transferase activity (outlined in **Figure 2.2**). First-strand synthesis therefore occurs in a SMART (Switching Mechanism At 5' End of RNA Transcript) fashion, whereby a few additional nucleotides ("overhang") are added to the 3' end of the cDNA as the RT approaches the 5' end of the mRNA. The RT then switches templates and continues replicating to the end, generating a full-length single-stranded cDNA which is then amplified. The usage of the "overhang" sequence ensures enrichment and synthesis of full-length cDNA, as cDNA without this sequence (i.e. prematurely terminated cDNA, cDNA from non-poly(A) RNA, contaminating genomic DNA) will not be exponentially amplified.[199] Detailed manufacturer's instructions of this kit can be found in **Appendix A.3**.

While this kit is advantageous in preferentially enriching for full-length cDNA sequences, it cannot differentiate between intact and truncated RNA, which is present in poor-quality samples and could be amplified as technical artefacts in the final cDNA library. A solution to circumvent this issue is to exploit the presence of the 5'-cap which is only present in intact RNA (5'-cap refers to 7-methylguanosine at the 5' end of mRNA which is added during transcription to protect nascent mRNA from degradation and assist in protein translation), using the Full-Length cDNA Amplification kit (Teloprime).[200] This kit relies on a double-stranded adapter that recognises and ligates to the 5' cap at the end of the first-strand synthesis step. I trialled this kit as part of my experiments, but was unable to generate sufficient cDNA.

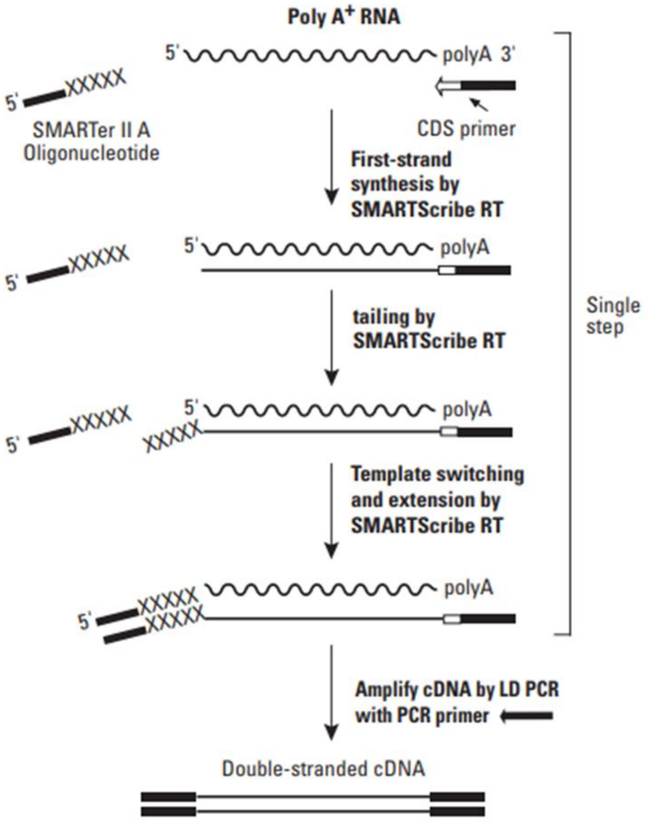

Poly A+ RNA
5'
polyA 3'
5'
XXXXX
SMARTer II A Oligonucleotide
CDS primer
First-strand synthesis by SMARTScribe RT
tailing by SMARTScribe RT
Template switching and extension by SMARTScribe RT
Single step
Amplify cDNA by LD PCR with PCR primer
Double-stranded cDNA

Figure 2.2: SMARTer cDNA synthesis. Shown is a flowchart of the SMARTer cDNA synthesis protocol to ensure the generation of full-length cDNA by leveraging the power of the enzyme's terminal transferase activity. Premature RT termination reduces the efficiency of the transferase activity, resulting in the absence of the overhang at the 3' end of the template for downstream amplification. cDNA synthesis is achieved in the following manner:

1. Oligo(dT) primer (3' SMART CDS Primer II A) primes the first-strand synthesis reaction by binding to the poly(A) tail and transcribes the RNA into single-stranded DNA.
2. As RT reaches the 5' end of the mRNA, the enzyme's terminal transferase activity adds a few additional nucleotides to the 3' end of the cDNA.
3. With a 3' end that is complementary to the added nucleotides, the SMARTer II A oligonucleotide (or the template switching oligonucleotide) base-pairs with it and creates an extended template.
4. RT then switches templates and continues transcribing to the end of the SMARTer oligonucleotide.
5. The resulting full-length, single-stranded cDNA contains the complete 5' end of the mRNA, as well as a 3' end that is complementary to the SMARTer oligonucleotide.
6. The SMARTer oligonucleotide and the poly(A) sequence then serve as universal priming sites for end-to-end cDNA amplification.

Of note, the SMARTer II A Oligonucleotide, 3' SMART CDS Primer II A, and 5' PCR Primer II A all contain a stretch of identical sequence.
Figure and legend are taken from the SMARTer PCR cDNA Synthesis Kit User Manual.[201]
RT - Reverse transcriptase.

To maximise throughput and minimise cost, my targeted experiments (**Chapters 6**) involved sequencing of multiple samples simultaneously in one sequencing run (i.e. in “multiplex”). To differentiate the samples, I used a unique barcoded oligo(dT) primer (**Table 2.2**) for each sample for cDNA synthesis instead of the standard oligo(dT) primer from the SMARTer cDNA synthesis kit (Clontech). The only difference between the barcoded oligo(dT) and the standard oligo(dT) primer is the addition of a unique 16-bp internal barcode, which does not interfere with priming and end-to-end cDNA amplification. The general structure of the barcoded oligo(dT) primer were as follows:

Primer sequence 16bp barcode oligo(dT)

5’AAGCAGTGGTATCAACGCAGAGTACtcagacgatgcgtcatTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN3’

Table 2.2: Barcoded oligo(dT) primers were used for multiplexing samples in targeted profiling. Tabulated is a list of barcoded primers that were used for targeted profiling of AD-risk genes in the rTg4510 cortex. Each of the barcoded primers contain the same 5' primer sequence and oligo(dT) for reverse transcription of first-strand cDNA synthesis using the SMARTer PCR cDNA Synthesis Kit (Clontech). The barcodes were provided from the official PacBio multiplex protocol.

Barcode	Sequence
Barcode 1	AAGCAGTGGTATCAACGCAGAGTACCACATATCAGAGTGCGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 2	AAGCAGTGGTATCAACGCAGAGTACACACACAGACTGTGAGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 3	AAGCAGTGGTATCAACGCAGAGTACACACATCTCGTGAGAGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 4	AAGCAGTGGTATCAACGCAGAGTACCACGCACACACGCGCGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 5	AAGCAGTGGTATCAACGCAGAGTACCACTCGACTCTCGCGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 6	AAGCAGTGGTATCAACGCAGAGTACCATATATATCAGCTGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 7	AAGCAGTGGTATCAACGCAGAGTACTCTGTATCTCTATGTGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 8	AAGCAGTGGTATCAACGCAGAGTACACAGTCGAGCGCTGCGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 9	AAGCAGTGGTATCAACGCAGAGTACACACACGCGAGACAGATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 10	AAGCAGTGGTATCAACGCAGAGTACACGCGCTATCTCAGAGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN

2.2.2 Polymerase chain reaction (PCR)

PCR is a well-established method for generating multiple copies of the same DNA sequence. Mimicking natural DNA replication, this relies on a thermostable DNA polymerase, a set of primers specific to the region of interest, and a cocktail of reagents required for polymerisation (i.e. deoxynucleotides (dNTPs) and buffers). This reaction is subjected to a series of heating and cooling steps:

1. Denaturation at 96°C to separate dsDNA.
2. Annealing, typically between 55°C and 65°C, for the binding of primers to the complementary sequences on the single-stranded DNA; the specific annealing temperature is dependent on the primer sequence.
3. Extension at 72°C to allow the polymerase to extend the primers, synthesising a new cDNA strand using dNTPs.

These three steps are then repeated multiple times, or "cycles", resulting in an exponential generation of the DNA template of interest.

2.2.3 Agarose gel electrophoresis

Agarose gel electrophoresis allows the separation of dsDNA molecules based on length, and works on the same principle as the Bioanalyzer and ScreenTape assays (described previously in **Section 2.1.4**). It is most commonly used to determine DNA quality and quantity, and assess the efficiency of molecular biology techniques such as PCR amplification by determining the number of optimal cycles. It is a well known phenomenon that an increased number of unnecessary PCR cycles can generate artefacts (strand invasion) and preferentially amplify shorter transcripts.[170,202] Instructions to prepare and run an agarose gel electrophoresis are detailed in **Appendix A.5**.

2.2.4 AMPure bead purification

At various stages of long-read library preparation, cDNA was purified using AMPure beads (**Figure 2.3A**). These are paramagnetic beads that reversibly bind to DNA in the presence of polyethylene glycol (PEG) and salt. The concentration of PEG, and consequent ratio of beads to DNA, determines the size of fragments that are bound and subsequently eluted (**Figure 2.3B**); the lower the concentration of beads to DNA, the greater the proportion of longer DNA fragments bound, due to the preferential binding of beads to larger molecular weight DNA with a higher negative charge - a 0.4X ratio would therefore preferentially retain the larger DNA fragments, whereas a 1X ratio would retain both long and short DNA fragments (**Figure 2.3B**). Briefly, AMPure bead purification was performed by thoroughly mixing and vortexing each sample with a pre-specified ratio of AMPure beads. The samples were then placed onto a magnet for clear separation of DNA-bound beads and solution, followed by two washes of 70% ethanol and DNA elution. Detailed instructions can be found in **Section A.2.0.1**.

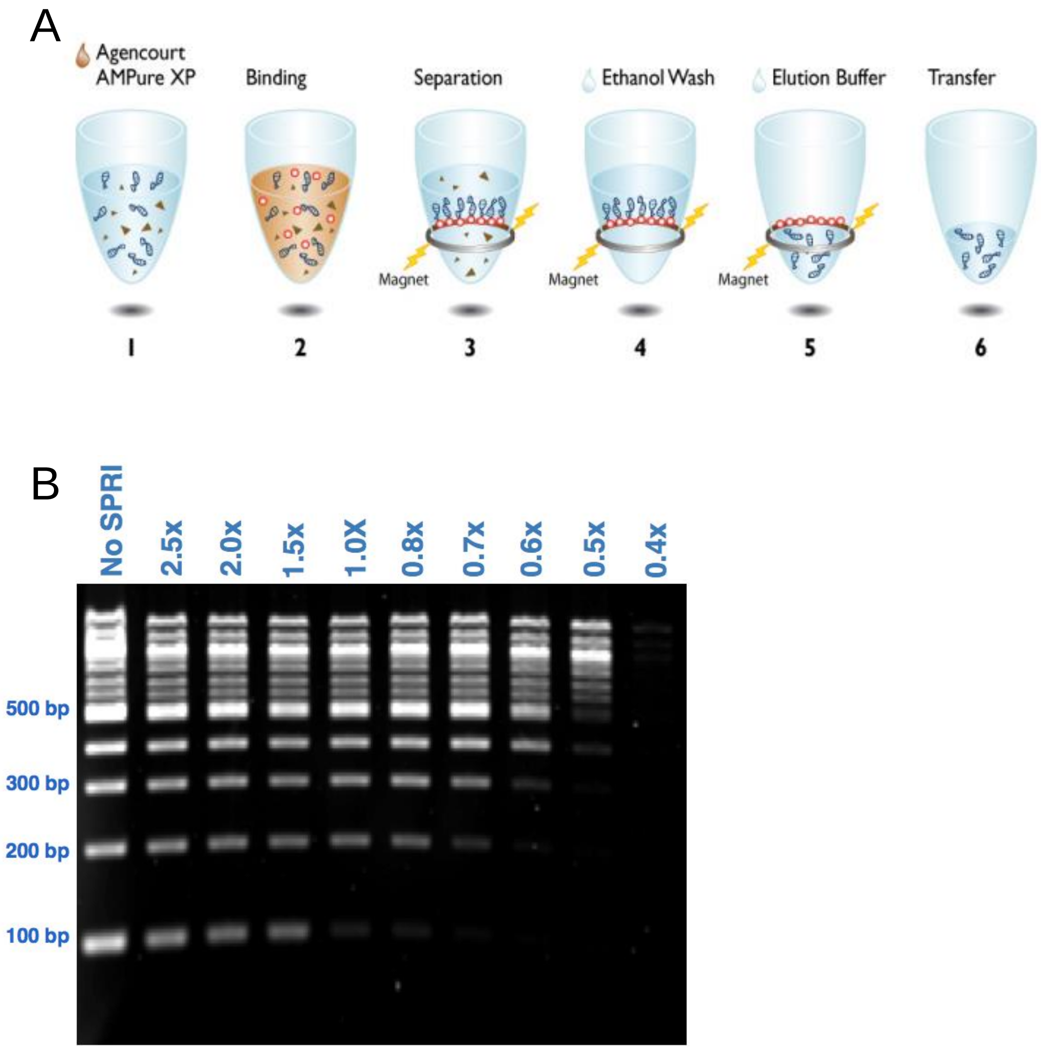


Figure 2.3: cDNA purification with AMPure beads. Shown is a schematic figure depicting the **(A)** steps of purifying DNA with AMPure beads, with initial binding of magnetic beads to negatively-charged DNA, followed by ethanol wash and elution. **(B)** An agarose gel image of DNA purified using a range of bead to DNA ratio for size selection; the lower the ratio, the greater the enrichment for longer fragments with displacement of the shorter fragments. Figures are taken from the Beckman Coulter website.

2.3 ERCC-RNA spike-in controls

A set of external RNA spike-in controls, generated by the External RNA Controls Consortium (hereby referred to as "ERCC"), was used to i) evaluate the performance of library preparation and the sequencing experiments, and to ii) validate the Iso-Seq bioinformatics pipeline in accurately characterising the transcriptome using long reads. ERCC consists of 92 polyadenylated synthetic transcripts (250 to 2000 nucleotides) of known sequences from the ERCC plasmid library, which were added in pre-determined amounts to the sample before first-strand cDNA synthesis.

The amount of ERCC added was determined using the below equation:[203]

$$mass_{RNAspike} = fraction_{spikedreads} * fraction_{targetRNA} * mass_{RNAinput}$$
$$concentration_{RNAspike} = mass_{RNAspike} * volume_{RNAspike}$$

where:

$mass_{RNAspike}$ = mass of ERCC to be added to sample
$concentration_{RNAspike}$ = final diluted concentration (ng/μL) of ERCC
$fraction_{spikedreads}$ = desired proportion of sequenced ERCC reads relative to total amount of sequenced reads *(3%)*
$fraction_{targetRNA}$ = expected proportion of target RNA, in this case mRNA relative to total RNA *(3%)*
$mass_{RNAinput}$ = input of total RNA *(200ng)*
$volume_{RNAspike}$ = volume of RNA spike-in *(0.1μL)*

Equation 2.1: Determining the amount of ERCC controls for sequencing runs. In determining the mass and final concentration of RNA-spike-in mix based on the above conditions, the stock ERCC RNA spike-in was diluted from the original concentration of 30ng/μL to 1.8ng/μL with a dilution factor of 1:16.8. The italicised parameters were taken from the "Wellcome Trust Advanced Course: RNA Transcriptomics (2018)"[203] (that I attended during my PhD) with the exception of total RNA input.

A separate pilot experiment (**Appendix C**) showed successful addition of ERCC with two main bands at ~600bp and ~1000bp (**Figure 2.4A**), reflecting significant enrichment of ERCC at these two respective lengths as expected (**Figure 2.4B**). However, the stark contrast of these two bands against the smear of cDNA suggests over-usage of ERCC - possibly due to the overestimation of assumed proportion of mRNA. A lower ERCC amount was therefore used across all the experiments to reduce unnecessary sequencing and coverage of ERCC (final concentration of 0.6ng/μL with a dilution factor of 1:50.5, **Figure 2.4C**).

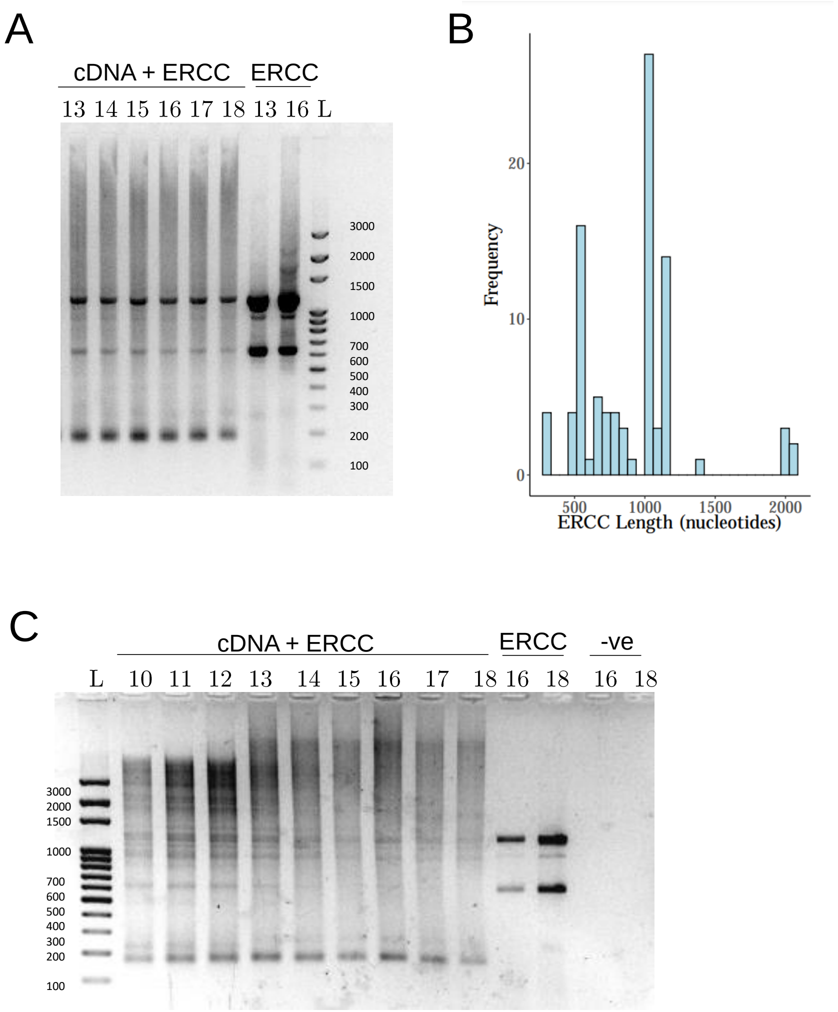


Figure 2.4: Usage of ERCC controls to evaluate performance of long-read sequencing runs. Shown is **(A)** an agarose gel image taken from PCR amplification of cDNA and ERCC (1.8ng/μL determined from **Equation 2.1**), and ERCC alone as a positive control. 5μL of PCR aliquots were taken every cycle (cycles 13 to 18) and then assessed on gel electrophoresis. The two bands at 600bp and 1000bp correspond to the enrichment of ERCC at these two lengths as expected. **(B)** Distribution of known ERCC length, with a significant proportion of transcripts sized at 500-600bp and 1000-1200bp. **(C)** An agarose gel image after a repeat of PCR amplification of cDNA and ERCC at a lower concentration (0.6ng/μL) with ERCC as positive and water as negative control, respectively. The numbers above the lanes refer to the number of PCR cycles. L denotes to 100bp Ladder.

The usage of ERCC guided me in the development of our bioinformatics pipeline by i) estimating and reducing the number of false negatives, and ii) ensuring that only one unique molecule per ERCC is detected. This resulted in optimisation of various parameters and addition of further filtering steps downstream (described later in **Section 3.1.4.5**).

Chapter 3

Long-read Sequencing

This chapter provides a detailed background into the lab workflow and bioinformatics pipelines developed during this PhD that were subsequently used to generate and analyse data from Pacific Bioscience (PacBio) SMRT sequencing (henceforth referred to as Iso-Seq) and Oxford Nanopore Technologies (ONT) nanopore cDNA sequencing in **Chapters 4-6**.

3.1 Pacific Biosciences: Isoform Sequencing

3.1.1 Introduction

Successful DNA polymerisation requires a high concentration of nucleotides for DNA polymerase processivity and accuracy. However, mimicking this in DNA sequencing results in a high background noise level, thereby reducing the sensitivity to detect base incorporation and fluorophore emission. Historically, “second-generation” sequencing technologies - such as that used in Illumina short-read RNA-Seq - have circumvented this issue by the step-wise addition, scan and wash of each set of labelled nucleotides, although at the expense of read lengths (as discussed in **Section 1.3.1**).

In 2013, PacBio pioneered the development of “third generation” long-read sequencing with the capability to generate substantially longer reads (reviewed in **Table 1.4**), due to its ability to mimic the natural, uninterrupted, processive DNA synthesis through three important innovations:[204]

1. The creation of a circular template, SMRTbell (**Figure 3.1A**), which is enclosed with hairpin adapters at the ends of the insert, allowing uninterrupted DNA polymerisation.[205]
2. The sequencing of each polymerase-bound SMRTbell at the bottom of a nanometre-wide well (zero-mode waveguide - ZMW)[206] (**Figure 3.1B**). The nanoscale size of the ZMWs and reduced volume allow sensitive detection of a single nucleotide incorporation event against the high background of labelled nucleotides, achieving a high-signal-to-noise ratio. PacBio currently offers two sequencers, which primarily differ in the number of ZMWs that can be sequenced: Sequel I and Sequel II with 1 Million and 8 Million ZMWs, respectively.
3. The addition of phospholinked nucleotides, each labelled with a different colour fluorophore that corresponds to the four different bases (A, C, G and T), allows for natural and processive DNA synthesis[207] (**Figure 3.1C**).

3.1.1.1 Mechanism

Due to the circular nature of the SMRTbell template, the polymerase is able to continually read through the insert in an uninterrupted fashion multiple times (or "passes"), resulting in the generation of a continuous read (known as a "polymerase read") (**Figure 3.1A**). By removing the hairpin-adapters that delineate the repeated insert sequence, this polymerase read is resolved to multiple reads (known as "subreads"), which are then merged to yield one high-quality and highly-accurate consensus read (known as a "circular consensus sequence" - CCS) (**Figure 3.1A**). The generation of these CCS reads can drastically improve a single-pass from 85%, due to random sequencing error, to 99% from alignment and correction of multiple subreads. Notably, the error rate is proportional to the number of "passes", and is dependent on the polymerase lifetime and insert length.[205]

Unlike the short reads produced by standard Illumina-based RNA-Seq, PacBio Iso-Seq reads are not of a set length, but a range of lengths that is reflective of the library size and polymerase activity.[208,209] Previous chemistries preferentially sequenced molecules of a certain length, due to biased loading of SMRTbell templates; "Diffusion Loading" favoured shorter molecules,[210] whereas "Magbead Loading" allowed proportional loading of DNA templates to the concentration rather than length, but prevented sequencing of templates < 1kb. However, recent improvements in both technology and chemistry have alleviated this sequencing read-length bias.[211]

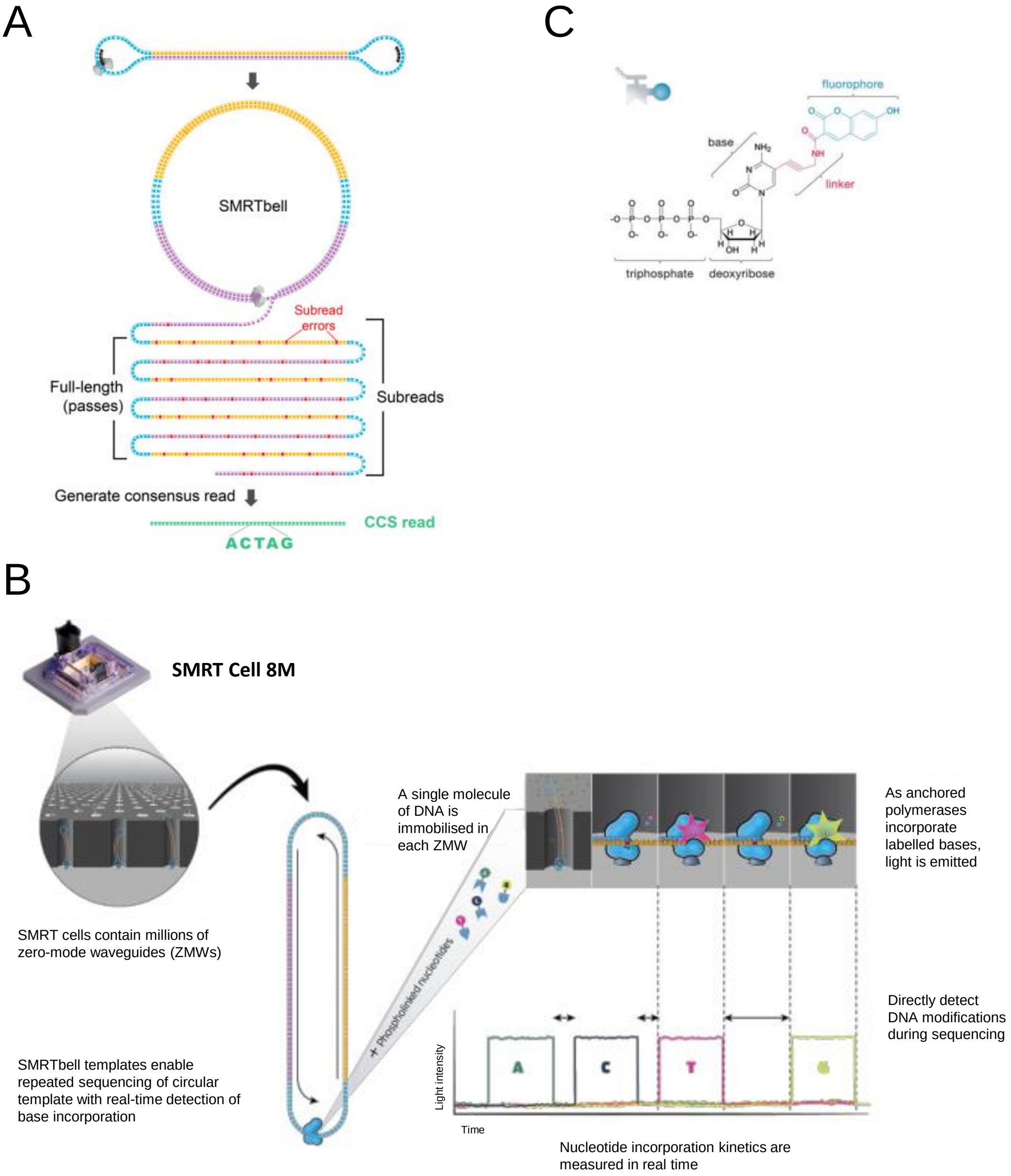

A
SMRTbell
Subread errors
Full-length (passes)
Subreads
Generate consensus read
CCS read
ACTAG
B
SMRT Cell 8M
SMRT cells contain millions of zero-mode waveguides (ZMWs)
SMRTbell templates enable repeated sequencing of circular template with real-time detection of base incorporation
A single molecule of DNA is immobilised in each ZMW
+ Phospholinked nucleotides
As anchored polymerases incorporate labelled bases, light is emitted
Directly detect DNA modifications during sequencing
Light intensity
Time
A
C
T
G
Nucleotide incorporation kinetics are measured in real time
C
fluorophore
base
NH2
NH
OH
linker
triphosphate
deoxyribose

Figure 3.1: Pacific Biosciences single-molecule real-time sequencing technology. Shown is an overview of Pacific Biosciences single-molecule real-time sequencing technology (SMRT), which is able to generate long reads > 10kb by **(A)** enclosing the cDNA fragment of interest within a circular template (SMRTbell) to allow uninterrupted DNA polymerisation, followed by the **(B)** sequencing of each SMRTbell with a bound polymerase at the bottom of ZMWs, enabling sensitive detection of polymerisation at the nucleotide level from **(C)** addition of phospholinked nucleotides with a differently-labelled fluorophore. ZMWs - Zero-mode waveguides.

3.1.2 Lab workflow

This section describes the lab workflow for PacBio Iso-Seq library preparation, which was subsequently implemented in the long-read sequencing experiments for global transcriptome and targeted profiling in **Chapters 4 and 6**, respectively. General methods pertaining to sample preparation, cDNA synthesis and amplification can be found in **Chapter 2**.

The Iso-Seq lab workflow for the global profiling of the transcriptome (**Chapter 4**), as outlined in **Figure 3.2**, involved three main steps: i) converting RNA to full-length cDNA using the Clontech SMARTer PCR cDNA synthesis kit (**Section 2.2.1**), ii) amplification (**Section 2.2.2**) and purification (**Section 2.2.4**) of double-stranded cDNA, and iii) the preparation of SMRTbell libraries.

The Iso-Seq lab workflow for the targeted profiling of the transcriptome (**Chapter 6**), outlined in **Figure 3.3**, involved the incorporation of barcode sequences during cDNA synthesis to allow sample multiplexing and an additional step of target enrichment for the genes of interest.

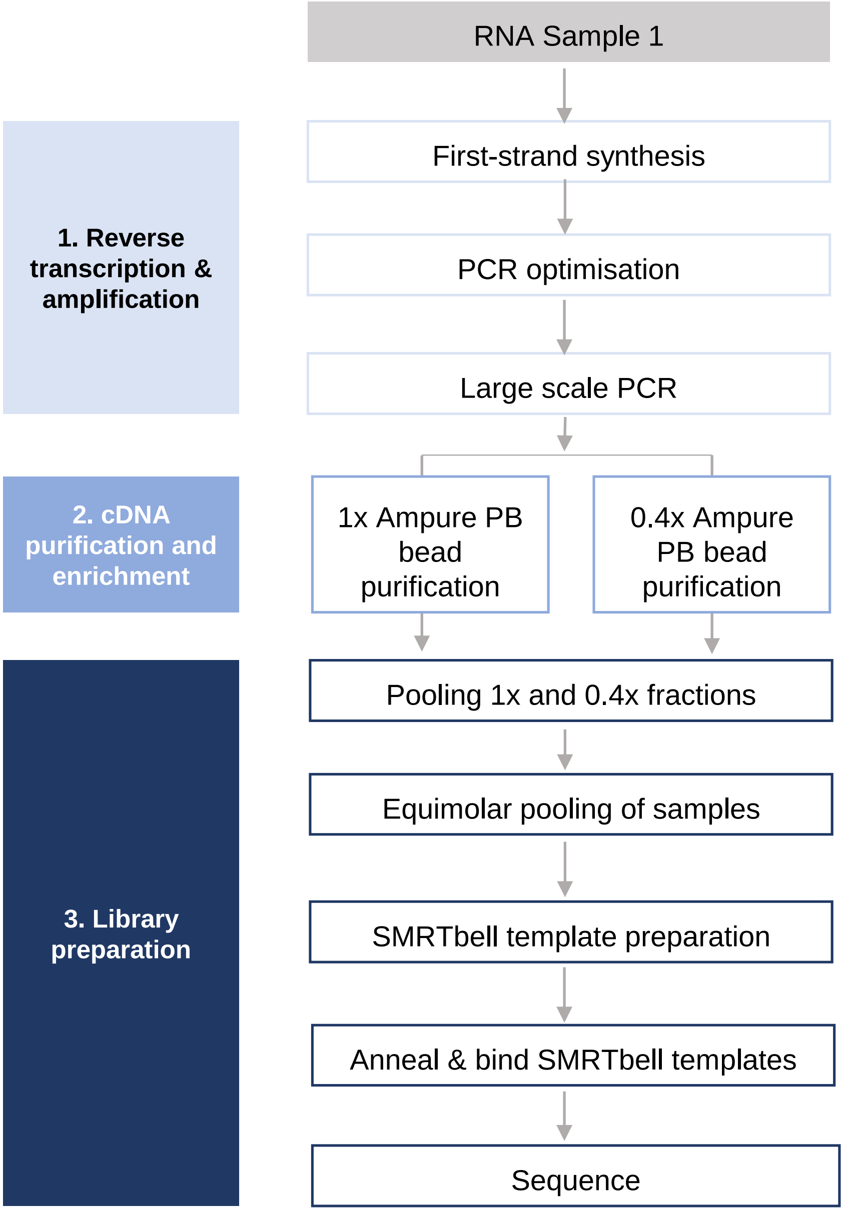

RNA Sample 1
1. Reverse transcription & amplification
First-strand synthesis
PCR optimisation
Large scale PCR
2. cDNA purification and enrichment
1x Ampure PB bead purification
0.4x Ampure PB bead purification
3. Library preparation
Pooling 1x and 0.4x fractions
Equimolar pooling of samples
SMRTbell template preparation
Anneal & bind SMRTbell templates
Sequence

Figure 3.2: Iso-Seq lab workflow for global transcriptome profiling. Shown is a flow diagram of the Iso-Seq lab workflow. Adapted from the official Iso-Seq protocol, it involves three main steps: 1) reverse transcription and amplification of cDNA (**Section 2.2.1**), 2) cDNA purification with AMPure beads (**Section 2.2.4**), and 3) library preparation involving the ligation of SMRTbell templates, and binding of the primer and polymerase (**Section 3.1.2.4**). Due to the usage of newer chemistries, size selection was not performed.

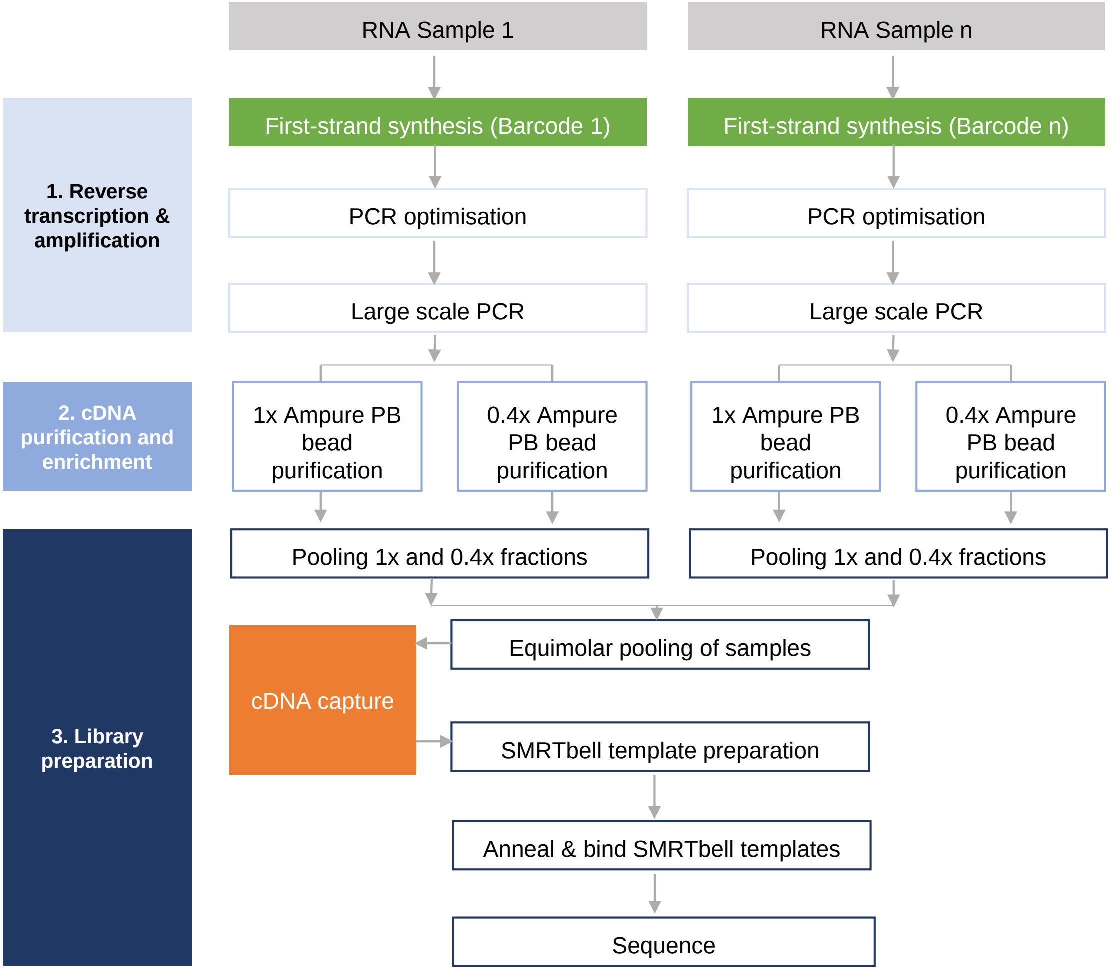

RNA Sample 1
RNA Sample n
1. Reverse transcription & amplification
First-strand synthesis (Barcode 1)
First-strand synthesis (Barcode n)
PCR optimisation
PCR optimisation
Large scale PCR
Large scale PCR
2. cDNA purification and enrichment
1x Ampure PB bead purification
0.4x Ampure PB bead purification
1x Ampure PB bead purification
0.4x Ampure PB bead purification
3. Library preparation
Pooling 1x and 0.4x fractions
Pooling 1x and 0.4x fractions
Equimolar pooling of samples
cDNA capture
SMRTbell template preparation
Anneal & bind SMRTbell templates
Sequence

Figure 3.3: Iso-Seq lab workflow for targeted profiling. Shown is a flow diagram of the Iso-Seq lab workflow for targeted profiling, which follows a similar workflow to that used in global transcriptome profiling (depicted in **Figure 3.2**), with the addition of a target cDNA capture step (boxed orange, described in **Section 3.1.2.3**) and the use of barcode sequences in cDNA synthesis (boxed green and denoted here as Barcode 1 and Barcode n) to allow sample multiplexing. The list of barcodes used can be found in **Table 2.2**.

3.1.2.1 PCR optimisation and DNA amplification

After cDNA synthesis, cDNA products were amplified using PCR to generate sufficient material for sequencing. To minimise PCR bias resulting in under- or over-representation of specific cDNA library sizes, the optimum number of PCR cycles for amplification was determined using the PrimeSTAR GXL DNA Polymerase (Clontech) (**Figure 3.4**). This involved collecting 5μL PCR aliquots every two cycles (cycles 10, 12, 14, 16, 18 and 20) followed by the visualisation of cDNA products with ethidium bromide on a 1.5% agarose gel. The optimum cycle number was determined by the cycle that generated sufficient amount of cDNA without compromising on the molecular weight, which is typically observed with PCR over-amplification of cDNA (**Figure 3.4**). Large-scale PCR amplification was then subsequently performed using the optimum number of cycles, which is typically 14 cycles (as illustrated later in **Figure 4.2** and **Figure 6.3**) .

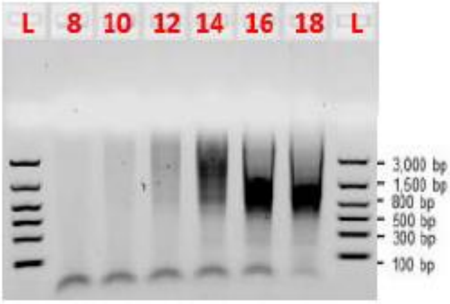


Figure 3.4: Example of an agarose gel for determining the optimum number of PCR cycles. Shown is an agarose gel of human brain total RNA after first-strand cDNA synthesis and PCR amplification through cycles 8 to 18, with PCR aliquots collected every two cycles. In this example, 10 cycles were determined to be the optimum number for large-scale amplification. While the smear distribution from 8 and 10 cycles looked similar, 10 cycles showed a slightly stronger smear, thereby generating more material for downstream pooling. Cycles above 12 showed signs of over-amplification, which would result in biased sequencing representation. Figure and legend are adapted from Iso-Seq protocol.

3.1.2.2 AMPure bead purification

After large-scale amplification, the resulting PCR products were divided into two fractions for purification with either 0.4X or 1X AMPure PB beads (PacBio). DNA purification with 0.4X AMPure beads was essential to ensure enrichment of longer fragments for sequencing (as described in **Section 2.2.4**). The quantity and size distribution of each fraction were then determined using the Qubit dsDNA High Sensitivity assay (Invitrogen) (**Section 2.1.5**) and Bioanalyzer assays on the 2100 Bioanalyzer (Agilent) (**Section 2.1.4**). The molarity of the two fractions were then calculated using **Equation 3.1**, and equal molar quantities of the two fractions were subsequently pooled for library construction with the SMRTbell Template Prep Kit v1.0 (PacBio).

$$\frac{concentration(\frac{ng}{\mu L}) \times 10^6}{660(\frac{g}{mol}) \times average\ library\ size\ in\ bp^*} = concentration\ in\ nM \tag{3.1}$$

* the average library size was determined by the start and end point of the cDNA smear on the Bioanalyzer

3.1.2.3 Target capture using IDT probes

Targeted profiling of the transcriptome was performed as part of the long-read sequencing experiments described in **Chapter 6**. This first involved equimolar pooling of uniquely-barcoded samples, followed by enrichment for target genes with a hybridisation-based capture approach (IDT). Through this approach, regions of interest within the library were captured ("hybridised") using pre-designed, 5' biotinylated, 120 nucleotide-long oligonucleotide baits (henceforth referred to as "probes", **Figure 3.5A**). Magnetic streptavidin beads were then used to isolate the hybridised library fragments for amplification (using Takara Hot-Start polymerase) and AMPure bead purification (outlined in **Figure 3.5B**). After assessing the quality and quantity of the target cDNA with Qubit and Bioanalyzer assays, SMRTbell library preparation was proceeded according to manufacturer's protocol.

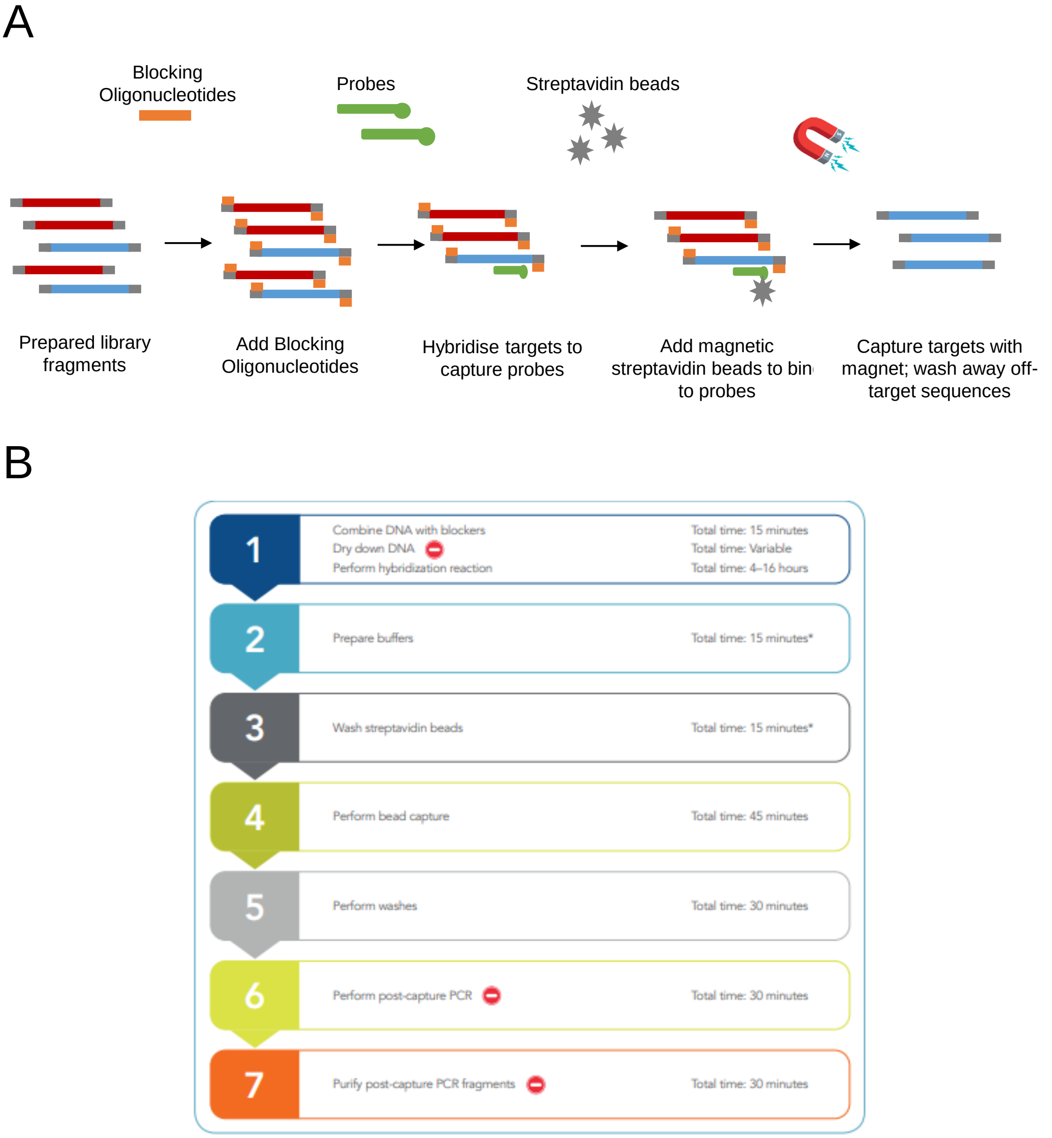
Probe Designs

Probes were designed to a selective panel of 20 AD-associated genes (henceforth referred to as "target genes"): *Abca1, Abca7, Ank1, Apoe, App, Bin1, Cd33, Clu, Fus, Fyn, Mapt, Picalm, Ptk2b, Rhbdf2, Snca, Sorl1, Tardbp, Trem2, Trpa1, Vgf* - the relevance of these genes in AD pathogenesis are detailed later in **Chapter 6** (**Table 6.2**). Two separate pools of equimolar probes were designed against the mouse (GRCm28/mm10) and human genome (GRCh37/hg19). While IDT provided a pre-designed set of probes to the exons of target genes (depicted in **Figure 3.6A**), the majority of exons were unnecessarily covered by contiguous probes, which can induce off-target binding and additional costs. Considering that previous targeted sequencing studies using the same hybridisation-based capture have achieved successful enrichment and sequencing with a few unique probes to the exonic region,[212] I manually assessed the list of probes for each target gene using the following criteria:

- All of the exons must be covered by at least one probe.

- Probes should be spaced 300 - 500bp within each exon (equivalent to 0.2x – 0.3x tiling density).
- Probes with the highest GC content (40 - 65% GC content) and lowest number of blast hits were selected from the contiguous cluster.
- Any probes covering the intronic regions were removed.

Examples of the initial set of probes provided by IDT and the final curated sets of probes after my filtering are illustrated in **Figure 3.6B,C**.


A
Blocking Oligonucleotides
Probes
Streptavidin beads
Prepared library fragments
Add Blocking Oligonucleotides
Hybridise targets to capture probes
Add magnetic streptavidin beads to bin to probes
Capture targets with magnet; wash away off-target sequences
B

1 Combine DNA with blockers — Total time: 15 minutes
Dry down DNA ⛔ — Total time: Variable
Perform hybridization reaction — Total time: 4–16 hours

2 Prepare buffers — Total time: 15 minutes*

3 Wash streptavidin beads — Total time: 15 minutes*

4 Perform bead capture — Total time: 45 minutes

5 Perform washes — Total time: 30 minutes

6 Perform post-capture PCR ⛔ — Total time: 30 minutes

7 Purify post-capture PCR fragments ⛔ — Total time: 30 minutes

Figure 3.5: Lab workflow for hybridisation-based cDNA capture for targeted profiling. Shown is **(A)** a schematic figure of the target gene enrichment step involving hybridisation of cDNA with probes and blocking oligonucleotides (such as oligonucleotides complementary to the poly(A) tail and cDNA synthesis primers), followed by isolation with streptavidin beads. The addition of blocking oligonucleotides prevents non-specific binding, and subsequently increases capture rate and target gene sequencing coverage. **(B)** An overview of the lab workflow.

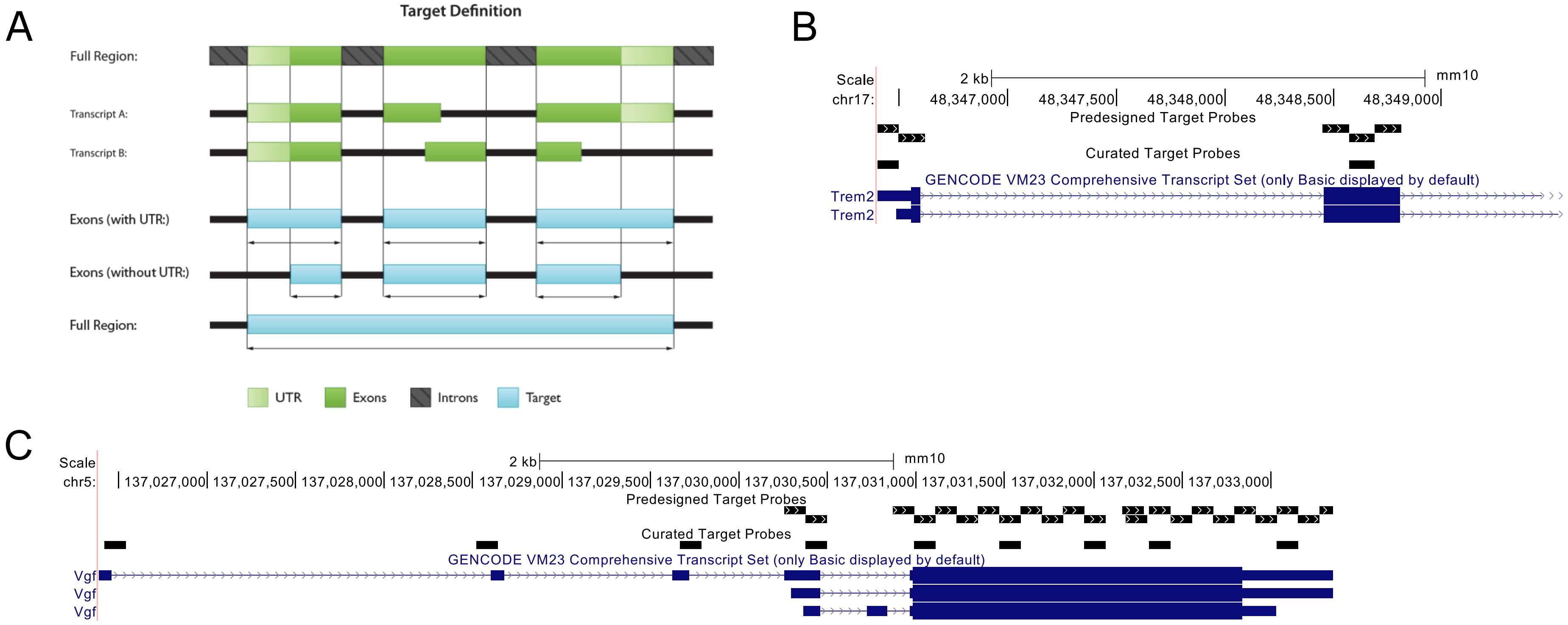


Figure 3.6: Manual curation of probes designed to 20 AD-associated target genes. **(A)** Pre-designed set of probes were provided to "Exons (with UTR)" of target genes. Shown are UCSC genome browser tracks of pre-designed and curated probes to **(B)** *Trem2*, and **(C)***Vgf* in the mouse genome (mm10). As shown, exons were unnecessarily covered by contiguous probes, which not only increase costs, but also induce off-target binding (referred in Figure B and C as "Pre-designed Target Probes"). Manual curation was therefore needed for each target gene (referred in Figure B and C as "Curated Target Probes") to ensure that exons were covered with one probe for every 500bp.

3.1.2.4 Library preparation, primer annealing & polymerase binding

After equimolar pooling of the two size fractions for global transcriptome profiling or multiple samples for targeted profiling (**Section 3.1.2.2**), SMRTbell template preparation was performed with the SMRTbell Template Prep Kit v1.0 (PacBio). This first involved repairing DNA damage and polishing the ends of fragments (**Figure 3.7**, Step 1, 2), which is essential for the generation of high-quality libraries of closed, continuous and circular SMRTbell templates. Abasic sites were filled-in, thymine dimers resolved, and deaminated cytosine alkylated. 3' overhangs were removed, whereas 5' overhangs were filled-in by T4 DNA Polymerase and phosphorylated by T4 PNK for ligation of blunt hairpin adapters. Following 1X AMPure bead purification of repaired dsDNA, hairpin adapters were ligated to the blunt ends for 24 hours (**Figure 3.7**, Step 3). Any templates failing to ligate were removed with exonuclease III and VII (**Figure 3.7**, Step 4). The repaired and ligated SMRTbell library was then purified with two rounds of 1X AMPure beads, and assessed for quality and quantity with Qubit and Bioanalyzer assays before proceeding to primer annealing and polymerase binding (**Figure 3.7**, Step 5, 6). Of note, the primer and polymerase to template ratio was key to successful loading of SMRTbell templates into ZMWs for sequencing, and was dependent on the final library molarity (as determined using **Equation 3.1**).

3.1.2.5 Loading and sequencing

All Iso-Seq experiments in this thesis were performed on the PacBio Sequel 1M SMRT cell. Samples were processed using either the v3 chemistry (Diffusion Loading at 5pM with a 4-hour pre-extension and a 20-hour capture time) or v2.1 chemistry (Magbead Loading at 50pM with a 2-hour pre-extension and a 10-hour capture time). As suggested in the name, Diffusion Loading involves immobilising polymerase-bound SMRTbell templates to ZMW by diffusion, whereas Magbead loading uses paramagnetic beads ("Magbeads") that roll across the ZMWs. Due to the different nature of loading, Diffusion Loading preferentially loads shorter transcripts, whereas Magbead Loading preferentially loads longer transcripts (> 1b). As a quality-control measure of loading and sequencing performance, a DNA internal control complex (PacBio) was added to each library before sequencing, the amount of which was dependent on the final library molarity. Mimicking SMRTbell templates, this internal control is composed of a 1966bp-insert with the SMRTbell adapters already ligated and the polymerase already bound.

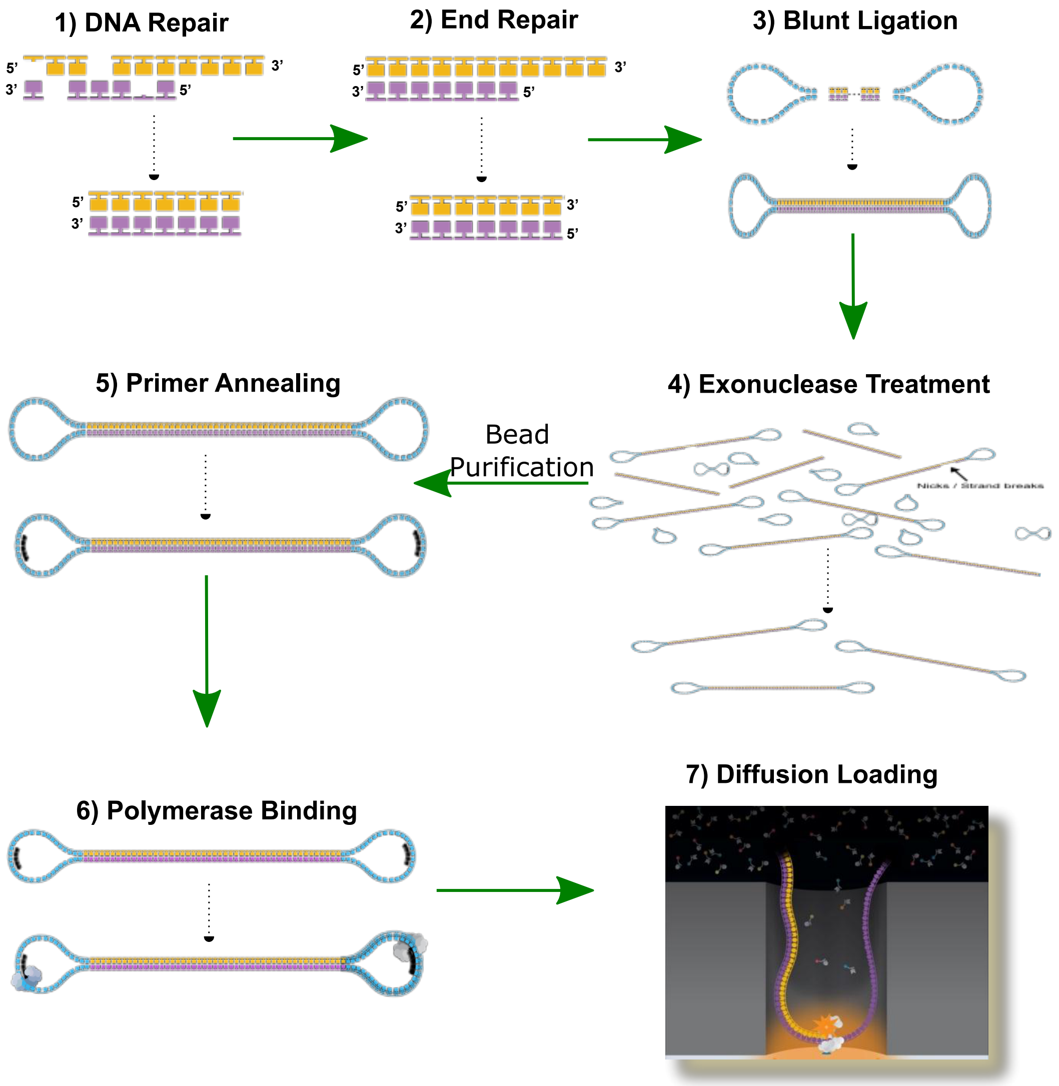


Figure 3.7: Detailed Iso-Seq lab workflow for SMRTbell library preparation. Shown is a flow diagram of the Iso-Seq lab workflow for library preparation:

1. Repair DNA by filling-in abasic sites, removing thymine dimers, oxidising guanines and deaminating cytosines. This is essential to ensure that there is a continuous sequence for uninterrupted polymerase processivity.
2. Repair ends for blunt ligation with removal of 3' hangs and addition of 5' hangs by T4 DNA polymerase, which is required for blunt ligation of SMRTbell adapters.
3. Blunt Ligation by adding hairpin SMRTbell adapters to repaired ends.
4. Exonuclease treatment to remove incomplete SMRTbell templates with Exonuclease III and IV to ensure optimal sequencing.
5. Annealing of primers to both ends of the SMRTbell templates to initiate sequencing.
6. Binding of polymerase to both ends of the SMRTbell templates for efficient loading into ZMWs.
7. Immobilisation of polymerase-bound SMRTbell templates to ZMW by diffusion.

Individual figures and legend are taken and adapted from "PacBio Sequel Library and Sequencing Preparation" presentation.

3.1.3 Run performance and quality metrics

Suboptimal PacBio sequencing performance can result from various causes including potential issues with the instrument and sequencing reagents to poor library preparation and incorrect loading. The performance of a sequencing run can be assessed by the performance of the DNA internal control and various productivity metrics.

DNA internal control

Sequencing metrics for the DNA internal control (described in **Section 3.1.2.5**) is provided in several ways: i) the number of control reads, ii) the mean control polymerase read length, and iii) the proportion of sequence identity match between the control raw reads and the reference control (concordance). Short control read lengths and/or low control read counts are suggestive of issues with the PacBio instrument and consumables, while a low concordance value (< 0.84) indicates overloading of SMRT cells. Conversely, normal control sequencing metrics in a run with overall low yield indicate sample-specific issues. The expected sequencing metrics from a correctly prepared control in an optimal Iso-Seq sequencing run are documented in **Table 3.1**.

Table 3.1: Iso-Seq DNA internal control sequencing metrics. Tabulated are the expected median values for the number of control reads (median count), the control polymerase read length (median length), and the identity match between control raw reads and reference sequence (median concordance). The expected values provided assume a sequencing run using the PacBio Sequel 1M SMRT cell with a 4-hour pre-extension and a 20-hour capture time. QR - Quantile range.

Metrics	Median count (QR)	Median length (kb) (QR)	Median concordance (QR)
Expected values	6900 (4000 - 10200)	46.9 (41.5 – 52.5)	0.862 (0.857 – 0.867)

Productivity metrics

Productivity or loading metrics provide a measure of the number of ZMWs that generated a positive signal that was then translated into useful sequencing data. Each ZMW is classified as either:

- P0 (Productivity 0): no active sequencing polymerase complex with no signal.
- P1 (Productivity 1): productive ZMWs with a high-quality (HQ) sequence within read.
- P2 (Productivity 2): detectable signal but no HQ sequence detected, possibly due to overloading of multiple inserts with multiple polymerases.

An optimal sequencing run would achieve a total run yield of 20 - 30Gb with ~70% of ZWMs in P1 (positive signal), and 20 - 30% ZMWs in P0 (empty ZMWs). A low P0 (< 20%) indicates over-loading of polymerase-bound templates, resulting in shorter P1 polymerase reads and poor sequencing yield (noisy basecalling). Conversely, a high P0 (> 40%) from under-loading would generate fewer P1 reads and result in a lower sequencing yield. A combination of high P0, low P1 and high P2 loading profiles indicates presence of contaminants (possibly from poor AMPure bead purification) that is interfering with productive polymerase activity. A good balance between P0, P1 and P2 is therefore key to achieving a good sequencing run with high yield and high-quality, long P1 polymerase reads. Multiple titrations of loading concentrations can be trialled to determine the optimum loading concentration essential for reaching this balance.

3.1.4 Bioinformatics pipeline

This section describes the bioinformatics pipeline that we established for analysing Iso-Seq data generated on the PacBio Sequel I following Iso-Seq library preparation (**Chapters 4 - 6**).

The bioinformatics pipeline, as depicted in **Figure 3.8**, involves three main steps: i) the processing and filtering of raw reads to generate HQ, full-length transcripts using the PacBio *IsoSeq3* suite,[141] ii) the alignment of HQ transcripts to the reference genome using *Minimap2*,[213] and iii) the clustering and collapsing of mapped transcripts to unique, annotated isoforms using *Cupcake*[214] and *SQANTI*.[215] Public annotations and short-read RNA-Seq data were used for validating Iso-Seq-derived isoforms.

While raw Iso-Seq data can be processed using the PacBio SMRT Link Suite, a web-based end-to-end user interface, we developed an end-to-end command line that allowed simultaneous parallel processing of multiple samples and streamlining of the analysis after raw read processing. The choice of parameters and packages was guided by a separate analysis on external RNA spike-in controls (ERCC) that were sequenced within the same runs (described in **Section 2.3**).

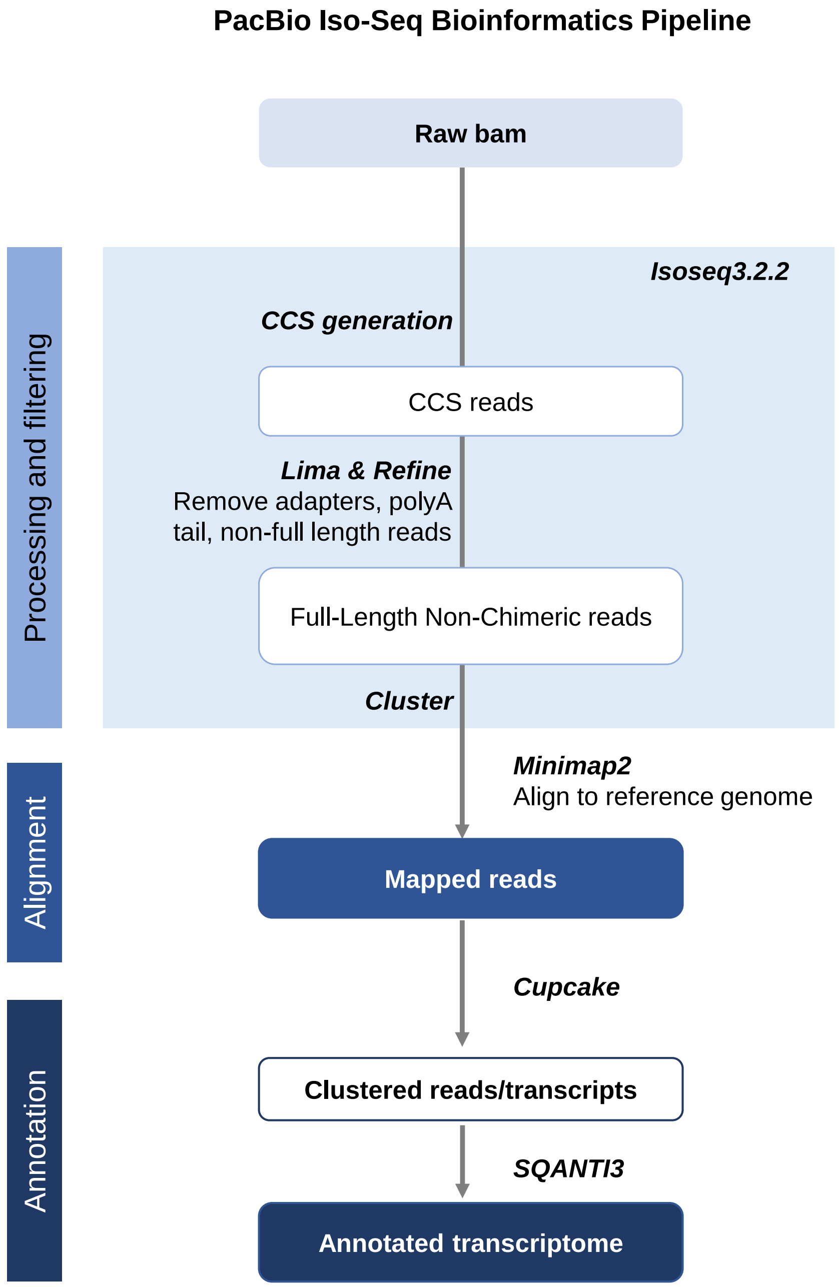

PacBio Iso-Seq Bioinformatics Pipeline
Raw bam
Isoseq3.2.2
CCS generation
CCS reads
Lima & Refine
Remove adapters, polyA tail, non-full length reads
Full-Length Non-Chimeric reads
Cluster
Minimap2
Align to reference genome
Mapped reads
Cupcake
Clustered reads/transcripts
SQANTI3
Annotated transcriptome
Processing and filtering
Alignment
Annotation

Figure 3.8: Iso-Seq bioinformatics pipeline. Shown is an overview of the Iso-Seq bioinformatics pipeline used in this thesis, which involves three main steps: i) processing and filtering of raw reads into high-quality, full-length transcripts using *IsoSeq3*, ii) alignment of HQ transcripts to the reference genome using *Minimap2* and iii) collapsing of mapped transcripts to unique, annotated isoforms using *Cupcake* and *SQANTI*. HQ - High-quality.

3.1.4.1 Processing of Iso-Seq raw reads

In response to the much higher experimental throughput of the PacBio Sequel compared to the older PacBio RSII sequencer, the official PacBio bioinformatics suite for processing Iso-Seq reads (henceforth referred to as *Iso-Seq3*) has been revised multiple times over the course of this PhD. Each subsequent version delivered a reduction in runtime coupled with an improvement in sensitivity and specificity to recover transcripts and reduce artefacts. A particularly noteworthy development was the forgoing of non-FL reads or RNA-Seq short-reads for error correction, due to the high throughput and subsequent generation of high-quality, accurate Iso-Seq reads.

Despite multiple major updates to *Iso-Seq3*, the core principles and processing steps have remained the same (depicted in **Figure 3.9**), namely: i) the generation of CCS reads from each sequencing ZMW, ii) the identification of full-length reads with the removal of cDNA primers and poly(A) tails, and iii) the grouping of full-length reads derived from the same transcript.

Generation of CCS reads with *CCS*

Raw Iso-Seq subreads from each productive ZMW were processed to generate one representative circular consensus read (**Figure 3.9A,B**) using *CCS* (v5.0.0) with the following parameters:

- minimum number of full “passes” for a ZMW to be considered. A full pass is defined by the presence of both SMRT adapters at both ends (default: 3 passes).
- minimum predicted read accuracy across all subreads (default: 99%).
- minimum and maximum length of subreads to generate a CCS (default: 10 and 21000 bases, respectively).
- quality of subreads predicted by the CCS model (default: -3.5 Z-score), and proportion of total subreads meeting the quality score (default: > 30%).

Removal of primers and barcodes with *lima*

After the successful generation of CCS reads, cDNA primers were identified and removed using *lima* (v2.0.0) to generate full-length (FL) reads (**Figure 3.9C**). Additional barcode sequences were also removed for targeted sequencing experiments to perform sample demultiplexing. Sequences were then orientated from 5’ to 3’, and any reads with unwanted combinations were removed. Of note, the ratio of recovered FL reads to CCS reads varies on the insert transcript size, but a good sequencing library with a distribution of 1kb - 3kb should recover 60 - 70% of CCS reads as FL reads.

Trimming of poly(A) tails and concatemer removal with *Iso-Seq Refine*

FL reads were further refined by the trimming of poly(A) tails (with a minimum length of 20 adenosine bases). Artificial concatemers were then removed to ensure a library of full-length non-chimeric (FLNC) reads (**Figure 3.9D**). Of note, artificial concatemers are cDNA sequences with internal runs of poly(A) and poly(T) sequences that were generated from using insufficient amount of blunt adapters during library preparation. The occurrence of these artefacts should be low (< 0.5%) in a standard library preparation, and the number of FLNC reads and FL reads should be similar. Any significant loss of reads at this stage implicates issues with SMRTbell library preparation.

Grouping of reads into transcripts with *Iso-Seq Cluster*

Using an iterative isoform-clustering algorithm, two or more FLNC reads were then grouped and considered to be the same transcript if they:

- differed < 100bp on the 5' end*.
- differed < 30bp on the 3' end.
- did not contain internal gaps with > 10bp.

* Greater leeway was given to the 5' end than the 3' end to account for 5' RNA degradation.

A minimum of two FLNC reads was required for clustering with the longest read chosen as the representative transcript, and any unique FLNC reads failing to cluster were discarded (**Figure 3.9E**). Transcripts generated from the *Iso-Seq Cluster* were therefore high-quality with a consensus accuracy $\geq$ 99% and a minimum of two FLNC read support (**Figure 3.9F**).

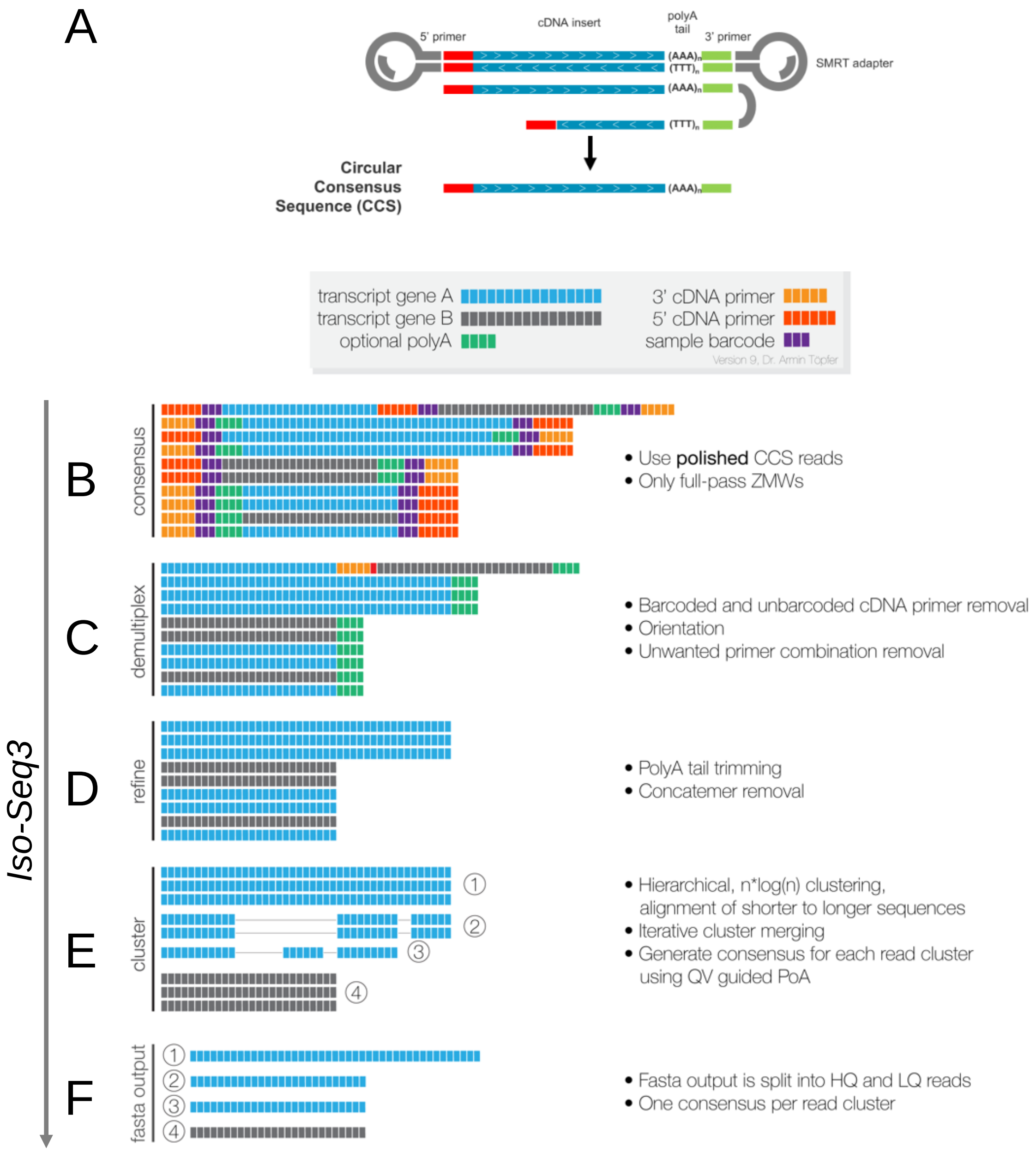

A
cDNA insert
polyA tail
5' primer
3' primer
SMRT adapter
(AAA)n
(TTT)n
Circular Consensus Sequence (CCS)
transcript gene A
transcript gene B
optional polyA
3' cDNA primer
5' cDNA primer
sample barcode
Version 9, Dr. Armin Töpfer
Iso-Seq3
B
consensus
• Use polished CCS reads
• Only full-pass ZMWs
C
demultiplex
• Barcoded and unbarcoded cDNA primer removal
• Orientation
• Unwanted primer combination removal
D
refine
• PolyA tail trimming
• Concatemer removal
E
cluster
① ② ③ ④
• Hierarchical, n*log(n) clustering, alignment of shorter to longer sequences
• Iterative cluster merging
• Generate consensus for each read cluster using QV guided PoA
F
fasta output
① ② ③ ④
• Fasta output is split into HQ and LQ reads
• One consensus per read cluster

Figure 3.9: PacBio *Iso-Seq3* bioinformatics suite for raw Iso-Seq read processing. An overview and flow diagram of the *Iso-Seq3* bioinformatics suite. **A)** The circular SMRTbell template allows uninterrupted, processive DNA synthesis to generate a polymerase read containing multiple subreads. **B)** *CCS* - A polymerase read, associated with each productive ZMW, with multiple "passes" are processed to generate a CCS read containing the 5' and 3' cDNA primer, poly(A) tail and barcode (if used). **C)** *lima* - Successfully generated CCS reads are then trimmed for cDNA primers and orientated to generate FL reads. **D)** *Refine* - FL reads are then trimmed for poly(A) tails and artificial concatemers are removed to generate FLNC reads. **E)** *Cluster* - FLNC reads considered to be derived from the same transcript are then clustered to generate unique transcripts. **F)** The primary output from this pipeline are accurate, high-quality transcripts. Of note, raw Iso-Seq reads are processed without using a reference genome or transcriptome, and the abundance for each transcript can be inferred from the number of associated FL reads (i.e number of ZMWs that sequenced the isoform of interest). CCS - Circular consensus sequence, FL- Full-length, FLNC - Full-length non-chimeric. Figure is taken from PacBio *Iso-Seq v3* GitHub repository.[216]

3.1.4.2 Alignment to reference genome

HQ transcripts generated from the *IsoSeq3* package were aligned to the reference genome using *Minimap2*[213] (v2.17), a splice-aware aligner that is faster, more precise and accurate than other mainstream mappers.[166, 217] Under the recommended parameters ("-ax splice -uf –secondary=no -C5 -O6,24 -B4"), *Minimap2* prioritises the known canonical junctions (GT[A/G]…[C/T]AG) over non-canonical splice junctions (GT[C/T]…[A/G]AG), and assumes that the read orientation is unknown in order to perform two rounds of alignment for greater accuracy.

3.1.4.3 Further transcript collapse to isoforms using *Cupcake*

Aligned HQ transcripts were filtered and further collapsed to unique, full-length, high-quality isoforms using a set of supporting scripts from *Cupcake* to reduce redundancy. Using the *collapse_isoforms_by_sam.py* script (parameters: "-c 85 -i 95 –dun-merge-5-shorter"), aligned transcripts with less than 85% coverage and 95% identity to the reference genome were removed. The number of associated FL reads associated with each isoform was then obtained as a proxy of isoform abundance with *get_abundance_post_collapse.py* script.

3.1.4.4 Transcriptome annotation with *SQANTI*

Isoforms were characterised using *SQANTI*[215] (v3), which i) performs a reference-based correction of sequences, ii) classifies isoforms based on splice junctions, iii) annotates the transcriptome with user-defined public annotations and matched RNA-Seq data, and iv) discards isoforms that are considered technical artefacts.

Isoform classification by splice junctions

Isoforms can be broadly classified as being either "known" or "novel" and annotated to a known gene, or a "novel gene" that is not currently present in existing reference genome annotations. Using *SQANTI*, known isoforms annotated to known genes were subclassified as "Full Splice Match" (FSM) if it fully aligned with the reference isoform with the same exonic structure and splice junctions, or "Incomplete Splice Match" (ISM) if it has fewer 5' exons than the reference isoform but is otherwise fully matching. Conversely, novel isoforms annotated to known genes were subclassified as "Novel in Catalogue" (NIC) if it contained a different exonic structure but from a combination of known donor or acceptor sites, or "Novel Not in Catalogue" (NNC) if there is at least one novel donor or acceptor site. Finally, novel genes were subclassified as either “antisense” or “intergenic” depending on the orientation. Depictions of RNA isoform classifications can be found in **Figure 3.10**. Splice junctions were defined by the two pairs of dinucleotides present at the exon-intron boundary, and any other combinations aside from GT-AG, GC-AG and AT-AC pairs were considered non-canonical.

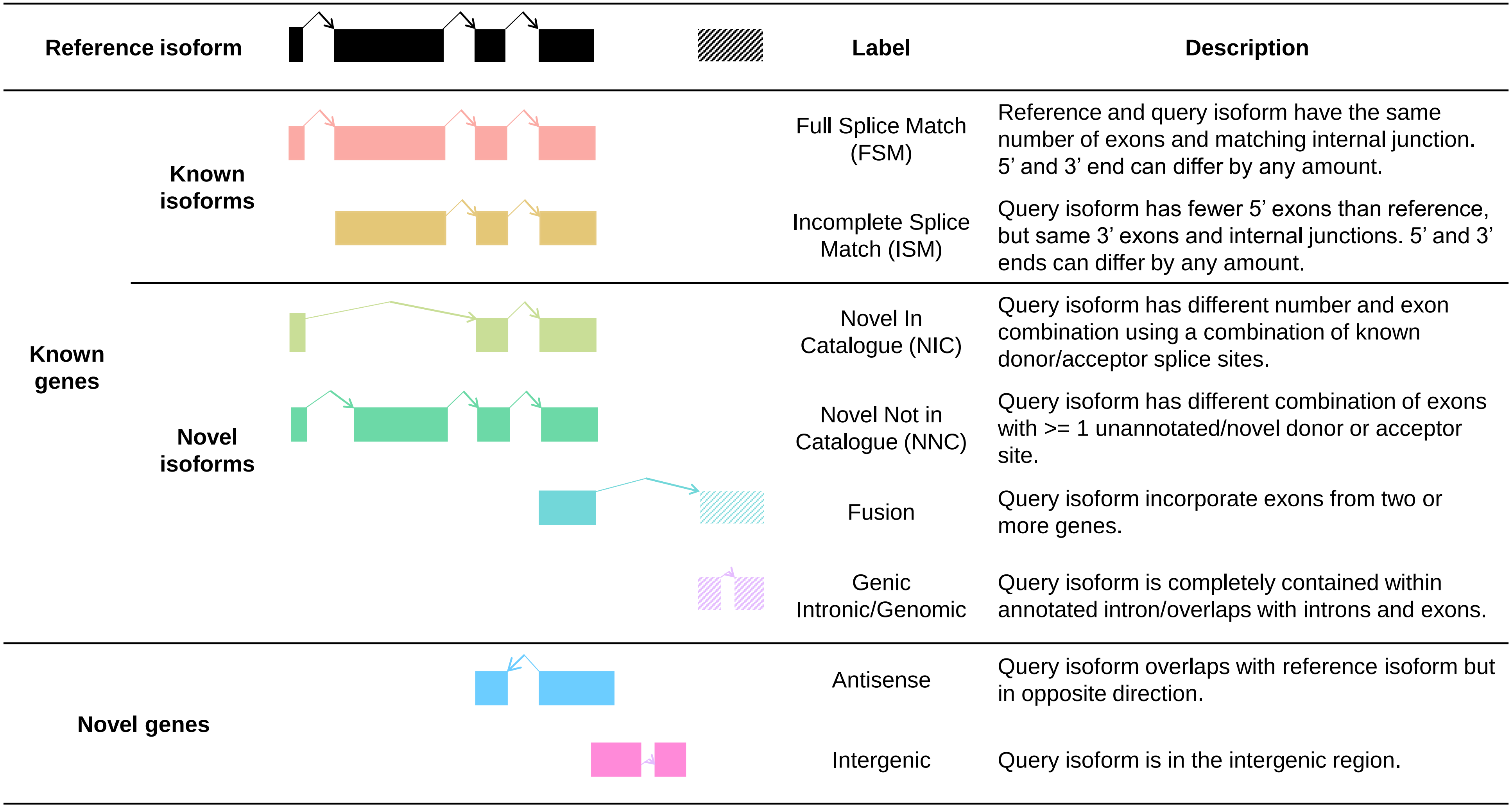
Isoforms were also classified as protein-coding, using the GeneMarkS-T algorithm,[218] if there is an open reading frame (ORF) with AUG as the initial codon. For incomplete isoforms (ISMs) with a shortened 5’ end, ORF was predicted from the first in-frame methionine. An isoform was predicted to undergo nonsense-mediated decay if there is a putative ORF and the coding sequence (CDS) ends at least 50bp from the last junction.

Usage of public annotations and matched RNA-Seq data

Various public annotations were imported into *SQANTI* for deeper characterisation of the transcriptome, including:

- the Cap Analysis of Gene Expression (CAGE) peaks derived from the FANTOM5 dataset,[219] which maps transcripts, transcription factors, transcriptional promoters and enhancers.
- the Intropolis junction bed file[220] from a comprehensive human RNA-Seq dataset.
- human and mouse poly(A) motifs provided by *SQANTI*.

RNA-Seq data from the same samples were also supplied to *SQANTI* in two forms: i) after alignment to the reference genome using *STAR*[221] (v1.9) to infer the number of RNA-Seq reads at splice junctions, and ii) after alignment to Iso-Seq-derived transcripts using *Kallisto*[222] (v0.46.0) for RNA-Seq expression.



Reference isoform		Label	Description
Known genes	Known isoforms	Full Splice Match (FSM)	Reference and query isoform have the same number of exons and matching internal junction. 5’ and 3’ end can differ by any amount.
		Incomplete Splice Match (ISM)	Query isoform has fewer 5’ exons than reference, but same 3’ exons and internal junctions. 5’ and 3’ ends can differ by any amount.
	Novel isoforms	Novel In Catalogue (NIC)	Query isoform has different number and exon combination using a combination of known donor/acceptor splice sites.
		Novel Not in Catalogue (NNC)	Query isoform has different combination of exons with >= 1 unannotated/novel donor or acceptor site.
		Fusion	Query isoform incorporate exons from two or more genes.
		Genic Intronic/Genomic	Query isoform is completely contained within annotated intron/overlaps with introns and exons.
Novel genes		Antisense	Query isoform overlaps with reference isoform but in opposite direction.
		Intergenic	Query isoform is in the intergenic region.

Figure 3.10: Isoform classifications using *SQANTI*. Shown are the isoform classifications from *SQANTI* with an isoform classified as being either “novel” or “known” and annotated to a “known” gene, or a “novel gene”.

Further filtering for technical artefacts

SQANTI Filter was used to filter the curated transcriptome for any technical artefacts introduced during library preparation, namely: i) RT template-switching events, which occur when RT transits within or across DNA templates without terminating cDNA synthesis, particularly if the original DNA template harbours two or more direct repeats[223] (**Figure 3.11A**), and ii) intra-priming events when oligo(dT) primer binds to other internal homo-polymeric adenine stretches (A's) located within the cDNA template[224] (**Figure 3.11B**). These events can generate chimeric, or short, incomplete and truncated cDNA that can otherwise be misinterpreted as isoforms generated from non-canonical splicing.[225]

It was therefore important to perform additional filtering using *SQANTI*, which identified RT-switch events by searching for direct repeats (given that RT switching is homology dependent). Intra-priming events were determined by measuring the proportion of genomic A's after the isoform 3' end within a 20-nucleotide window, and any isoforms with > 60% A's were discarded.

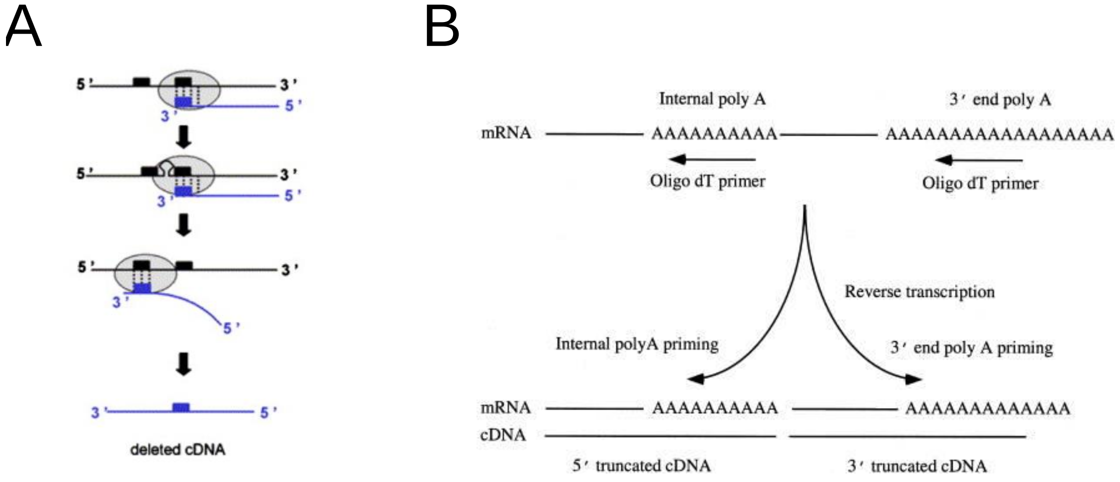


Figure 3.11: Examples of technical artefacts generated during cDNA synthesis. Shown are schematic figures of **(A)** a reverse transcription template-switching event. The black and blue lines represent the original cDNA and synthesising cDNA from RT, respectively. The black box and light grey sphere represent the direct repeats and RT enzyme, respectively. As exemplified, RT template switching is further facilitated by RNA secondary structures that could bring the repeats into proximity.[223] Figure is taken from Cocquet et al. (2006).[223] **(B)** Intra-priming events from priming of oligo(dT) to an internal poly(A) sequence rather than the 3'end poly(A) tail during cDNA synthesis, generating two truncated cDNA templates. Figure is taken from Nam et al. (2002).[226]

Under these filtering criteria (depicted in **Figure 3.12**), an isoform classified as FSM was always retained unless the 3' end was unreliable (i.e. > 50bp from reference TTS), implicating the occurrence of intra-priming events. Conversely, much more stringent filters were applied to other isoforms not classified as FSM, and such isoforms were only retained if the 3' end was reliable, if they did not contain a junction detected as RT switching and all the junctions were either canonical or supported by at least three RNA-Seq reads (if matched RNA-Seq data was provided). Of note, long-read sequencing data generated from targeted profiling experiments (**Chapter 6**) were not filtered by RNA-Seq data, due to the relatively low sequencing coverage and sensitivity of matched RNA-Seq data, which would have otherwise resulted in filtering of true novel transcripts.

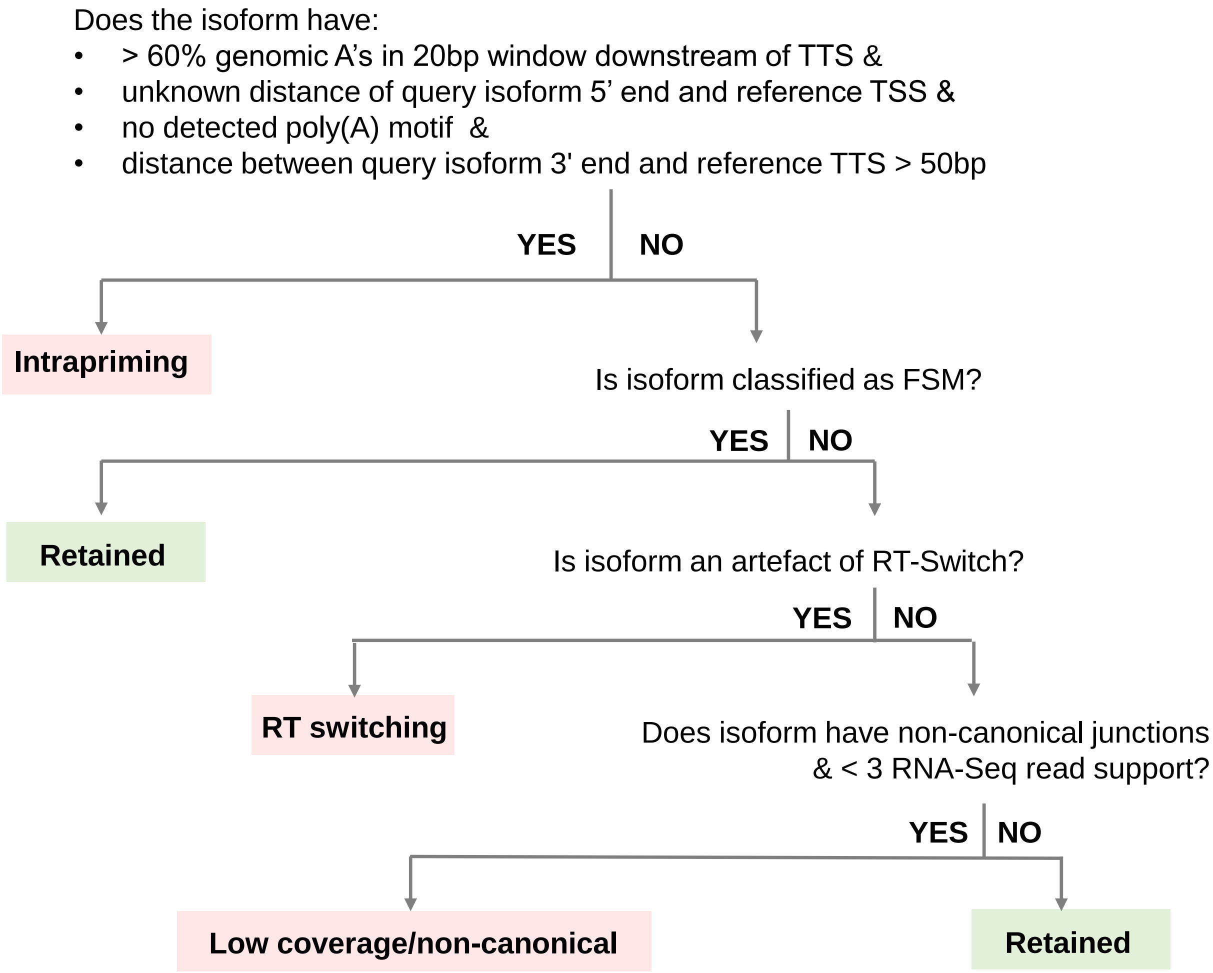

Does the isoform have:
• > 60% genomic A's in 20bp window downstream of TTS &
• unknown distance of query isoform 5' end and reference TSS &
• no detected poly(A) motif &
• distance between query isoform 3' end and reference TTS > 50bp
YES
NO
Intrapriming
Is isoform classified as FSM?
YES
NO
Retained
Is isoform an artefact of RT-Switch?
YES
NO
RT switching
Does isoform have non-canonical junctions & < 3 RNA-Seq read support?
YES
NO
Low coverage/non-canonical
Retained

Figure 3.12: Filtering of technical library artefacts using *SQANTI*. Shown is a binary decision tree for filtering of isoforms using *SQANTI Filter*. An isoform classified as "Full Splice Match" (FSM) is always retained unless the 3' end is unreliable (i.e. no detected poly(A) motif and > 50bp from reference TTS). Conversely, all the other classified isoforms are only retained if the 3' end is reliable, and do not contain any junctions that are either predicted RT switching or are not supported by matched RNA-Seq data. Coloured boxes indicate the output from the decision tree with red and green box indicating isoforms being removed and retained, respectively.

3.1.4.5 Methodological contribution: Usage of ERCC to inform Iso-Seq bioinformatic analysis

A set of 92 synthetic spike-in ERCC controls was added to the global transcriptome profiling experiments (described in **Section 2.3**) to assess the sensitivity of the Iso-Seq approach and validate our downstream bioinformatics pipeline. After processing of Iso-Seq raw reads (described in **Section 3.1.4.1**), HQ transcripts were aligned to ERCC reference sequences in parallel to the reference genome. ERCC-aligned and reference-aligned transcripts were then collapsed using *Cupcake* scripts under default parameters (-c 95 -i 99) and annotated using *SQANTI* annotations - the standard bioinformatics pipeline that has been recommended by the PacBio research community.

The application of this pipeline to our data, however, resulted in the detection of only a proportion of the individual ERCC molecules (n = 37, 40.22%). Furthermore, several ERCC molecules were annotated with more than one molecule (n = 8, 8.7%), contrary to the fact that there should only be one synthetic molecule sequenced for each ERCC. These "multiple-isoformic ERCC" molecules were generally more abundant, suggesting that more highly-expressed genes are likely to be associated with more isoforms that have failed to collapse properly. Visualisation and BLAST analysis of these "isoforms" revealed them to be shorter fragments of the original ERCC sequence, generated as technical artefacts either from fragmentation of the originals molecule or incomplete PCR synthesis. Application of *Tama-remove-fragment-models.py* script from *TAMA*[227] successfully removed these partial, redundant isoforms, while retaining the longer, intact isoforms.

Deeper investigation into the low coverage of ERCC identified 20 additional less-abundant ERCC molecules that were discarded from *Cupcake* due to an imperfect reference alignment with a shorter 5' end - a likely result of 5'degradation. Lowering the coverage threshold (the amount of sequence overlap) from 99% (default) to 95% rescued these ERCC molecules and increased the total number of ERCC molecules detected by 20% (n = 57 unique number of ERCC, 61.96%), subsequently strengthening the correlation between FL Iso-Seq read count and the actual amount of ERCC used (95% coverage: corr = 0.98, $P = 1.41 \text{ x } 10^{-41}$; 99% coverage: corr = 0.82, $P = 4.89 \text{ x } 10^{-10}$, illustrated in **Figure 4.5**). This finding highlights the limitation of our current Iso-Seq approach in failing, to i) differentiate between intact and truncated RNA, resulting in reduced confidence about isoform TSS, and ii) detect lowly-expressed genes and transcripts where the deleterious impact of RNA degradation is more significant.

3.2 Oxford Nanopore Technologies: cDNA Sequencing

3.2.1 Introduction

Following the success of PacBio SMRT for generating long sequencing reads in real-time (reviewed in **Table 1.4**), Oxford Nanopore Technologies (ONT) introduced an alternative long-read, single-molecule real-time sequencing technology with the commercial release of the MinION in 2014. In contrast to all existing sequencing applications which rely on a "sequencing-by-synthesis" approach (including PacBio SMRT sequencing), ONT pioneered the approach of directly reading a single DNA strand using a protein nanopore rather than by measuring the incorporation events on the template strand[228] (**Figure 3.13A**). Partly owing to the relatively lower cost and portability of ONT technology, nanopore sequencing has been widely used for transcriptome profiling (reviewed in **Table 1.5**), with theoretically no upper limit to read length[229] (the longest read to date is over 150kb), and with no bias towards length or GC content.[230, 231]

3.2.1.1 Mechanism

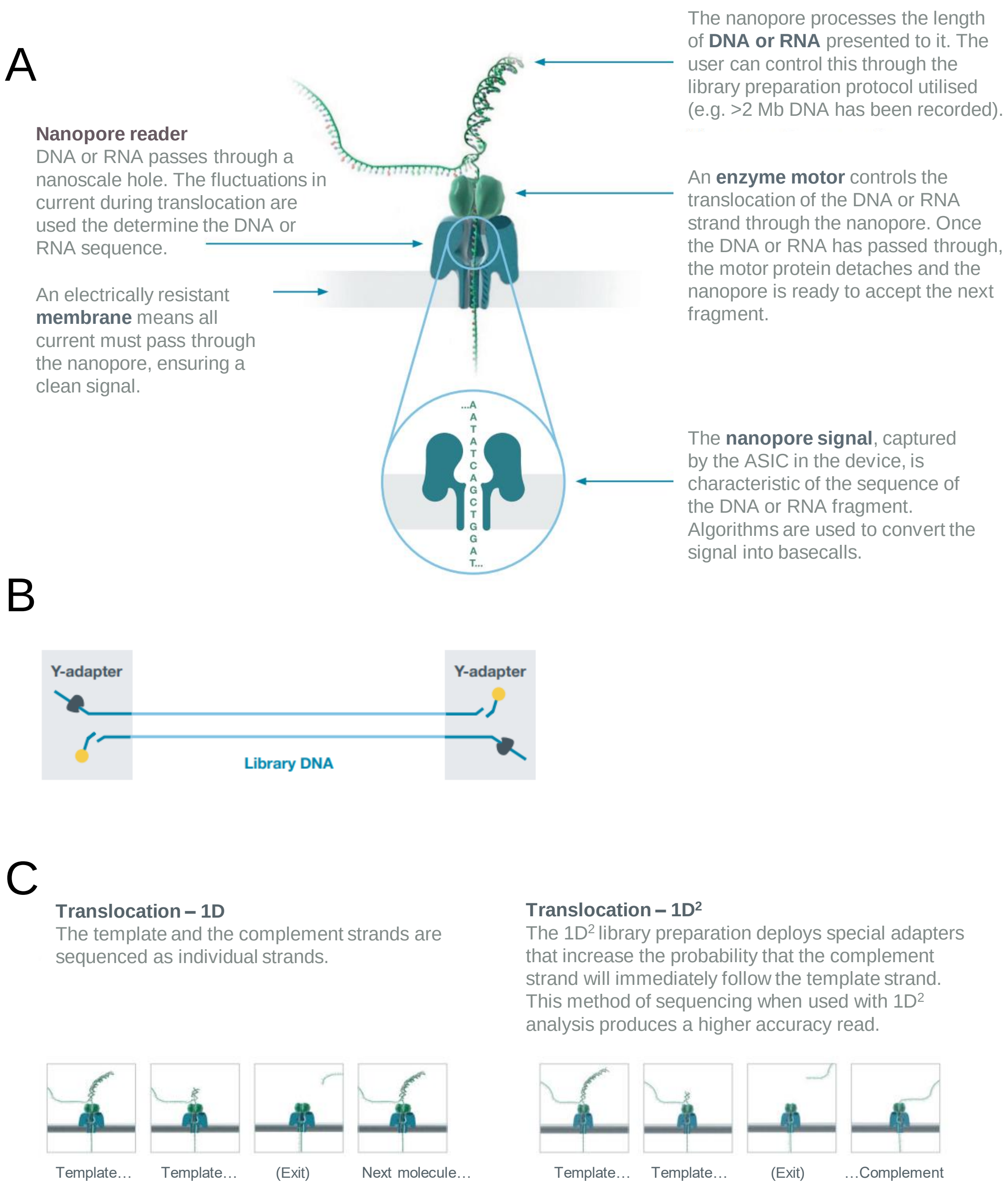
The MinION is a hand-held portable USB-powered device. At its centre is a flow cell that contains a sensor array, which houses a total of 2048 individual nanopores that are controlled in four groups of 512 channels; this allows up to 512 independent DNA molecules to be sequenced simultaneously.[228] As a voltage is applied across the nanopore, the single-stranded DNA sequence translocates through the nanopore and subsequently interrupts the current in a nucleotide-dependent manner, generating a unique signal of electric current perturbations that acts as a proxy of the underlying nucleotide sequence (**Figure 3.13A**).

Successful nanopore sequencing requires the efficient capture and threading of single-stranded DNA (ssDNA) into the pore, followed by the ability to identify individual DNA bases in a time-resolved manner. This was achieved through several key innovations:[232]

1. Generation of an internal positive charge within the protein nanopore to induce capture of negatively-charged DNA: each nanopore is embedded into an electrical resistant membrane that is immersed in an electrolyte solution.
2. Discovery and usage of biological pore proteins: the *Staphylococcus aureus* αHL pore was first implemented for sequencing,[233] followed by the *Mycobacterium smegmatis*

MspA pore.[234] Current ONT nanopore systems use a modified *Escherichia coli* CsgG pore, which contain a short and narrow channel constriction site, enabling detection of distinct ionic currents at a single-nucleotide resolution.

3. Ratcheting the DNA through the pore for time-resolved base identification: this involves a processive enzyme (henceforth referred to as a "motor protein", **Figure 3.13A,B**), which facilitates DNA movement and reduces the translocation speed of the molecule for improved signal (average speed of 450bp/s).[235] This processive enzyme is ligated to the 5' end of both strands during library preparation.



Nanopore reader
DNA or RNA passes through a nanoscale hole. The fluctuations in current during translocation are used the determine the DNA or RNA sequence.

An electrically resistant **membrane** means all current must pass through the nanopore, ensuring a clean signal.

The nanopore processes the length of **DNA or RNA** presented to it. The user can control this through the library preparation protocol utilised (e.g. >2 Mb DNA has been recorded).

An **enzyme motor** controls the translocation of the DNA or RNA strand through the nanopore. Once the DNA or RNA has passed through, the motor protein detaches and the nanopore is ready to accept the next fragment.

The **nanopore signal**, captured by the ASIC in the device, is characteristic of the sequence of the DNA or RNA fragment. Algorithms are used to convert the signal into basecalls.

Translocation – 1D
The template and the complement strands are sequenced as individual strands.

Translocation – 1D^2
The 1D^2 library preparation deploys special adapters that increase the probability that the complement strand will immediately follow the template strand. This method of sequencing when used with 1D^2 analysis produces a higher accuracy read.

Figure 3.13: ONT nanopore cDNA Sequencing. Shown is an overview of Oxford Nanopore Technologies (ONT) nanopore platform. **(A)** ONT nanopore sequencing involves the translocation of a DNA sequence through a biological nanopore, which is controlled by the enzyme motor protein and causes nucleotide-sensitive perturbations in the electric current. **(B)** The structure of the library DNA after ligation of the sequencing adapters, containing the motor protein (brown circle), to the template and complementary strand. **(C)** Two sequencing translocation modes are currently offered, generating either 1D or 1D^2 reads. Figures are taken from the Oxford Nanopore Product Brochure July 2018.

3.2.2 Lab workflow

This section describes the library preparation for ONT cDNA sequencing experiments used in **Chapter 6** for the targeted profiling of AD genes in the mouse cortex. At the time of my PhD research, the ONT technology was significantly less advanced that the PacBio technology with only basic protocols. Nanopore sequencing was therefore conducted on a subset of mouse samples as a source of validation and technology comparison using methods optimised during my research.

For a fairer and more direct comparison, all steps prior to the ONT library preparation were adopted from the Iso-Seq protocol, including the conversion of RNA to cDNA using the SMARTer PCR cDNA synthesis kit (**Section 2.2.1**), large-scale DNA amplification using the GXL DNA Polymerase and target enrichment with hybridisation-based capture (workflow is depicted in **Figure 3.14**). Consequently, nanopore reads were generated with the same cDNA primers and barcode sequences as Iso-Seq reads (refer to **Table 2.2** for sequences). Post cDNA synthesis and amplification, the ONT library preparation was broadly similar to the Iso-Seq library preparation (also outlined in **Figure 3.14**), with the exception that the motor enzyme is pre-bound to the adapters in ONT reads whereas the polymerase is only loaded after adapter ligation in Iso-Seq.

3.2.2.1 ONT MinION library preparation

After obtaining high-quality and full-length cDNA sequences, nanopore library preparation was performed using the SQK-LSK109 1D Sequencing by Ligation protocol (outlined in **Figure 3.15**). The ONT library preparation was relatively simple: cDNA ends were first repaired and dA-tailed using the NEBNext End Repair/dA-tailing Module, followed by 1X AMPure bead purification and adapter ligation. The library was then subjected to a final round of 0.4X AMPure Bead Purification before loading into the ONT MinION for sequencing. The ONT adapters (depicted in **Figure 3.13B**) contained a dT overhang for ligation to the dA-ends of cDNA, the pre-bound motor protein, and a cholesterol moiety which facilitates DNA capture by tethering the molecule to the flow cell's lipid membrane.

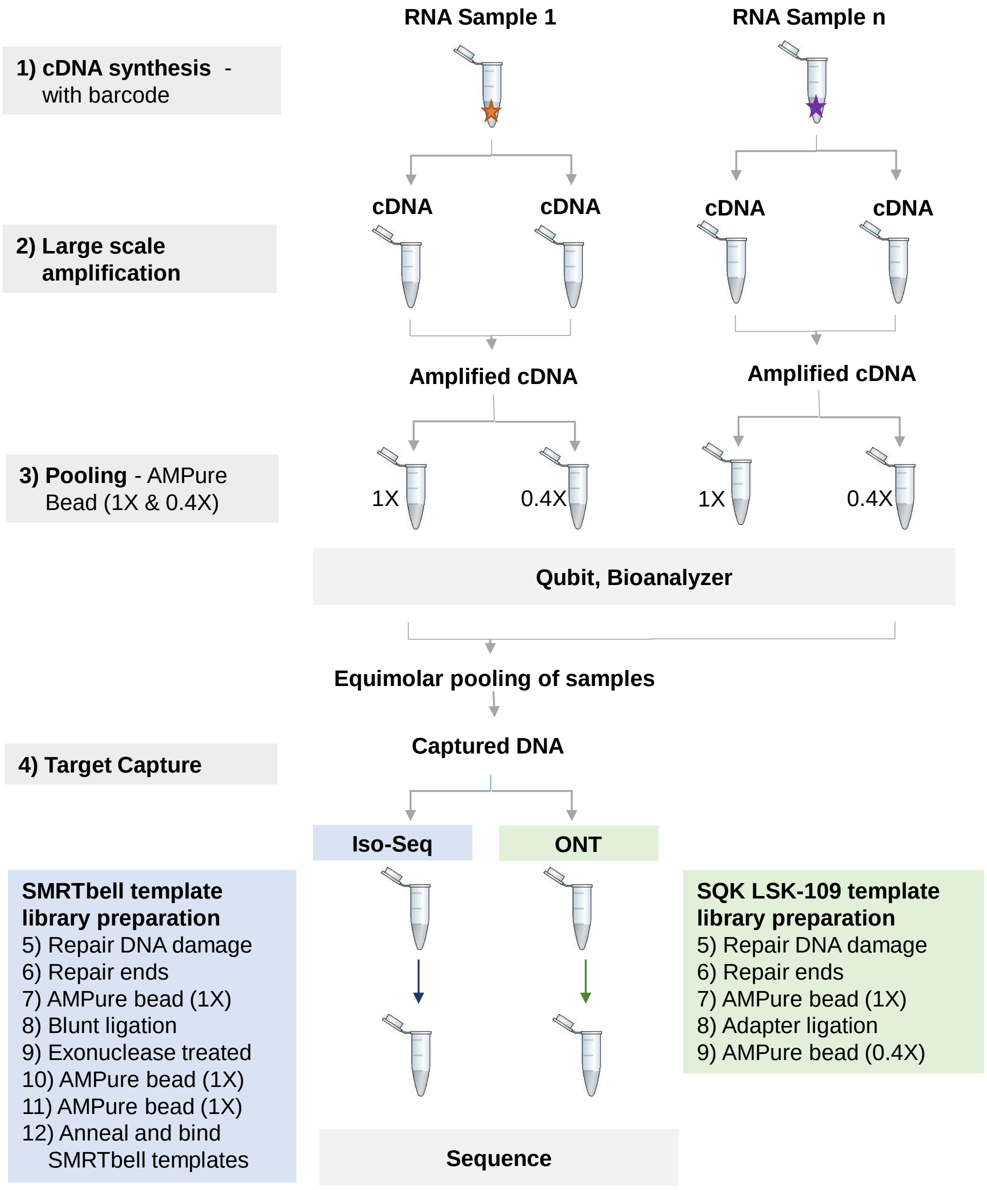


Figure 3.14: Comparison of the ONT and Iso-Seq lab workflow for targeted profiling. Shown is a flow diagram of the ONT lab workflow in parallel with the Iso-Seq lab workflow for a fair and direct comparison of targeted profiling. RNA was barcoded (denoted by the orange and purple star) and converted to cDNA using the SMARTer PCR cDNA synthesis kit. Amplified and purified cDNA was then pooled in equimolar quantities across multiple fractions and samples, followed by target gene enrichment using the hybridisation-based capture (IDT), respective library preparation and sequencing.

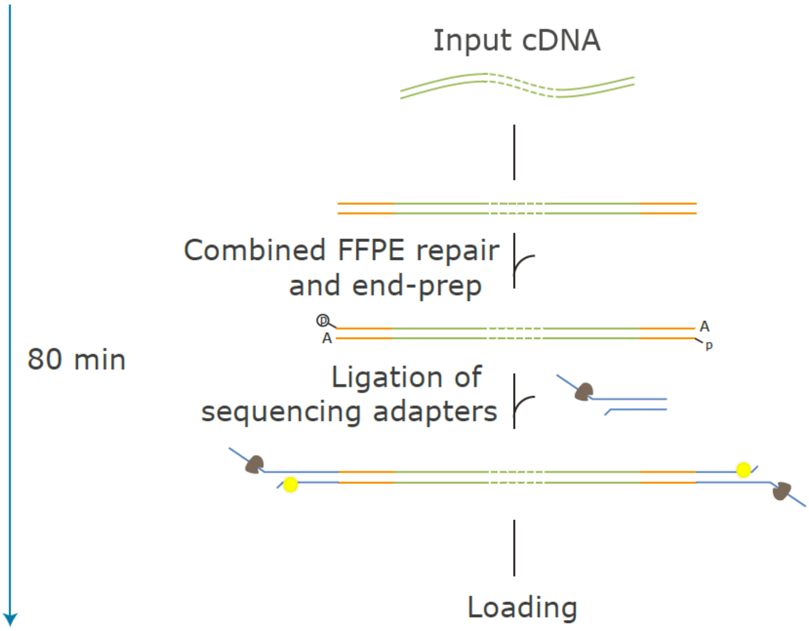

Input cDNA
Combined FFPE repair and end-prep
Ligation of sequencing adapters
Loading
80 min
A
A
p

Figure 3.15: ONT library preparation with 1D ligation sequencing kit. Shown is a flow diagram of the ONT library preparation with the ONT ligation sequencing kit (SQK-LSK109), which primarily involved repairing cDNA ends and dA-tailing followed by ligation of sequencing adaptors. The motor protein and cholesterol moiety are represented by the brown and yellow circle, respectively. Figure is adapted from ONT Nanopore Protocol 1D amplicon/cDNA by Ligation (SQK-LSK109).

3.2.2.2 Priming the Flow Cell and Sequencing

Nanopore sequencing was performed on the MinION using a Min106D Flow cell, which contains the R9 nanopore (as shown in **Figure 3.16**). Prior to sequencing, the flow cells were tested for the total number of functional pores present and were only used if > 800 pores were available (as recommended by ONT). The flow cell was subsequently primed for sequencing using a "Running Buffer with Fuel mix" (RBF), which contained the substrate cofactor essential for efficient motor protein activity (i.e. ATP for the ATPase activity of the helicase component of the translocation motor). The library was then loaded onto the MinION with "Library Loading Beads" (LLB), which are sepharose beads that work on a principle similar to that of Iso-Seq MagBeads by immobilising the library to the lipid membrane.

3.2.3 Run performance and quality metrics

One major drawback of nanopore sequencing is the relatively high error rate compared to short-read RNA sequencing. This can arise from random and systematic error during sequencing or during translation of the raw electric signal into a DNA sequence (a process known as "basecalling").[235] The first error is exacerbated by the fact that i) several nucleotides occupy the pore at any given time point, resulting in multiple effects on the signal, and that ii) the signal does not change with translocation of homopolymers (stretches of identical bases).

However, major advances in the basecalling algorithms, the chemistry and nanopore itself have drastically increased the accuracy of single-pass sequencing reads from 60%,[228] to 98.3% (i.e. vR.9.4.1 and *Bonito*) (**Figure 3.16**), and more recently at the beginning of 2021, 99% (Q20+).[236] These developments include the ability to sequence the complementary strand immediately after the template strand, thereby attaining a more accurate consensus read ($1D^2$) that increases the accuracy of template reads (1D) alone by 5%[235] (**Figure 3.13C**), though at the expense of throughput.[237] Of note, earlier releases of nanopore sequencing offered 2D sequencing which involved ligation of both strands with a hairpin adapter, though this has been largely replaced by $1D^2$ sequencing. However, the error rate still falls slightly short of the 99.9% achieved by PacBio CCS reads and short-read platforms. Notably, errors near the splice sites can result in spurious alignments and incorrect clustering of reads. Other approaches to mimic the PacBio circular consensus approach have been proposed (i.e. INC-seq[238] and R2C2[188]), with accuracy approaching 97.5%. However, such methods are laborious and not commonly used.

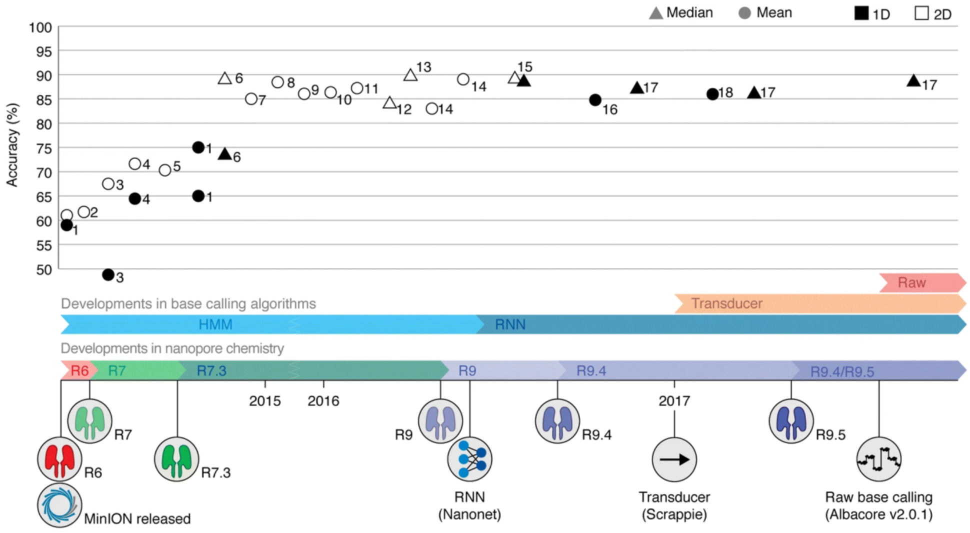


Figure 3.16: Advances in ONT nanopore sequencing read accuracy. Shown is a timeline of the improvement of read accuracy as a consequence of the development of ONT nanopore chemistry and basecalling algorithms. The R7 and R9 nanopore series are based on the MspA and CsgG protein, respectively. Of note, the figure does not include the latest chemistry (1D^2) or nanopore development (R.10.4, which involves a longer barrel and two pinch points to provide a better resolution of homo-polymer sequences). Figure is taken from Rang et al.(2018).[235]

In contrast to PacBio SMRT sequencing, real-time feedback and progress of the nanopore sequencing run are provided with information given on the run statistics (i.e. the total number of reads generated at any time) and the channel states over time. The channel state is an indication of the pore occupancy and is classified as being either: i) sequencing (active with current DNA translocation), ii) pore (active but without DNA translocation), iii) recovering, iv) inactive and v) unclassified (channels are divided into four groups and used sequentially to maximise throughput, and unclassified channels are those that not currently used). The duty time plot provides a good assessment of the current performance of the run, and an early indication whether to continue or stop the run (examples of successful and suboptimal runs are given in **Figure 3.17**).

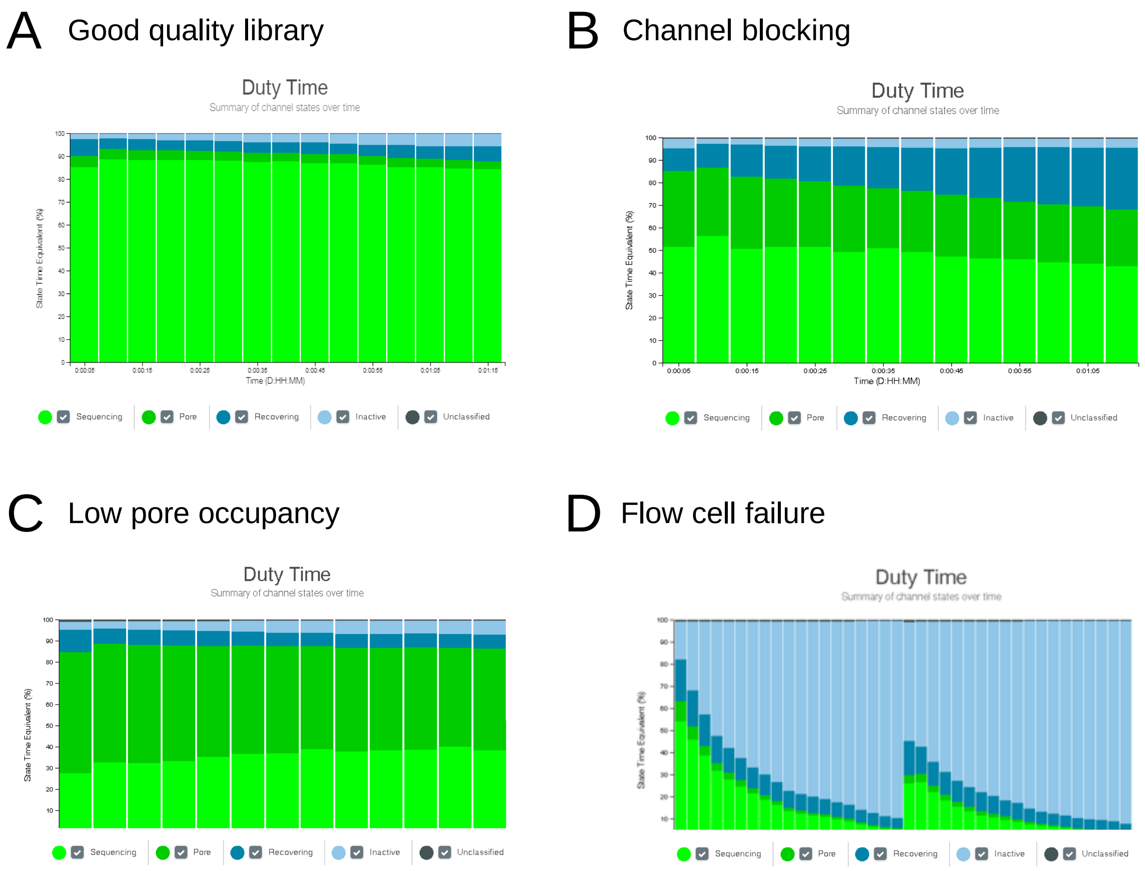


Figure 3.17: Examples of successful and suboptimal ONT nanopore sequencing runs. Shown are duty time plots from **(A)** a good quality run indicated by the majority of pores in the "sequencing" state (bright green), **(B)** a suboptimal run with channels being blocked as indicated by an accumulation of pores in the "recovering" state (dark blue), **(C)** a suboptimal run with low pore occupancy as indicated by the high ratio of "pore" (dark green) to "sequencing" state, and **(D)** a suboptimal run with flow cell failure indicated by the majority of pores in "inactive state" (light blue).

Channel blocking typically occurs when there are contaminants in the library. Conversely, low pore occupancy suggests insufficient loading material or poor library preparation (poor ligation reaction). Flow cell failure indicates damaged channels or membranes, which can be caused by multiple factors (air bubbles, osmotic imbalance, presence of detergents in library, among others). Channel states are classified as sequencing (bright green), pore (dark green), recovering (dark blue), inactive (light blue) and unclassified (grey). Figures are taken from the "Wellcome Trust Advanced Course: RNA Transcriptomics (2018)" that I attended during my PhD.

3.2.4 Bioinformatics pipeline

This section describes our developed bioinformatics pipeline for processing and analysing ONT cDNA sequencing data generated on the MinION following ONT library preparation (**Chapter 6**).

Unlike the Iso-Seq bioinformatics pipeline which was largely established by PacBio (described in **Section 3.1.4** and outlined in **Figure 3.8**), the bioinformatics pipeline for processing ONT raw reads was less defined and streamlined when I undertook this work. While significant improvements in bioinformatic tools have been released by ONT over recent years, many of the new tools were only applicable to sequencing data generated from ONT-specific protocols and primers. Given that the ONT dataset in this thesis was generated using the same primers and barcodes as the Iso-Seq dataset (illustrated in **Figure 3.14**), the initial stage of the bioinformatics pipeline was adapted from a protocol from the "Wellcome Trust Advanced Course: RNA Transcriptomics (2018)" (provided by J.Ragoussis and henceforth referred to as WTAC), which I attended during my PhD, and refined using ERCC control oligonucleotides as a benchmark. Many of the downstream tools initially developed for Iso-Seq were then similarly applied for the latter stages of the ONT bioinformatics pipeline, with the exception of the use of *TALON* in place of *Cupcake* for collapsing transcripts. Consequently, the bioinformatics pipeline for processing and analysing the ONT targeted cDNA sequencing data was broadly similar to the pipeline previously tailored for the Iso-Seq targeted dataset (refer to **Figure 3.18** for comparison).

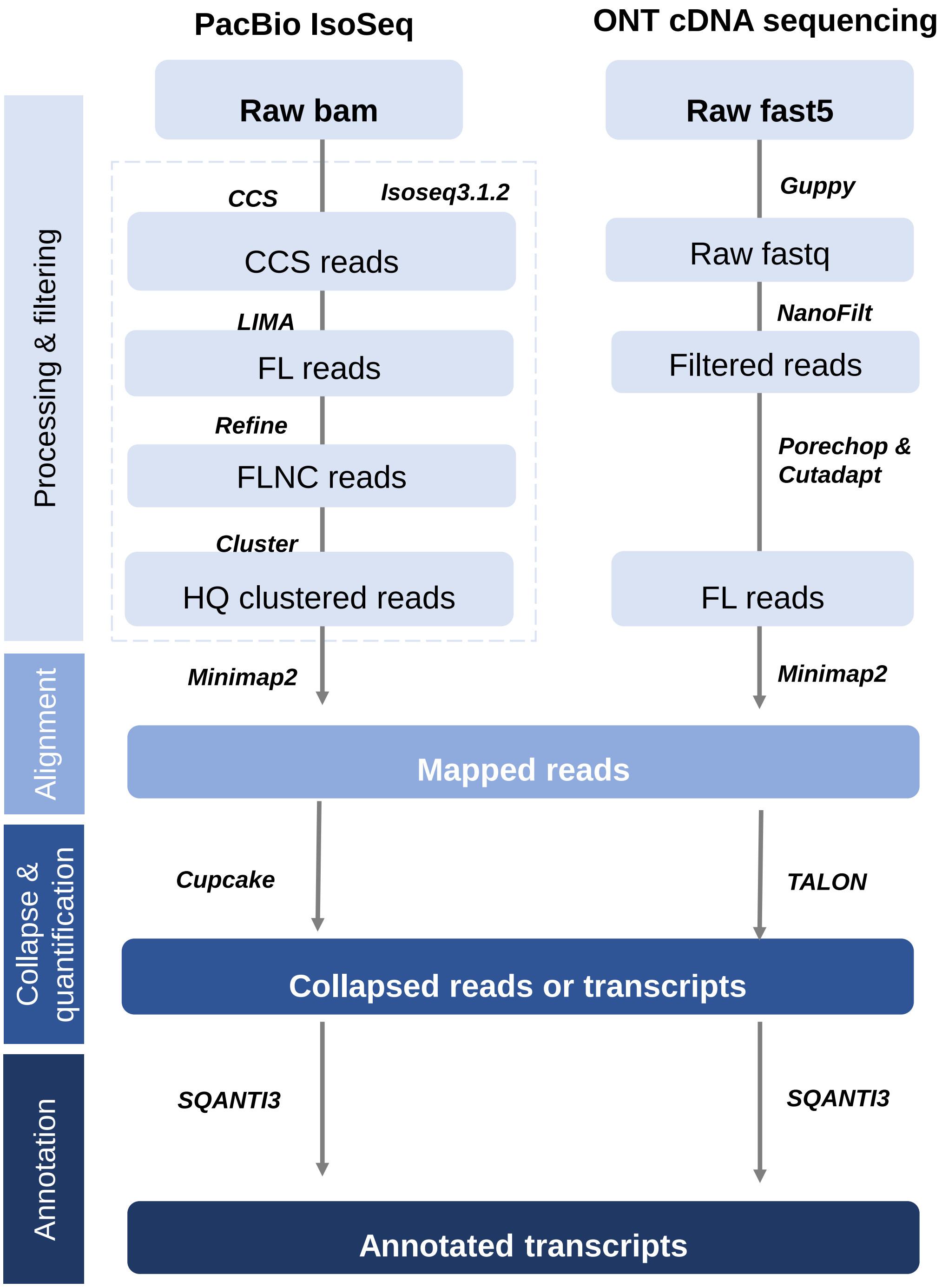


Figure 3.18: Comparison of the bioinformatics pipeline for processing PacBio Iso-Seq and ONT 1D-reads. Shown is a side-by-side comparison of the bioinformatics pipeline used to process PacBio Iso-Seq and ONT 1D-reads from initial processing of raw reads, alignment to the reference genome using *Minimap2*, to collapsing of reads to transcripts and the annotations of the long-read-derived transcriptome using *SQANTI*. The bioinformatic pipelines adopted are largely similar between Iso-Seq and ONT with the difference primarily in the initial processing of raw reads; raw Iso-Seq reads were processed using the PacBio bioinformatics suite (*Iso-Seq3*), whereas raw ONT reads were processed using various community-based packages.

3.2.4.1 QC of run performance, base-calling and filtering of basecalled reads

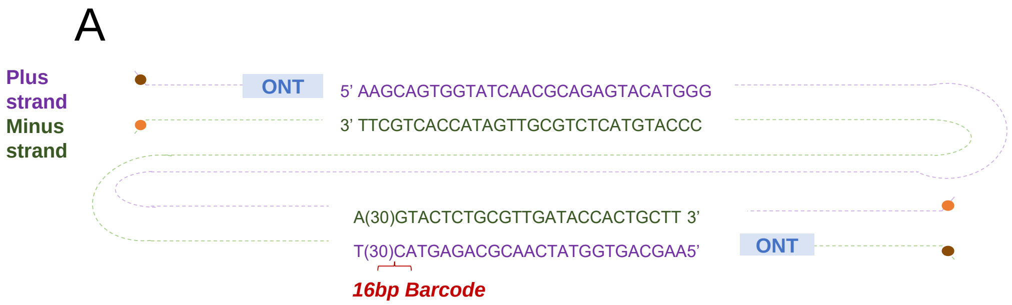
The performance of each nanopore sequencing run was assessed using *PycoQC*[239] and the official Nanopore QC tutorial,[240] by evaluating i) the number of active pores during the run, ii) the number of reads generated over time, and iii) the length and quality score distribution of basecalled reads. ONT raw reads were then basecalled using *Guppy*, the latest released ONT basecaller that converts the raw electrical signal to DNA sequence and is superior to other available basecallers with higher read accuracy and faster basecalling.[241] Basecalled reads with read quality score < 7 (recommended by ONT) were discarded using *Nanofilt*[242] (v2.3.0) with default parameters.

3.2.4.2 Removal of Nanopore and cDNA sequencing adapters

cDNA primer sequences and nanopore sequencing adaptors were removed to prevent spurious alignment using *Porechop*[243] (v0.2.4). Under recommended parameters (–end_size=100 –adapter_threshold=90 –end_threshold=75 –min_trim_size=15, –discard_middle –extra_end_trim=1), a window of 100 nucleotides from the end of each read was searched for a set of adaptors, which must have a minimum 90% identity to be considered present for trimming and a minimum 75% identity at the end of the reads; alignments smaller than 15bp or those found within the middle of the reads were considered chimeric and discarded.

Notably, *Porechop* has been unsupported since 2018 and has been largely replaced by the ONT official tool, *Pychopper*.[244] Despite being recommended for ONT-specific barcode demultiplexing, *Pychopper* failed to differentiate and orientate reads from the plus and minus strand without unique sequences, rendering all the ONT reads in the targeted dataset as being "unclassified". Given that the ONT cDNA reads were generated using the SMARTer cDNA synthesis kit (Clontech) (described in **Section 2.2.1**, depicted in **Figure 3.19A**), the 5' end of the plus and minus strands are reverse complements of each other with a few nucleotide differences (plus strand ends with ATGGG whereas the minus strand ends with poly(T), **Figure 3.19B**). Conversely, *Porechop* was able to differentiate the strand orientation with input of the unique set of adaptors that includes the cDNA primers, ONT adaptors and corresponding poly(A/T) tail (provided in **Table 3.2**). Sample demultiplexing was also performed by including the 16bp barcode sequence, and reads were assigned to the sample with the highest identity.

Trimmed reads with adaptors present at both ends were retained, and reads corresponding to the minus strand were reverse complemented. Using *Cutadapt*[245] (v2.9, -a "A60"), the poly(A) sequence was then trimmed 60 nucleotides from the 3' end.


A
Plus strand
Minus strand
ONT
5' AAGCAGTGGTATCAACGCAGAGTACATGGG
3' TTCGTCACCATAGTTGCGTCTCATGTACCC
A(30)GTACTCTGCGTTGATACCACTGCTT 3'
T(30)CATGAGACGCAACTATGGTGACGAA5'
ONT
16bp Barcode

B

Strand	Sequence
Plus start	AATGTACTTCGTTCAGTTACGTATTGCTAAGCAGTGGTATCAACGCAGAGTACATGGG
Plus end	AAAAAAAA<B>GTACTCTGCGTTGATACCACTGCTT
Minus start	AATGTACTTCGTTCAGTTACGTATTGCTAAGCAGTGGTATCAACGCAGAGTAC<B>TTTTTTTT
Minus end	CCCATGTACTCTGCGTTGATACCACTGCTT

Figure 3.19: Structure of the ONT cDNA template. Shown is the **(A)** final structure of the cDNA molecules for ONT sequencing, after cDNA synthesis and adaptor ligation, and the corresponding differentiating start and end sequence for the plus and minus strands, and **(B)** the adaptor sequences for comparison between the different read strands. The original cDNA molecules are outlined in purple and green, and the ONT boxes indicate the position of the ONT sequencing adaptors. The barcode location of sample demultiplexing is indicated in red (see **Table 2.2** and **Table 3.2** for barcode sequences). The coloured text in **Figure B** correspond to the coloured sequences in **Figure A**, with ONT adaptor sequences denoted in blue.

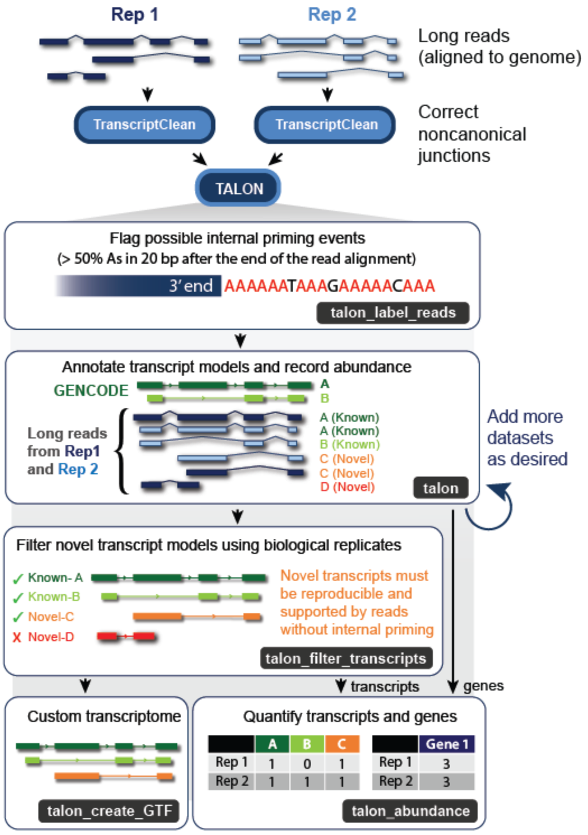
As illustrated, the barcode is only present in the 3' end of the plus strand and 5' end of the minus strand, as part of the oligo(dT) primer during cDNA synthesis (**Table 2.2**). The differing nucleotides between the plus start and the minus start is highlighted in yellow. The brown and orange circle refer to the motor protein and cholesterol moiety, respectively. The start and end of the strand is defined by the 5' and 3' end, respectively.

Table 3.2: ONT adapter sequences used to discriminate sample-specific plus and minus ONT reads. Tabulated are the sequences used in *Porechop* for sample demultiplexing and identification of the plus and minus strands. As depicted in **Figure 3.19**, only the plus strand end sequences and the minus strand start sequences contain the sample-specific barcode sequence (reverse complementary of one another). BC - Barcode.

Barcoded samples	Plus strand		Minus strand	
	Start sequence	End sequence	Start sequence	End sequence
BC1	TTGCTAAGCAGTGGTATCAACGCAGAGTACATGGG	AAAAAACGCACTCTGATATGTGGCA	CACATATCAGAGTGCGTTTTTT	CCCATGTACTCTGCGTTGATACCACTGCTTAGCAATACGTAACT
BC2		AAAAAAACTCACAGTCTGTGTGTGCA	ACACACAGACTGTGAGTTTTTTT	
BC3		AAAAAAACTCTCACGAGATGTGTGCA	ACACATCTCGTGAGAGTTTTTTT	
BC4		AAAAAAACGCGCGTGTGTGCGTGGCA	CACGCACACACGCGCGTTTTTTT	
BC5		AAAAAAAACGCGAGAGTCGAGTGGCA	CACTCGACTCTCGCGTTTTTTTT	
BC6		AAAAAAAACAGCTGATATATATGGCA	CATATATATCAGCTGTTTTTTTT	
BC7		AAAAAAACACATAGAGATACAGAGCA	TCTGTATCTCTATGTGTTTTTTT	
BC8		AAAAAAACGCAGCGCTCGACTGTGCA	ACAGTCGAGCGCTGCGTTTTTTT	
BC9		AAAAAAATCTGTCTCGCGTGTGTGCA	ACACACGCGAGACAGATTTTTTT	

3.2.4.3 Genome Alignment and Transcript Collapse

Trimmed reads from each sample were then aligned to the reference genome using *Minimap2*[213] (v2.17-r941, parameters: -ax splice) and were processed using *TALON*[246] (v5.0) for simultaneous transcript discovery and quantification (depicted in **Figure 3.20**). After trialling various bioinformatic tools, including *TAMA*[227] and *FLAIR*[166] (the results of these comparisons are documented in **Appendix D**), we found that *TALON* superseded the other tools for a number of reasons, namely it i) allows reference-based error-correction of ONT reads, which was essential for improving the confidence of splice junctions and recovering rare, novel transcripts, ii) performs quantification-led filtering of novel transcripts, retaining only transcripts that are reproducibly detected in biological replicates, iii) generates an abundance output file documenting the number of associated full-length read count for each transcript per sample, thereby facilitating downstream isoform-level analysis, and vi) requires less computing memory and time than other computational tools.



Rep 1
Rep 2
Long reads (aligned to genome)
TranscriptClean
TranscriptClean
Correct noncanonical junctions
TALON
Flag possible internal priming events
(> 50% As in 20 bp after the end of the read alignment)
3' end AAAAAATAAAGAAAAACAAA
talon_label_reads
Annotate transcript models and record abundance
GENCODE
A
B
Long reads from Rep1 and Rep 2
A (Known)
A (Known)
B (Known)
C (Novel)
C (Novel)
D (Novel)
talon
Add more datasets as desired
Filter novel transcript models using biological replicates
✓ Known- A
✓ Known-B
✓ Novel-C
X Novel-D
Novel transcripts must be reproducible and supported by reads without internal priming
talon_filter_transcripts
transcripts
genes
Custom transcriptome
talon_create_GTF
Quantify transcripts and genes
A B C
Rep 1 1 0 1
Rep 2 1 1 1
Gene 1
Rep 1 3
Rep 2 3
talon_abundance

Figure 3.20: Transcript discovery and quantification of ONT reads using *TALON*. Shown is a schematic figure of *TALON*, which was used for processing and analysing aligned ONT-derived transcripts. Figure is taken from Wyman et al. (2020).[246]

3.3 Differential expression and splicing analyses

This section describes the statistical expression and splicing analyses that were performed following the pre-processing and filtering of our long-read sequencing data in **Chapters 5 and 6**. The aim was to identify statistically significant differences in expression, splicing and usage of genes and transcripts between experimental groups. Parameters that are specific to individual results chapters can be found in the Method section of the relevant chapter.

3.3.1 Gene and isoform quantification

Any differential expression analysis first requires an estimation of the gene and/or transcript expression. In handling the short-read nature of RNA-Seq data, previous bioinformatic tools and computational models have determined this primarily from the number of reads that align to each transcript sequence from a reference genome annotation[224] (**Figure 3.21A**). While such approaches can accurately determine gene expression, it becomes much more challenging to estimate transcript expression due to the overlapping exonic structure of related transcripts, resulting in ambiguous read assignment (previously illustrated in **Figure 1.12**). Several sophisticated algorithms have been developed, including the Expectation-Maximization (EM) algorithm (adopted in *Kallisto*[222] and *RSEM*[247]), which assign reads to multiple genomic loci and work without a reference genome.[224]

The advent of long-read transcript sequencing data and the availability of matched short-read RNA-Seq data enabled gene and isoform expression to be estimated in two ways: i) a hybrid approach that involves mapping RNA-Seq data to Iso-Seq-derived or ONT-derived transcripts (**Figure 3.21B**), or ii) directly using the normalised full-length read count from long reads as a proxy of gene and transcript expression (**Figure 3.21C**); notably, the gene expression is estimated from the summation of full-length read counts from associated transcripts. While the former hybrid approach still suffers from ambiguous alignment to a degree, usage of the long-read-defined transcriptome in place of the reference genome would minimise misalignment, and improve mapping to condition-specific transcripts and other novel transcripts that are otherwise missing in the reference annotations.[145] Conversely, the latter approach does not rely on transcript assembly and is thus not impeded by misalignment. However, long-read sequencing data from global transcriptome profiling is still often considered semi-quantitative in most instances, due to the insufficient coverage required to detect expression changes.

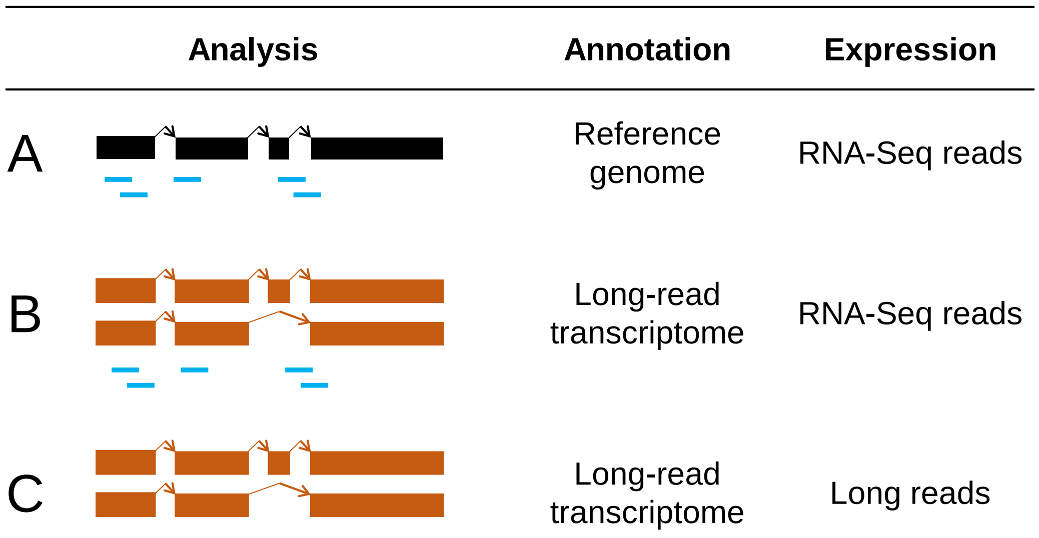


Figure 3.21: Strategies for isoform quantification. Shown is a schematic diagram of the three strategies adopted for determining isoform abundance: **(A)** RNA-Seq reads (blue lines) aligned to the reference genome (black boxes, approach adopted in previous transcriptome profiling studies), **(B)** a hybrid approach that involves aligning RNA-Seq reads (blue lines) to the long-read-defined transcriptome (orange boxes), or **(C)** directly using normalised full-length read counts from long reads. Long reads refer to both Iso-Seq and ONT reads.

3.3.2 Differential gene and transcript expression analyses

Differential gene expression (DGE) or transcript expression analysis (DTE) identifies genes or transcripts that have a statistically significant change in abundance across biological conditions (i.e. identifying features that are "differentially expressed") (**Figure 3.22A**). To facilitate unbiased comparisons across samples and experimental groups, raw read counts are normalised to eliminate feature-length and library-size effects - longer transcripts and samples sequenced at a higher depth would accumulate more reads - to a standard metric, namely TPM (Transcripts per million). Full-length reads from long-read sequencing are thus normalised to TPM using the following equation:

$$FL\ \ TPM(x_{sample}, y_{sample}) = \frac{Raw\ \ FL\ \ count(x_{isoform}, y_{sample})}{Total\ \ FL\ \ count(y_{sample})} * 10^6 \tag{3.2}$$

Between-sample normalisation methods, such as TMM[248] (Trimmed mean of M-values), are also used to account for differences in sample RNA library composition. This is particularly important when comparing samples between differential experimental groups with varying library composition.

While significant computational advances have been made in processing long-read data for transcriptome annotations, methods to harness such data for downstream differential expression analyses have been limited. Current differential expression analyses of long-read sequencing data typically rely on existing tools originally developed for short-read RNA-Seq,[249] such as *DESeq*, *maSigPro*, *edgeR*, among others. Highlighting the challenges of performing such analyses, various benchmarking studies have demonstrated that the choice of tool can affect the outcome considerably and no single method performs favourably across all datasets. Notably, tools based on negative binomial modelling performed better with higher specificity and sensitivity.[250] Recent methods, such as *FLAIR* and *LIQA*,[251] have emerged specifically for isoform expression analysis of long-read data. However, such methods have not been systematically assessed and are challenging to use for time-series data analyses; our targeted experiments in **Chapters 5 and 6** include data from two different conditions and across four time points.

3.3.3 Differential splicing analysis

A change in alternative splicing can be assessed in two ways: i) differential transcript expression (DTE), as described above, defined by a change in *absolute* expression of a transcript, and ii) differential transcript (or isoform) usage (DTU) defined by a change in the *relative* expression of a transcript, manifesting to a change in the proportions of the transcript (or isoform) of a gene (**Figure 3.22B**). As shown in **Figure 3.22**, DTU always implies DTE whereas the reverse is not necessarily true; e.g. a two-fold increase of two associated isoforms results in a change in the absolute but not the relative expression (**Figure 3.22A**), indicating a transcription-related mechanism. Conversely, any change in relative abundance of isoforms indicate a splicing-related mechanism. We observed examples of these mechanisms, notably in *Trem2* (later described in **Section 6.3.10.1**) and *Bin1* (later described in **Section 6.3.10.3**) from targeted profiling of the rTg4510 cortex.

One phenomenon characterised in differential splicing analysis is the significant altering of isoform proportions (also known as "Isoform Fraction"), resulting in the detection of a different dominant isoform. This phenomenon is known as "major isoform switching" (**Figure 3.22C**). In this circumstance, the same isoform is predominantly expressed in one condition (where it is the major isoform), but is lowly expressed in another (where it is the minor isoform). Notably, up-regulation of one isoform could be compensated by down-regulation of another, resulting in no net change at the gene-level (**Figure 3.22D**). Transcriptomic profiling studies at the gene-level would thus fail to capture such nuances, highlighting the complexity of gene regulation and the importance of performing differential expression analysis at a transcript-level.

Despite the limited utility of short-read RNA-Seq data for elucidating differential splicing events, a number of computational methods have been developed (reviewed in **Table 3.3**), and are based around two major strategies: i) isoform-based and ii) count-based methods, which are further subdivided into exon-based and event-based. Isoform-based methods aim to reconstruct the transcripts from sequencing reads and estimate the relative abundance in each sample, followed by statistical testing to identify transcripts with significant expression differences across experimental groups.[253] Conversely count-based methods dissect genes into counting units and document the number of reads falling within those units;[253] exon-based methods assign reads into exonic and junction regions, whereas event-based methods quantify transcripts by measuring the inclusion of individual splicing events with a percent splicing index (PSI) value for each event (i.e. proportion of associated isoforms that contain the splicing event of interest).

However, similar to transcript quantification, there is no clear consensus about the optimal tool or pipeline for such analysis. Benchmarking studies have similarly revealed that the choice of tools can directly impact the sensitivity and precision to detect differential transcripts, which are influenced by the number of replicates and the conditions heterogeneity.[254] Exon-based methods (i.e. *DEXSeq, edgeR, limma*) were found to overall perform better (superior precision and sensitivity) than other methods, with *edgeR* recommended for faster performance and reduced memory requirements.[253]

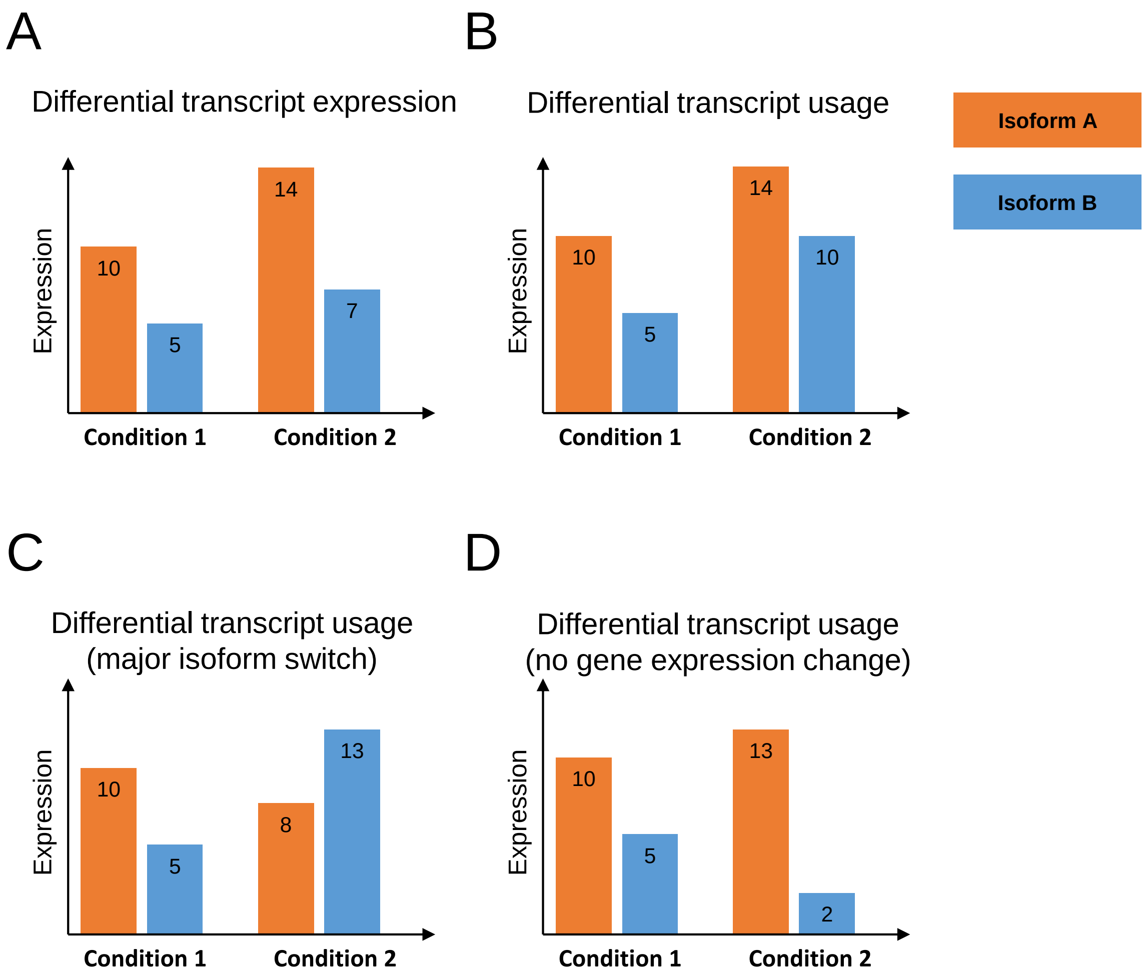


Figure 3.22: Scenarios of differential splicing. Shown is a schematic illustration of four different scenarios envisioned under differential splicing of a gene with two isoforms between conditions 1 and 2:

- **A)** Differential transcript expression (DTE) indicates an expression change for at least one transcript between conditions 1 and 2. However, the expression proportion of each transcript (defined as percentage of the total expression of all associated transcripts, in this case 50% for Isoform B) remains constant.
- **B)** Conversely in differential transcript usage (DTU), the relative expression of the isoforms is changed across conditions - in this case, Isoform B has a relative expression of 33.3% in condition 1 (5/15), but a relative expression of 41.6% in condition 2 (10/24).
- **C)** Differential transcript usage can occur with a switch of the major isoform - in this case, the more abundantly expressed isoform is switched from Isoform A in condition 1 to Isoform B in condition 2.
- **D)** Differential transcript usage can result in no overall change in gene expression if the change of transcript expression occurs in opposite directions.

Figures and legends were adapted from Soneson et al. (2016).[252]

Table 3.3: Bioinformatic approaches and tools to perform differential splicing analysis. Tabulated is an overview of some of the more commonly used bioinformatic approaches and tools for differential splicing analysis.
AS - Alternative splicing, AF - Alternative first exon, IR - Intron retention, MX - Mutually exclusive, ES - Exon skipiping. Table is adapted from Mehmood et al. (2020)[253] and is by no means comprehensive.

Approach	Method	Annotation	Designs	Model
Isoform-based	*Cufflinks* /*cuffdiff2*	Yes, *de novo*	2 groups	• Following transcript assembly, transcript abundance is estimated by maximising the likelihood score across all possible combinations of relative abundances of each associated isoform. • Variability between replicates and uncertainty in abundance are accounted with a beta negative binomial model.
	DiffSplice	*Ab initio*	2 groups	• Reconstructs a graph of the transcriptome based on reads, from which the abundance is estimated from alternative paths and alternative splicing modules. • Abundance of modules is compared using a non-parametric permutation test.
Exon-based	*DEXSeq*	Yes	Complex	• Applies a generalised linear model to exon-level expression data to model differential usage of exons across experimental groups, assuming that read counts follow a negative binomial distribution.
	edgeR	Yes	Complex	• Fits a negative binomial generalised log-linear model to exon-level expression data to test differential exon usage by comparing the log-fold-change of an exon to that of the gene.
	JunctionSeq	Yes, *de novo*	Complex	• Uses a similar statistical method as *DEXSeq* with added features to include novel exon junctions in differential exon usage analysis.
	limma	Yes	Complex	• Fits a linear model to exon-level expression data for differential exon usage between experimental groups
Event-based	*dSpliceType*	Yes	2 groups	• For each AS event type (ES, RI, MX, A3SS, A5SS), it calculates the read coverage signal for each base and the normalised logarithmic ratios of PSI between groups. Differential splicing events are then identified using a parametric test on the PSI.

MAJIQ	Yes, *de novo*	2 groups	• Uses local splicing variations, which denote splits in a splice graph mapping to the edges of a reference exon to calculate PSI. • Changes in PSI are then quantified using Bayesian modelling and bootstrapping.
rMATS	Yes	2 groups, paired samples	• Calculates PSI for each AS event after applying a hierarchical framework to account for within-sample uncertainty and between-sample variability. • Mean PSI across each AS event is then tested between experimental conditions using a likelihood ratio.
SUPPA2	Yes	2 groups, paired samples	• Determines transcript abundance using RSEM to estimate PSI for each AS event

3.3.4 TappAS: Integrated framework for differential expression and splicing analyses

After trialling various methods, we selected *tappAS*[255] (v1.0.0) as a framework for the differential expression and splicing analyses of long-read sequencing data across biological conditions (i.e. AD vs non-AD) in this thesis (**Chapters 5 and 6**). To date, it is the only tool that allows integration of isoform-level, long-read-derived annotations with public databases to comprehensively understand the functional implications of alternative splicing. Accessible as a user-friendly Java application, it provides the flexibility to incorporate expression derived from short-reads or long-reads, and supports complex design experiments: i) case-control, ii) time-course single series, and iii) time-course multiple series. Developed by the same authors as *SQANTI*,[215] it was recommended as an extension to the Iso-Seq bioinformatics pipeline for downstream isoform-level analyses.

The following sections detail specific analyses from *tappAS* in investigating differential expression and splicing changes associated with progressive tau pathology in rTg4510 mice at a global (**Chapter 5**) and targeted level (**Chapter 6**). All details are summarised from Lorena de la Fuente et al. (2020).[255]

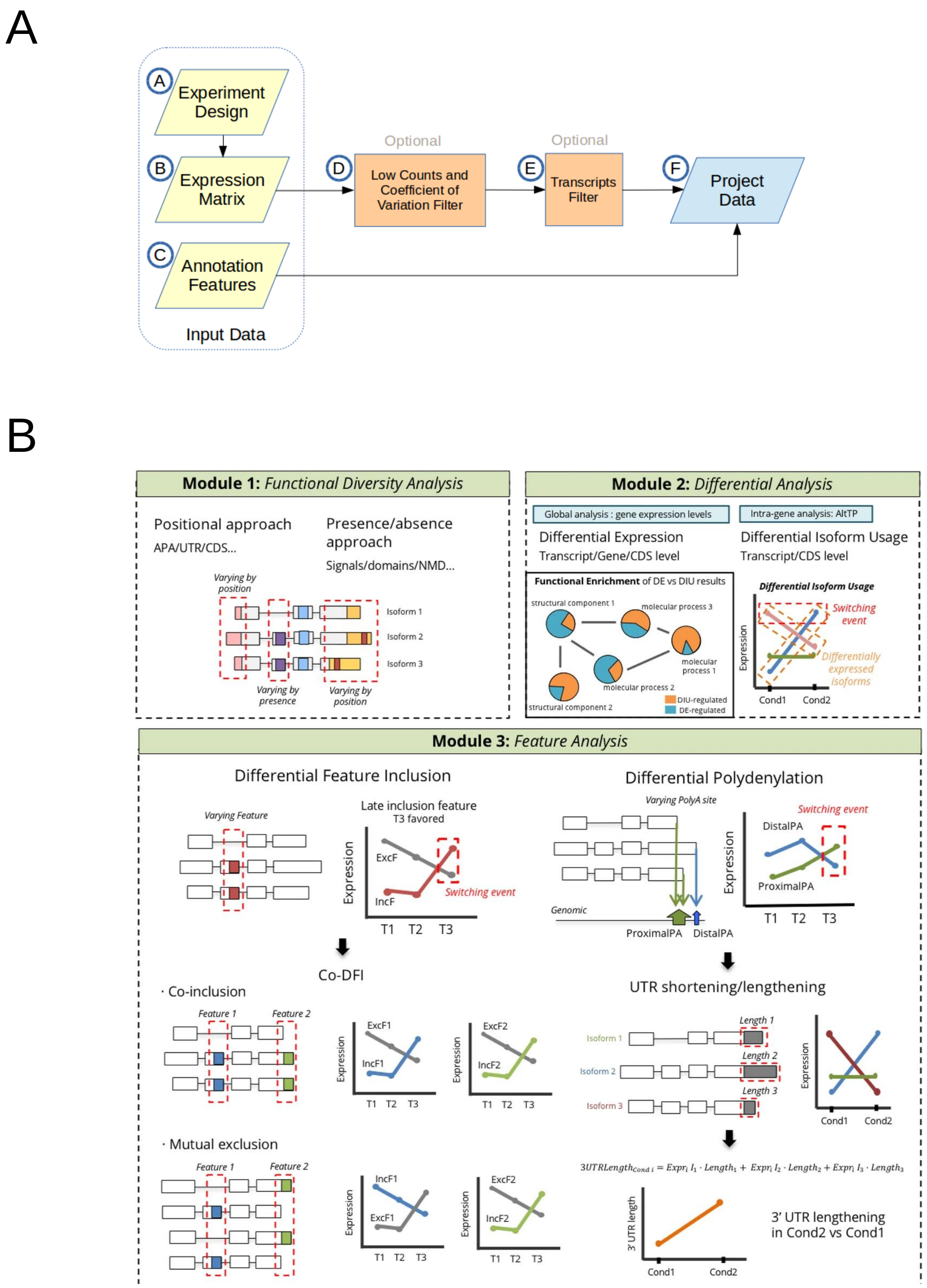


Figure 3.23: Differential expression and splicing analyses of long-reads using *tappAS*. Shown is **(A)** the *tappAS* project creation workflow, which requires three input files: a transcript-level expression matrix, an experimental design file and a transcript-level functional annotation file. The expression matrix can be obtained from mapping RNA-Seq data to a long-read-defined transcriptome (**Figure 3.21B**) or using full-length read counts directly from long-read sequencing (**Figure 3.21C**). **(B)** Overview of *tappAS* modules for functional isoform annotation and implications of alternative splicing. Figures and legends are adapted from Fuente et al. (2020).[255]

3.3.4.1 Functional annotations of long-read-derived isoforms

tappAS requires three inputs (**Figure 3.23A**):

1. An experimental design file to enable comparisons between two or more groups and/or over a time-course.
2. A transcript-level functional annotation file, which is generated post *SQANTI* using *IsoAnnot* (https://isoannot.tappas.org), as a “scaffold” for transcript-level annotations. For the purpose of this thesis, the annotation file was provided from a conglomerate, long-read-derived transcriptome of all the samples merged. The annotations incorporate feature elements from public annotations at both the transcript and protein level.
3. A transcript level expression matrix, which can either be derived directly from the full-length long-read transcript counts, or from mapping short-reads to the long-read-derived transcriptome using *Kallisto*.[222] Raw transcript counts were tabulated per sample.

3.3.4.2 Isoform pre-filtering and normalisation

Lowly-expressed transcripts with a sum of expression value less than 1 CPM (Counts per million) before normalisation or a large variance (> 100 coefficient of variation) across all the samples were removed to reduce noise. The raw transcript counts were then normalised using TMM normalisation[248] to account for differences in library size (sequencing depth) and sample RNA library composition. Of note, TMM assumes that the majority of the transcripts are not differentially expressed. Gene abundance was then deduced from the sum of normalised counts of associated isoforms, after removing transcripts with low or highly-varied expression values.

3.3.4.3 Differential gene and transcript expression analyses

TappAS uses *maSigPro* to perform differential gene and transcript expression analyses.[256] Briefly, it performs a two-step regression strategy to first define a negative binomial generalised linear model[256] for each gene or transcript, accounting for both condition and longitudinal effects, and identify differentially expressed genes. A stepwise regression is then applied to identify the conditions for which the differentially expressed genes have statistically significant profiles, determined by a user-defined threshold for the R^2 of the regression model; the R^2 defines the proportion of deviance that was explained by the linear regression model (“goodness of fit”), whereby a recommended threshold of 0.5 was used to identify differentially expressed genes with meaningful biological implications.[257] P-values were adjusted for multiple testing, by controlling the false discovery rate (FDR) with the Benjamin and Hochberg correction; an FDR-adjusted *P* value < 0.05 was considered as significant.

Following the identification of statistically significant gene models, the specific conditions (phenotype or time-associated changes) for which the genes show statistically significant profile changes (the significant variables) were identified by using an iterative backward stepwise approach.[258] As such, the procedure starts with *all* the variables (predictors, which are, in this case, the phenotype and age at different time points) imputed (hence "backwards") and determines the *P* value associated with each variable at each iteration as the variables are removed individually. This iteration continues until all the remaining variables are statistically significant (*P* < 0.05).

3.3.4.4 Differential transcript usage

In addition to absolute expression changes across conditions, the relative expression, and as such the usage, of isoforms can also change (described in **Section 3.3.3**). To identify genes that exhibit differential transcript usage, *tappAS* calculates the proportion (or fraction) of the associated isoform using the following equation:

$$IF_{cig} = \frac{\bar{E}_{cig}}{\sum_{i=1}^{n} \bar{E}_{cig}} \tag{3.3}$$

where:

$\bar{E}_{\text{cig}}$ = mean normalised expression for isoform *i* associated to gene *g* under condition *c*.
n = total number of isoforms associated with gene *g*.

TappAS then implements *maSigPro* to perform differential transcript usage in a similar manner to differential expression analyses (as described in **Section 3.3.4.3**) by fitting a generalised linear model and testing the significance of each variable.

Notably, *TappAS* recommends filtering of minor isoforms before performing differential transcript usage. While there is abundant evidence of widespread isoform diversity,[93] most protein-coding genes have been reported to typically express a few dominant isoforms,[259,260] with other isoforms being very lowly expressed and unlikely to be main contributors to the proteome.[259] As such, filtering of these minor isoforms can reduce the number of “false positives" from genes that are associated with differential transcript usage due to the "flat" behaviour of these minor isoforms.[255]

tappAS provides two strategies to filter lowly-expressed isoforms, by: i) proportion, whereby an isoform is only retained if its proportion relative to other isoforms is greater than the pre-specified threshold (default: proportion $>$ 10%) in at least one sample, or ii) fold-change, if its proportion relative to the major isoform is below a pre-specified threshold (default: fold-change = 2). A major isoform is defined as the isoform with the highest expression across all the conditions, with the remaining isoforms denoted as minor. Notably, the former strategy is more sensitive and reliant on the power to detect all isoforms at a deep coverage, whereby the latter strategy is dependent on the expression of only the major isoform. Consequently, we selected the latter strategy of using fold-change to exclude lowly-expressed isoforms, given long-read sequencing is considered semi-quantitative. Finally, we also removed lowly-expressed genes for this analysis given the reduced confidence in measuring their isoform usage.

Chapter 4

Global characterisation of isoform diversity in the mouse cortex

This chapter is an abridged and modified version of a peer-reviewed manuscript, which I co-authored and has been accepted for publication in *Cell Reports* (Leung et al. 2021)[261] (the complete published paper is presented in **Appendix F**). Additional lab QC output and run reports are included.

4.1 Introduction

Characterisation of the full complement of isoforms across tissues and development is important for understanding the role of transcriptional variation in health and disease. Previous transcriptomic profiling studies of AD post-mortem brain tissue and AD mouse models have revealed significant variation in transcript expression and splicing (reviewed in **Table 1.2** and **Table 1.3**, respectively), implicating the role for transcriptomic dysregulation and aberrant splicing in AD pathogenesis.[92] Despite significant advances in sequencing technology, the accurate detection of alternative splicing events and isoform diversity remains a challenge due to technical limitations with standard RNA-Seq approaches (explained in **Section 1.3.1**). Recent advances in long-read cDNA sequencing address these limitations - Pacific Biosciences (PacBio) Isoform Sequencing (Iso-Seq) and Oxford Nanopore Technologies (ONT) nanopore sequencing (described in **Section 3.1** and **Section 3.2**, respectively) - enabling direct assessment of alternatively-spliced transcripts.[249]

Given the importance to comprehensively characterise the full complement of isoforms, this chapter aimed to characterise the global isoform diversity and splicing events in the mouse cortex. The objectives of this chapter were as follows:

1. To perform PacBio Iso-Seq profiling (as described in **Section 3.1**) of the mouse cortex and generate full-length cDNA sequences.
2. To comprehensively annotate the mouse transcriptome and identify novel transcripts, novel genes and fusion genes.
3. To compare the performance of long-read Iso-Seq and short-read RNA-Seq for transcriptome annotation.
4. To comprehensively characterise splicing events in the mouse cortex.

4.2 Methods

4.2.1 Samples

Entorhinal cortex tissue was dissected from 6 female rTg4510 transgenic mice and 6 wild-type mice, aged 2 and 8 months (n = 3 mice per group) (**Table 4.1**). Additional details on mouse breeding conditions and animal procedures can be found in **Section 2.1.2**. For each mouse sample, RNA was isolated using the AllPrep DNA/RNA Mini Kit (Qiagen) from ~5mg tissue and quantified using the Bioanalyzer 2100 (Agilent) (described in **Section 2.1.4**).

4.2.2 Iso-Seq library preparation and SMRT sequencing

RNA from each mouse sample was prepared for Iso-Seq library preparation and SMRT sequencing following the Iso-Seq lab workflow, as detailed and described in **Section 3.1.2**. Briefly, first-strand cDNA synthesis was performed on 200ng RNA using the SMARTer PCR cDNA Synthesis Kit (Clontech) (described in **Section 2.2.1**), with the addition of External RNA Controls Consortium (ERCC) standards to the majority of mouse cortex samples (n = 10) (described in **Section 2.3**), followed by PCR amplification of 14 cycles with PrimeSTAR GXL DNA Polymerase (Clontech) (described in **Section 2.2.2**); an example of an agarose gel image taken after PCR amplification is provided in **Figure 4.1**. The resulting amplicons were then divided into two fractions and purified using 0.4X and 1X AMPure PB beads (as shown in **Figure 4.2**). The two fractions were then recombined at equimolar quantities and library preparation was performed using the SMRTbell Template Prep Kit v1.0 (shown in **Figure 4.2**). Each sample was then sequenced on the PacBio Sequel 1M SMRT cell with the v3 chemistry (Diffusion Loading at 5pM with a 4-hour pre-extension and a 20-hour capture time) (described in **Section 3.1.2.5**).

Table 4.1: Phenotype information of the mouse samples used for global transcriptome profiling. Tabulated is a summary of the phenotype information of the rTg4510 mouse samples sequenced using Iso-Seq and RNA-Seq.
ECX - Entorhinal cortex, ONT - Oxford Nanopore Technologies, RIN - RNA integrity number, TG - Transgenic rTg4510 mice, WT - Wild-type.

Sample ID	Tissue	Sex	Genotype	Age (months)	RIN	Iso-Seq	RNA-Seq
Mouse 1	ECX	F	WT	2	9.2	✓	✓
Mouse 2	ECX	F	TG	2	8.8	✓	✓
Mouse 3	ECX	F	WT	8	9.1	✓	✓
Mouse 4	ECX	F	TG	8	9.2	✓	✓
Mouse 5	ECX	F	TG	8	8.7	✓	✓
Mouse 6	ECX	F	WT	2	9.2	✓	✓
Mouse 7	ECX	F	TG	2	8.9	✓	✓
Mouse 8	ECX	F	WT	8	9	✓	✓
Mouse 9	ECX	F	TG	8	8.6	✓	✓
Mouse 10	ECX	F	WT	2	9.2	✓	✓
Mouse 11	ECX	F	TG	2	8.9	✓	✓
Mouse 12	ECX	F	WT	8	9.1	✓	✓

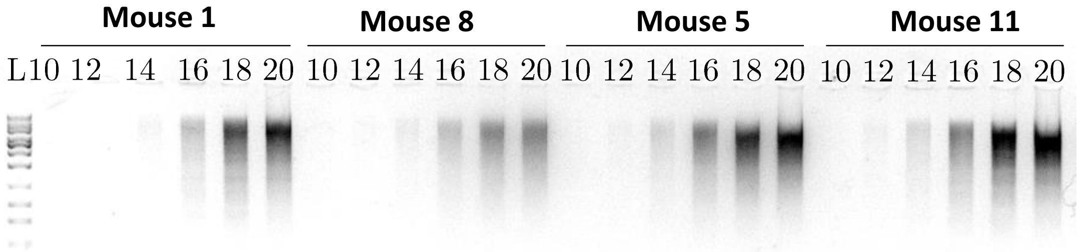

Mouse 1
Mouse 8
Mouse 5
Mouse 11
L10 12 14 16 18 20 10 12 14 16 18 20 10 12 14 16 18 20 10 12 14 16 18 20

Figure 4.1: Samples were amplified using 14 PCR cycles. Shown is an example of an agarose gel image from PCR cycle optimisation of four mouse samples after cDNA synthesis. PCR aliquots were collected every two cycles (10, 12, 14, 16, 18, 20) and then assayed using agarose gel electrophoresis. 14 cycles were determined to be optimum for large-scale amplification, as cycles below showed insufficient amplification whereas cycles above showed signs of over-amplification, which could result in a biased sequencing representation. Ladder (L) denotes to a 1kb DNA ladder.

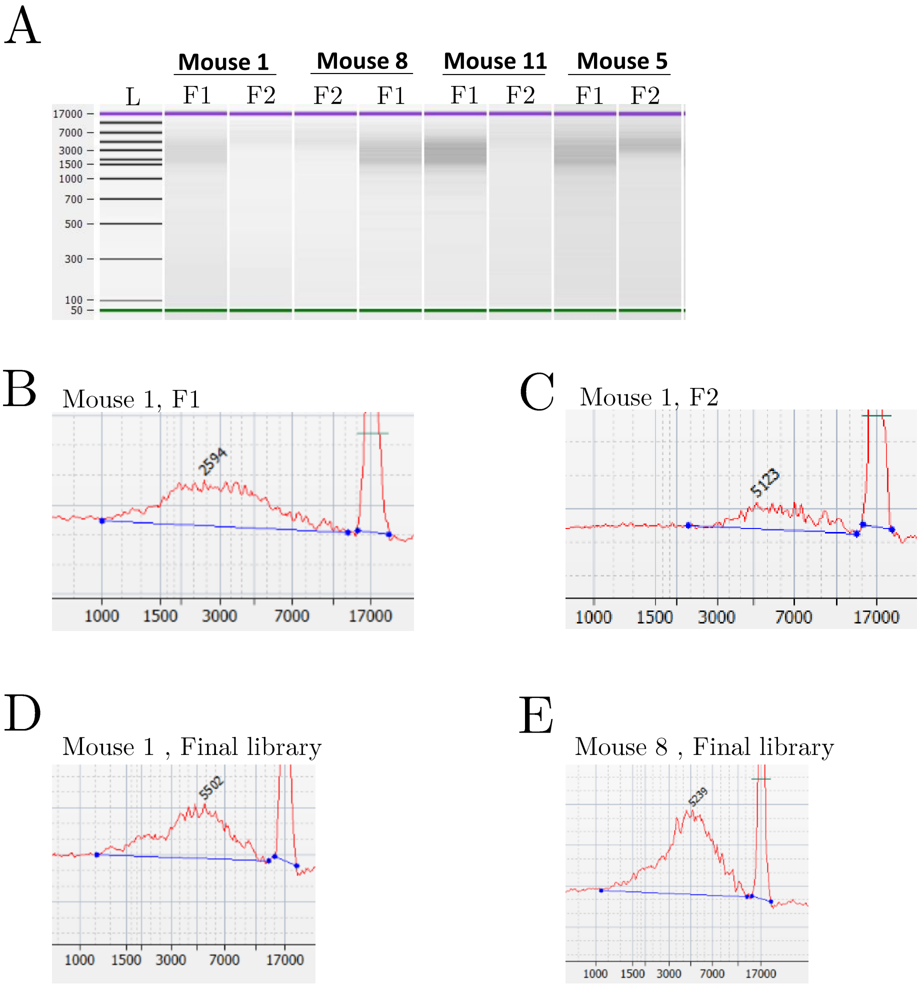


Figure 4.2: Library preparation was performed for each sample with successful cDNA purification and ligation with SMRT bell templates. Following large-scale amplification using the optimal cycle number (as determined from **Figure 4.1**), the resulting cDNA was divided into two fractions (denoted here as F1 and F2) and purified using 1X (F1) and 0.4X (F2) AMPure beads. Shown is **(A)** a Bioanalyzer gel of the purified cDNA from the two fractions, and zoomed-in Bioanalyzer electropherograms of **(B)** Mouse 1 Fraction 1, **(C)** Mouse 1 Fraction 2, **(D)** Mouse 1 and **(E)** Mouse 8 after library preparation. The x-axis of the Bioanalyzer electropherogram represents the molecular size. Size distribution for each fraction was determined from the start to the end point of the smear.

Of note, cDNA in Fraction 2 has a significantly higher molecular weight across all the samples as expected (shown in Figures A and C). Pooling of both fractions enriched for higher molecular weight cDNA molecules, which were intact after performing SMRTbell template preparation (as seen in Figures D and E). Despite the fact that samples were prepared sequentially, Bioanalyzer electropherogram profiles were fairly consistent across samples.

4.2.3 RNA-Seq library Preparation and Illumina sequencing

RNA from the same mouse samples (n = 12) was processed in parallel using the TruSeq Stranded mRNA Sample Prep Kit (Illumina) and subjected to 125bp paired-end sequencing using the HiSeq2500 (Illumina).[90] Briefly, cDNA libraries were prepared from ~450ng of total RNA plus ERCC spike-in synthetic RNA controls (Ambion, dilution 1:100), purified using AMPure XP magnetic beads (Beckman Coulter) and profiled using the D1000 ScreenTape System (Agilent).

4.2.4 SMRT sequencing QC and Iso-Seq data processing

Raw Iso-Seq reads were evaluated using the SMRT Link Portal v7.0 and analysed using the optimised Iso-Seq bioinformatics pipeline, as depicted in **Figure 4.3**. Further details are provided in **Section 3.1.4**. Briefly, CCS reads were generated from a minimum of 1 pass using *Iso-Seq3 CCS* (v3.4.1). Primers and SMRT adapters were then removed using *Lima* (v1.9) to generate full-length reads, followed by removal of artificial concatemers reads and trimming of poly(A) tails in *Iso-Seq3 Refine*. Full-length, non-chimeric (FLNC) reads were then collapsed to high-quality transcripts using *Iso-Seq3 Cluster*, and mapped to the mouse reference genome (mm10) using *Minimap2* (v2.17) with the following parameters "-ax splice -uf –secondary=no -C5 -O6,24 -B4". *Cupcake collapse-isoforms-by-sam.py* script was subsequently applied with the following parameters "-c 0.85 -i 0.95 –dun-merge-5-shorter" to reduce redundancy.

4.2.5 Transcriptome annotation and filtering

After filtering for partial isoforms (such as 5' degradation products) using *TAMA* with default parameters, isoforms detected using SMRT sequencing were characterized and classified using *SQANTI2* (v7.4) in combination with mouse reference gene annotations (mm10, GENCODE, vM22) , FANTOM5 CAGE peaks, poly(A) motifs, STAR output junction file, full-length read counts (abundance file), and *Kallisto*-derived counts from RNA-Seq data (described below in **Section 4.2.6**). Additional details are provided in **Section 3.1.4.4**. Potential artefacts such as reverse transcription jumps or intra-priming of intronic lariats were filtered out using the *SQANTI2* filter script with an intra-priming rate of 0.6 (the fraction of genomic A's above which the isoform will be filtered, as detailed in **Section 3.1.4.4**). The occurrence of mutually exclusive exons (MX) and exon skipping (ES) were assessed using *SUPPA2*[262] with the parameter "–f ioe", intron retention (IR) using *SQANTI2*, and alternative first exons (AF), alternative last exons (AL), alternative 5' splice sites (A5), and alternative 3' splice sites (A3) using custom scripts based on splice junction coordinates.

4.2.6 RNA-Seq QC and data processing

Raw RNA-Seq reads were filtered (removal of ribosomal sequences, quality threshold of Q20, minimum sequence length of 35bp) and trimmed using *fastqmcf* (v1.0), yielding a mean trimmed read depth of ~20 million reads per sample. Reads were then mapped to the mouse reference genome (mm10) using *STAR*[221] (v1.9). Gene and transcript expression were determined by aligning merged RNA-Seq reads to the Iso-Seq derived annotations (*Cupcake* collapsed) using *Kallisto*[222] (v0.46.0) with default parameters. An RNA-Seq derived annotation was also generated from RNA-Seq reads and the mouse reference annotation (mm10, GENCODE, vM22) using *Stringtie*[263] (v2.1.4), which was subsequently annotated and filtered using *SQANTI2* under default parameters.

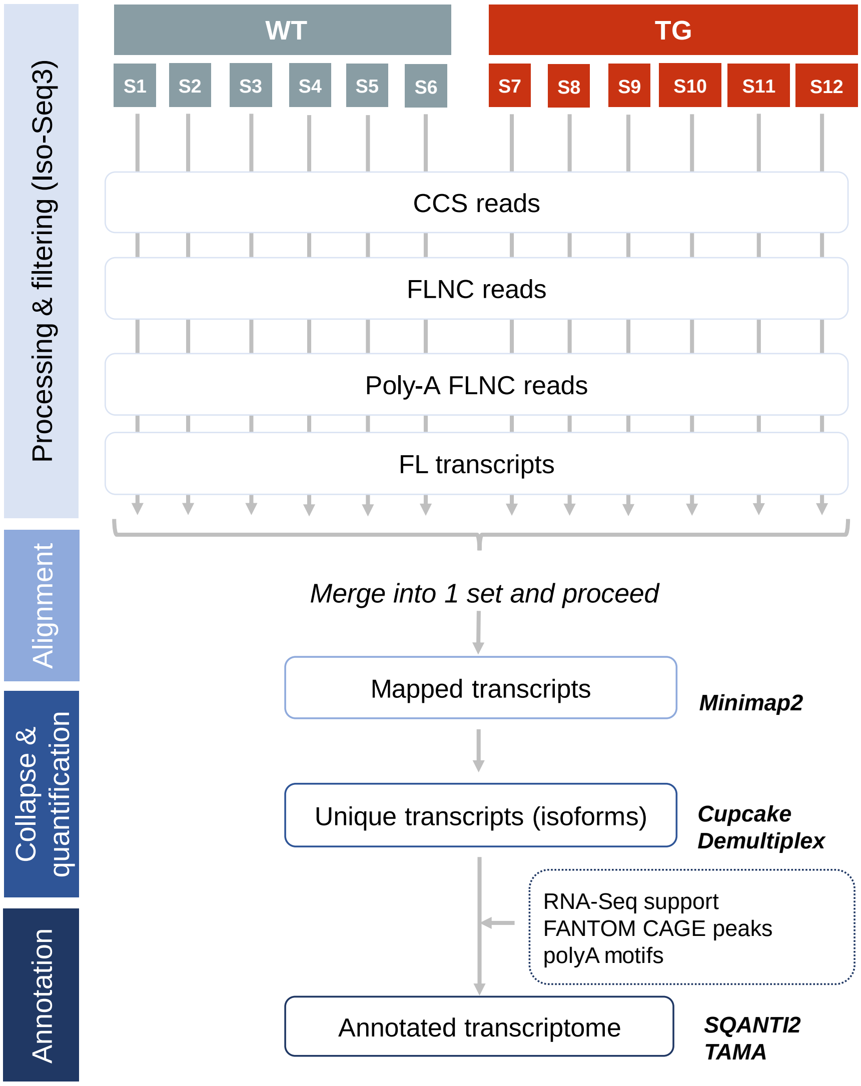


Figure 4.3: Iso-Seq reads from each mouse sample were processed individually before merging into one unified dataset. Shown is an overview of the bioinformatics pipeline used to generate full-length transcript annotations of the mouse entorhinal cortex (n = 12, WT = 6, TG = 6). Briefly, polymerase reads generated from PacBio Sequel I for each sample were processed using *Iso-Seq3* (v3.1.2) and *Cupcake* to generate high quality, full-length transcripts, which were then mapped to the mouse reference genome using *Minimap2*. Isoforms were then collapsed and merged to generate one complete dataset, which was annotated using *SQANTI2*. Additional details can be found in **Section 3.1.4**. CCS - Circular consensus sequence, FLNC - Full-length non-chimeric, FL - Full-length.

4.3 Results

4.3.1 PacBio Iso-Seq run performance and sequencing metrics

Following library preparation and SMRT sequencing, we generated a total of 371Gb (s.d = 4.35Gb, range = 22.5Gb - 38.7Gb) and 8,082,647 polymerase reads (s.d = 63,013 reads, range = 530,974 - 733,495 reads) (**Table 4.2**). Following the Iso-Seq bioinformatics pipeline, raw reads were processed and clustered to unique consensus transcripts, which were then mapped to the genome. A total of 5.66 million CCS reads (sample mean = 471K, s.d = 46.8K, range = 353K - 512K) and 4.5 million FLNC reads were successfully generated (sample mean = 379K, s.d = 47.0K, range = 270K - 412K). Clustering of these reads yielded a total of ~273K high-quality full-length transcripts (97% of all FL transcripts, mean = 32.7K, s.d = 1.25K, range = 30.3K - 34.4K), which were mapped to 278K loci of the mouse reference genome (5K had multi-mapping). After filtering for alignment length and identity (described in **Section 4.2.4**), 266K transcripts were retained. Rarefaction curves confirmed that the dataset approached saturation, indicating that our coverage of the isoform diversity was representative of the true population of transcripts (**Figure 4.4A**).

4.3.2 Widespread isoform diversity in the mouse cortex

Following stringent quality-control, Iso-Seq reads mapped to 14,482 known genes with expression patterns reflecting those expected for the cortex; using the Mouse Gene Atlas database, the 500 most abundantly-expressed genes were most significantly enriched for the "cerebral cortex" (odds ratio = 6.07, adjusted P = 6.8 x 10^{-17}). We identified 46,626 isoforms (mean length = 3.18kb, s.d = 1.68kb, range = 0.083 − 15.9kb), which were enriched near Cap Analysis Gene Expression (CAGE) peaks from the FANTOM5 dataset (median distance from CAGE peak = -1bp, 35,262 (75.6%) transcripts located within 50bp of a CAGE peak), and were also located proximal to annotated transcription start sites and transcription termination sites. A significant proportion of isoforms (n = 20,621, 45%) were sized 2 - 4kb in length (median length = 2.96kb, mean length = 3.18kb, s.d = 1.68kb, range = 0.083kb - 15.9kb) (**Figure 4.4B**), corresponding to the mean length of mRNA in the mouse reference genome, with a wide range in the number of exons (range = 1 - 89) observed per isoform (mean number of exons = 10.8). A wide range in the number of multi-exonic RNA isoforms was also identified per gene (range = 1 - 86), and longer genes with more exons were typically annotated with more isoforms (Pearson's correlation between isoform number and gene length: corr = 0.25,

P = 1.33 x 10^{-197}; Pearson's correlation between isoform number and exon number: corr = 0.25, P = 4.02 x 10^{-193}). Notably, only 10% of isoforms (n = 4,641) were detected across all the samples (**Figure 4.5A**), with about half (47.8%) detected in 2 - 3 samples with very low transcript expression (**Figure 4.5B**).

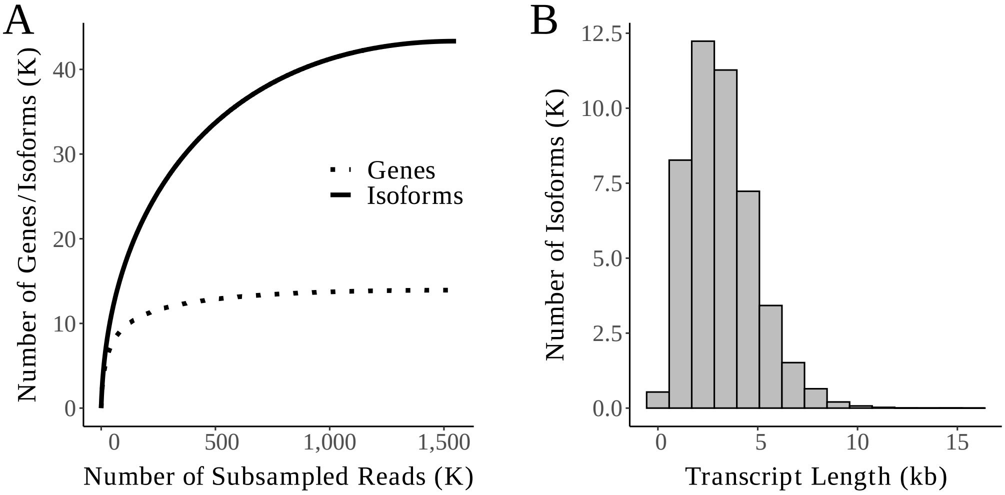


Figure 4.4: Saturation was reached at the gene and isoform level with the majority of transcripts sized ~3kb. Shown is **(A)** a rarefaction curve of the number of subsampled reads against the number of unique genes and isoforms detected, and **(B)** a distribution of the transcript length from merging all the Iso-Seq datasets. K - Thousand.

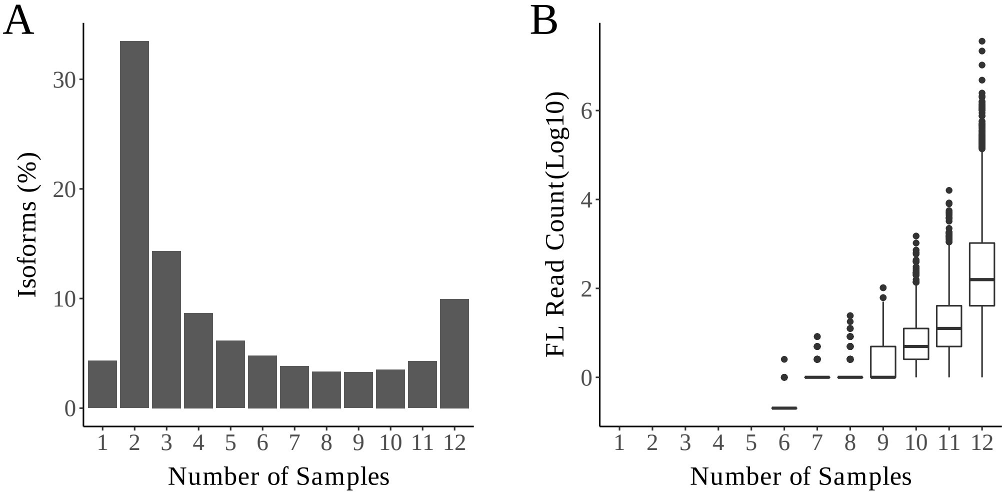


Figure 4.5: Highly-expressed isoforms were more likely to be sequenced across samples. Shown is **(A)** the distribution of isoforms detected in the number of mouse samples, with a third detected in any two of the total 12 samples. However, **(B)** quantification of these isoforms had very low expression (1 - 2 FL reads), whereas those that were commonly detected across all 12 samples were more abundant. FL - Full-length.

Table 4.2: Iso-Seq sequencing yield from global transcriptome profiling. Tabulated is a summary of the Iso-Seq sequencing metrics from global transcriptome profiling of the rTg4510 mouse cortex. Sequencing runs appeared optimal with > 20Gb achieved per sample, expected subread lengths, good productivity ratios and normal control metrics. Further details on evaluation of the performance of Iso-Seq sequencing runs are provided in **Section 3.1.3**. K - Thousand, Pol - Polymerase. N50 is defined as the sequence length of the shortest read at 50% of all reads.

Sample ID	Total bases (GB)	Pol reads (K)	Read length (kb)						Productivity			Control				Local base rate	Template	
			Polymerase		Subread		Insert		P0	P1	P2	Total reads	Length (kb)	Concordance			Adapter dimer	Short insert
			Mean	N50	Mean	N50	Mean	N50						Mean	Mode			
Mouse 1	29.56	674	43.86	90.56	1.25	2.02	3.34	4.75	10.55% (107K)	67.42% (682K)	22.73% (230K)	7036	34.7	0.85	0.89	2.72	0.08	0.06
Mouse 2	31.1	566	54.89	101.22	1.26	1.78	2.86	3.66	29.77% (300K)	57.25% (577K)	14.05% (142K)	10707	44.6	0.87	0.89	3.05	0	0
Mouse 3	34.60	698	49.56	98.80	1.70	2.67	3.78	4.78	16.1% (164K)	69.2% (704K)	14.7% (150K)	5951	40.5	0.85	0.89	2.78	0	0
Mouse 4	34.61	711	48.68	97.02	1.71	2.49	3.83	5.02	14.22% (145K)	70.49% (718K)	15.28% (156K)	6762	38.4	0.85	0.87	2.671	0.01	0.01
Mouse 5	38.74	675	57.37	112.63	1.87	2.87	3.90	4.79	17.41% (176K)	68.08% (686K)	15.575 (157K)	10647	44.2	0.86	0.89	2.96	0.01	0
Mouse 6	30.45	661	46.08	91.63	2.23	2.75	3.95	4.73	16.6% (169K)	65.9% (671K)	17.5% (179K)	10301	38.7	0.85	0.87	2.79	0.01	0.01
Mouse 7	22.53	531	42.42	85.33	2.61	3.15	3.44	4.08	41.8% (426K)	52.6% (536K)	5.5% (56.4K)	5415	49.8	0.86	0.85	2.05	0	0
Mouse 8	31.25	731	42.77	89.37	1.49	2.35	3.61	4.88	9.37% (94.5K)	73.33% (740K)	18.19% (184K)	8908	35.0	0.85	0.89	2.56	0.06	0.04
Mouse 9	33.16	715	46.36	92.52	2.00	2.93	3.98	4.95	11.51% (117K)	70.91% (722K)	17.58% (18.0K)	6855	38.0	0.85	0.87	2.6	0.01	0.01
Mouse 10	24.52	733	33.43	70.75	2.56	3.29	3.71	4.75	15.9% (162K)	72.1% (735K)	12.0% (122K)	1668	44.2	0.85	0.85	1.99	0.00	0.01
Mouse 11	30.41	683	44.55	90.04	1.44	2.04	3.28	4.40	11.98% (121K)	68.45% (692K)	20.35% (206K)	7881	36.5	0.86	0.89	2.85	0.11	0.07
Mouse 12	30.28	704	42.99	89.16	1.35	2.02	3.27	4.38	7.02% (71.1K)	70.18% (710K)	23.39% (237K)	6019	35.2	0.85	0.89	2.57	0.01	0.01

4.3.3 Detection of many novel isoforms with novel splice junctions

Among the isoforms annotated to known genes, 50% (n = 23,096) were novel and not present in existing reference annotations (**Table 4.3**). Compared to known isoforms, these novel isoforms were less abundant (Mann-Whitney-Wilcoxon test: W = 3.66 x 10^8, P < 2.23 x 10^{-308}, **Figure 4.6A,B**), longer (W = 2.37 x 10^8, P = 2.13 x 10^{-42}, **Figure 4.6C,D**) and had more exons (W = 1.94 x 10^8, P < 2.23 x 10^{-308}, **Figure 4.6E,F**), suggesting that they would have been harder to detect using RNA-Seq due to the difficulty in assembling transcripts with limited read coverage. These novel isoforms were also more likely to be associated with novel transcription start sites (TSS) with more novel isoform TSSs detected > 1kb of an annotated TSS (n = 1,454 novel isoform TSSs, n = 1,154 known isoform TSSs, Fisher's Test: P = 6.16 x 10^{-12}, odds ratio = 1.32). Similarly, novel isoforms were more likely to be associated with novel termination sites (TTS) with more novel isoform TTSs detected within 1kb of an annotated TTS (n = 21,506 novel isoform TTSs, n = 21,434 known isoform TTSs). Assessing the reliability of novel isoforms against known isoforms, there was no difference in the number of isoforms supported within 50bp of a CAGE peak (n = 17,252 novel isoforms, 75.4%; n = 17,842 known isoforms, 75.8%; Fisher's Test: P = 0.31, odds ratio = 0.978). While novel isoforms had a lower RNA-Seq expression (mean RNA-Seq expression: novel isoforms = 1.99 TPM, known isoforms = 8.95 TPM, two-tailed unpaired t-test: t(46401) = 14.8, P = 1.37 x 10^{-49}), this is a likely reflection of the relatively lower expression of novel isoforms and RNA-Seq's lack of power to detect them.

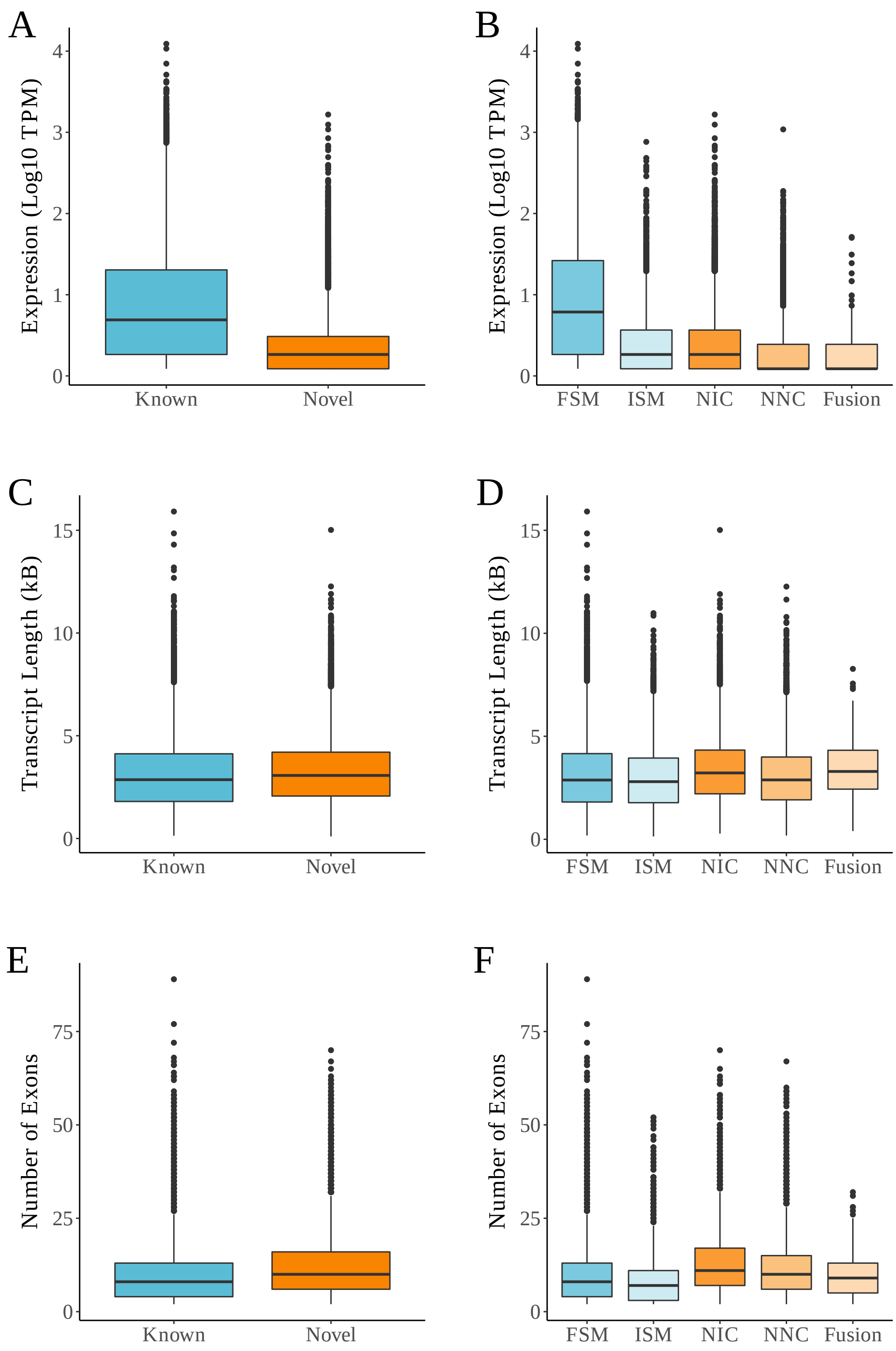


Figure 4.6: Novel isoforms were less expressed, longer and had more exons than known isoforms. Shown are box-plots of the **(A,B)** Iso-Seq transcript expression (log10 TPM), **(C, D)** transcript length, and **(E,F)** number of exons of known and novel isoforms. The **(B)** Iso-Seq transcript expression, **(D)** transcript length and **(F)** exon number are also shown for isoforms further classified using *SQANTI* annotations. Known isoforms were subdivided into FSM and ISM, whereas novel isoforms were subdivided into NIC, NNC, and fusion. FSM – Full Splice Match, ISM – Incomplete Splice Match, NIC – Novel In Catalogue, NNC – Novel Not in Catalogue.

Table 4.3: Transcriptome annotations from global transcriptome profiling of the mouse cortex. Tabulated is an overview of the Iso-Seq transcriptome annotations in the mouse cortex (n = 12). Classifications were performed using *SQANTI2* (**Figure 3.10**). FSM - Full Splice Match, ISM - Incomplete Splice Match, NIC - Novel In Catalogue, NNC - Novel Not in Catalogue.

Description	Number	Isoform definition
Number of genes	14684	
Number of isoforms	46626	
Known genes	14482 (98.62%)	
Known isoforms	23530 (50.47%)	
FSM	19803 (42.47%)	exact alignment as reference
ISM	3727 (7.99%)	exact alignment as reference but fewer 5' exons
Novel isoforms	23096 (49.53%)	
NIC	13763 (29.52%)	a combination of known donor/acceptor sites
NNC	8751 (18.77%)	at least one novel donor/acceptor site
Fusion	297 (0.64%)	
Genic Genomic	62 (0.13%)	overlaps with introns and exons
Novel Genes	202 (1.38%)	
Intergenic	104 (0.22%)	located in the intergenic region
Antisense	119 (0.26%)	opposite-strand orientation to known gene

4.3.4 Comparisons with matched RNA-Seq data confirms the sensitivity of the Iso-Seq approach

Although Iso-Seq is accurate at characterising RNA diversity,[264] its sensitivity for quantifying gene expression has not been systematically explored. Generating highly-parallel RNA-Seq data on the same samples (n = 12), we found a strong correlation between gene-level expression quantified using both methods (n = 13,923 genes; Pearson's correlation: corr = 0.71, $P < 2.23 \times 10^{-308}$). To further assess the quantitative accuracy of Iso-Seq, we included ERCC spike-in control molecules. Among the detected ERCC molecules (n = 57, 62%) we found a near-perfect correlation between the full-length Iso-Seq reads and the actual amount of control used (Pearson's correlation: corr = 0.98, $P = 1.42 \times 10^{-41}$), highlighting the power of Iso-Seq to accurately quantify the abundance of highly-expressed transcripts. The vast majority of unique splice junctions identified in our Iso-Seq data were supported by RNA-Seq (n = 152,872 junctions, 98.1%). For transcripts that could be recapitulated in the matched RNA-Seq data, there was a significant correlation between transcript expression levels quantified using both sequencing approaches (n = 41,488 transcripts; Pearson's correlation: corr = 0.48, $P < 2.23 \times 10^{-308}$) further highlighting that transcript abundance can be reliably quantified using Iso-Seq.

Using our Iso-Seq data as a scaffold, we generated a reference-guided transcriptome assembly from our mouse cortex RNA-Seq data using *Stringtie*.[263] Many of the isoforms reconstructed from RNA-Seq reads appeared to represent incomplete fragments of full-length transcripts identified in Iso-Seq. Overall, isoforms assembled using RNA-Seq reads had a significantly shorter mean length (RNA-Seq: mean length = 2.31kb; Iso-Seq: mean length = 3.18kb; two-tailed unpaired t-test: t = 71.9, $P < 2.2 \times 10^{-16}$), lower average number of exons (RNA-Seq: mean n = 7.30; Iso-Seq: mean n = 10.8; two-tailed unpaired t-test: t = 76.7, $P < 2.2 \times 10^{-16}$) and were less likely to be located within 50bp of a CAGE peak (RNA-Seq: 34.0% vs Iso-Seq: 71.9%, Fisher's Test: odds ratio = 4.97, $P < 2.2 \times 10^{-16}$, **Figure 4.8B**). Importantly, more than 50% of isoforms robustly detected using Iso-Seq could not be readily recapitulated using standard RNA-Seq (**Figure 4.8C**), highlighting the advantage of long-read sequencing for characterizing isoform diversity.

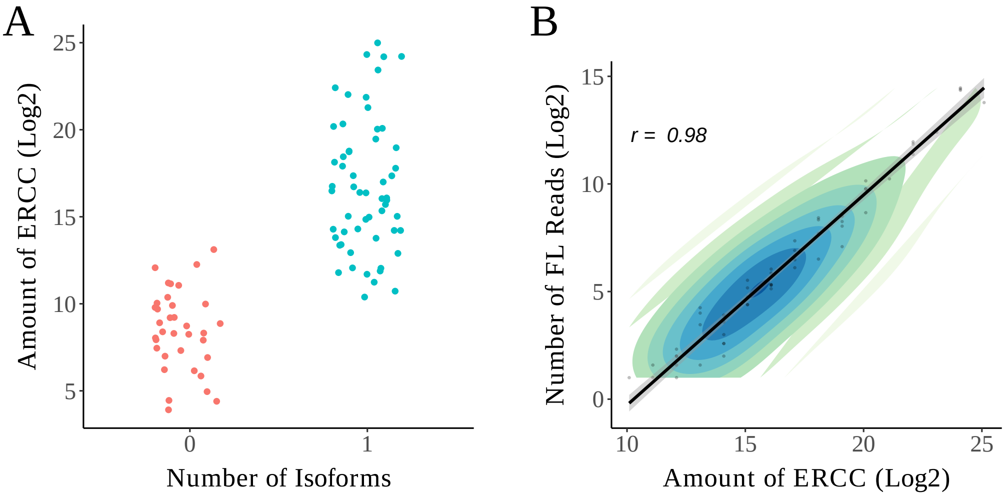


Figure 4.7: Over 60% of ERCC controls were detected with highly accurate quantification. Shown is **(A)** a scatter plot of the number of isoforms detected per ERCC control. As expected, highly-concentrated ERCC controls were detected as single molecules. **(B)** A density plot of the number of full-length reads associated for each detected ERCC control against the known amount used. FL - Full-length. The Iso-Seq bioinformatics pipeline was optimised to ensure only one unique molecule was detected per ERCC.

4.3.5 Detection of transcripts with fusion events across genes

Transcriptional read-through between two or more adjacent genes can produce "fusion transcripts" that represent an important class of mutation in several types of cancer.[265] Although fusion events are thought to be rare,[266] we found evidence of fusion transcripts (n = 297 fusion transcripts, 0.64% of total transcripts) associated with 218 genes (1.48% of total genes), **Figure 4.9A,B**) with a quarter of these genes associated with more than one fusion transcript (n = 53 genes, 24.3% of fusion genes).

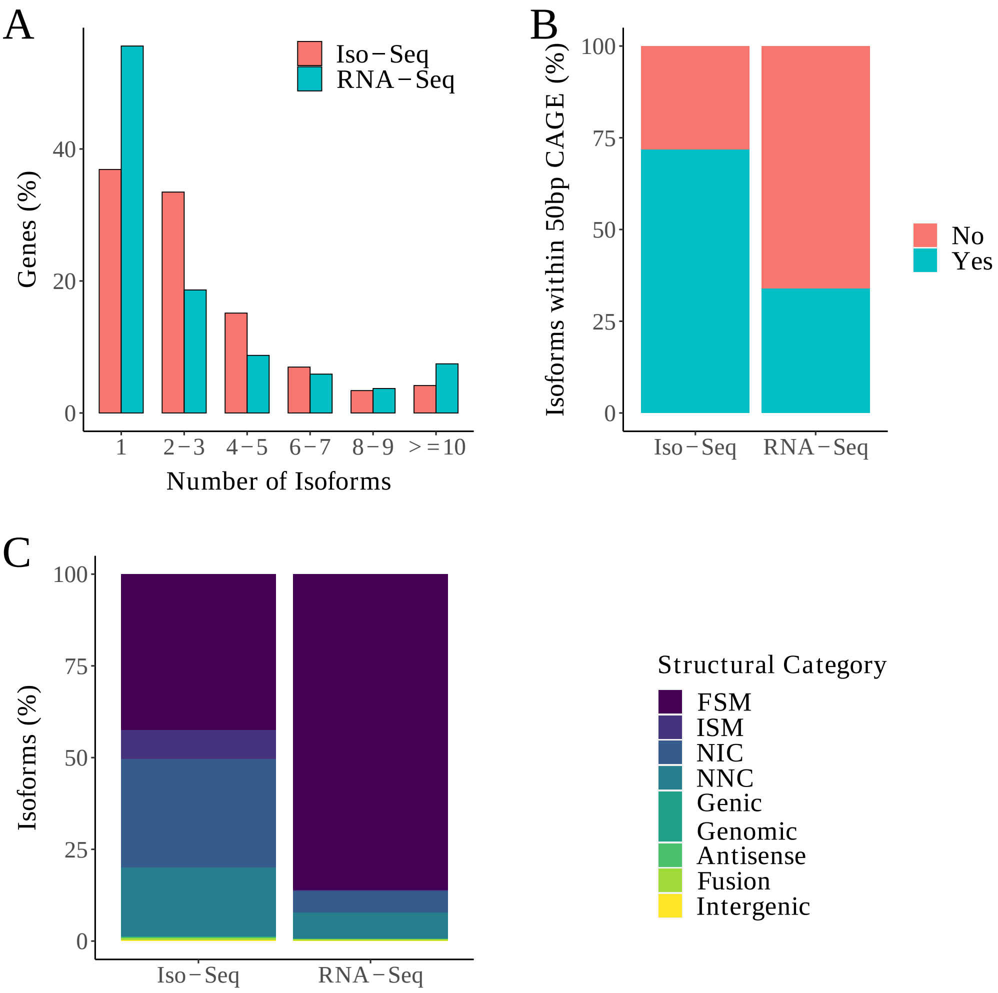


Figure 4.8: Iso-Seq identified more novel isoforms per gene that were more likely to be located within a CAGE peak. A reference-guided transcriptome using only RNA-Seq data (RNA-Seq-defined transcriptome) was generated. Shown are barplots of **(A)** the isoform diversity in Iso-Seq- and RNA-Seq-defined transcriptome, **(B)** the number of isoforms located within 50bp of a CAGE peak, and **(C)** the number of isoforms classified as novel and known using *SQANTI*. FSM - Full Splice Match, ISM - Incomplete Splice Match, NIC - Novel In Catalogue, NNC - Novel Not in Catalogue.

4.3.6 Iso-Seq identifies "novel" cortex-expressed genes

Although the vast majority of isoforms were annotated to known genes, a small number represented expression from potentially novel genes (n = 223 transcripts mapping to 202 novel genes). These novel genes were all multi-exonic (mean length = 1.75kb, s.d = 1.21kb, range = 0.098kb - 6.86kb, mean number of exons = 2.5), and more than half of the transcripts from these novel genes were predicted to be non-coding (n = 143, 64.1% novel-gene transcripts). These transcripts were generally shorter and less abundant than transcripts of known genes (length: W = 7.79 x 10^6, *P* = 5.22 x 10^{-45}; expression: W = 2.29 x 10^6, *P* = 1.5 x 10^{-73}), and a quarter of these novel-gene transcripts were enriched near CAGE peaks (n = 58, 26.0% novel-gene transcripts).

Interestingly, over half of these novel-gene transcripts were antisense to known genes (n = 119 transcripts, 53.4%, mapping to 97 novel genes, **Figure 4.9C**), with the majority of them found within the body of an annotated gene (n = 95, 97.9% of antisense novel genes). A relatively large proportion of these further shared exonic regions (exon-exon overlap, n = 72, 74.2%) reflecting sense-antisense (SAS) pairs.

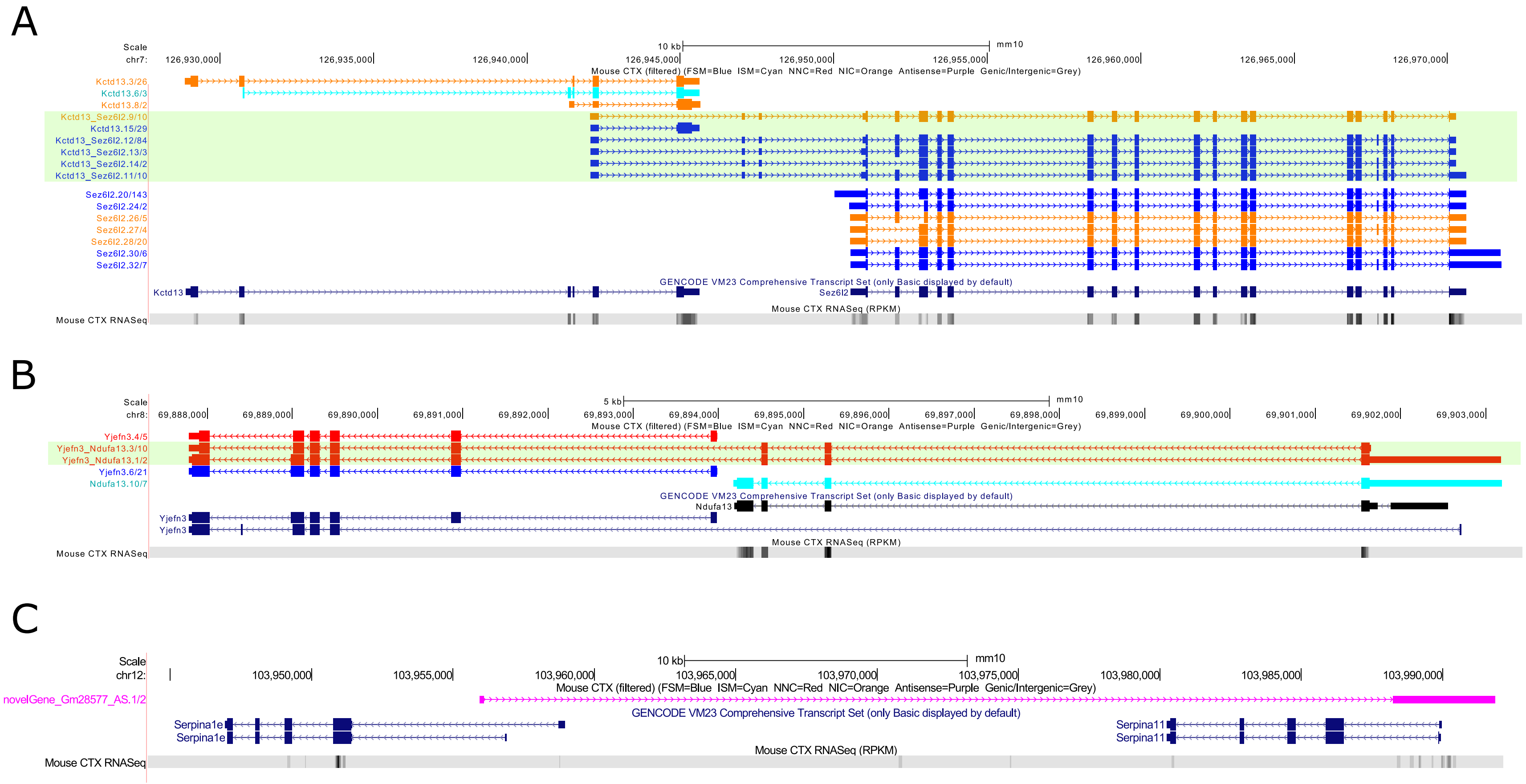


Figure 4.9: Examples of fusion transcripts and novel genes identified in the mouse cortex. Shown are UCSC genome browser tracks of **(A)** Five read-through "fusion" transcripts incorporating exons from *Kctd13* (SZ-associated) and *Sez6l2* (SZ-associated), **(B)** Two read-through "fusion" transcripts incorporating exons from *Yjefn3* and *Ndufa13* (SZ-associated), and **C)** a novel antisense transcript spanning across *Serpina1e* and *Serpina11* in the mouse cortex. Transcripts are coloured based on *SQANTI2* classification categories (blue = FSM, cyan = ISM, red = NNC, orange = NIC). Fusion transcripts are boxed in green. SZ - Schizophrenia.

4.3.7 Many transcripts map to long non-coding RNA genes

Although the majority of transcripts (n = 43,450, 93.6% of total transcripts) were classified as protein-coding by the presence of an open reading frame (ORF), a relatively large number of transcripts were annotated as encoding long non-coding RNAs (lncRNAs) (n = 1,141 transcripts associated with 734 genes). These lncRNA transcripts were shorter than non-lncRNA transcripts (lncRNA transcripts: mean length = 2.22kb, s.d = 1.36kb, range = 0.148kb - 8.49kb; non-lncRNA transcripts: mean length = 3.21kb, s.d = 1.68kb, range = 0.083kb - 15.9kb; W = 3.52 x 10^{7}, P = 8.24 x 10^{-98}, **Figure 4.10A**), and contained fewer exons[267] (W = 4.56 x 10^{7}, P < 2.23 x 10^{-308}, **Figure 4.10B**) with a dramatic enrichment of mono-exonic molecules[227] (n = 273, 23.9% of lncRNA transcripts) compared to non-lncRNA transcripts (n = 914, 2.02% of non-lncRNA transcripts). They were also characterised by lower transcript expression than non-lncRNA transcripts[267,268] (W = 3.16 x 10^{7}, P = 5.67 x 10^{-40}, **Figure 4.10C**), and detected with fewer isoforms (lncRNA transcripts: mean n = 1.55; non-lncRNA transcripts: mean n = 3.29; W = 7.40 x 10^{6}, P = 5.76 x 10^{-107}, **Figure 4.10E**). A small proportion of these annotated lncRNA transcripts further contained a putative ORF (n = 153, 13.4%, **Figure 4.10D**) supporting recent observations that some lncRNAs have potential protein coding capacity,[269] although the majority of such ORFs are unlikely to code for proteins;[270] of note, these ORFs were shorter than those identified in non-lncRNA transcripts (non-lncRNA ORF: mean length = 139bp; lncRNA ORF: mean length = 519bp; W = 1.75 x 10^{7}, P = 8.33 x 10^{-195}).

4.3.8 AS strongly contributes to cortical isoform diversity

In total, 40,249 alternative splicing events were identified in known genes with AF (Alternative first exon) (n = 2,853 (31.9%) associated with 6,476 (44.1%) genes) and ES (Exon skipping, n = 8,686 (21.6%) events associated with 4,570 (31.1%) genes) being the most prevalent (**Figure 4.11A**). Splicing events and frequency were also compared between known and novel isoforms. Except for AF and AL, all the other splicing events, particularly intron retention, were more likely to be observed in novel isoforms. This highlights the power of Iso-Seq to recapitulate the usage of complex splicing events that would have otherwise been underestimated using RNA-Seq data (Fisher's Test, A3: P = 7.78 x 10 $^{-14}$, odds ratio = 1.34; A5: P = 1.21 x 10^{-13}, odds ratio = 1.45; IR: P < 2.23 x 10^{-16}, odds ratio = 4.92; MX: P = 4.18 x 10^{-11}, odds ratio = 1.81; ES: P < 2.23 x 10^{-16}, odds ratio = 1.57, **Figure 4.11A**). For the majority of genes characterised by splicing, only 1 - 2 AS events were observed (n = 10,708, 81.8% of AS genes, **Figure 4.11B**), suggesting that AS events are often mutually independent.

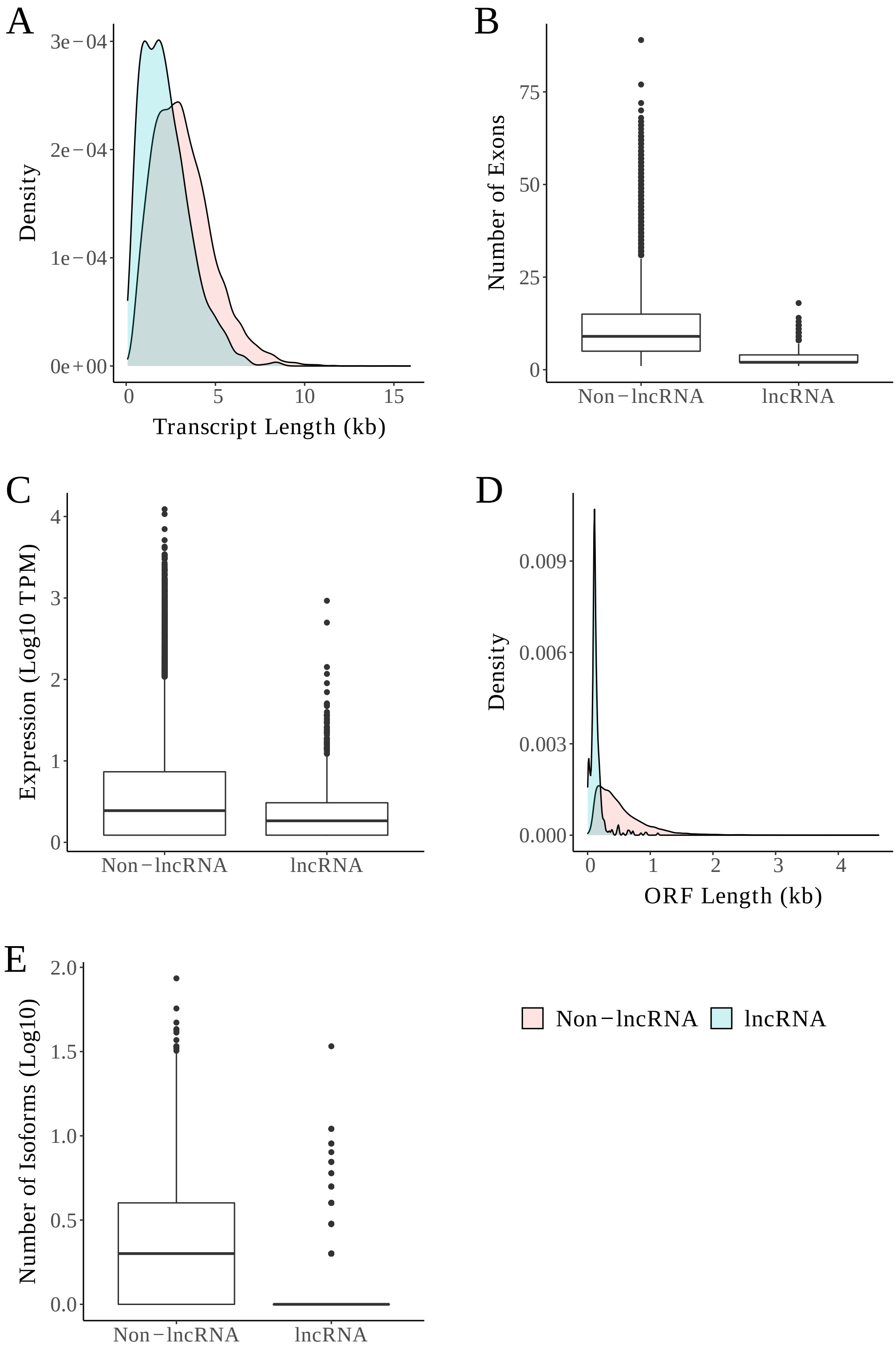


Figure 4.10: LncRNA transcripts were more lowly expressed and typically longer than non-lncRNA transcripts, despite containing fewer exons. Shown are distributions of the **(A)** transcript length, **(B)** number of exons, **(C)** transcript expression, **(D)** open reading frame (ORF) length and the **(E)** number of isoforms annotated to lncRNAs and non-lncRNAs. LncRNAs – Long non-coding RNAs.

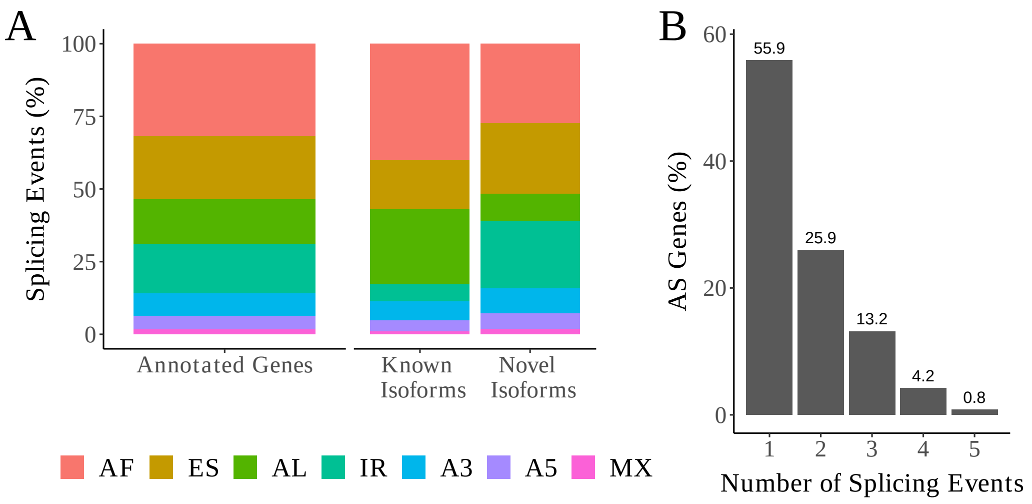


Figure 4.11: Alternative first is the most prevalent AS event, and novel isoforms were more likely to be characterised with complex AS events. Shown are bar-plots of the **(A)** proportion of AS events in known genes, known and novel isoforms, and the **(B)** proportion of alternatively-spliced genes with varying number of splicing events. AF – Alternative first exon, AL – Alternative last exon, A5' – Alternative 5' splice site, A3' – Alternative 3' splice site, IR – Intron retention, MX – Mutually exclusive, ES – Exon skipping.

4.3.9 Intron retention is associated with reduced expression and nonsense-mediated decay

Nonsense-mediated decay (NMD) acts to reduce transcriptional errors by degrading transcripts containing premature stop codons[271] and is one mechanism by which intron retention can influence gene expression[272] (described in **Section 1.2.1**). Overall, > 10% of transcripts mapping to annotated genes were predicted to undergo NMD (NMD transcripts) characterized by the presence of an open reading frame and a coding sequence (CDS) end motif before the last junction (n = 6,014 (13.0%) transcripts associated with 2,945 (20.3%) of annotated genes). These NMD transcripts were less abundant than non-NMD transcripts (NMD transcripts: mean expression = 11.2 TPM, s.d = 85.0 TPM; non-NMD transcripts: mean expression = 23.1 TPM, s.d = 143.1 TPM; W = 8.72 x 107, P = 6.15 x 10^{-156}). NMD was particularly enriched among transcripts that contained an IR event (IR transcripts) and were also predicted to be protein-coding (n = 2,341 (36.2%) IR transcripts associated with 1,380 (9.53%) genes), and transcripts with both IR and predicted NMD were particularly lowly expressed (W = 7.50 x 10^{6}, P = 1.67 x 10^{-42}, **Figure 4.12A**). Only a small number of genes had transcripts where IR and NMD were mutually exclusive (n = 277 genes, 1.91%, **Figure 4.12B**), providing additional support for the hypothesized relationship between these two transcriptional control mechanisms.[273]

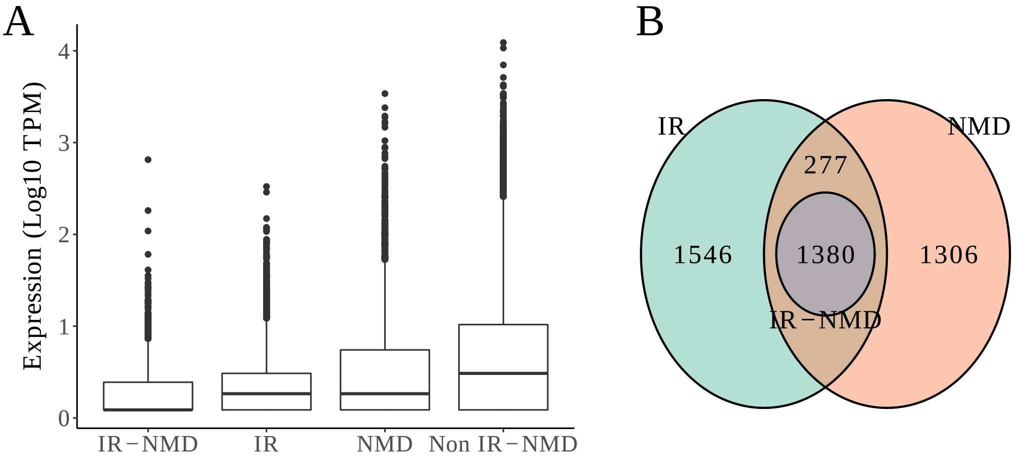


Figure 4.12: Intron retention is associated with nonsense-mediated decay and reduced expression. Shown is **(A)** a bar-plot of the expression of transcripts characterised with intron retention (IR) and nonsense-mediated decay (NMD), and **(B)** a Venn diagram of the number of genes associated with transcripts characterised with intron retention (IR), nonsense-mediated decay (NMD), or both (IR-NMD). As shown in Figure B, 1380 genes were associated with IR transcripts that were predicted for NMD, and 277 genes were associated with mutually exclusive IR transcripts and NMD transcripts.

4.4 Conclusions

We used long-read isoform sequencing to characterize full-length cDNA sequences and generate a detailed map of alternative splicing in the mouse cortex. To our knowledge, this study represents the most comprehensive characterization of cortical isoform diversity yet undertaken.

Several findings are particularly notable. First, we highlight that existing gene annotations are incomplete and that novel transcripts are likely to exist for a large proportion of expressed genes. Our data show examples of novel exons and even entire genes not currently annotated in existing databases. Second, we show that read-through transcripts (or gene fusion transcripts) occur naturally[274] and at detectable levels in the cortex. Although many of these fusion transcripts appear to be associated with NMD, some have the potential to be translated into proteins or may have a regulatory effect at the RNA level. Third, we are able to highlight the significant extent to which alternative splicing events contribute to isoform diversity in the cortex. In particular we show that IR is a relatively common form of AS in the cortex that is associated with reduced expression and NMD. Finally, our findings highlight the power of long-read sequencing approaches for transcriptional profiling. We show that transcriptional profiles generated using Iso-Seq reflect the cerebral cortex as expected, and our findings were validated using complementary approaches (i.e. RNA-Seq, and by comparison to existing genomic databases). Despite long-read sequencing often assumed to be less quantitative than standard short-read RNA sequencing methods,[275] we observed a strong correlation between expected and detected levels of ERCC spike-in control molecules, highlighting the power of Iso-Seq to accurately quantify the abundance of highly-expressed transcripts.

Our results should be interpreted in the context of several limitations. First, we profiled tissue from a relatively small number of mouse samples. Although rarefaction curves confirmed our sequencing dataset was close to saturation, we were unable to explore inter-individual variation in alternative splicing. Future work will aim to extend our analyses to larger numbers of samples to explore population-level variation in transcript abundance in the mouse cortex and differences associated with AD pathology. Second, despite the advantages of long-read sequencing approaches for the characterization of novel full-length transcripts, we implemented a stringent QC pipeline and undertook considerable filtering of our data. Many true transcripts from our final dataset, particularly lowly-expressed transcripts, are likely to have been filtered out. Our analyses are likely to represent an underestimation of the extent of RNA isoform diversity in the cerebral cortex. Future work will aim to sequence samples at a deeper coverage to explore gene-specific splicing differences associated with AD. Third, our analyses were performed on "bulk" cortex tissue containing a heterogeneous mix of neurons, oligodendrocytes and other glial cell-types. A recent study using a combination of long-read and single-cell sequencing identified cell-type-specific transcript diversity in the mouse hippocampus and prefrontal cortex[192] (described in **Table 1.6**). However, we were limited in exploring these differences in our data. Finally, although we explored the extent to which novel transcripts contained ORFs, the extent to which they are translated and contribute to cortical proteomic diversity is unknown.

In summary, our data confirm the importance of alternative splicing and alternative first exon usage in the mouse entorhinal cortex, dramatically increasing transcriptional diversity and representing an important mechanism underpinning gene regulation in the brain. We highlight the power of long-read sequencing for completing our understanding of mouse gene annotation. The transcript level data is provided as a resource to the scientific community (http://genome.exeter.ac.uk/BrainIsoforms.html).

Chapter 5

Splicing signatures of progressive tau pathology in AD mouse model

5.1 Introduction

There is increasing evidence on the role of transcriptional dysregulation and aberrant splicing in the development and pathogenesis of AD (described in **Section 1.2.3**). Recent transcriptome profiling studies have identified changes in splicing and transcript expression in both human AD post-mortem brain tissue and AD mouse models (reviewed in **Table 1.2** and **Table 1.3**, respectively). However, to date, these studies have relied on short-read RNA sequencing approaches, which cannot reliably detect specific isoforms (as discussed in **Section 1.3.1**) and have broadly ignored identifying differences at the transcript level. In contrast, we have illustrated the power of long-read sequencing to identify full-length transcripts and improve our annotations of alternatively-spliced isoforms in the cortex of the rTg4510 mouse model of AD tauopathy (**Chapter 4**).

While long-read sequencing approaches are currently considered to be only semi-quantitative, recent studies have delivered promising strategies for transcript-based analysis by using a hybrid approach:[276] the alignment of short-read RNA-Seq data to improved transcriptome annotations derived from long-read data (as depicted in **Figure 3.21**). This has enabled the identification of differentially expressed isoforms and analysis of differential transcript usage between experimental groups.[276]

Following on from the results presented in **Chapter 4**, this chapter aimed to exploit the cortical long-read sequencing datasets (hereby referred to as "Iso-Seq global dataset") generated from rTg4510 transgenic (TG) and wild-type (WT) mice to identify transcriptional and splicing alterations associated with progressive tau pathology. The objectives of this chapter were as follows:

1. To assess global variation in splicing patterns between rTg4510 TG and WT mice.
2. To perform differential gene expression analysis and validate differences in gene expression associated with tau pathology from previous RNA-Seq studies.
3. To perform differential transcript expression analysis to identify differences in transcript expression associated with tau pathology.
4. To perform differential transcript usage analysis to identify genes with significant alterations in isoform proportions between rTg4510 TG and WT mice.

5.2 Methods

5.2.1 Datasets

All analyses presented in this chapter follow on from **Chapter 4** and use the same Iso-Seq long-read datasets generated from 12 female mice (n = 6 WT, n = 6 TG, aged 2 and 8 months, **Table 4.1**). Briefly, RNA was prepared for Iso-Seq library preparation and SMRT sequencing on the PacBio Sequel (**Section 4.2.2**), followed by QC and data processing (**Section 4.2.4**). Reads from individual samples were processed separately with *IsoSeq3* and merged for transcript collapse using *Cupcake*. High-quality, full-length transcripts from the merged dataset were then aligned to the mouse reference genome (mm10, GENCODE) using *Minimap2* (v2.17) and re-annotated using *SQANTI3* with no splice junction filtering from short-read RNA-Seq data. ISM transcripts with only the 3' fragment matching reference transcript (3' ISM) were considered technical artefacts resulting from 5' degradation and thus removed.

5.2.2 Quantification of human *MAPT* transgene expression

As described in **Section 2.1.1**, rTg4510 mice recapitulate AD tauopathy through the overexpression of the human tau transgene, MAPT^{P301L}. The presence of the human-specific *MAPT* sequence was therefore determined in the Iso-Seq datasets by using the *grep* Unix command, as QC of sample identity. A 2kb region present in the 3' UTR was chosen as the representative human *MAPT* sequence.[90]

5.2.3 Characterisation of alternative splicing events

Alternative splicing events were examined using a range of packages and custom scripts (as described and implemented in **Section 4.2.5**), to assess whether there was a difference in splicing patterns associated with progressive tau pathology in the rTg4510 mouse model.

5.2.4 Gene and isoform quantification

Gene and isoform expression were estimated using two approaches (as described in **Section 3.3.1**). Briefly, these were: i) the alignment of short-read RNA-Seq reads to the Iso-Seq-derived transcriptome (hybrid approach) using *Kallisto*[222] (v0.46.0), and ii) the use of normalised Iso-Seq full-length read counts as a proxy for expression. Full-length read counts for each sample were taken from the *read_stat.txt* file generated using the *collapse_isoforms_by_sam.py* script (*Cupcake*) with the sequencing run ID as identifiers.

5.2.5 Differential expression analysis

Differential expression analysis was performed using *tappAS* (fully described in **Section 3.3.4**). Briefly, *tappAS* filters out lowly-expressed isoforms, normalises read counts using the TMM approach, and implements *maSigPro*[256–258] to elucidate the effects of genotype and age with the following model:[257]

Let *i* denote the genotype group (WT - wild-type mice, TG - rTg4510 transgenic mice), *j* the age (2 or 8 months) and *r* the replicate number (assuming that gene or transcript expression is measured in replicated samples).

$$\begin{aligned} y_{ijr} = \beta_0 + \beta_1 D_i \\ + \delta_0 T_{ijr} + \delta_1 T_{ijr} D_{ijr} \end{aligned}$$

where

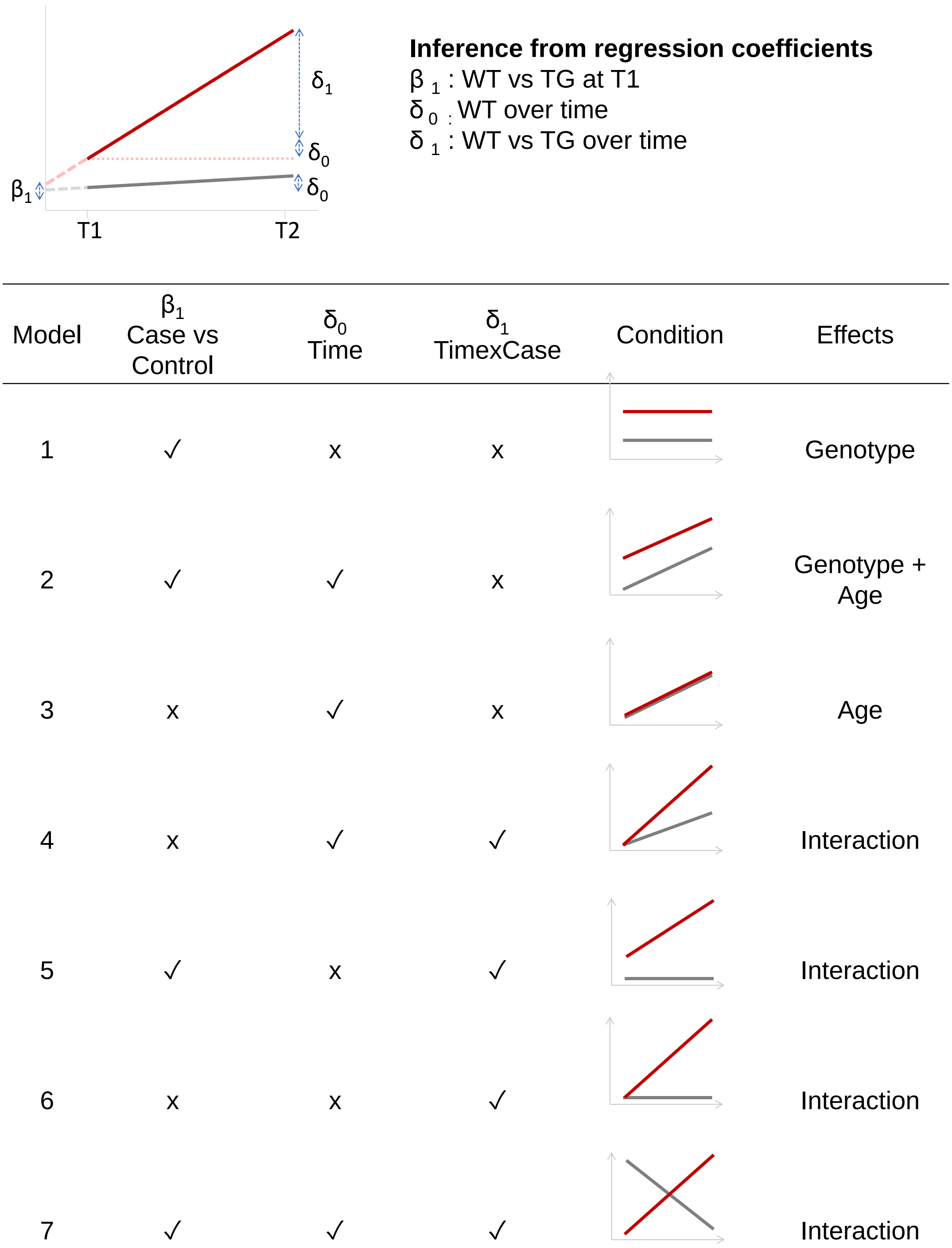
y_{ijr} = normalised expression value for each gene or transcript in the situation *ijr* (genotype group *i* at age *j* of replicate *r*)
D = dummy binary variable to distinguish between the genotype groups, whereby 0 and 1 refers to reference (WT) and experimental group (TG), respectively
T = age at 2, 8 months described using a polynomial model (degree of 1)
β_0, δ_0 = regression coefficients for reference group (WT) relating to the age
β_1, δ_1 = regression coefficients for the difference between experimental group (TG) and reference group (WT) at each age

therefore, if:

$FDR(\beta_1) < 0.05$ = significant expression difference between WT & TG at 2 mos
$FDR(\delta_0) < 0.05$ = significant expression difference in WT across 2 and 8 mos

Equation 5.1: Linear regression model to determine differential gene and transcript expression. The model is adapted from *MaSigPro* and implemented as part of *tappAS*. It identifies differences in gene and transcript expression between two groups (WT - wild-type mice, TG - rTg4510 transgenic mice) at different time points (age in months). FDR - False discovery rate. mos - Months.

Under this model, a differentially expressed gene or transcript between WT and TG mice across age was defined by a statistically significant regression coefficient (adjusted $P < 0.05$) and a regression model with $R^2 > 0.5$ (i.e. the amount of variance explained by the model) (**Figure 5.1**).



Inference from regression coefficients
β $_1$: WT vs TG at T1
δ $_0$: WT over time
δ $_1$: WT vs TG over time

Model	β$_1$ Case vs Control	δ$_0$ Time	δ$_1$ TimexCase	Condition	Effects
1	✓	x	x		Genotype
2	✓	✓	x		Genotype + Age
3	x	✓	x		Age
4	x	✓	✓		Interaction
5	✓	x	✓		Interaction
6	x	x	✓		Interaction
7	✓	✓	✓		Interaction

Figure 5.1: Different conditions modelled for rTg4510 genotype and age effects. Shown is a linear model implemented in *maSigPro* to dissect genotype and age effects using **Equation 5.1** between two experimental groups (WT - Wild-type/Control, TG - Transgenic/Case) and across two time points (T1, T2).

The regression coefficients from **Equation 5.1** - β_1, δ_0, δ_1 - refer to the different variables modelled, the significance of which can be used to infer whether there is a genotype, age or interaction effect. The significance is symbolised by the tick and cross, which refers to adjusted P (FDR) < 0.05 and > 0.05 respectively. A significant value of β_1 denotes to a statistically significant difference between WT and TG at T1 (genotype effect), δ_0 to a difference in WT over time (age effect), and δ_1 to a difference between WT and TG across age (interaction effect).

5.3 Results

5.3.1 PacBio Iso-Seq run performance and sequencing metrics

No significant difference in sequencing yield was identified between WT and TG mice (n = 12 samples, two-tailed unpaired t-test, t(10) = -0.636, *P* = 0.539, **Figure 5.2A**), and no significant correlation was observed between run yield and RIN across samples (n = 12 samples, Pearson's correlation, corr = -0.296, df = 10, *P* = 0.350, **Figure 5.2B**). No significant difference was also observed in the number of reads (**Figure 5.2C**) and transcripts generated between WT and TG mice (n = 12 samples, two-tailed unpaired t-test, t = -0.005, df = 10, *P* = 0.996, **Figure 5.2D**) or by age (n = 12 samples, t = -1.58, df = 10, *P* = 0.15). Notably, a similar read profile was attained for all the samples except the first two samples, which were sequenced using an older chemistry and had a relatively lower throughput. Nonetheless, all the samples were successfully sequenced with optimal runs, as indicated by the high throughput and the similar number of full-length, full-length non-chimeric (FLNC) and poly(A) FLNC reads recovered. ERCC alignment and annotations similarly revealed no difference in the number of ERCC control molecules detected between WT and TG (mean number of ERCC controls: WT = 32.4, 35%; TG = 32.2, 35.22%).

5.3.2 *MAPT* transgene is only expressed in rTg4510 TG mice

As expected, human-specific *MAPT* sequences were only detected in reads from TG mice, confirming stable activation of the human *MAPT* transgene (**Figure 5.3A**) and supporting our previous analysis using short-read RNA-Seq data.[90] In line with previous results, we also observed a decrease in transgene expression associated with age - a likely reflection of progressive neuronal loss, given that the transgene expression is largely restricted to excitatory neurons under the CAMK2a promoter. Alignment of these human-specific transcripts to the mouse genome were either mapped to the mouse prion protein gene (*Prnp*) with high identity but low alignment length (i.e similar in nucleotide sequence but low overlap, **Figure 5.3B,C**) or to the mouse *Mapt* gene with low identity but high alignment length (i.e not similar in nucleotide sequence but high overlap, **Figure 5.3B,D**). This is reflective of the transgene sequence in rTg4510 mouse model, given it contains exons 2 and 3 of mouse *Prnp*[195] and is homologous to the mouse *Mapt* gene. Applying filter thresholds (85% alignment identity and 95% alignment length) for downstream analysis removed these human-specific *MAPT* transcripts (**Figure 5.3B**).

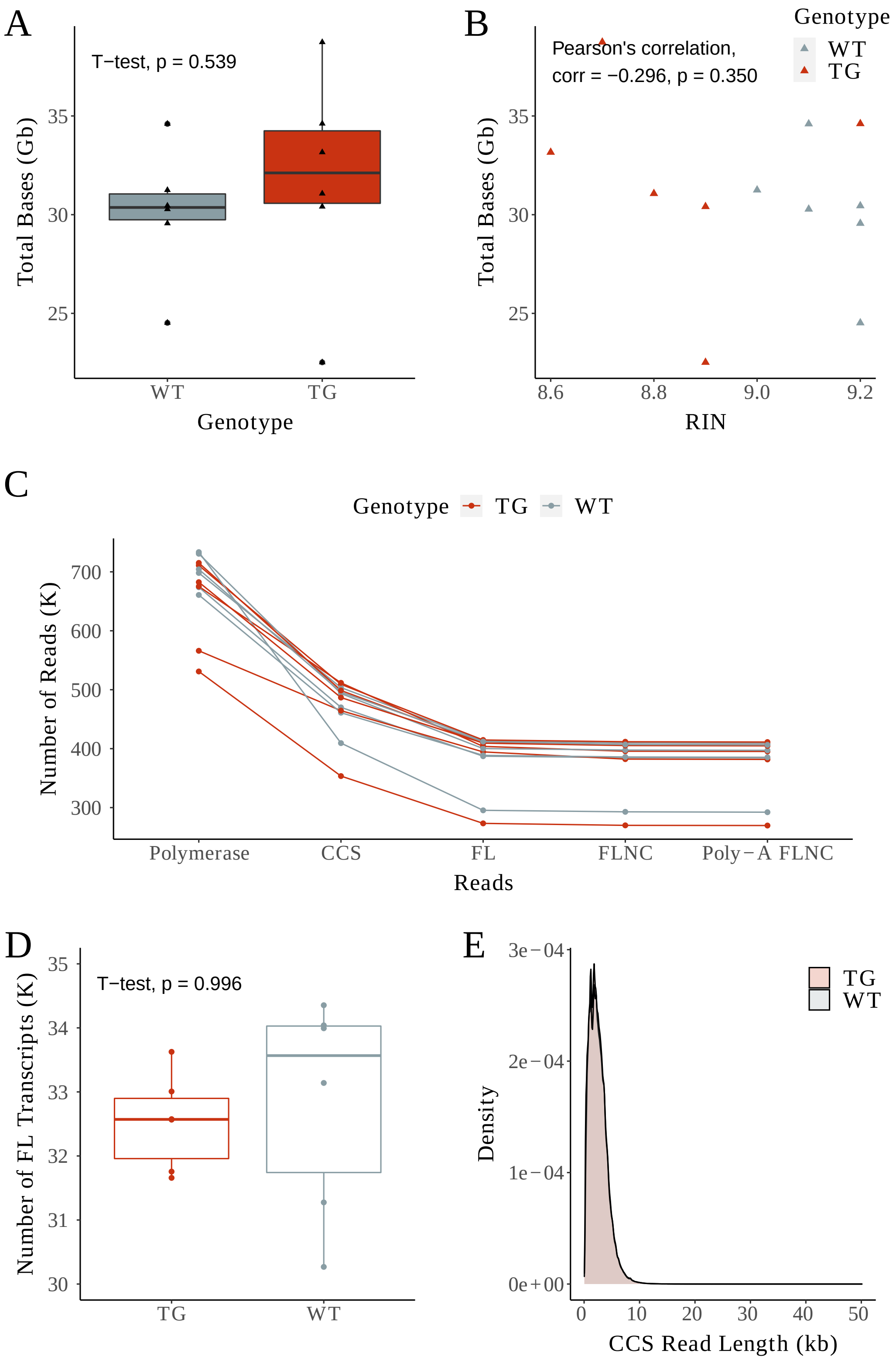


Figure 5.2: No significant difference in sequencing metrics, number of transcripts and read length were observed between WT and rTg4501 TG mice: *Legend continues on the following page.*

Figure 5.2: Shown is **(A)** a box plot of the total yield generated from Iso-Seq sequencing of rTg4510 WT (n = 6) and TG mice (n = 6). Full details of all runs are provided in **Table 4.2**. **(B)** A scatter plot of the total yield generated and the RIN attained for each sample (RIN refers to the quality of RNA used for library preparation). **(C)** The number of reads generated through the Iso-Seq bioinformatics pipeline from initial generation of CCS reads, full-length reads with primer removal to poly(A) FLNC reads with removal of artificial concatemers and trimming of poly(A) tails. Note, the first two samples with lower throughput were sequenced using an older chemistry. **(D)** A box plot of the total number of full-length transcripts generated for WT and TG mice. **(E)** Distribution of CCS read length. CCS - Circular consensus sequence, FL - Full-length, FLNC - Full-length non-chimeric, Gb - Gigabases, K - Thousand, kb - Kilobases, TG - rTg4510 transgenic mice, WT - Wild-type mice.

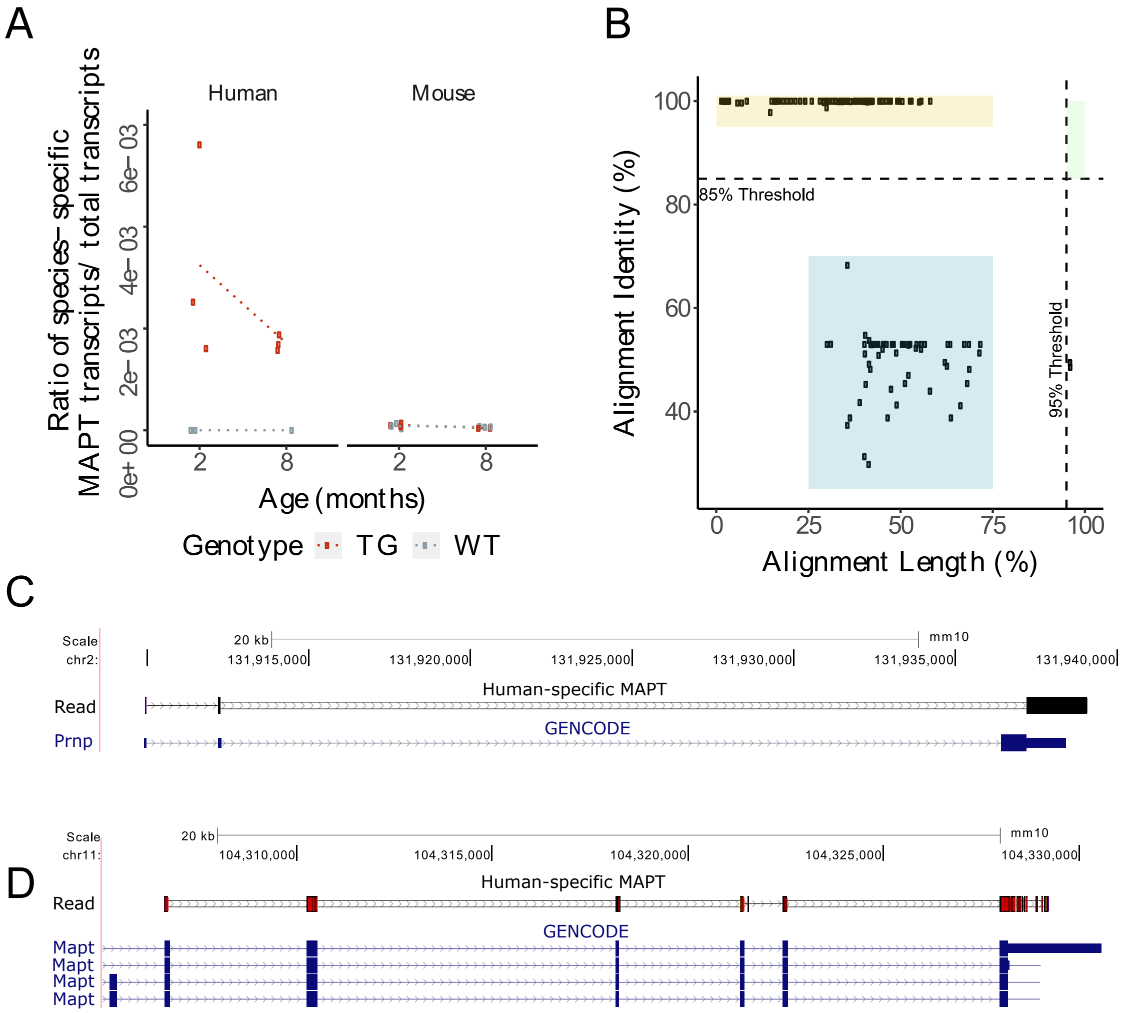


Figure 5.3: Human-specific *MAPT* sequences were only present in transgenic mice with relatively low homology to mouse *Prnp* and *Mapt* gene. The presence of human- and mouse-specific *MAPT*/*Mapt* sequences was measured in full-length transcripts generated from Iso-Seq merged dataset. Shown is **(A)** a scatter plot of the ratio of full-length transcripts that were mapped to human-specific *MAPT* and mouse-specific *Mapt* sequences. Dotted lines represent the mean paths across ages. **(B)** A scatter plot of the alignment metrics of human-specific *MAPT* transcripts to the mouse genome. Transcripts were either aligned to mouse *Prnp* gene (boxed yellow) with high identity but low length (given that the transgene contains only exon 2 and 3 of mouse *Prnp* gene[195]) or mouse *Mapt* gene (boxed blue) with low alignment identity but relatively high length. Green box refers to the transcripts retained after applying identity and length threshold. **(C)** UCSC genome browser tracks of human-specific (black) *MAPT* transcripts (transgene) and mouse *Prnp* gene and **(D)** mouse *Mapt* gene. Blue tracks represent known transcripts from mouse reference genome (mm10). The double horizontal lines indicate unalignable sequences and the red lines indicate bases that differ between the genome and transcript. Tracks were cropped and modified to remove irrelevant genes within the same locus. UTR - Untranslated region.

5.3.3 rTg4510 WT and TG mice were characterised with a similar global transcriptomic profile

Despite identifying widespread RNA isoform diversity amongst genes expressed in the mouse entorhinal cortex (**Chapter 4**), the global transcriptomic profile between rTg4510 WT and TG mice were very similar. No difference was observed in the number of genes (mean n = 13,572 genes) or isoforms (mean n = 53,833 isoforms). Further characterisation of the transcriptome revealed similar profile of isoform diversity across genotype and age (**Table 5.2**), with half of the isoforms annotated as known and FSM (mean n = 30,018 isoforms, 55.8%, as also shown in **Section 4.3.3**) and with a similar distribution of isoform length and exon number (median = 8, range = 1 - 89). Splicing patterns were also very similar across genotype and age with usage of alternative first exons (AF) (mean n = 12,564 AF splicing events, 35%) (**Table 5.1**) as the most prevalent AS event across all the datasets, in line with previous findings (**Section 4.3.3**).

Table 5.1: Alternative splicing events associated with tau pathology and age. Tabulated is the number of splicing events detected for wild-type and rTg4510 transgenic mice aged 2 and 8 months (n = 12 samples, 3 biological replicates per group). Refer to **Figure 1.9** for depiction of the different types of splicing events.

Splicing events	Wild-type		Transgenic	
	2 months	8 months	2 months	8 months
A3	2164 (6.58%)	2571 (6.61%)	2388 (6.77%)	2388 (6.5%)
A5	1369 (4.16%)	1589 (4.09%)	1473 (4.18%)	1488 (4.05%)
AF	12048 (36.61%)	13073 (33.61%)	12514 (35.48%)	12622 (34.36%)
AL	8140 (24.73%)	9688 (24.91%)	8641 (24.5%)	9287 (25.28%)
IR	3611 (10.97%)	5404 (13.9%)	4293 (12.17%)	4774 (13%)
MX	299 (0.91%)	392 (1.01%)	331 (0.94%)	329 (0.9%)
SE	5278 (16.04%)	6174 (15.88%)	5632 (15.97%)	5846 (15.91%)

Table 5.2: Transcriptome annotations from global transcriptome profiling of the rTg4510 cortex by genotype and age. Tabulated is an overview of the Iso-Seq global transcriptome datasets generated from the rTg4510 mouse model, subsected by phenotype and age. Annotations from wild-type mice (n = 6 samples) and rTg4510 transgenic mice (n = 6 samples) were generated from merging Iso-Seq datasets from mouse aged 2 and 8 months of the respective phenotype. Novel genes refer to genes that were not currently present in existing genome annotations (mm10). Isoforms can be further classified as known (FSM, ISM) or novel (NIC, NNC, Genic Genomic, Antisense, Fusion, Intergenic, Genic Intron), as described in **Section 3.1.4.4**. FSM – Full Splice Match, ISM – Incomplete Splice Match, NIC – Novel In Catalogue, NNC – Novel Not in Catalogue.

	Wild-type (n = 6)	Transgenic (n = 6)	Wild-type, 2 months (n = 3)	Wild-type, 8 months (n = 3)	Transgenic, 2 months (n = 3)	Transgenic, 8 months (n = 3)
Total number of genes	14118	14213	13191	13312	12985	13616
Known genes	13932 (98.68%)	14031 (98.72%)	13081 (99.17%)	13168 (98.92%)	12874 (99.15%)	13474 (98.96%)
Novel genes	186 (1.32%)	182 (1.28%)	110 (0.83%)	144 (1.08%)	111 (0.85%)	142 (1.04%)
Total number of isoforms	62533	63038	48516	50278	45903	52730
FSM	33239 (53.15%)	33563 (53.24%)	27878 (57.46%)	28689 (57.06%)	26825 (58.44%)	29916 (56.73%)
ISM	4927 (7.88%)	4864 (7.72%)	3426 (7.06%)	3841 (7.64%)	3279 (7.14%)	3764 (7.14%)
NIC	15305 (24.48%)	15595 (24.74%)	11012 (22.7%)	11407 (22.69%)	10214 (22.25%)	12369 (23.46%)
NNC	8518 (13.62%)	8484 (13.46%)	5838 (12.03%)	5953 (11.84%)	5259 (11.46%)	6282 (11.91%)
Genic Genomic	63 (0.1%)	61 (0.1%)	44 (0.09%)	44 (0.09%)	32 (0.07%)	47 (0.09%)
Antisense	97 (0.16%)	104 (0.16%)	52 (0.11%)	77 (0.15%)	68 (0.15%)	75 (0.14%)
Fusion	276 (0.44%)	268 (0.43%)	200 (0.41%)	186 (0.37%)	167 (0.36%)	196 (0.37%)
Intergenic	108 (0.17%)	99 (0.16%)	66 (0.14%)	81 (0.16%)	59 (0.13%)	81 (0.15%)
Genic Intron	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Isoform length (bp)	Median: 2691, Range: 82-15016	Median: 2698, Range: 82-15913	Median: 2740, Range: 88-15016	Median: 2614, Range: 82-14850	Median: 2548, Range: 88-14302	Median: 2754, Range: 82-15913
Number of exons	Median: 8, Range: 1-89	Median: 8, Range: 1-89	Median: 9, Range: 1-89	Median: 8, Range: 1-89	Median: 8, Range: 1-77	Median: 9, Range: 1-89
Number of isoforms within 50bp CAGE	52096 (83.31%)	52633 (83.49%)	40589 (83.66%)	42378 (84.29%)	38227 (83.28%)	44729 (84.83%)

5.3.4 Iso-Seq confirms widespread gene expression differences associated with tau pathology in rTg4510 mice detected using short-read RNA-Seq

Although long-read sequencing is often assumed to be less quantitative than traditional short-read RNA-Seq approaches, we previously demonstrated the power of Iso-Seq to accurately quantify the abundance of highly-expressed transcripts (as described in **Section 4.3.4**). Subsequently, we sought to evaluate the utility of full-length Iso-Seq read counts as a proxy of abundance to identify differences in gene expression associated with progressive tau pathology. Of note, a recent RNA-Seq study by our group identified extensive gene expression differences in the same mouse model using short-read RNA-Seq data mapped to the mouse reference genome annotation.[90]

Using Iso-Seq read counts as a proxy of abundance (as detailed in **Section 3.3.1**), we identified 483 genes differentially expressed at a stringent FDR < 0.05. Using *MasigPro* to differentiate genotype and age effects (as illustrated in **Figure 5.1**), we identified evidence for differential gene expression associated with the rTg4510 genotype (**Figure 5.4A,B**) and age (**Figure 5.4B,C**), and interactions between genotype and age (**Figure 5.4D,E,F,G**). Classifying differentially expressed genes by effects, we identified 18 (3.73%) differentially expressed genes that were associated with genotype effect, and 356 (73.7%) genes whose expression significantly altered with tau pathology progression in rTg4510 mice (i.e interaction effect) (**Figure 5.5**). Among these, there was a significant (Exact bionomial test: n = 356 genes, P = 1.91 x 10^{-44}) enrichment of up-regulated genes (n = 304 genes (85.3%) with increased expression in TG compared to WT; n = 52 (14.6%) genes with decreased expression in TG). Using *EnrichR*, the differentially expressed genes were found to be highly enriched in the lysosome (GO Cellular Component: adjusted P = 4.19 x 10^{-4}, odds ratio = 3.06) and in particular, the TGF-β signalling pathway (WikiPathway 2021 Human: adjusted P = 2.92 x 10^{-2}, odds ratio = 17.16). Further in line with previous findings, a third of the differentially expressed genes were enriched in pathways involved in immune system activation (n = 140 genes, 34.4% of genes identified in the "turquoise" co-expression module,[90] **Figure 5.5B**).

Our previous RNA-Seq study[90] was more powered with a bigger sample size (RNA-Seq: n = 30 WT, n = 29 TG; Iso-Seq: n = 6 WT, n = 6 TG) and a deeper sequencing coverage (RNA-Seq: mean number of reads = 18.8M; Iso-Seq: mean number of CCS reads = 5.7M reads) and unsurprisingly identified a larger number of gene expression differences (n = 1,916 differentially expressed genes). However, 116 (6.05%) of these genes were also detected as differentially expressed using normalised Iso-Seq read counts as a proxy of expression, illustrating the utility of long-read sequencing for gene quantification and gene-level analyses. Recapitulating findings from our previous RNA-Seq study, we also identified *Gfap* and *C4b* as top-ranked differentially expressed genes associated with progressive tau pathology. Upregulation of *Gfap* (**Figure 5.6A,B**) - which encodes for the glial fibrillary acidic protein, a cytoskeletal protein that acts as a marker for astrocyte activation - and *C4b* (**Figure 5.6C,D**) - a member of the complement immune system - have also been previously observed in human AD post-mortem brain tissue and other AD mouse models.[277–279] Other top-ranked differentially expressed genes, whose expression differences were also recapitulated using Iso-Seq full-length read counts (**Table 5.3**), included: i) *Slc14a1*[280] encoding the urea transporter 1, ii) *Tgfbr1* encoding the TGF-βreceptor protein (**Figure 5.6E,F**), and iii) *Unc93b1*,[281] a transmembrane protein required for the toll pathway.

Table 5.3: Top-ranked differentially expressed genes associated with rTg4510 genotype. Tabulated is a summary of the top-ranked genes identified as differentially expressed in rTg4510 mice using *maSigPro* with Iso-Seq-derived transcriptome for annotation and Iso-Seq FL read count for quantification. Gene expression is determined from the sum of normalised expression of associated transcripts.

Gene	FDR[a]	$R^{2,b}$	$\log_2 FC_{genotype}$[c]	Mean gene expression			
				Wild-type		Transgenic	
				2 months	8 months	2 months	8 months
C4b	1.6×10^{-41}	0.945	4.38	4.94	2.73	4.97	103
Gfap	6.04×10^{-36}	0.933	3.12	82.8	70.5	118	1030
Tgfbr1	7.9×10^{-24}	0.892	2.95	0.663	3.38	2.03	15.7
Slc14a1	4.31×10^{-22}	0.899	2.95	9.55	14.7	6.16	47.7
Pros1	1.05×10^{-17}	0.894	2.08	8.17	9.32	6.26	26.4
Unc93b1	1.46×10^{-16}	0.863	1.61	3.59	5.04	6.47	19.8

[a] False discovery rate

[b] R^2 is a statistical measure that represents the amount of variance explained by the model

[c] $\log_2$ fold change of TG aged 8 months vs WT aged 8 months

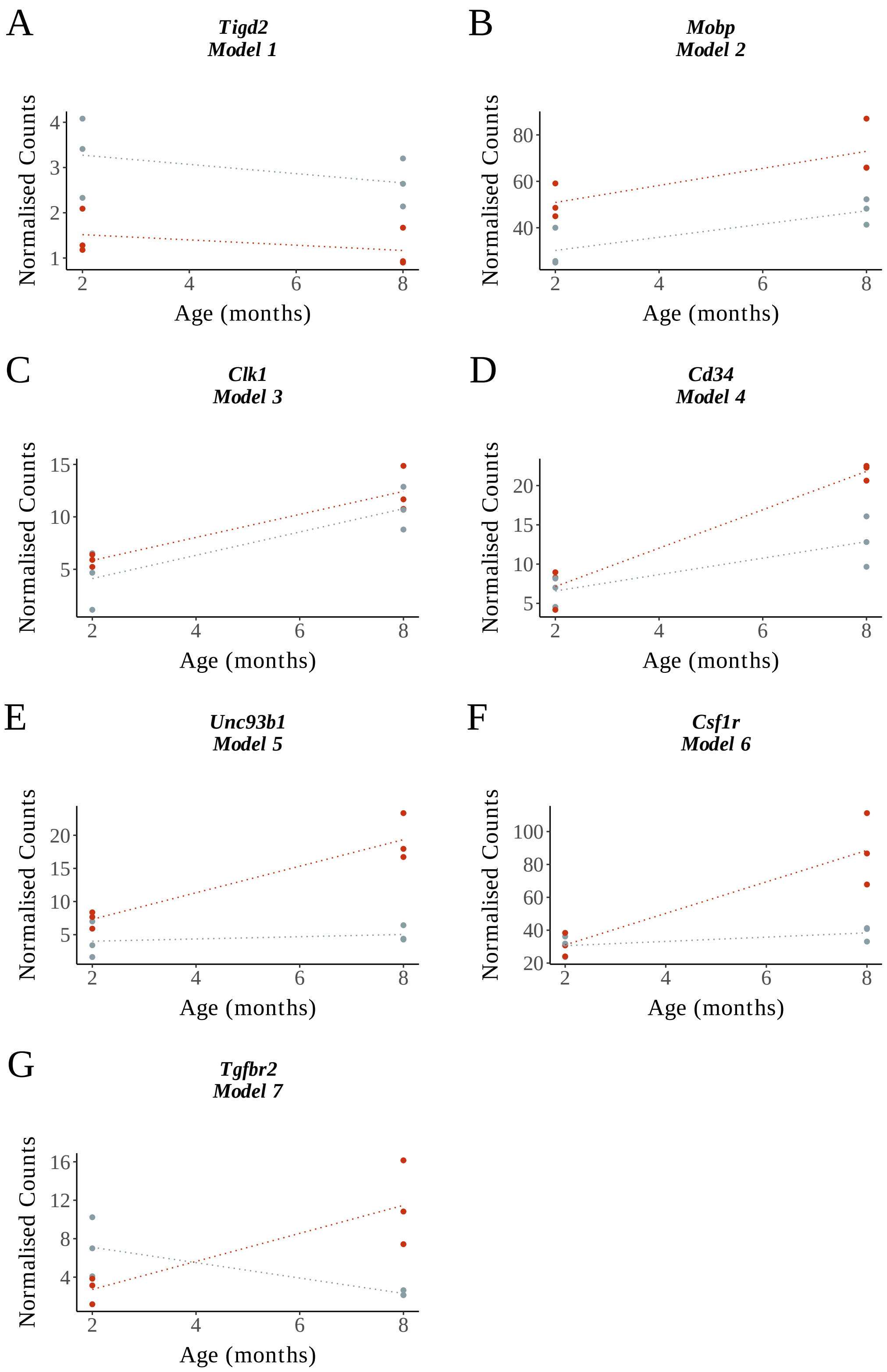


Figure 5.4: Differentially expressed genes exhibiting genotype, age and interaction effects. Shown are examples of differentially expressed genes classified under the different models using the Iso-Seq global transcriptome profiling (n = 6 WT, n = 6 TG, across age 2 and 8 months) with Iso-Seq full-length read counts for quantification: **(A)** *Tigd2* with a genotype effect, **(B)** *Mobp* with a genotype and age effect, **(C)** *Cik1* with an age effect, and **(D)** *Cd34*, **(E)** *Unc93b1*, **(F)** *Csf1r* and **(G)** *Tgfbr2* with an interaction effect. Dashed lines represent mean paths across age groups. Wild-type and rTg4510 transgenic mice are denoted by red and grey, respectively. The models are defined using **Equation 5.1** and depicted in **Figure 5.1**.

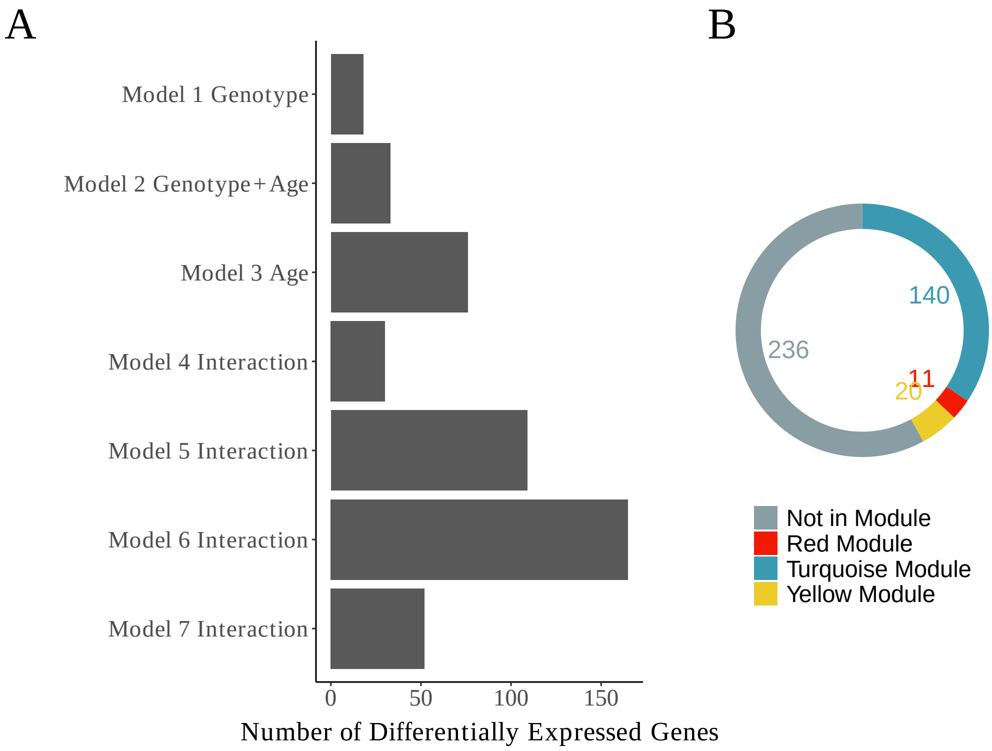


Figure 5.5: Differentially expressed genes were identified across all the different conditions with a number of differentially expressed genes exhibiting an interaction effect of rTg4510 genotype and age. (A) A bar chart of the number of differentially expressed genes (n = 483), determined from Iso-Seq FL read count as a proxy of expression and classified by rTg4510 genotype, age, and interaction effect (n = 6 WT, n = 6 TG, across 2 and 8 months). **(B)** A pie chart of the number and proportion of differentially expressed genes with genotype and interaction effect (n = 407 genes) identified in discrete co-expression network modules taken from our previous RNA-Seq study;[90] all three modules were significantly associated with progressive tau pathology: the "Red" module was down-regulated in TG mice and enriched for synaptic transmission, the "Turquoise" module was up-regulated in TG mice and enriched for immune system activation, and the "Yellow" module was down-regulated in TG mice and enriched for mitochondrial and synpatic processes. These modules refer to clusters of highly-correlated genes, which were determined using weighted gene correlation network analysis (WGCNA)[282] and functionally-annotated using gene ontology (GO) analyses.[283]

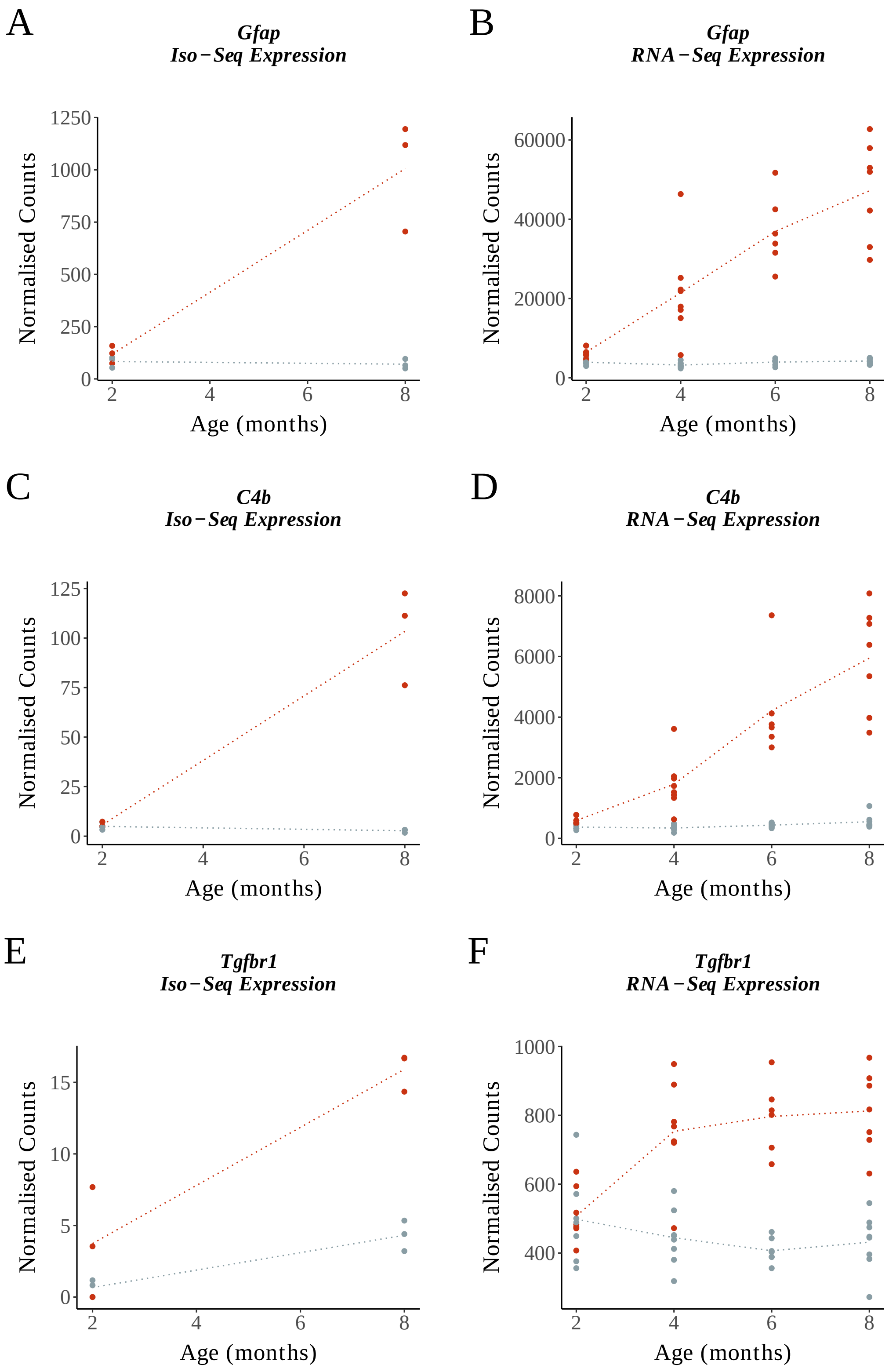


Figure 5.6: *Gfap* and *C4b* were the top-ranked differentially expressed genes associated with progressive tau pathology in the rTg4510 mice. Shown are scatter plots of the gene expression for **(A, B)** *Gfap*, **(C, D)** *C4b* and **(E, F)** *Tgfbr1* using either Iso-Seq full-length read count or RNA-Seq reads for transcript quantification. Dashed lines represent mean paths across age groups. Wild-type and rTg4510 transgenic mice are denoted by red and grey, respectively.

5.3.5 rTg4510 mice characterised by expression differences in novel, antisense genes

Highlighting the power of long-reads to comprehensively annotate the transcriptome, we previously detected novel genes in our Iso-Seq dataset that were not present in existing genome annotations (**Section 4.3.6**). These genes were often lowly-expressed and typically antisense to known genes with overlap at the UTR or gene body (as illustrated in **Figure 4.9**). Given the improved transcript annotation afforded by our Iso-Seq data, we next sought to test for expression differences in these novel genes associated with rTg4510 genotype. This was achieved by quantifying levels of expression by mapping RNA-Seq reads to our improved Iso-Seq-derived transcriptome annotation.

We identified three of these novel genes with evidence for differential expression associated with rTg4510 genotype. The most significant differentially expressed novel gene was located on chromosome 10 (PB.1799.1, **Figure 5.7A**) and was characterised by progressive down-regulation in TG mice (**Figure 5.8A**). The other two differentially expressed novel genes were found antisense to known genes: *Fgfr1op* (PB.6616.1, **Figure 5.7B**) within the gene-body and *Htra1* at the 5' UTR (PB.15002.1, **Figure 5.7C**). Both genes were up-regulated with progressive tau pathology in TG mice (**Figure 5.8B,D**). Notably, while *Fgfr1op* was not identified as differentially expressed (**Figure 5.8C**), *Htra1* was also found to have a higher expression in rTg4510 TG compared to WT mice (**Figure 5.8E**).

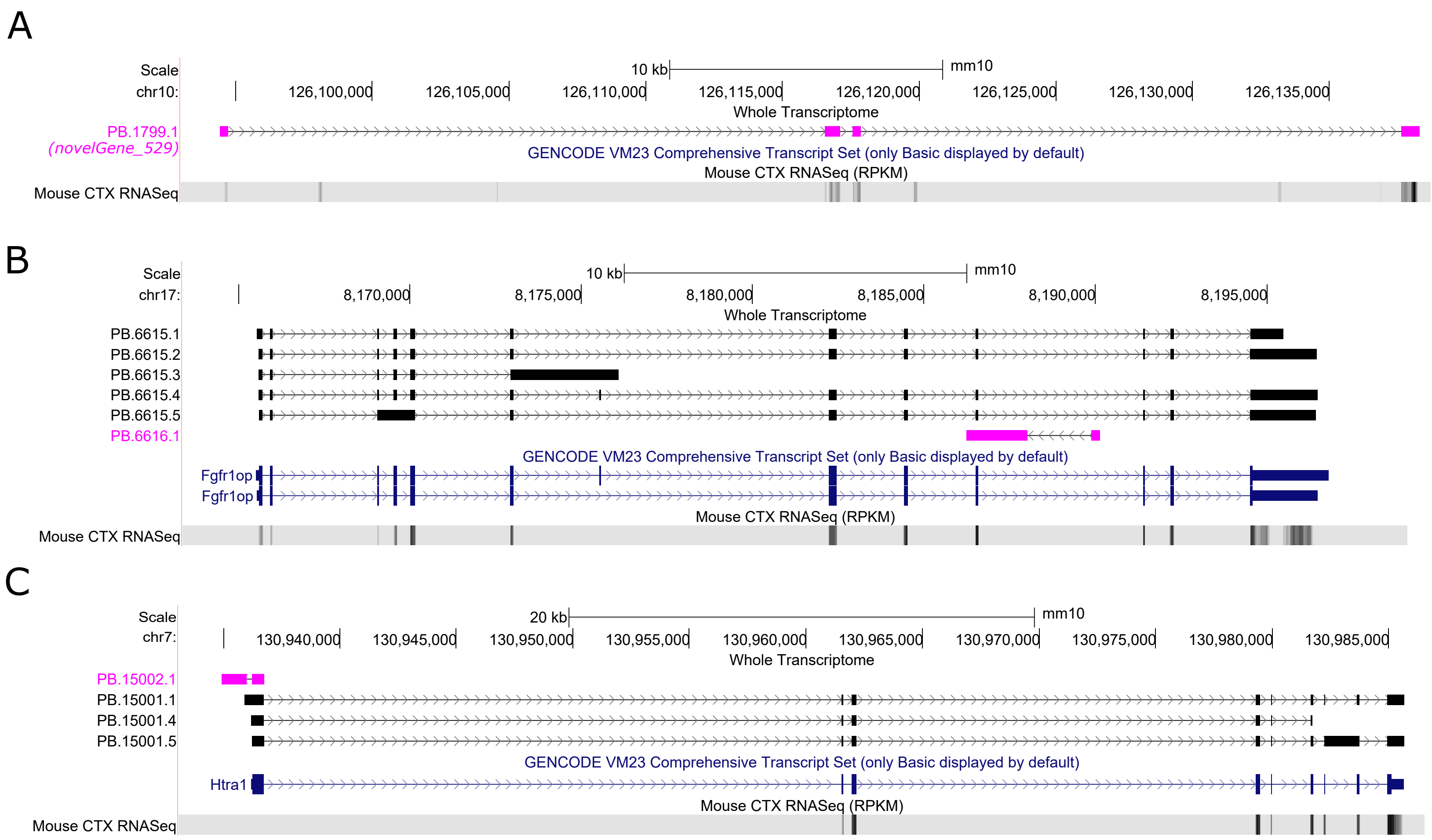


Figure 5.7: Visualisation of novel genes that were differentially expressed. Shown are UCSC genome browser tracks of three differentially expressed novel genes (coloured pink): **(A)** novel gene on chromosome 10, **(B)** novel gene antisense to *Fgfr1op*, and **(C)** novel gene antisense to *Htra1*. Shown are also mouse reference genome annotations (mm10) and RNA-Seq data from matched samples.

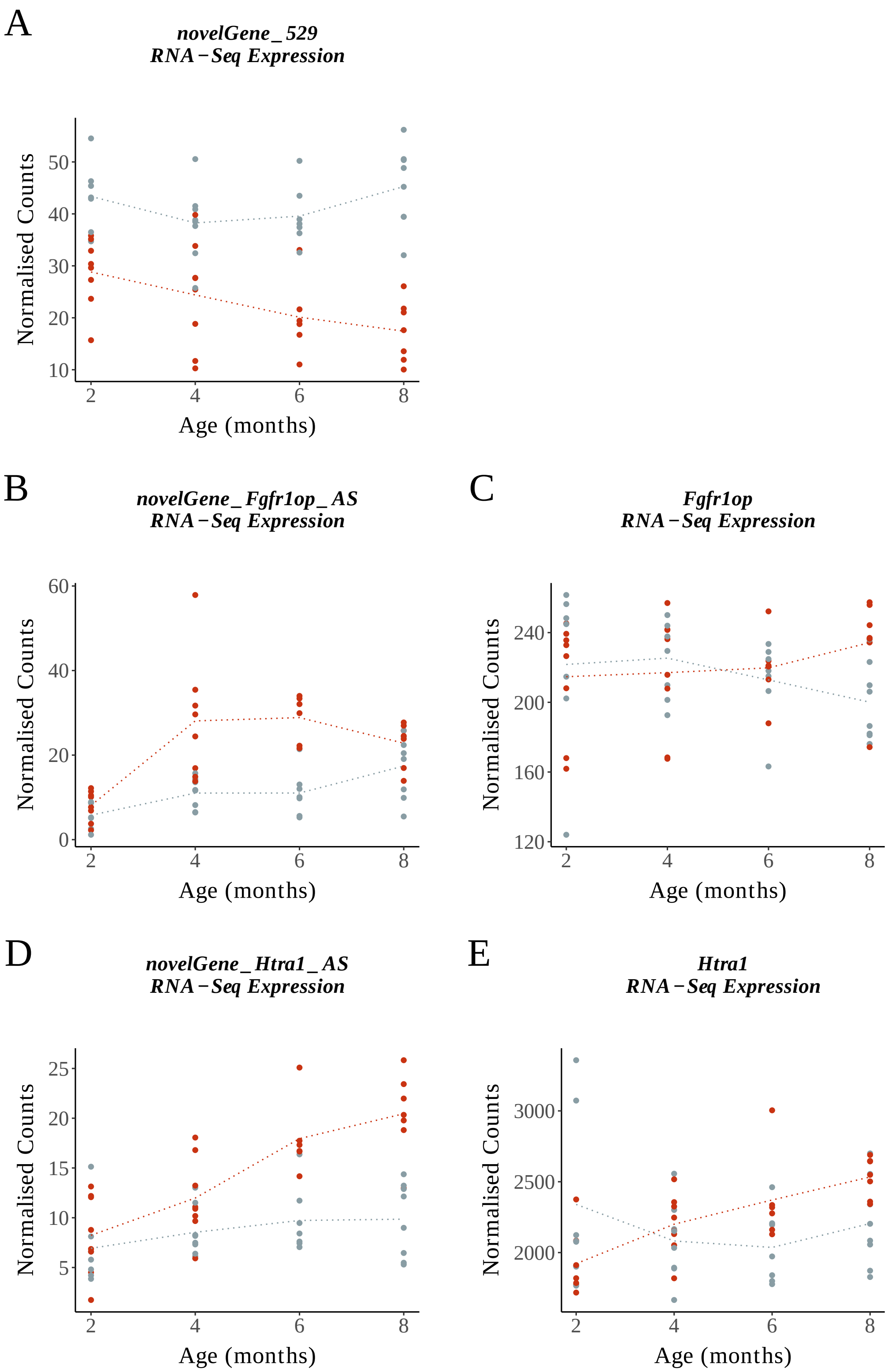


Figure 5.8: Three novel genes were found differentially expressed in rTg4510 mice. Shown are scatter plots of three differentially expressed novel genes, located **(A)** in chromosome 10 (PB.1799.1, **Figure 5.7A**), **(B)** antisense to *Fgfr1op* (PB.6616.1, **Figure 5.7B**) and **(D)** to *Htra1* (PB.15002.1, **Figure 5.7C**). Gene expression for the two known genes, **(C)** *Fgfr1op* and **(E)** *Htra1*, are also shown. Gene expression was determined from mapping RNA-Seq reads to Iso-Seq-derived annotations.

5.3.6 Gene expression differences in rTg4510 mice were primarily driven by the differential expression of dominant isoform

One of the added advantages of long-read sequencing is the improved confidence to reliably identify isoforms with significant expression differences across experimental conditions. Given that we were able to reliably detect tau-associated differentially expressed genes in rTg4510 TG mice using normalised full-length long-read read counts (described in **Section 5.3.4**), we subsequently sought to identify differentially expressed *transcripts* using the same approach.

By performing differential transcript expression analysis using *tappAS* with Iso-Seq reads for annotation and quantification, we identified 886 differentially expressed transcripts. Among these, 673 (75.9%) transcripts were associated with progressive tau pathology (interaction effect), 43 (4.85%) transcripts with tau pathology (genotype effect) and 170 (19.2%) transcripts with age. Similar to the gene-level analyses (**Section 5.3.4**), there was a significant (Exact bionomial test: n = 673 transcripts, P = 1.72 x 10^{-42}) enrichment of up-regulated transcripts (n = 510 transcripts (75.8%) increased in TG compared to WT mice). Using *EnrichR*, the differentially expressed transcripts were found to be highly enriched in the lysosome (GO Cellular Component: adjusted P = 1.65 x 10^{-4}, odds ratio = 3.25), and in several molecular functions including protein kinase binding (adjusted P = 0.048, odds ratio = 2.02) and ATPase binding (adjusted P = 0.048, odds ratio = 4.23).

Two of the most significant differentially expressed transcripts associated with the progression of tau pathology (**Table 5.4**) were annotated to *Gfap* (**Figure 5.9**) and *C4b* (**Figure 5.10**), the top two most differentially expressed genes (**Table 5.3**). Both genes were characterised by a dominant known isoform in rTg4510 mice: Gfap-201 (ENSMUST00000077902.5, PB.2972.16) (**Figure 5.9A,B**) and C4b-201 (ENSMUST00000069507.8, PB.7004.8) (**Figure 5.10A,B**). The expression of these two known isoforms were significantly higher than that of the novel isoforms, and were strongly up-regulated with progressive tau pathology (**Figure 5.9D**, **Figure 5.10D**). This suggests that increased *Gfap* (**Figure 5.6A**) and *C4b* gene expression in aged rTg4510 TG mice were primarily driven by the up-regulation of their respective dominant isoform. This corroborates with previous studies that reported a differential increase in the human-equivalent Gfap-201 transcript in AD temporal cortex,[284] and qPCR studies in AD mouse models that similarly showed up-regulation of *Gfap*-associated isoforms.[285] Of note, all the other minor novel isoforms were more abundant in the aged rTg4510 transgenic mice (**Figure 5.9C**, **Figure 5.10C**).

These findings were validated with usage of normalised RNA-Seq read counts, after alignment to the improved Iso-Seq-derived transcriptome annotation, as a proxy of expression; both Gfap-201 (**Figure 5.9E**) and C4b-201 (**Figure 5.10E**) were found to be dramatically up-regulated with progressive tau pathology. However, some of the minor novel isoforms annotated to *Gfap* were also found to change with rTg4510 genotype with a more pronounced up-regulation than Gfap-201 (PB.2972.8, PB.2972.17, **Figure 5.9E**). Visualisation of these novel differentially expressed isoforms revealed that they were generally very similar (**Figure 5.9A**), with an almost identical internal exonic structure; for example, PB.2972.8 and Gfap-201 only differed by the presence of exon 3, whereas PB.2972.17 contained an alternative splice site at exon 8. The sensitivity of RNA-Seq reads to differentiate these almost-identical isoforms is therefore questionable, given that there are only a few loci that can be used for unambiguous assignment of short RNA-Seq reads.

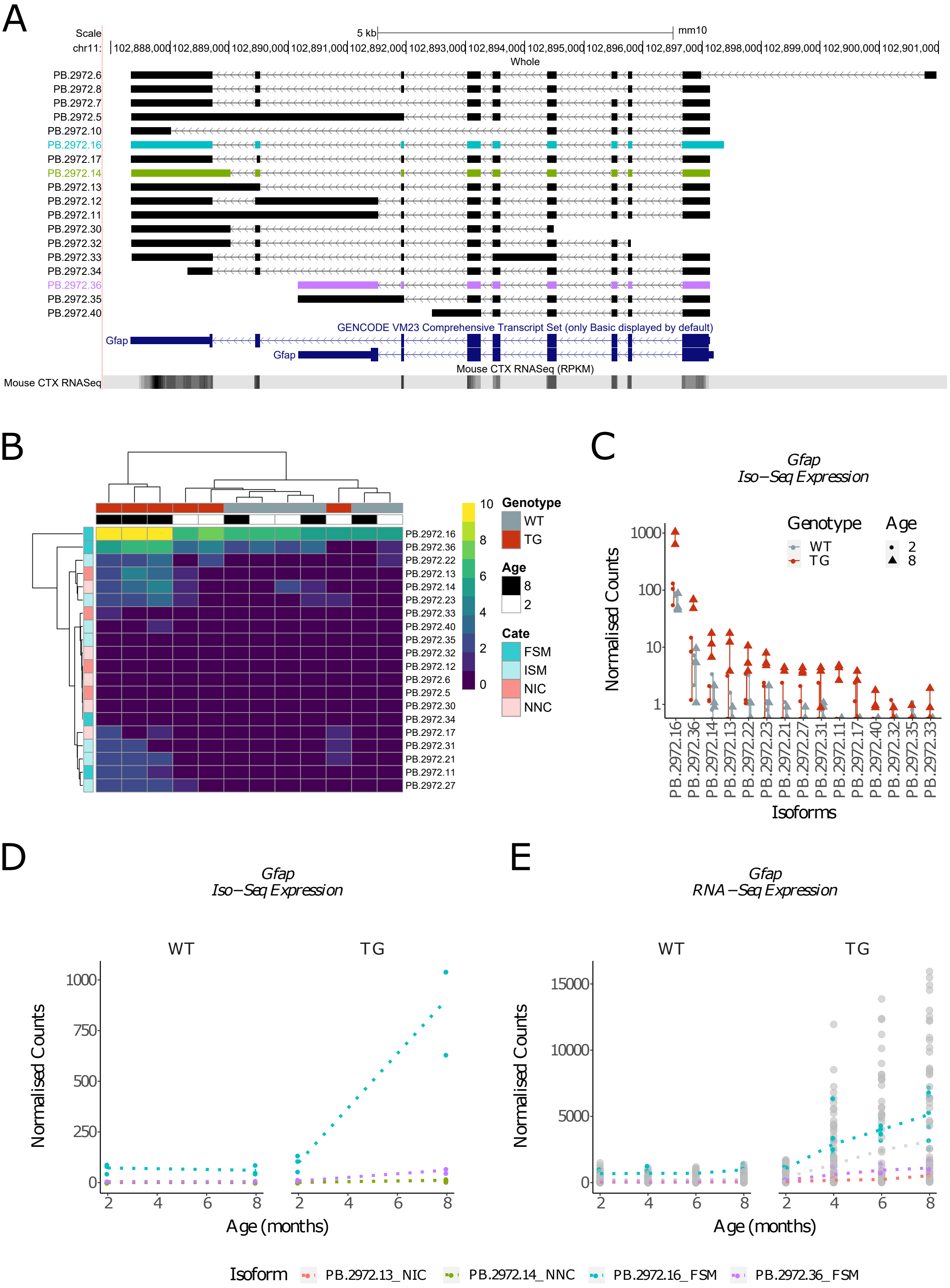


Figure 5.9: Significant up-regulation of the known isoform of *Gfap* with progression of tau pathology in rTg4510 mice. Shown are **(A)** UCSC genome browser tracks of isoforms annotated to *Gfap*, **(B)** hierarchical clustering of *Gfap*-associated isoforms by abundance (Iso-Seq FL read count, log2), **(C)** Normalised Iso-Seq FL read count of the top 15 most abundant isoforms, and **(D)** differentially expressed transcripts identified using normalised Iso-Seq read and **(E)** RNA-Seq read counts. Grey dots denote to differentially expressed transcripts identified using RNA-Seq but not Iso-Seq reads for quantification. FSM - Full Splice Match, ISM - Incomplete Splice Match, NIC - Novel In Catalogue, NNC - Novel Not in Catalogue. WT - Wild-type mice, TG - rTg4510 transgenic mice. Dotted lines represent the mean paths across age.

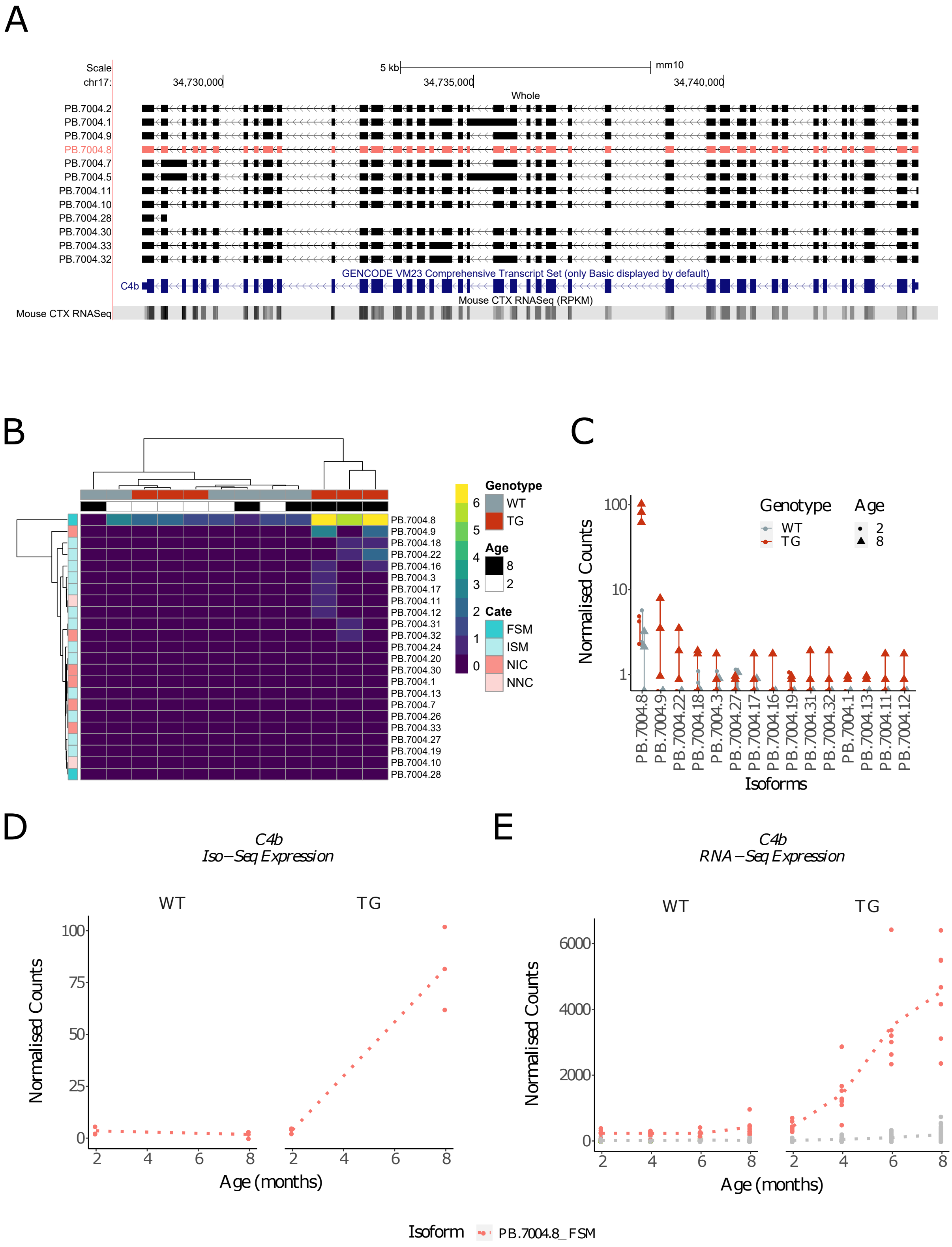


Figure 5.10: Significant up-regulation of the known isoform of *C4b* with progression of tau pathology in rTg4510 mice. Shown are **(A)** UCSC genome browser tracks of isoforms annotated to *C4bp*, **(B)** hierarchical clustering of *C4b*-associated isoforms by abundance (Iso-Seq FL read count, log2), **(C)** Normalised Iso-Seq FL read count of the top 15 most abundant isoforms, and **(D)** differentially expressed transcripts identified using normalised Iso-Seq read and **(E)** RNA-Seq read counts. Grey dots denote to differentially expressed transcripts identified using RNA-Seq but not Iso-Seq reads for quantification. FSM - Full Splice Match, ISM - Incomplete Splice Match, NIC - Novel In Catalogue, NNC - Novel Not in Catalogue. WT - Wild-type mice, TG - rTg4510 transgenic mice. Dotted lines represent the mean paths across age.

5.3.7 rTg4510 mice were characterised by the differential expression of transcripts of genes implicated in AD

The list of transcripts progressively altered in rTg4510 transgenic mice were annotated to genes previously implicated in AD development and pathology (**Table 5.4**). This included: i) Padi2-201 (ENSMUST00000030765.6) annotated to *Padi2*/*Pad2* (**Figure 5.11**), which encodes for an enzyme that is abnormally activated in astrocytes from AD patients,[286] ii) H2-D1-202 (ENSMUST00000172785.7) annotated to *H2-D1* (**Figure 5.12**), which encodes for major histocompatibility complex (MHC) class 1, an immune-related gene that is also up-regulated in microglia isolated from a neurodegenerative mouse model with AD-like phenotypes,[178] iii) Gatm-201 (ENSMUST00000028624.8) annotated to *Gatm* (**Figure 5.13**), encoding a mitochondrial protein recently revealed as a key AD protein signature,[287] and iv) Ctsd-202 (ENSMUST00000151120.8) annotated to *Ctsd* (**Figure 5.14**), encoding Cathepsin D, a lysosomal protease involved in Aβ[288] and tau[289] degradation, and a key regulator of A$\beta_{42/40}$ ratio,[290] among others. Drawing parallels to *Gfap* and *C4b*, these genes were characterised by a dominant known isoform that was significantly up-regulated with progressive tau pathology, which was validated using RNA-Seq reads mapped to our improved Iso-Seq-derived transcriptome annotation.

Table 5.4: Differentially expressed transcripts associated with rTg4510 genotype. Tabulated are the top-ranked differentially expressed transcripts between wild-type and rTg4510 transgenic mice using *maSigPro*. Iso-Seq reads were used for both annotation and quantification.

Rank[a]	Gene	Isoform	Isoform ID	FDR[b]	$\log_2 \mathrm{FC}_{\mathrm{genotype}}$[c]	Mean WT transcript expression		Mean TG transcript expression	
						2 months	8 months	2 months	8 months
1	*Ubqln1*	ENSMUST00000058735.11	PB.4255.13	7.43E-42	0.96	0.969	33.9	43.4	22.8
2	*C4b*	ENSMUST00000069507.8	PB.7004.8	5.9E-40	0.942	4.41	3.49	1.78	3.86
3	*Gfap*	ENSMUST00000077902.5	PB.2972.16	1.11E-35	0.933	3.19	72.3	60.9	99
4	*Tgfbr1*	ENSMUST00000007757.14	PB.10959.1	1.09E-19	0.841	3.16	0.66	1.31	1.21
5	*Cd34*	ENSMUST00000016638.7	PB.1036.2	8.43E-18	0.894	1.99	1.31	5.4	2.71
29	*Padi2*	ENSMUST00000030765.6	PB.11607.2	5.2E-11	0.792	2.13	22.6	24.3	26.7
76	*H2-D1*	ENSMUST00000172785.7	PB.7039.1	8.47E-08	0.697	1.5	30.6	28.1	40.3
79	*Gatm*	ENSMUST00000028624.8	PB.9298.1	9.79E-08	0.738	0.869	29.1	34.5	34.6
175	*Ctsd*	ENSMUST00000151120.8	PB.15108.6	0.00000592	0.598	1.03	89.7	91.8	127

[a] The order of differentially expressed transcripts (n = 886) by FDR

[b] False discovery rate

[c] $\log_2$ fold change of TG aged 8 months vs WT aged 8 months

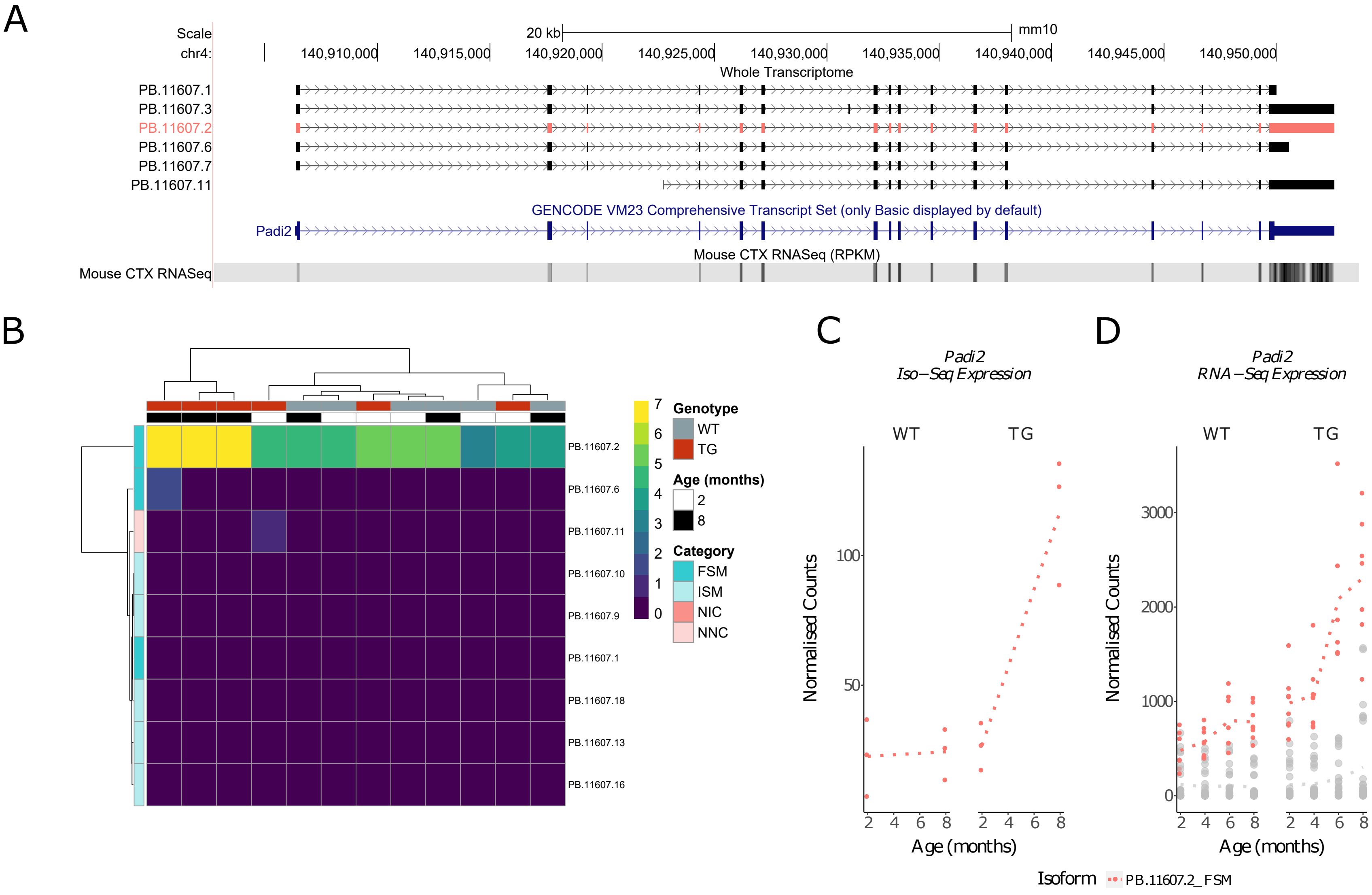


Figure 5.11: Significant up-regulation of *Padi2-201* with progression of tau pathology in rTg4510 mice. Shown are **(A)** UCSC genome browser tracks of isoforms annotated to *Padi2*, **(B)** hierarchical clustering of *Padi2*-associated isoforms by abundance (Iso-Seq FL read count, log2), **(C)** differentially expressed transcript (Padi2-201, ENSMUST00000030765.6, PB.11607.2) identified using normalised Iso-Seq read and **(E)** RNA-Seq read counts. Grey dots denote to differentially expressed transcripts identified using RNA-Seq but not Iso-Seq reads for quantification. Dotted lines represent the mean paths across age.

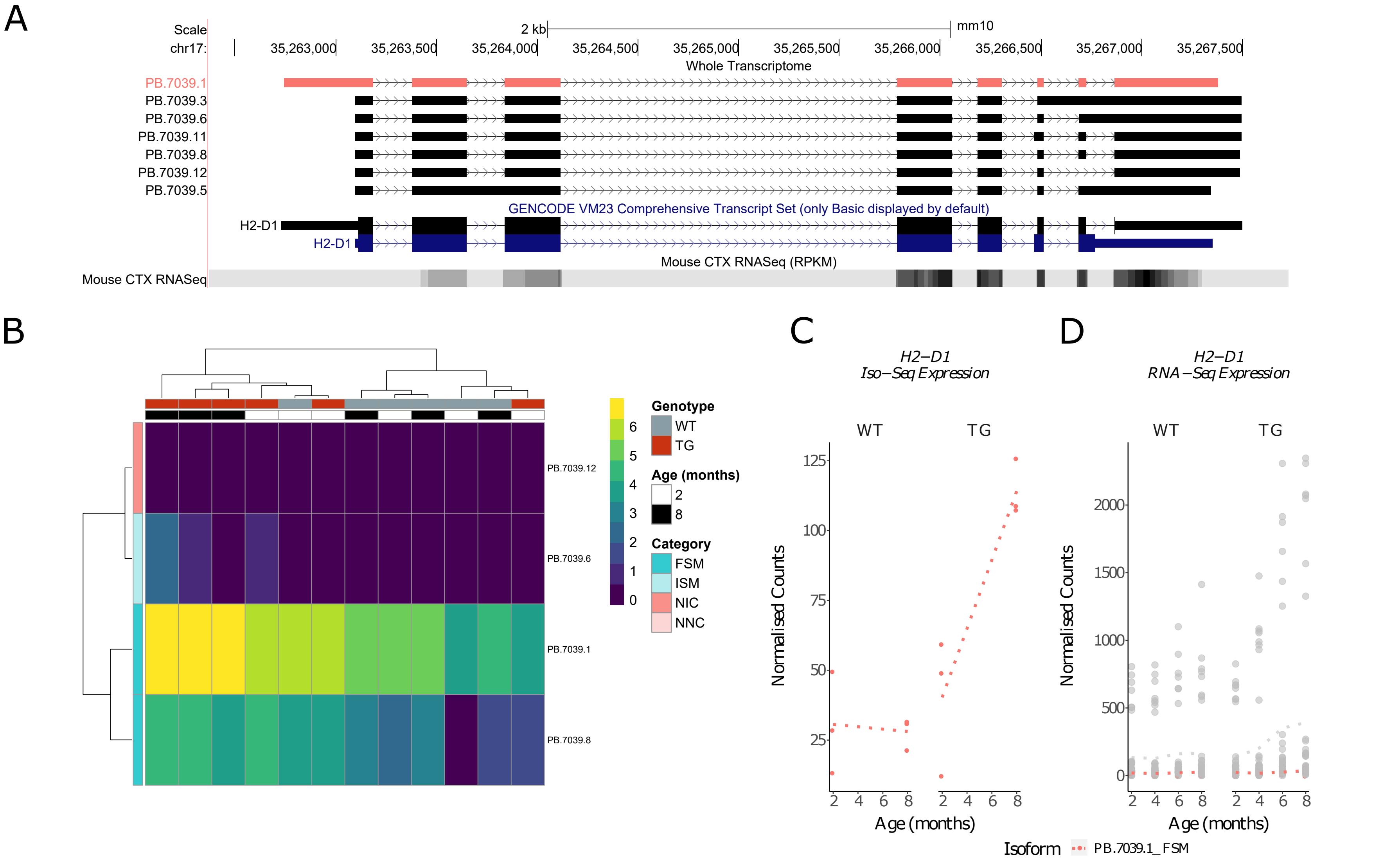


Figure 5.12: Significant up-regulation of *H2-D1-202* with progression of tau pathology in rTg4510 mice. Shown are **(A)** UCSC genome browser tracks of isoforms annotated to *H2-D1*, **(B)** hierarchical clustering of *H2-D1*-associated isoforms by abundance (Iso-Seq FL read count, log2), **(C)** differentially expressed transcript (H2-D1-202, ENSMUST00000172785.7, PB.7039.1) identified using normalised Iso-Seq read and **(E)** RNA-Seq read counts. Grey dots denote to differentially expressed transcripts identified using RNA-Seq but not Iso-Seq reads for quantification. Dotted lines represent the mean paths across age.

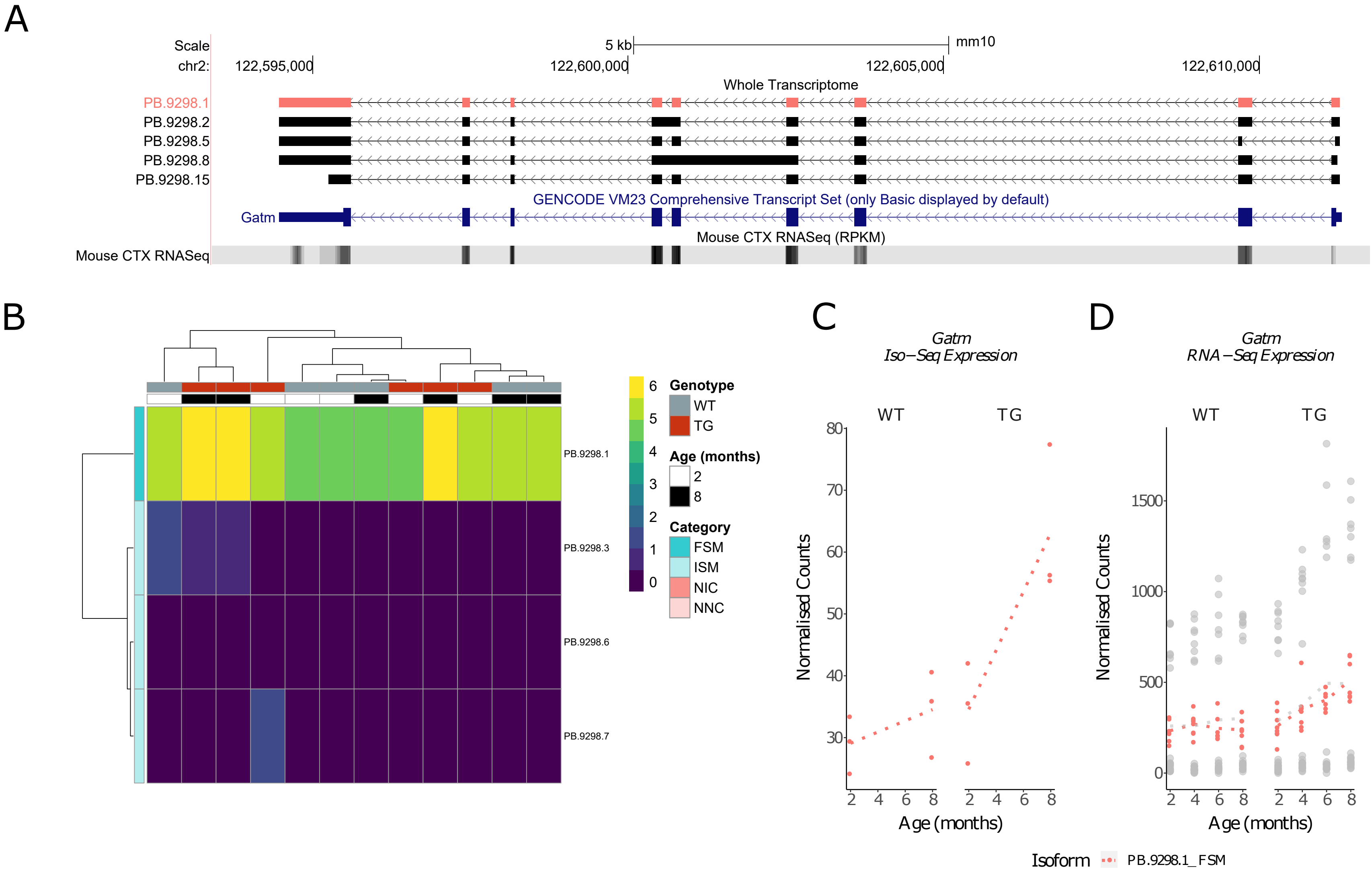


Figure 5.13: Significant up-regulation of *Gatm-201* with progressive tau pathology in rTg4510 mice. Shown are **(A)** UCSC genome browser tracks of isoforms annotated to *Gatm*, **(B)** hierarchical clustering of *Gatm*-associated isoforms by abundance (Iso-Seq FL read count, log2), **(C)** differentially expressed transcript (Gatm-201, ENSMUST00000028624.8, PB.9298.1) identified using normalised Iso-Seq read and **(E)** RNA-Seq read counts. Grey dots denote to differentially expressed transcripts identified using RNA-Seq but not Iso-Seq reads for quantifications. Dotted lines represent the mean paths across age.

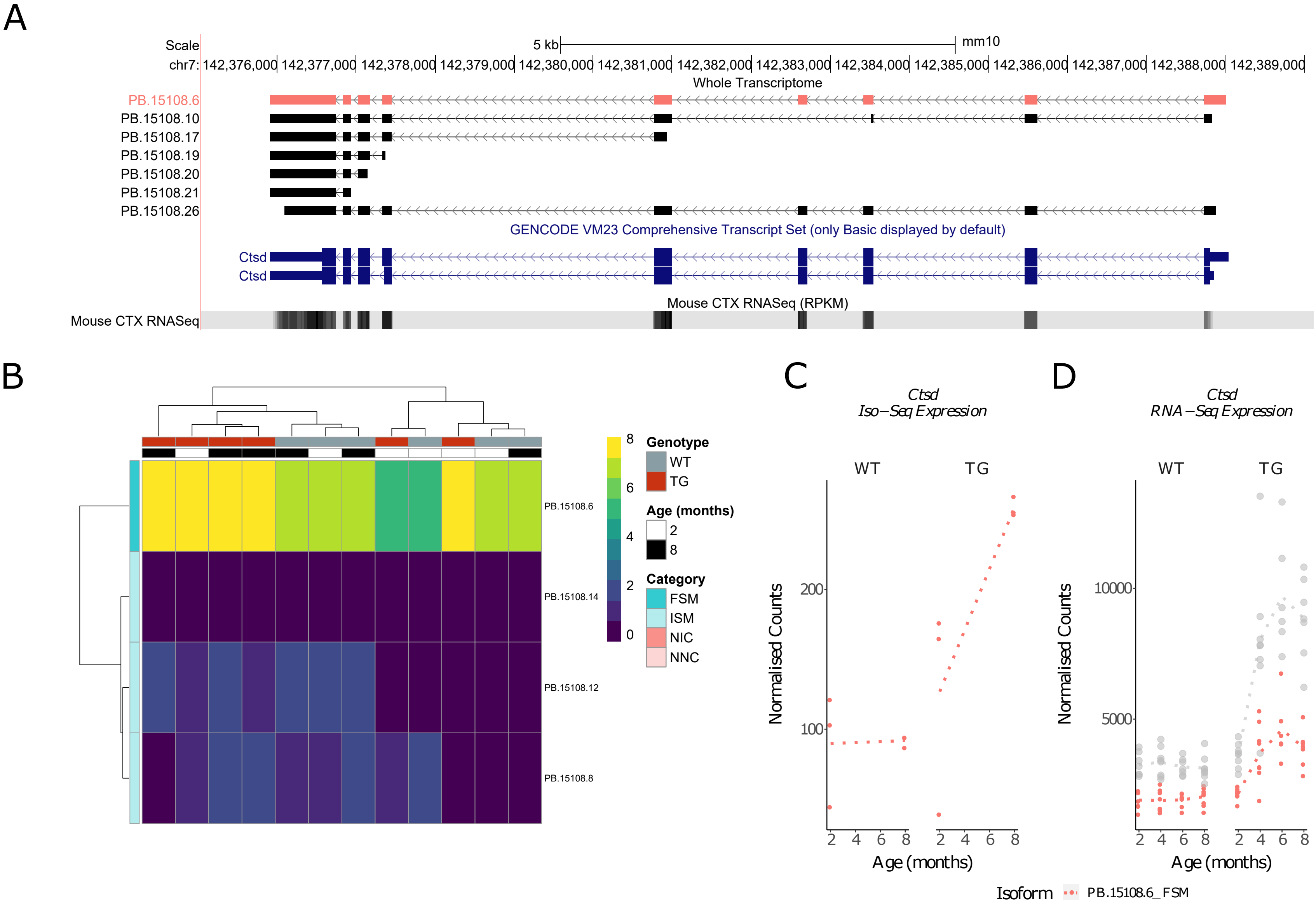


Figure 5.14: Significant up-regulation of *Ctsd-202* with progressive tau pathology in rTg4510 mice. Shown are **(A)** UCSC genome browser tracks of isoforms annotated to *Ctsd*, **(B)** hierarchical clustering of each *Ctsd*-associated isoform based on abundance (Iso-Seq FL read count, log2), **(C)** differentially expressed transcript (Ctsd-202, ENSMUST00000151120.8, PB.15108.6) identified using normalised Iso-Seq read and **(E)** RNA-Seq read counts. Grey dots denote to differentially expressed transcripts identified from RNA-Seq but not Iso-Seq reads. Dotted lines represent the mean paths across age.

Despite the demonstrated utility of using long reads for differential expression analysis, we found that the expression differences for the majority of Iso-Seq-identified differentially expressed transcripts (n = 545, 90.6%) were not recapitulated with normalised RNA-Seq read counts. This included the top ranked transcript, Ubqln1-201 (ENSMUST00000058735.11, PB.4255.13) annotated to *Ubqln1* and Cd34-201 (ENSMUST00000016638.7, PB.1036.2) annotated to *Cd34* (**Table 5.4**). Although both transcripts were up-regulated with progressive tau pathology in rTg4510 mice using Iso-Seq FL read counts, no significant transcript expression differences were identified with normalised RNA-Seq read counts (**Figure 5.15**). We suspected that this could be partially due to the relatively low sensitivity of RNA-Seq reads to differentiate these almost-identical transcripts. Deeper examination of the Iso-Seq expression profiles, however, further revealed that while there was a difference in mean expression, there was also a large variance due to the relatively small number of samples profiled. We further noted that the majority of Iso-Seq-identified differentially expressed transcripts (n = 497 transcripts, 82.1%) were very lowly-expressed (< 24 mean normalised FL reads, n = 12 samples).

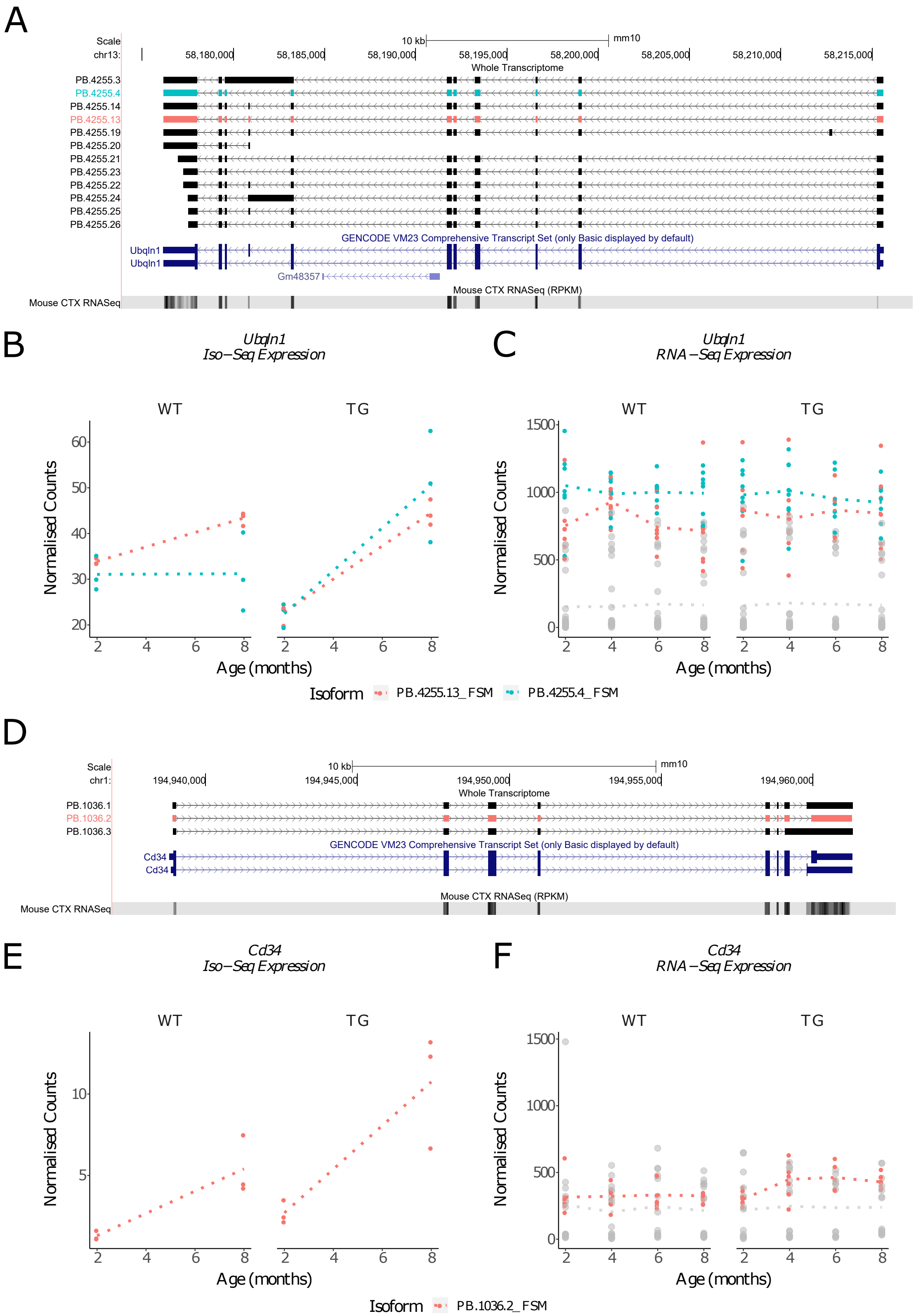


Figure 5.15: Disparities in differential transcript expression analysis. Shown are **(A)** UCSC genome browser tracks of the isoforms annotated to *Ubqln1*, with two isoforms identified as differentially expressed using **(B)** normalised Iso-Seq read counts but not using **(C)** RNA-Seq read counts for quantification. **(D)** UCSC genome browser tracks of isoforms annotated to *Cd34*, **(E)** *Cd34* Iso-Seq transcript expression profile and **(F)** *Cd34* RNA-Seq transcript expression profile are also shown. The colours between the tracks and scatter plots refer to the respective Iso-Seq-identified differentially expressed transcript.

5.3.8 Hybrid approach identifies tau-pathology associated differential transcript usage with major isoform switching events

Further complexity in transcriptional regulation is reflected in the fact that the expression of a gene may remain constant between conditions, but the *relative* expression of individual isoforms (and thus isoform *proportions*) may differ; this phenomenon is known as differential transcript usage (DTU) and is described in detail in **Section 3.3.3**. We therefore assessed whether the relative isoform abundance for each gene altered with rTg4510 genotype and/or age, and whether there was switching of the dominant major (highest expressed) isoform between experimental groups (major isoform switching).

Using Iso-Seq reads, we were not able to identify genes with differential transcript usage, likely reflecting the relatively low sequencing coverage and small sample size of our Iso-Seq experiments. In contrast, we identified 671 DTU genes (**Table 5.5**) when using normalised RNA-Seq read counts aligned to our Iso-Seq-derived transcriptome annotation (using the hybrid approach described in **Section 3.3.1**). Strikingly, the majority of these genes (n = 519, 77.3%), while characterised by a change in isoform proportions, were not differentially expressed at the gene level between WT and TG mice. We further identified 61 genes that were not differentially expressed but were identified with both DTU and major isoform switching. This indicates that a significant degree of post-transcriptional regulation was independent of gene expression regulation. These genes (n = 580) were enriched as targets for a number of transcription factors (listed in **Table 5.6**), particularly TAF1 (adjusted P = 1.10 x 10^{-6}, odds ratio = 1.77), which is a key component of the pre-initiation complex that initiates RNA polymerase II transcription.[291]

Table 5.5: Summary of differential expression and splicing analyses. A summary of the number of genes identified as differentially expressed, and characterised with differential transcript usage and major isoform switching. Expression was determined using normalised RNA-Seq counts after alignment to the Iso-Seq-derived transcriptome.

Conditions			Number of genes
Differential gene expression	Differential transcript usage	Major isoform switching	
✓	✓	✓	16
✓	✓	x	75
x	✓	✓	61
x	✓	x	519
Total number of genes			671

Table 5.6: Transcription factor terms for differentially spliced genes. Tabulated is a list of the top 10 transcription factor terms ("ENCODE and ChEA Consensus TFs from ChIP-X") for genes (n = 580) identified with differential isoform usage but no differential gene expression. Gene ontology analysis was performing using *Enrichr.*

Term	Adjusted *P*	Odds ratio
TAF1	1.10 x 10^{-6}	1.771
MAX	5.78 x 10^{-6}	1.866
UBTF	7.76 x 10^{-5}	1.843
YY1	9.44 x 10^{-5}	1.645
RCOR1	1.68 x 10^{-4}	2.219
RUNX1	3.10 x 10^{-4}	1.834
BRCA1	3.10 x 10^{-4}	1.544
E2F1	3.10 x 10^{-4}	2.021
SP2	1.28 x 10^{-3}	1.846
NFYA	1.72 x 10^{-3}	1.549

The top gene characterised by a major isoform switch was *Cisd3* (FDR = 3.26 x 10^{-26}, **Figure 5.16A**), a mitochondrial iron-sulphur domain-containing protein involved in regulating iron homeostasis essential for mitochondrial function.[292] While there was little difference in overall gene expression (**Figure 5.16B**), a major isoform switch was observed between rTg4510 genotype that was consistent across all ages (**Figure 5.16C-E**). The two known isoforms (Cisd3-201 and Cisd3-202) involved in this switch only differed at the 5' end by the presence of an intron retention (IR) event occurring between exons 1 and 2. ORF predictions revealed that this IR event generated a shortened reading frame, which could translate to a different N-terminal peptide sequence (**Figure 5.16A**). Cisd3-201 (ENSMUST00000107583.2, PB.2833.2), which contained this IR event, was up-regulated with progressive tau pathology, while Cisd3-202 (ENSMUST00000107584.7, PB.2833.1) was down-regulated.

Another gene with significant DTU but no difference in gene expression was *Shisa5* (FDR = 1.26 x 10^{-11}, **Figure 5.17A**), a transmembrane that modulate both Wnt and FGF signalling by inhibiting their maturation and trafficking to the cell surface.[293] An example of the compensatory mechanism that is sometimes observed with differential transcript expression, we identified a gradual isoform shift associated with progressive tau pathology (**Figure 5.17D,E**). The two isoforms of interest differed significantly in length due to the usage of an alternative promoter (**Figure 5.17A**); the longer isoform (Shisa5-201, ENSMUST00000026737.11, PB.16934.2) spanned the full-length of the gene across all six exons, whereas the shorter isoform (Shisa5-203, ENSMUST00000154184.4, PB.16934.9) lacked the first three upstream exons but contained an alternative first exon. Unsurprisingly, ORF prediction revealed a significant disparity in the ORF length with the longer isoform containing the whole Shisa Pfam domain. While the shorter isoform was the dominant isoform in wild-type mice across all ages, we observed a down-regulation of this isoform coupled with an up-regulation of the longer transcript in rTg4510 TG mice (**Figure 5.17C,D**), resulting in a zero net change in gene expression (**Figure 5.17B**).

In addition to identifying genes with DTU but no overall differences in gene expression, we also identified genes with evidence for altered gene expression accompanied with differential transcript expression and major isoform switching. This included *Fblim1* (FDR = 1.87 x 10^{-13}, **Figure 5.18A**) - a gene that encodes for a filamin-binding protein involved in actin filament assembly and cell adhesion[294] - which was up-regulated with progressive tau pathology in rTg4510 transgenic mice. Drawing parallels to *Gfap* and *C4b*, increased *Fblim1* gene expression was also primarily driven by one isoform (**Figure 5.18B**). However, this isoform was the less abundant, minor known isoform, Fblim1-203 (ENSMUST00000105785.8, PB.11626.1) found in wild-type mice rather than the dominant known isoform, Fblim1-202 (ENSMUST00000105784.7, PB.11626.2) (**Figure 5.18D,E**). Given that detection of Fblim1-203 was negligible in wild-type mice, a strong increase in this isoform paralleling tau accumulation resulted in a robust isoform switch (**Figure 5.18C**). Characterisation of these two isoforms revealed that they only differed by the presence of an alternative first exon, with the reading frame being broadly similar (**Figure 5.18A**).

Finally, we identified a novel fusion gene (this phenomenon is described in **Section 4.3.5**), *Arpc4-Ttll3* that was characterised by altered transcript usage and major isoform switching (FDR = 1.28 x 10^{-4}). Absent in the mouse reference genome annotation, three read-through transcripts (PB.13540.3, PB.13540.7, PB.13540.8) were annotated across the full-length of *Arpc4* - which encodes one of the subunits of the Arp2/3 protein complex involved in actin polymerisation - and *Ttll3* - which encodes the tubulin tyrosine ligase-like 1 (**Figure 5.19A**). The three transcripts differed by only 45bp at the first exon and the presence of exon 13 (exon 8 in *Ttll3*) in PB.13540.7, which was skipped in the other two isoforms (**Figure 5.19A**). While there was no difference in overall gene expression (**Figure 5.19B**), we observed an isoform switch with PB.13540.7 up-regulation and PB.13540.3 down-regulation in rTg4510 TG mice (**Figure 5.19C**), over time (**Figure 5.19D,E**). ORF predictions showed that skipping of exon 13 maintained the reading frame, although the predicted frame for all three transcripts only covered *Ttll3* rather than spanning across both genes (**Figure 5.19A**).

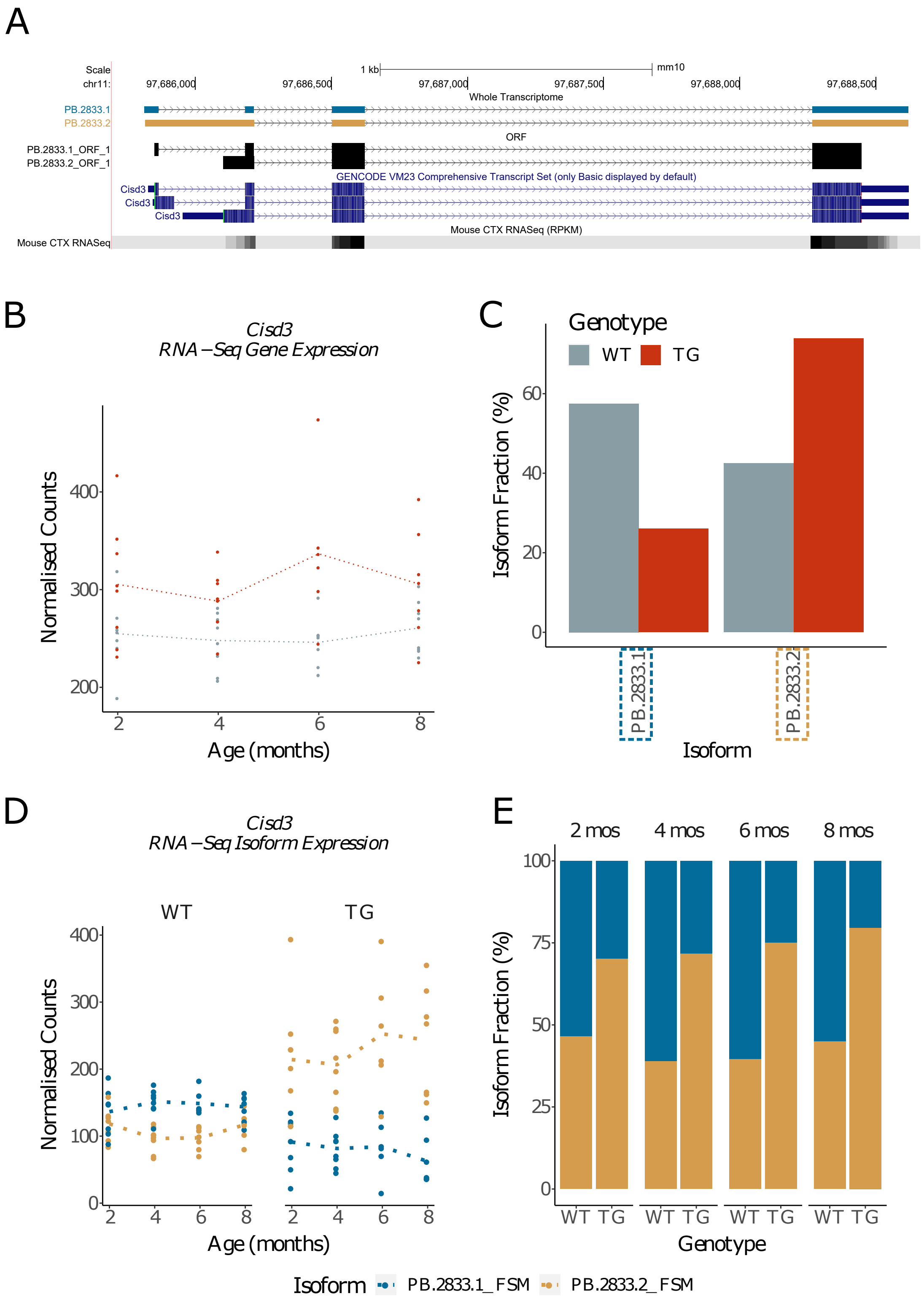


Figure 5.16: Differential transcript expression and usage of *Cisd3* with progression of tau pathology in rTg4510 mice. Shown are **(A)** UCSC genome browser tracks of the isoforms annotated to *Cisd3* with the two differentially expressed isoforms colour-coded and their respective predicted open reading frame (black), **(B)** *Cisd3* gene expression, **(C)** proportion of isoform usage with rTg4510 genotype, independent of age, **(D)** *Cisd3* transcript expression, and **(E)** proportion of isoform usage by age and genotype. Expression is determined from normalised RNA-Seq read counts after alignment to Iso-Seq-derived transcriptome.

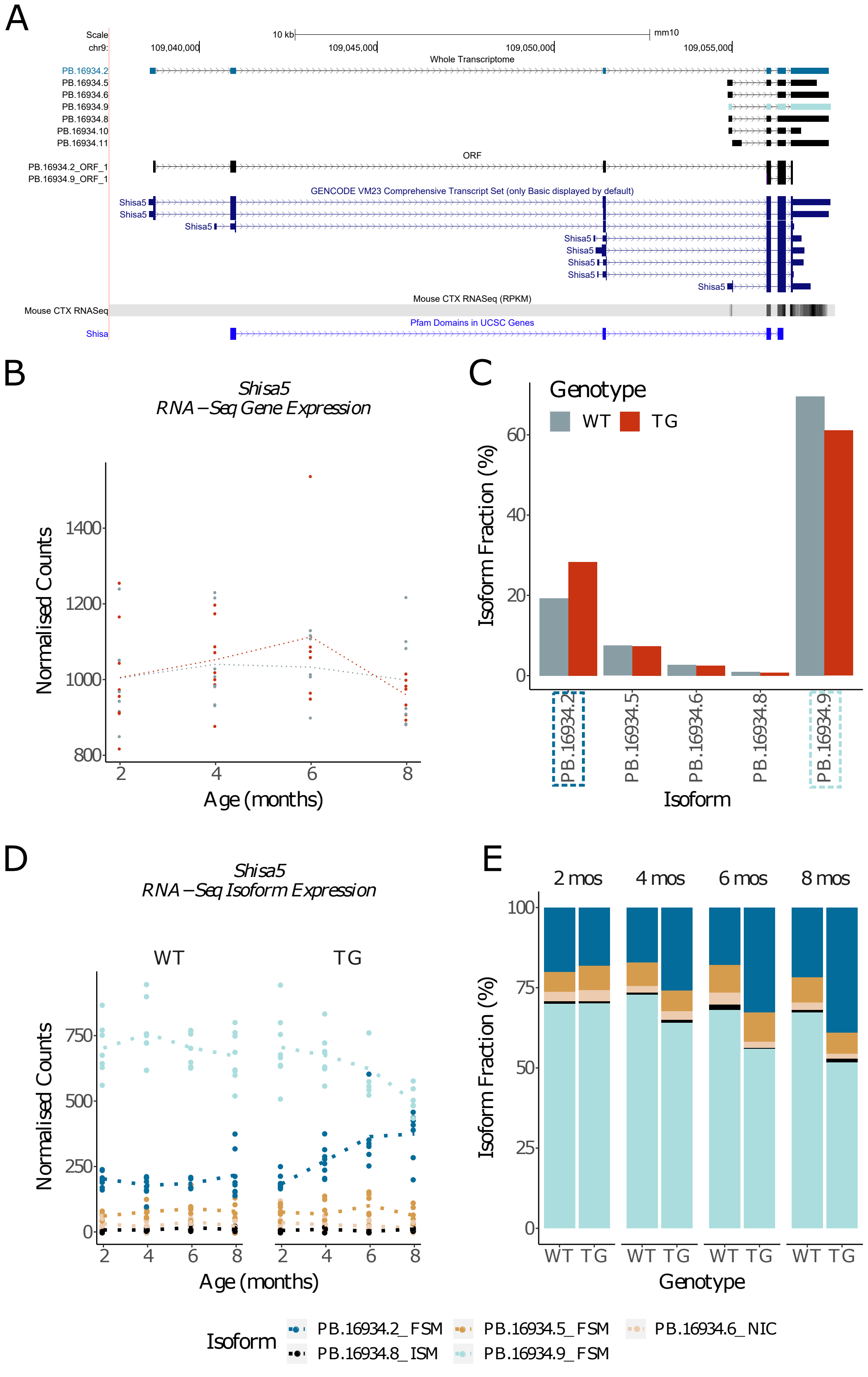


Figure 5.17: Differential transcript expression and usage of *Shisa5* with progression of tau pathology in rTg4510 mice: Shown are the **(A)** UCSC genome browser tracks of the isoforms annotated to *Shisa5* with the two differentially expressed isoforms colour-coded, **(B)** *Shisa5* gene expression, **(C)** proportion of isoform usage with rTg4510 genotype, independent of age, **(D)** *Shisa5*-associated transcript expression and **(E)** proportion of isoform usage by age and genotype. Expression is determined from normalised RNA-Seq read counts after alignment to Iso-Seq-derived transcriptome.

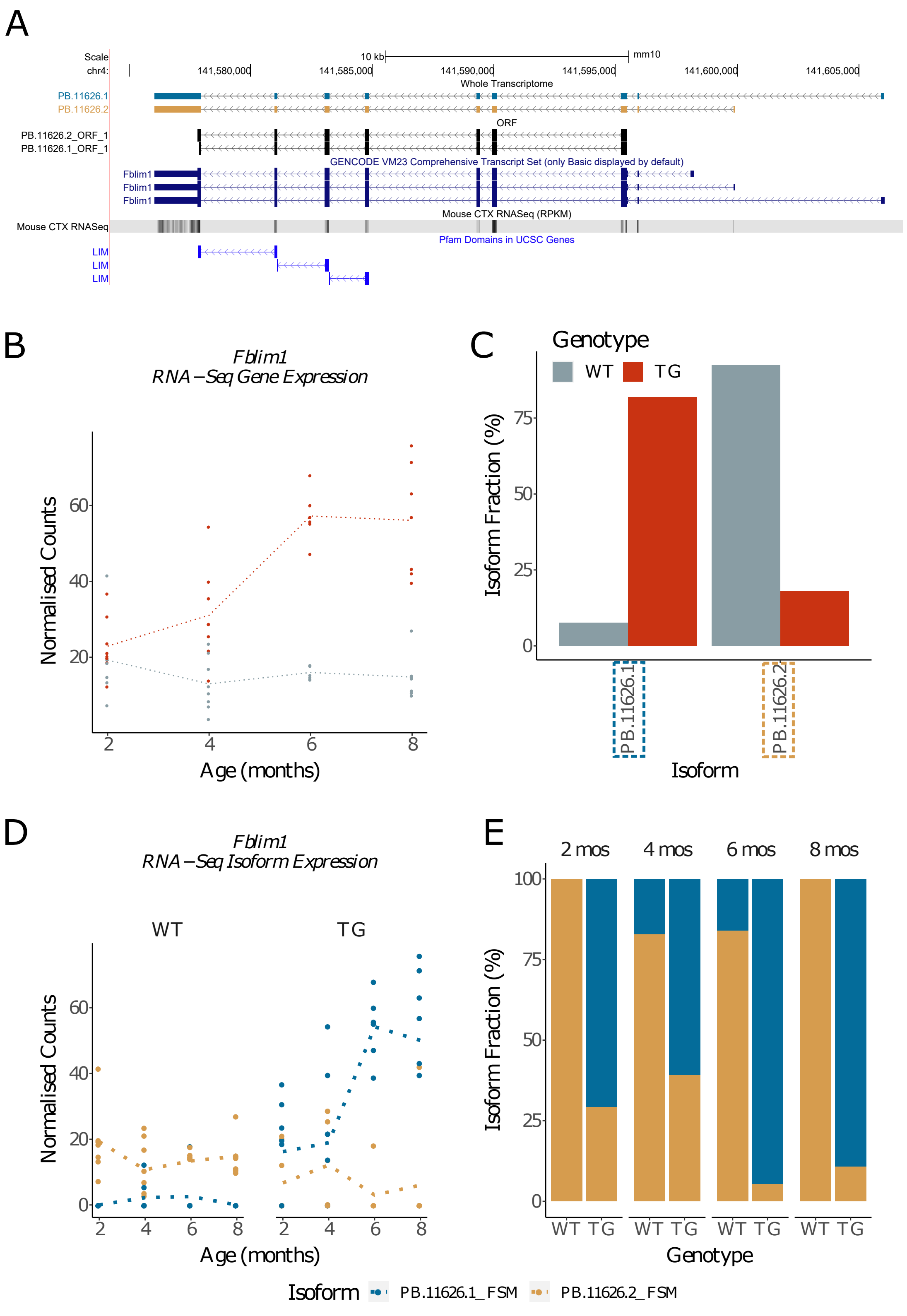


Figure 5.18: Differential transcript expression and usage of *Fblim1* with progression of tau pathology in rTg4510 mice: Shown are the **(A)** UCSC genome browser tracks of the isoforms annotated to *Fblim1* with the two differentially expressed isoforms colour-coded and their respective predicted open reading frame (black), **(B)** *Fblim1* gene expression, **(C)** proportion of isoform usage with rTg4510 genotype, independent of age, **(D)** *Fblim1*-associated transcript expression and **(E)** proportion of isoform usage by age and genotype. Expression is determined from normalised RNA-Seq read counts after alignment to Iso-Seq-derived transcriptome.

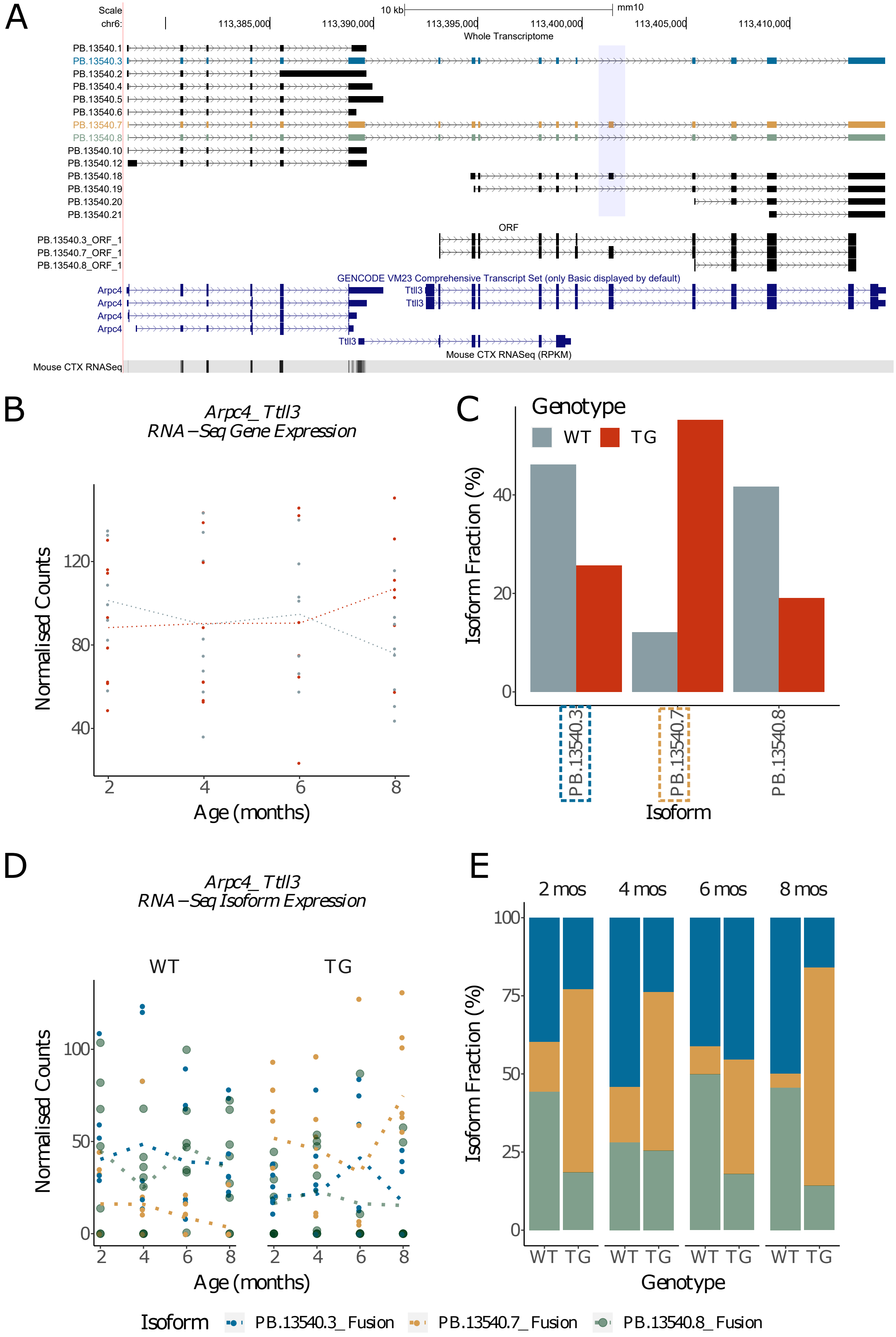


Figure 5.19: Differential transcript expression and usage of *Arpc4-Ttll3* with progressive of tau pathology in rTg4510 mice. Shown are **(A)** UCSC genome browser tracks of the three fusion transcripts annotated to *Arpc4-Ttll3* with exon skipping (purple) and their respective predicted open reading frame (black), **(B)** *Arpc4-Ttll3* gene expression, **(C)** proportion of isoform usage with rTg4510 genotype **(D)** expression of the fusion transcripts and **(E)** proportion of isoform usage by age and genotype. Expression is determined from normalised RNA-Seq read counts after alignment to Iso-Seq-derived transcriptome.

5.4 Discussion

In this chapter, we leveraged the power of long-read sequencing to identify transcriptional and splicing differences associated with progressive tau pathology in a transgenic mouse model. To our knowledge, this represents the first comprehensive long-read sequencing dataset generated on a mouse model of tau pathology, facilitating the accurate interrogation of cortical expression alterations at the gene and transcript level.

5.4.1 Overview of results

Demonstrating the dual utility of long reads for isoform annotation and quantification, we identified widespread gene expression differences paralleling the development of tau pathology in rTg4510 mice. At the gene-level, these results broadly recapitulated findings from our previous RNA-Seq study,[90] and implicate the role of transcriptional dysregulation in AD development. With the capacity to detect full-length transcripts, our study was further powered to identify robust genotype-associated differences in isoform expression. Notably, we found that the differential expression of two well-established AD-associated genes, *Gfap* and *C4b*, were primarily driven by the robust up-regulation of their dominant isoform.

Although previous studies of gene expression have provided key insights into the molecular mechanisms driving AD pathogenesis,[90, 124, 126, 137] these studies fail to capture the dynamics in the expression of specific isoforms, particularly for genes where there are no overall gene expression differences. By complementing the improved isoform annotation provided by the long-read sequencing data with the deep sequencing depth achieved from short-read RNA-Seq data, we revealed a number of genes characterised with differential transcript usage ("isoform switching") but no significant difference in global (gene-level) expression. Of note, these genes were involved in key pathways implicated in AD pathology, and included i) *Shisa*, a modulator of the FGF and Wnt signalling pathways, which is essential for neuronal survival and is known to be suppressed in AD brains,[295] ii) *Cisd3*, a member of the same CDGSH domain-containing family as *Cisd2*, which was recently identified as a promising new target in AD due to its neuroprotective role of the mitochondria against Aβ accumulation,[296] and iii) *Arpc4*, which was recently found to co-aggregate with phosphorylated tau in NFTs extracted from AD patients[297] and in synaptosomes isolated from AD mouse models.[298] Among these genes, we detected altered splicing and isoform switches with potential functional consequences.

5.4.2 Limitations

Our results should be interpreted in the context of several limitations. Firstly, we performed long-read sequencing on a relatively small number of mouse samples. Although we found strongly consistent patterns of alternative splicing across biological replicates, we were unable to achieve the depth required to fully recapitulate the tau-associated transcriptional differences without relying on the deep RNA-Seq expression generated from matched samples. Notably, this hybrid approach still suffers from a degree of ambiguous alignment. We also observed that the relatively low sequencing depth afforded by global transcriptome profiling limits the power to perform differential splicing analysis. While we were able to detect differentially expressed isoforms using full-length read counts, we were not powered to identify changes in isoform usage. This analysis requires reliable quantification of the relative proportion of isoforms, which is dependent on detecting all the isoforms, including the rare novel isoforms, present in an RNA sample. Future work will aim to extend our analyses by sequencing larger numbers of samples and at a deeper sequencing depth. Profiling of the same samples using another long-read sequencing platform will also be useful to comprehensively investigate and validate the transcriptional variation associated with AD pathology.

Secondly, our analyses were performed on bulk entorhinal cortex tissue, comprising a heterogeneous mix of neurons, oligodendrocytes and other glial cell-types. Despite compelling evidence from recent studies reporting cell- and disease-specific transcriptional signatures, we were unable to explore these differences in our datasets. This challenge can be addressed using a combined approach of single-cell sequencing and long-read sequencing, as shown in recent studies (reviewed in **Table 1.6**) - however, this strategy is currently limited to achieve the depth required to detect reliable disease- and cell-specific splicing variations. Future work would build on recent methodological developments by our group that facilitate the purification of nuclei from different neural cell types prior to genomic profiling.[299]

Furthermore, while isoform expression was normalised for the library sequencing depth, our analyses did not account for differences in cellular composition between WT and rTg4510 TG mice. Given neuronal loss and astrogliosis are prominent hallmarks of AD pathogenesis, we were unable to discern whether transcriptional variations (i.e. up-regulation of astrocyte markers) are a direct consequence of AD-associated transcriptional regulation or a reflection of changes in cell composition. Future work should aim to use single-cell RNA-Seq data generated on similar samples (if available) or publicly-available single-cell datasets[192] to infer cell type proportions in our bulk transcriptomic datasets.

Finally, we only profiled entorhinal cortex tissue from female mice in order to minimise heterogeneity in our analyses. While the entorhinal cortex is an invariant focus of studying AD pathogenesis, as one of the first regions of the brain to be affected, tissue-specific differences in splicing and isoform usage of AD-risk genes have been previously reported.[300] A number of sex differences have also been reported, with female mice exhibiting earlier and more severe cognitive and behavioural impairments than transgenic male mice.[194] Future work should cross-examine results from our study with transcriptional variation in other tissue types and male mice for a more comprehensive understanding of the development of tau pathology in the AD brain.

5.4.3 Conclusion

In summary, our study revealed transcriptional and splicing differences in the entorhinal cortex associated with tau accumulation. Importantly, we identified changes in key isoforms that drive the altered expression of AD-associated genes, with evidence of isoform switching events that could have important functional consequences. Our results demonstrate the utility of long-read sequencing data for isoform-level analyses, facilitating the detection of splicing alterations underlying the development of AD pathology.

Chapter 6

Isoform landscape of AD-associated genes from targeted profiling of tau mouse model

6.1 Introduction

Long-read sequencing of the whole transcriptome at a global level provides valuable insights into the role of splicing and RNA isoforms in health and disease.[301] In generating unambiguous, full-length isoforms, we have demonstrated the dual utility of long-reads for comprehensive isoform annotation and identification of differentially expressed genes (**Section 5.3.4**) and isoforms (**Section 5.3.7**) in a mouse model of AD tauopathy, rTg4510. However, in comparison to short-read sequencing platforms, the relative low sequencing depth associated with this approach results in lower sensitivity to quantify (or even detect) lowly-expressed transcripts.[147] This was demonstrated by the missed detection of lowly-abundant ERCC molecules in **Chapter 4** (n = 30, 32.6% of ERCC molecules, **Figure 4.7**), despite saturation of sample size (n = 12 samples), indicating a biased sampling of the more abundant molecules. Increasing the sample size without increasing the number of reads per sample is therefore unlikely to make any difference in detecting the more lowly-expressed transcripts on a per sample basis. However, the number of reads per sample is currently limited to the technology, particularly the number of wells available for sequencing (1M ZMWs in PacBio Sequel I, as described in **Section 3.1**).

One established solution to circumvent this low sequencing coverage of rare transcripts is to target or enrich for transcripts associated with a gene of interest (i.e. the “target gene”), and perform targeted sequencing.[212] This can be achieved primarily in two ways:[301] i) Amplicon sequencing, which involves long-range PCR across target genes with primers designed to the 5’ and 3’ UTR (**Figure 6.1A**), and ii) CaptureSeq, which utilises a pool of oligonucleotide probes designed to sequences unique to the target genes for hybridisation-based enrichment (**Figure 6.1B**). While amplicon sequencing enables extremely deep coverage of target genes, including longer isoforms, this approach generates a lower throughput, is associated with increased PCR amplification bias, and is typically applied to a small number of genes (1 - 2 genes from previous profiling studies[152, 153]). In contrast, CaptureSeq can be applied to multiple genes of interest in parallel (theoretically unlimited, although the number of target genes is negatively correlated to the sequencing coverage achieved per target gene), reducing cost and easing library preparation. Of note, the CaptureSeq is incorporated into the official Iso-Seq protocol and is recommended by PacBio as an approach for targeted sequencing.[148]

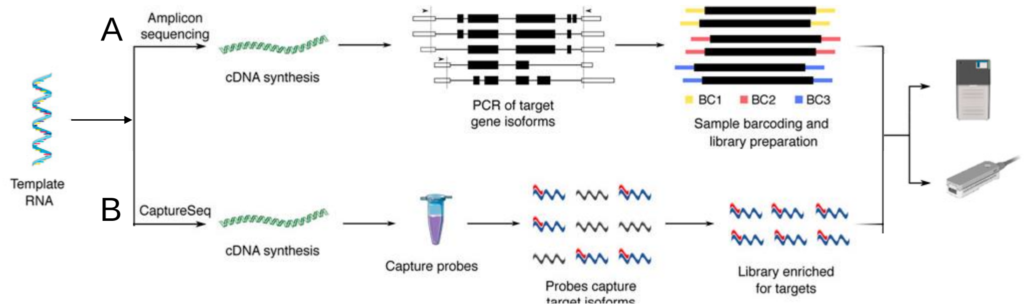

A
Amplicon sequencing
cDNA synthesis
PCR of target gene isoforms
BC1
BC2
BC3
Sample barcoding and library preparation
Template RNA
B
CaptureSeq
cDNA synthesis
Capture probes
Probes capture target isoforms
Library enriched for targets

Figure 6.1: Lab approaches for targeted profiling. Shown is a schematic figure describing two commonly used methods for targeted long-read sequencing: **(A)** Amplicon sequencing and **(B)** CaptureSeq. Due to greater flexibility, we used CaptureSeq (hybridisation-based enrichment with custom designed IDT probes) to enrich and sequence 20 AD-associated genes in the rTg4510 cortex. More details can be found in **Section 3.1.2.3**. Figure is taken and adapted from De Paoli-Iseppi et al. (2021).[301]

.

Both targeted sequencing approaches have been implemented in recent studies[148, 152, 165] to comprehensively survey the isoform landscape of disease-associated genes, including *CACNA1C* (schizophrenia-associated risk gene),[165] *NRXN1*[152] (implicated in several neurodevelopmental disorders), and *SNCA*,[148] with notable success. Nanopore sequencing of *CACNA1C* further identified a pronounced isoform switch in the cerebellum compared to other cortical brain regions from using normalised full-length read counts,[165] highlighting the power of targeted sequencing to achieve sufficient depth required for detectable differential isoform usage.

Given the demonstrated success of targeted long-read sequencing to identify disease-gene specific isoforms, this chapter focuses on comprehensively characterising the isoform landscape of 20 AD-associated genes (**Figure 6.2**, **Table 6.1**). Of note, this significantly covers a wider scope than previous studies (as described above), which only focused on characterising a single gene. These 20 AD-associated genes (hereby also referred to as "target genes") have been previously implicated in various molecular mechanisms underpinning AD pathogenesis, with evidence for altered splicing (detailed and reviewed in **Table 6.2**). By performing targeted profiling of these 20 well-known AD-associated genes using the CaptureSeq approach (as described in **Section 3.1.2.3**), we aimed to comprehensively characterise the transcriptional and splicing changes of these genes, and to test for associations with progressive tau pathology in the rTg4510 mouse model. The objectives of this chapter were as follows:

1. To enrich and sequence transcripts from 20 AD-associated genes in the rTg4510 mouse model at four time points (2, 4, 6, and 8 months) with PacBio long-read sequencing (hereby referred to as "Iso-Seq targeted dataset").
2. To validate the Iso-Seq targeted dataset by sequencing a subset of samples using targeted ONT nanopore cDNA sequencing (hereby referred to as "ONT targeted dataset").
3. To compare the isoform landscape of AD-associated genes observed in the Iso-Seq global transcriptome dataset (generated in **Chapter 4**, hereby referred to as "Iso-Seq global dataset") with Iso-Seq targeted dataset from the same samples.
4. To compare the isoform landscape of the AD-associated genes in the Iso-Seq and ONT targeted datasets.
5. To comprehensively characterise isoform diversity and splicing events for AD-associated genes in the rTg4510 mouse model.
6. To perform differential isoform-based analysis (differential transcript expression and differential transcript usage) for target genes to identify transcriptional and splicing differences between rTg4510 WT and TG mice.

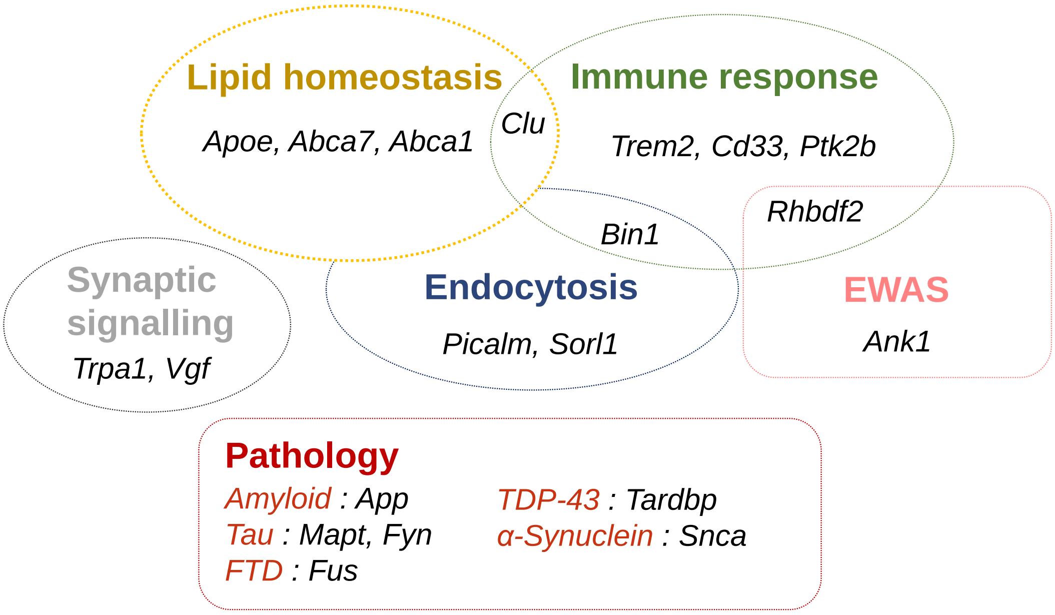


Figure 6.2: Targeted profiling of 20 AD-associated genes in the rTg4510 cortex. In this chapter, we performed targeted sequencing of 20 AD-associated genes (classified here by molecular pathway) in the rTg4510 mouse model. EWAS - Epigenome-wide association study.

Apoe, Abca7, Abca1, Picalm, Sorl1, Clu, Bin1, Trem2, Cd33, Fus and *Ptk2b* were selected for enrichment as they are well-known AD-risk genes identified from various GWAS and TWAS (depicted in **Figure 1.3**). *App, Mapt, Snca* and *Tardbp* were chosen due to their role and relevance in key hallmarks of AD pathology (described in **Section 1.1.3**). *Trpa1, Fus* and *Vgf* were requested by Eli Lilly & Company Ltd. given there is increasing evidence for their role in AD pathology (described later in **Table 6.2**). Finally, we included *Ank1* and *Rhbdf2*, which have been consistently identified as differentially-methylated from EWAS; inclusion of these two genes would allow us to further study the interactions between epigenetics and splicing in AD.

Table 6.1: Transcriptional features of AD-associated genes selected for targeted profiling. Tabulated is a list of the 20 AD-associated genes selected for targeted sequencing, with a summary of the respective gene length, transcript length, number of isoforms and exons taken from the mouse reference genome (mm10, GENCODE, vM22).

Abca1 - ATP-binding cassette subfamily A member 1, *Abca7* - ATP-binding cassette subfamily A member 7, *App* - Amyloid precursor protein, *Bin1* - Bridging integrator 1, *Clu* - Clusterin, *Fus* - Fused in sarcoma, *Mapt* - Microtubule-associated protein tau, *Picalm* - Phosphatidylinositol binding clathrin assembly protein, *Ptk2b* - Protein-tyrosine kinase 2-beta, *Rhbdf2* - Rhomboid 5 homolog 2, *Sorl1* - Sortilin related receptor 1, *Tardbp* - TAR DNA-binding protein, *Trem2* - Triggering receptor expressed on myeloid cells 2, *Trpa1* - Transient receptor potential ankyrin 1, *Vgf* - VGF nerve growth factor inducible.

Target gene	Genome co-ordinates[a]	Gene length (kb)	Number of known isoforms[a]	Transcript length[a] (min - max, kb)	Number of exons[a] (min - max)	Expression[b]
Abca1	chr 4 : 53030670 - 53160014	129.108	2	0.769-10.212	1-50	983.29
Abca7	chr 10 : 79997615 - 80015572	19.078	3	6.544-6.649	1-47	319.14
Ank1	chr 8 : 22974836 - 23150497	175.612	17	0.325-8.321	1-46	1038.15
Apoe	chr 7 : 19696125 - 19699285	3.079	11	0.453-1.404	1-5	27354.1
App	chr 16 : 84954317 - 85173826	224.081	11	0.654-8.149	1-18	21953.7
Bin1	chr 18 : 32377217 - 32435740	58.492	6	0.533-2.676	1-19	2556.89
Cd33	chr 7 : 43528610 - 43533290	16.054	6	0.4-5.722	1-8	93
Clu	chr 14 : 65968483 - 65981545	13.064	9	0.553-1.801	1-9	17854.89
Fus	chr 7 : 127967479 - 127982032	18.244	16	0.35-5.521	1-15	2214.98
Fyn	chr 10 : 39369799 - 39565381	195.617	9	0.329-3.548	1-14	2077.36
Mapt	chr 11 : 104231436 - 104332096	100.7	12	0.289-5.243	1-13	7739.63
Picalm	chr 7 : 90130232 - 90209447	83.25	15	0.33-8.188	1-21	1695.2
Ptk2b	chr 14 : 66153138 - 66281171	127.795	8	1.278-4.003	1-31	5506.74
Rhbdf2	chr 11 : 116598082 - 116627138	28.85	4	0.81-3.836	1-19	23.4
Snca	chr 6 : 60731454 - 60829974	98.28	4	0.565-1.403	1-6	3716.69
Sorl1	chr 9 : 41968370 - 42124408	159.577	4	0.221-10.667	1-48	3372.8
Tardbp	chr 4 : 148612263 - 148627115	14.637	30	0.356-7.471	1-10	1337.29
Trem2	chr 17 : 48346401 - 48352276	7.707	4	0.523-4.78	1-5	301.92
Trpa1	chr 1 : 14872529 - 14918981	46.214	2	3.262-4.236	1-27	5.77
Vgf	chr 5 : 137030295 - 137033351	6.959	4	0.579-2.829	1-5	1849.63

[a] According to the mouse reference genome (mm10, GENCODE vM22)

[b] Gene expression in rTg4510 cortex (n = 30 WT, n = 29 TG) derived from normalised RNA-Seq data[90]

Table 6.2: Role and relevance of 20 AD-associated target genes in AD pathogenesis. Tabulated is a detailed review of the role and relevance of the 20 selected AD-associated genes, which were enriched for targeted sequencing in the rTg4510 mouse model.
* Details of mouse models can be found in **Table 1.1**, where stated.
CSF - Cerebrospinal fluid , EWAS - Epigenome-wide association study, KPI - Kunitz-type protease inhibitor domain, GWAS - Genome-wide association study, Lof - Loss-of-function, NMD - Nonsense-mediated decay, TG - Transgenic mice, TWAS - Transcriptome-wide association study, WT - Wild-type mice.

Gene	Pathway	Function	Role and relevance in AD pathogenesis from human and mouse model studies*
Abca1	Lipid homeostasis	Transmembrane protein for cholesterol efflux to apolipoprotein	• **Genetics**: Identification of rare non-synonymous variants in controls vs AD cases, including LoF mutation Asp1800His (N1800H) which is strongly associated with increased AD risk.[302] • **Pathology**: *ABCA1* expression linked to ApoE isoform-specific and Aβ clearance; *ABCA1* deletion in amyloid mouse model resulted in decreased ApoE and increased Aβ accumulation, whereas overexpression prevented Aβ aggregation.[303] *ABCA1* haploinsufficiency in APP/PS1 mice significantly exacerbated memory deficits and reduced Aβ clearance in Apoe-E4 expressing mice but not in Apoe-E3.[304]
Abca7	Lipid homeostasis	Transmembrane protein for cholesterol efflux to apolipoprotein	• **Genetics**: Rare LoF variants associated with aberrant mRNA splicing, including generation of intron-retained transcripts predicted for NMD,[305–307] aberrant 14bp extension of exon 41 in human AD brains[305,308] and a 44bp deletion predicting a frameshift mutation from rs142076058 SNP.[309]
Ank1	EWAS	Scaffolding proteins for linking membrane proteins to cytoskeleton	• **Epigenetics**: *ANK1* hypermethylation in AD post-mortem brain tissues.[310,311] • **Expression**: 4-fold increase in mRNA expression in microglia, but not in neurons or astrocytes suggesting an immune-based function.[312]
Apoe	Lipid homeostasis	Lipoprotein-mediated lipid transport	• **Genetics**: *APOE*ϵ2 and ϵ4 are associated with lower (i.e. protective) and higher AD risk, respectively. • **Pathology**: ApoE exhibit isoform-dependent Aβ binding affinity and clearance of Aβ; astrocytic overexpression of *APOE ϵ4* expression (but not *APOE ϵ2* or *ϵ3*) increased phosphorylation and aggregation of tau oligomers in mouse model.[313] • **Splicing**: All ApoE isoforms consist of 299 amino acids differing only at two key residues (Cys-112, Arg-158).

App	Amyloid pathology	Transmembrane glycoprotein	• **Genetics**: Identified causative mutations for EOAD. • **Pathology**: Posited as the amyloid cascade hypothesis, cleavage of APP produces longer Aβ that accumulate and form insoluble fibrils and plaques characteristic of AD pathology (**Section 1.1.3**). • **Splicing**: Expression of KPI-containing APP isoforms is reported to be differentially expressed in AD brain and associated with Aβ accumulation.[314] No differential isoform expression in AD frontal lobe vs controls.[315]
Bin1	Endocytosis	Adaptor protein	• **Genetics, Epigenetics**: GWAS AD-associated variants do not alter coding sequence but localised to regulatory region upstream of promoter; rs59335482 SNP is associated with increased *BIN1* expression in AD brain.[316] • EWAS reveal differential methylation of *BIN1* in AD. • **Pathology**: Levels of BIN1 positively correlated with NFTs whereas no change in Aβ deposition in *BIN1*-haploinsufficient 5xFAD,[317] indicating a role in tau clearance.[318] • **Splicing**: Decreased expression of BIN1 isoform 1 (exon 7 inclusion) was associated with tau accumulation and AD-related traits;[319] whereas, increased expression of isoform 9 correlated with up-regulation of astrocytic and microglial markers,[319] and favoured tau release through extracellular vesicles.[318] • No change in neuronal BIN1 isoform 1 expression in AD post-mortem brains, but an increase in phospho-BIN1(T348):BIN1 ratio, postulating that increased BIN1 T348 phosphorylation is involved in protective effect of interacting and subsequently blocking accumulation of phosphorylated tau.[320]

Cd33	Immune response	Transmembrane receptor for cell signalling	• **Genetics**: Multiple AD-associated SNPs identified from GWAS, including rs12459419[28, 29, 32] located within exon 2, encoding the IgV domain involved in sialic acid binding.[321] • **Pathology**: CD33 inactivation in mouse models result in reduced Aβ $_{42}$ production with enhanced phagocytosis.[322] Differential gene expression in microglia lacking CD33 depended on the presence of TREM2, suggesting TREM2 acts downstream of CD33.[323] • **Splicing**: Short CD33 isoform preferentially encoded by the AD-protective variant (rs12459419) revealed to have a gain of function variant that enhances Aβ phagocytosis.[324] • **Expression**: CD33 expression is elevated in AD microglia and infiltrating macrophages.[322]

Clu	Lipid homeostasis	Secreted glycoprotein (apolipoprotein) with chaperone-like activity	• **Genetics**: AD-associated SNP, rs2279590, is identified within *CLU* enhancer element and associated with increased *CLU* expression.[325] • **Pathology**: Multiple *CLU* mutations (frameshift mutation, mutations in disulphide bride region, rare-coding mutations in CLU β-chain) deregulate secretion and lead to protein degradation.[326] • Percentage of synapses containing clusterin is higher in APOE4 carriers than APOE3 carriers.[327] • **Splicing**: Up-regulation of 2 major isoforms (CLU1, CLU2) in AD brains, generating similar-sized secreted proteins.[328] • Identification of a novel isoform (mitoCLU) localised to the mitochondrial matrix. Mouse mitoCLU is translated from start site exon 3, which coincides with start site in human. • Cell-type specific *CLU* expression profile observed: mRNA with exons 1B, 2, 3, 4 detected in both neurons and astrocytes, whereas exons 1A and 1C unique to astrocytes and neurons, respectively.[329] • Intracellular form of *CLU* (iCLU) was up-regulated in rTg4510 mice, but not in Tg2576 mice. iCLU contains a coiled-coil motif that interacts with tau and Bin1 isoforms (1 - 3).[330] • Various isoforms generated with isoform-specific function and localisation (nucleus: 49kDa, mitochondria: 53kDa, endoplasmic reticulum/Golgi: 80kDa). • **Expression**: mRNA expression up-regulated in AD brains vs controls.[331]
Fus	FTD genetic association	RNA-binding protein	• **Pathology**: Disease-associated *FUS* mutations result in altered splicing of tau with disproportional increase of the 4R/3R-tau ratio, and eventually neurodegeneration in ALS/FTLD-FUS, ALS/FTLD-TDP but not in AD.[332]

Fyn	Tau pathology	Tyrosine protein kinase for cell signalling	• **Pathology**: Fyn phosphorylates tau tyrosine residues and interacts with tau through the SH3 domain.[333] • Fyn overexpression in hAPP mice accelerated synaptic loss and reduced memory retention.[334] • **Splicing**: FynB and FynT predominantly expressed in the brain and haematopoietic cells respectively; FynT, with exon 7 skipping and different linker region, exhibited enhanced kinase activity. • **Expression**: Increased Fyn expression in AD post-mortem brains[335] and in AD TG mice,[336] with up-regulation of FynT expression and isoform switching (reduced FynB expression).[335]
Mapt	Tau pathology	Microtubule assembly and stability	• **Pathology**: *MAPT* encodes tau, which aggregates into neurofibrillary tangles characteristic of AD pathology. • **Splicing**: Altered splicing of exon 10; tauopathy-associated intronic mutations result in exon 10 inclusion and subsequent increased 4R (4R tau, E10+)/ 3R (3R tau, E10-) ratio.[337] • Exon 2 inclusion; differential expression of exon 2 splicing regulators in AD brains.[337] • **Expression**: Regional distribution of *MAPT* expression with highest tau protein levels observed in frontal cortex.[54]
Picalm	Endocytosis	Adaptor protein involved in clathrin-mediated endocytosis	• **Genetics**: Identified multiple SNPs from GWAS, including protective rs3851179 SNP which is associated with modest increase in *PICALM* expression. • rs592297 SNP, located in exon 5, is associated to exons 2–4 skipping.[338] • **Pathology**: *Picalm* haploinsufficiency in tau mouse model resulted in increased & accelerated tau phosphorylation and autophagy deficits,[339] whereas *Picalm* up-regulation reversed disruptive effects of ApoE4 on early endocytosis.[340] • **Splicing, Expression**: Decreased PICALM expression in AD brains vs controls.[341]
Ptk2b	Tau pathology	Calcium-activated non-receptor tyrosine kinase	• **Genetics**: Altered splicing reported as a direct mechanism for the effects of *PTK2B* susceptibility alleles from AD TWAS; a G-to-A mutation was associated with increased intron retention in AD.[92] • **Pathology**: *Ptk2b* deletion did not markedly alter mouse 5xFAD phenotype, whereas overexpression corrected deficits in synaptic proteins. Decreased Ptk2b phosphorylation level observed in aged 5xFAD mice. • **Expression**: Pt2kb protein levels were not altered in AD hippocampus or mouse model.[342]

Rhbdf2	EWAS	Serine protease involved in TNFα secretion	• **Epigenetics**: Most significant differentially-methylated region from meta-analysis of AD EWAS resided in *Rhbdf2* intronic region between exons 3 and 4.[343–345] • **Pathology**: *Rhbdf2* deletion in mice inhibited release of TNFα, a major inflammatory cytokine involved in AD neuroinflammation.[346]
Snca	α-Synuclein pathology	Presynaptic protein	• **Splicing**: Altered *SNCA* splicing generated isoforms with different post-translational modifications and varying propensity for aggregation: α-Synuclein 112 (exon 6 skipping) with C-terminus truncation more likely to aggregate than α-Synuclein 140 (FL and major *SNCA* isoform) and α-Synuclein 126 (exon 4 skipping).[347,348]
Sorl1	Endocytosis, Lipid homeostasis	APOE receptor	• **Genetics**: Multiple AD-associated rare LOF *SORL1* variants from nonsense, frameshift and splice site mutations.[349] • **Pathology**: *SORL1*-deficient hiPSC neurons exhibited early endosome enlargement (not seen in microglia), accompanied with altered APP localisation in early endosome, suggesting altered APP trafficking.[350] • **Splicing**: Down-regulation of full-length SORL1 isoform in AD brains, whereas no change in expression of the shorter isoform (exon 2 skipping). Isoform with exon 19 skipping resulted in NMD.[308] • **Expression**: Total SORL1 expression was reduced in AD and in rTg4510[351] with down-regulation of truncated SORL1 isoform in AD cerebellum.[300]
Tardbp	TDP-43 pathology	Heterogeneous nuclear ribonuclear protein involved in gene regulation and splicing	• **Pathology**: *Tardbp* encodes TDP-43, the major constituent of neuronal inclusions characteristic of FTLD pathology.[352] • Up to 60% of AD patients are characterised with TDP-43 deposits from inheritance of a AD-associated mutation.[352] • *Tardbp* overexpression in AD mouse model resulted in decreased Aβ plaque burden but increased abnormal tau aggregation.[353] • ApoE4 associated with increased risk of developing TDP-43 pathology in AD. • **Expression**: TDP-43 pathology is associated with severe AD pathology with significant increase in TDP-43 levels in late stage AD patients.[354]

Trem2	Immune response	Receptor for cell signalling pathways	• **Genetics**: Most LOAD-associated risk variants are located in exon 2 (Ig-like V domain), which do not impact expression or folding but reduce ligand binding affinity,[356] modulate TREM2 signalling and result in partial LoF;[357] rs75932628 SNP (encoding p.R47H) induces a small conformational change resulting in decreased stability.[356] • **Pathology**: TREM2 is essential for microglia recruitment and phagocytosis of Aβ plaques; TREM2-deficient or -haploinsufficient mice exhibit reduction of plaque-associated microglia and defective Aβ removal.[355] • **Splicing**: Identification of a novel isoform lacking exon 2 (10% of Trem2 mRNA);[358] Human isoform (ENST00000373122) expression was lower in TREM2- p.R62H carriers than in AD cases, whereas expression of canonical transcript (ENST00000373113) was two fold higher.[359] • **Expression**: Increased mRNA expression in TgCRND8 TG mice[357] & in Tg4510 microglia.[351]
Trpa1	Synaptic signalling	Transmembrane calcium channel for cell signalling pathways	• **Pathology**: TRPA1 induced astrocyte hyperactivity, whereas inhibition of channel activity normalised astrocyte activity and reduced plaque expansion in AD mouse model.[360] Deletion of TRPA1 in mice showed reduced morphological damage and memory loss after Aβ injection, implicating detrimental role of TRPA1 receptors in AD.[361] • **Expression**: Higher TRPA1 protein level in hippocampal astrocytes from APP/PS1 TG mice than WT.[360]
Vgf	Synaptic signalling	Neurosecretory protein cleaved into peptides	• **Pathology**: 9 VGF peptides were repeatedly found to decrease in AD CSF samples vs controls, representing a reliable diagnostic AD biomarker.[362] • *Vgf* overexpression rescued cognitive deficits in 5xFAD mice.[363] • **Expression**: *VGF* was the most significantly down-regulated gene in AD brains vs controls,[364]whereas *Vgf* expression was stable between 5xFAD WT and TG mice.[363]

6.2 Methods

6.2.1 Samples

Entorhinal cortex was dissected from 12 female rTg4510 TG and 12 female WT mice, aged 2, 4, 6 and 8 months (n = 3 mice per group) (**Table 6.3**). Additional details on mouse breeding conditions and animal procedures are provided in **Section 2.1.2**. For each mouse sample, RNA was isolated using the AllPrep DNA/RNA Mini Kit (Qiagen, UK) from ~5mg tissue and quantified using the Bioanalyzer 2100 (Agilent, UK) (described in **Section 2.1.4**).

6.2.2 Library preparation and sequencing

Following the Iso-Seq lab workflow (depicted in **Figure 3.3**), first-strand cDNA synthesis was performed on 200ng RNA using the SMARTer PCR cDNA Synthesis Kit (Clontech) with specific oligo(dT) barcodes (listed in **Table 2.2**) for multiplexing. Large-scale PCR amplification was subsequently performed using 14 cycles (**Figure 6.3**, **Section 2.2.2**), and the resulting amplicons were divided into two fractions and purified with 0.4X and 1X AMPure PB beads (PacBio). Quantification and size distribution of each fraction was then determined using the Qubit dsDNA High Sensitivity assay (Invitrogen) and Bioanalyzer 2100 (Agilent). The two fractions were recombined at equimolar quantities and subjected to targeted enrichment (described in **Section 3.1.2.3**) using custom-designed probes (summarised in **Table 6.4**). Following successful enrichment for target genes (listed in **Figure 6.2**), Iso-Seq library preparation was performed using the SMRTbell Template Prep Kit v1.0 (PacBio) for subsequent sequencing on the PacBio Sequel 1M SMRT cell (results from successful library preparation are provided in **Figure 6.4**).

ONT library preparation was performed on a subset of the mouse samples (n = 8 WT, n = 10 TG) using the Ligation Sequencing Kit (SQK-LSK109, ONT), after target enrichment (depicted in **Figure 3.14**). Sequencing was then performed on the ONT MinION using a FLO-Min106D flow cell (as described in **Section 3.2.2.1**).

Finally, RNA from matched samples (n = 12 WT, n = 12 TG) were prepared with the TruSeq Stranded mRNA Sample Prep Kit (Illumina) and subjected to 125bp paired-end short-read RNA sequencing on the HiSeq2500 (Illumina).[90]

Table 6.3: Phenotype information for targeted profiling of the rTg4510 cortex. Tabulated is an overview of the phenotype information of the rTg4510 mouse samples sequenced using PacBio Iso-Seq and ONT nanopore cDNA sequencing. While global transcriptome profiling was performed by sequencing each sample separately, targeted profiling was performed in batches after sample barcoding. ECX - Entorhinal cortex, RIN - RNA integrity number, TG - rTg4510 transgenic mice, WT - Wild-type mice.

Sample demographics						Sequencing platform and approach		
						PacBio		ONT
Sample	Phenotype	Age (months)	RIN	Concentration (ng/μL)	Batch (Barcodes)[a]	Global transcriptome	Targeted profiling	Targeted profiling
Mouse 1	WT	4	8.8	236	1 (PB_BC_1)		X	
Mouse 2	WT	8	9.1	143	1 (PB_BC_2)	X	X	
Mouse 3	WT	6	9	138	1 (PB_BC_3)		X	
Mouse 4	TG	2	8.8	136	1 (PB_BC_4)	X	X	
Mouse 5	TG	4	9.1	80.4	1 (PB_BC_5)		X	
Mouse 6	WT	2	9.2	77.1	1 (PB_BC_6)	X	X	
Mouse 7	WT	4	9.1	84.9	2 (PB_BC_1)		X	X
Mouse 8	TG	8	9.2	65.4	2 (PB_BC_2)	X	X	X
Mouse 9	TG	8	8.7	68.6	2 (PB_BC_3)	X	X	X
Mouse 10	WT	2	9.2	72.3	2 (PB_BC_4)	X	X	X
Mouse 11	TG	2	8.9	115	2 (PB_BC_5)	X	X	X
Mouse 12	WT	8	9	91.8	2 (PB_BC_6)	X	X	X
Mouse 13	TG	6	9.1	83.5	2 (PB_BC_7)		X	X
Mouse 14	WT	6	8.9	92.2	2 (PB_BC_8)		X	X
Mouse 15	TG	6	9	68.7	2 (PB_BC_9)		X	X
Mouse 16	TG	8	8.6	99.7	3 (PB_BC_1)	X	X	X
Mouse 17	WT	2	9.2	83.3	3 (PB_BC_2)	X	X	X
Mouse 18	TG	2	8.9	115	3 (PB_BC_3)	X	X	X
Mouse 19	WT	8	9.1	95.5	3 (PB_BC_4)	X	X	X
Mouse 20	TG	6	8.8	87.2	3 (PB_BC_5)		X	X
Mouse 21	WT	6	8.7	85.8	3 (PB_BC_6)		X	X
Mouse 22	TG	4	8.8	145	3 (PB_BC_7)		X	X
Mouse 23	WT	4	9	70.8	3 (PB_BC_8)		X	X
Mouse 24	TG	4	9	85	3 (PB_BC_9)		X	X

[a] Samples were multiplexed and sequenced in batches for targeted profiling.

Table 6.4: Mouse probes for target profiling of AD-associated genes. Tabulated is a summary of the pre-designed probes used to enrich for 20 AD-associated genes (as illustrated in **Figure 3.6** and detailed in **Section 3.1.2.3**). bp - base pairs.

Target gene	Number of probes	Genome co-ordinates	Strand	Full region (bp)	Exons (bp)
Abca1	56	chr 4 : 53030670 - 53160014	-	129,107	10,260
Abca7	47	chr 10 : 79997615 - 80015572	+	17,958	6,594
Ank1	52	chr 8 : 22974836 - 23150497	+	175,662	9,018
Apoe	5	chr 7 : 19696125 - 19699285	-	2,923	1,251
App	20	chr 16 : 84954317 - 85173826	-	219,272	3,357
Bin1	20	chr 18 : 32377217 - 32435740	+	58,524	2,455
Cd33	9	chr 7 : 43528610 - 43533290	-	5,716	2,571
Clu	9	chr 14 : 65968483 - 65981545	+	13,063	1,808
Fus	16	chr 7 : 127967479 - 127982032	+	14,554	1,845
Fyn	18	chr 10 : 39369799 - 39565381	+	195,583	3,692
Mapt	23	chr 11 : 104231436 - 104332096	+	100,661	5,387
Picalm	24	chr 7 : 90130232 - 90209447	+	79,216	4,174
Ptk2b	32	chr 14 : 66153138 - 66281171	-	127,796	4,034
Rhbdf2	21	chr 11 : 116598082 - 116627138	-	28,855	3,934
Snca	7	chr 6 : 60731454 - 60829974	-	98,283	1,463
Sorl1	48	chr 9 : 41968370 - 42124408	-	155,801	6,938
Tardbp	15	chr 4 : 148612263 - 148627115	-	14,615	7,454
Trem2	5	chr 17 : 48346401 - 48352276	+	5,876	1,146
Trpa1	28	chr 1 : 14872529 - 14918981	-	46,215	4,263
Vgf	9	chr 5 : 137030295 - 137033351	+	3,057	2,553
	Total: 464				

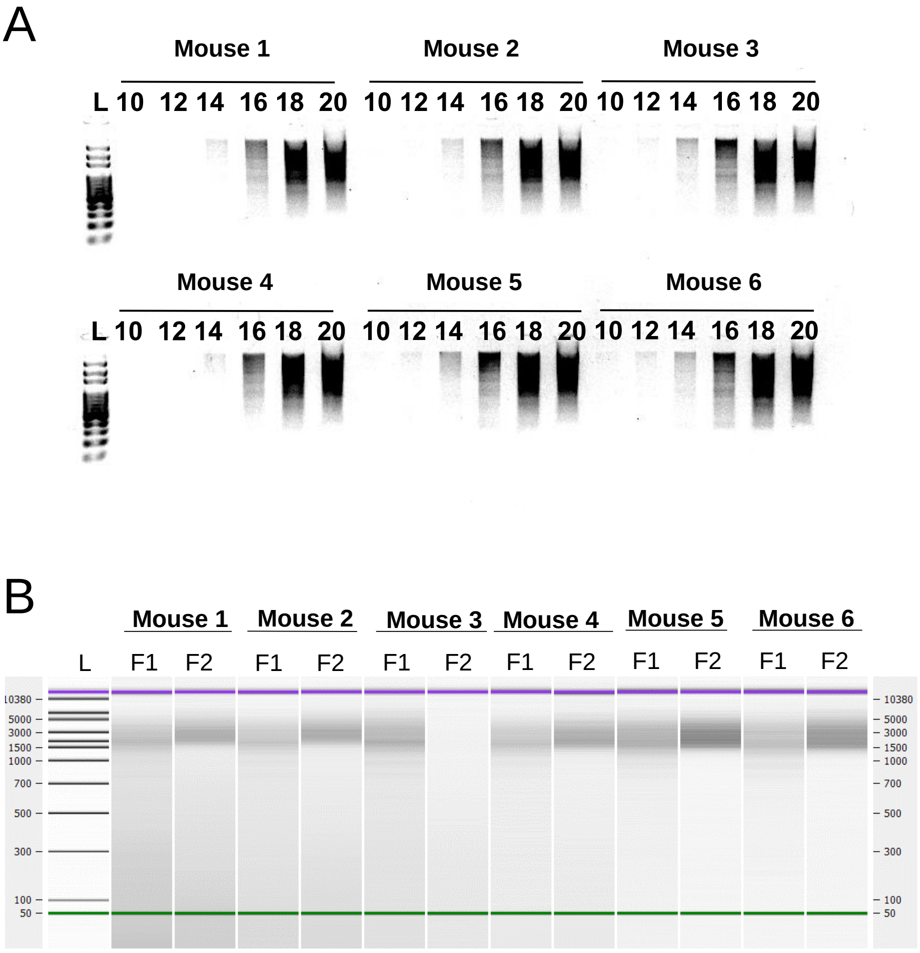

A
Mouse 1
Mouse 2
Mouse 3
L 10 12 14 16 18 20 10 12 14 16 18 20 10 12 14 16 18 20
Mouse 4
Mouse 5
Mouse 6
L 10 12 14 16 18 20 10 12 14 16 18 20 10 12 14 16 18 20
B
Mouse 1
Mouse 2
Mouse 3
Mouse 4
Mouse 5
Mouse 6
L F1 F2 F1 F2 F1 F2 F1 F2 F1 F2 F1 F2
10380
5000
3000
1500
1000
700
500
300
100
50

Figure 6.3: Samples for targeted profiling were amplified with 14 PCR cycles. Shown is **(A)** an example of an agarose gel image from PCR cycle optimisation of six mouse samples after cDNA synthesis. Analogous to global transcriptome profiling (**Figure 4.1**), 14 cycles were determined to be optimal for large-scale PCR amplifications. Ladder (L) denotes to a 100bp DNA ladder. **(B)** A Bioanalyzer gel of amplified cDNA after purification with 1X (F1) and 0.4X (F2) AMPure beads. Size distribution for each fraction was determined from the start to the end point of the smear. Ladder (L) denotes to a 12kb DNA ladder, whereby the green and purple line represent the lower marker at 50bp and the upper marker at 12kb, respectively.

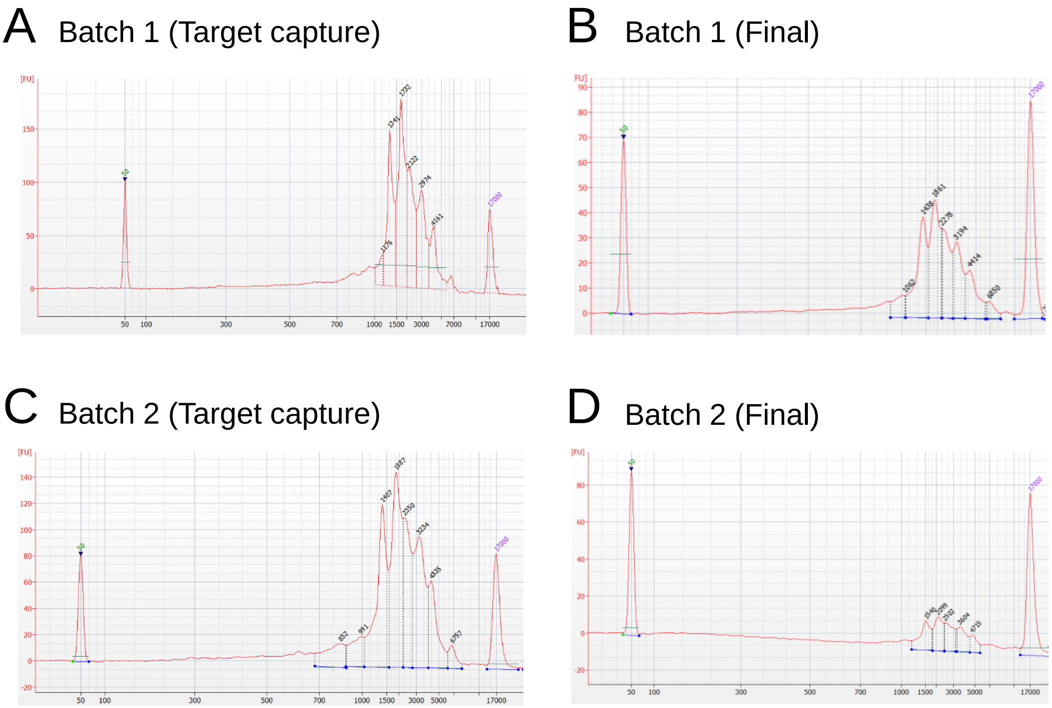


Figure 6.4: Successful target capture and Iso-Seq library preparation. Shown are Bioanalyzer electropherograms of **(A)** batch 1 (n = 6 samples) after target enrichment and **(B)** Iso-Seq library preparation, and **(C)** batch 2 (n = 9 samples) after target enrichment and **(D)** Iso-Seq library preparation.

Illustrating successful target enrichment, we observed peaks that correspond to enriched transcript lengths from genes of interest, which notably differ from the broad peaks seen in the Iso-Seq global datasets (**Figure 4.2**). Iso-Seq library preparation (Figure B and D) retained these target transcripts with detection of similar peaks, albeit less pronounced due to a lower cDNA input for Bioanalyzer assays (in order to maximise the amount of cDNA available for sequencing).

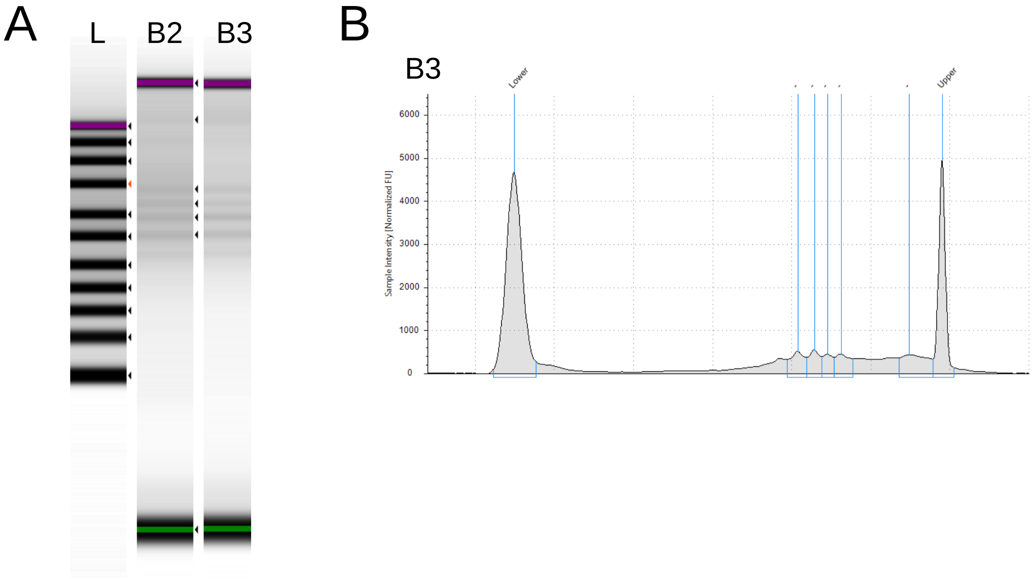


Figure 6.5: Successful target capture and ONT library preparation. Shown is the **(A)** ScreenTape gel image of batch 2 (n = 9) and batch 3 (n = 9) after target enrichment and ONT library preparation, and **(B)** ScreenTape electropherogram of batch 3. Further details of ScreenTape assays are provided in **Section 2.1.4**.

Illustrating successful target enrichment analogous to **Figure 6.4**, we observed peaks that correspond to enriched transcript lengths from genes of interest. Lower cDNA input was used for ScreenTape assays to maximise the amount of cDNA available for sequencing. L - Ladder, B2 - batch 2, B3 - batch 3.

6.2.3 SMRT sequencing QC and data processing

The processing of raw Iso-Seq reads was performed using the optimised Iso-Seq bioinformatics pipeline (outlined in **Section 3.1.4**), in an approach analogous to that used in the processing of the Iso-Seq global dataset. The only difference was an additional step for sample demultiplexing using barcode-specific sequences. Briefly, CCS reads were generated from a minimum of 1 pass (*Iso-Seq3 CCS*, v5.0.0) for each batch followed by removal of primers and barcode sequences using *Lima* (v1.9) to generate full-length (FL) reads for each sample. After removing artificial concatemer reads and trimming of poly(A) tails using *Iso-Seq3 Refine*, full-length reads were merged and collapsed to high quality transcripts using *Cupcake* (parameters: -c 0.85 -i 0.95 –dun-merge-5-shorter), which were then mapped to the mouse reference genome (mm10) using *Minimap2* (v2.17). Full-length Iso-Seq read counts from each individual sample were extracted from the "read_stat.txt" file, generated from the *collapse_isoforms_by_sam.py* script (*Cupcake*), with the CCS read ID as sample identifiers.

6.2.4 ONT nanopore sequencing QC and data processing

The QC of raw ONT reads was performed using *PycoQC*[239] (v2.2.3) followed by subsequent analysis using the optimised ONT bioinformatics pipeline (details are provided in **Section 3.2.4**). Briefly, raw ONT reads were basecalled using *Guppy* (v4.0) and reads with Phred (Q) < 7 were discarded. Primers and ONT adapters were then removed using *Porechop* (v0.2.4) to generate full-length reads for each sample. After trimming of poly(A) tails using *Cutadapt* (v2.9), full-length reads were then mapped to the mouse reference genome (mm10) using *Minimap2* (v2.17, parameters: "-ax splice"). Owing to the high error rate of ONT nanopore sequencing, artefactual non-canonical splice junctions from mapped reads were corrected using *TranscriptClean*.[365] Corrected reads were then processed using *TALON*[246] (v5.0) for annotation, quantification and filtering for intra-priming events (–maxFracA = 0.5, the fraction of genomic A's above which the isoform will be filtered, as detailed in **Section 3.1.4.4**). Novel transcripts were only retained if they were covered by more than 5 full-length reads and detected in at least 2 samples. *TALON* was chosen as the preferred tool for ONT processing after trialling multiple tools (more details are provided in **Appendix D**).

6.2.5 Comparison of Iso-Seq and ONT datasets

The Iso-Seq targeted dataset (n = 24 samples) was examined with other datasets using *GffCompare*[366] (v0.12.2); such datasets included Iso-Seq-derived transcripts identified from global transcriptome profiling (n = 12 samples, **Table 4.2**, Iso-Seq global dataset) and ONT-derived transcripts from targeted profiling (n = 18 samples, **Table 6.3**, ONT targeted dataset). For a fair comparison, the Iso-Seq global dataset was re-annotated with *SQANTI3* with no splice junction filtering from short-read RNA-Seq data, and only transcripts derived from matched samples were used for comparison. Conversely, all processed but unfiltered ONT reads were used for a comprehensive comparison between the two technologies, with Iso-Seq derived transcripts as reference.

6.2.6 Merged annotation and quantification

For a comprehensive characterisation of the target genes enriched in the rTg4510 cortex, full-length transcripts from the Iso-Seq and ONT targeted datasets were merged using *GffCompare* (depicted in **Figure 6.7**). A custom python script ("identify_common_targeted_transcripts.py") was then applied to: i) identify transcripts detected using both PacBio Iso-Seq and ONT nanopore sequencing, which were defined as a "complete exact match" in the *GffCompare* output (class code: "="), ii) retain ONT-derived novel transcripts that did not pass *TALON* filtering (> 5 reads and detected in > 2 samples), but were detected in the Iso-Seq targeted dataset, iii) retain all transcripts unique to the Iso-Seq targeted dataset given the stringent processing and high accuracy of Iso-Seq reads, and iv) generate an abundance file for each sample and transcript, either tabulating counts from *Cupcake* for Iso-Seq-derived-transcripts, counts from *TALON* for ONT-derived-transcripts or the count summation for commonly-detected transcripts. The merged dataset was then annotated with *SQANTI3* in combination with the mouse reference gene annotations (mm10, GENCODE, vm22), FANTOM5 CAGE peaks and *STAR*-aligned RNA-Seq junctions. Isoform were subsequently classified as either FSM, ISM, NIC, NNC, antisense, fusion, and intergenic (described in **Section 3.1.4.4**). Isoforms classified as ISM with 3' fragment were assumed to be partial 5' degraded products and hence removed.

6.2.7 Quantification of human *MAPT* transgene expression

The presence of human- and mouse-specific *MAPT/Mapt* sequences was determined in full-length transcripts as QC check of sample identity (as detailed in **Section 5.2.2**).

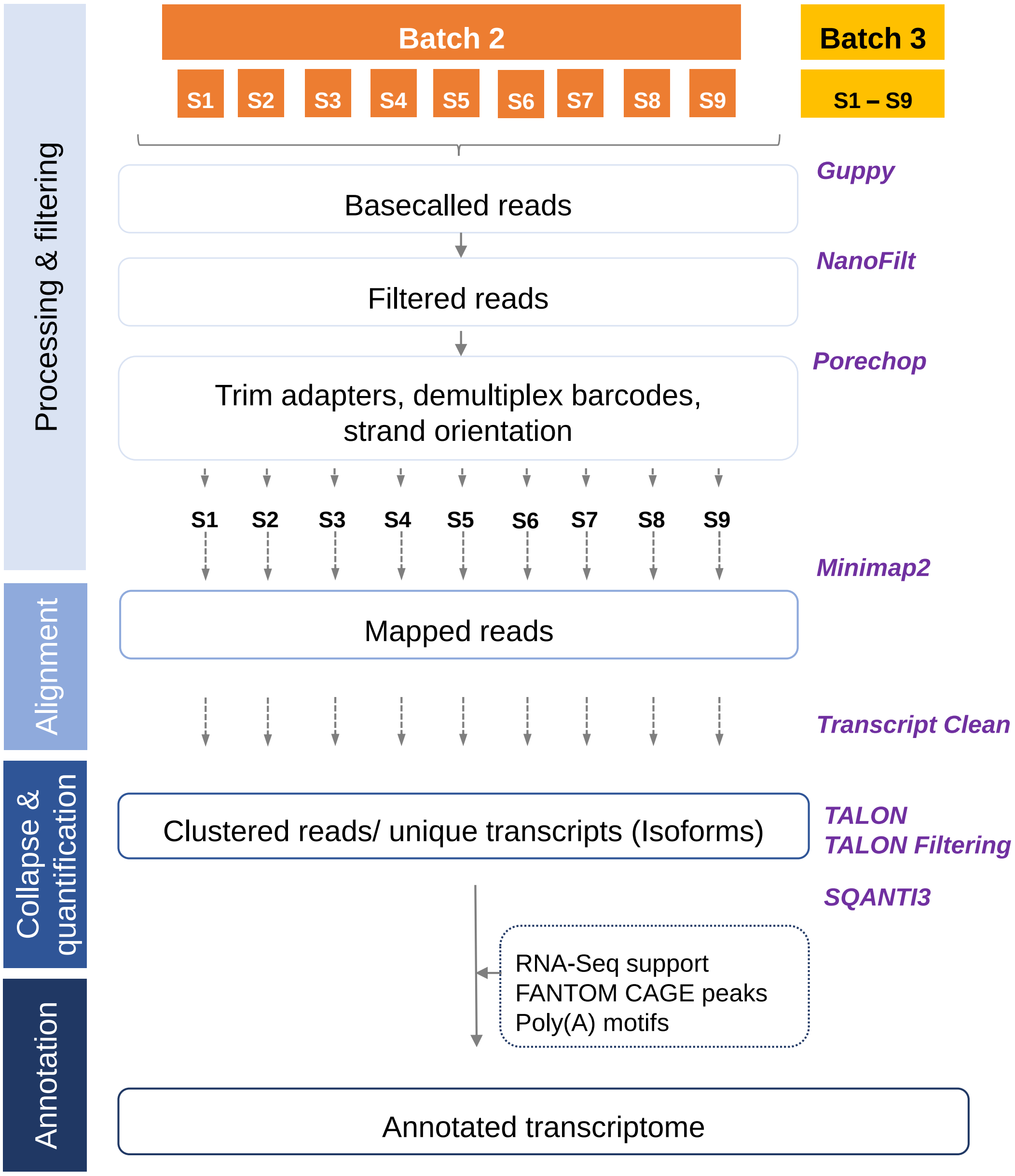


Figure 6.6: Detailed ONT bioinformatics pipeline for targeted profiling. Shown is a detailed bioinformatics pipeline for processing ONT reads from targeted profiling of the rTg4510 cortex (n = 18 samples) on two flow cells (referred as batch 2 and batch 3 of the Iso-Seq targeted dataset, summarised in **Table 6.3**). Supplementing **Figure 3.18**, raw ONT reads from each flow cell were processed and demultiplexed using *Porechop* to generate sample-specific reads, which were subsequently processed independently for collapse and transcript quantification. Samples from both batches were then merged into one dataset, while retaining sample-specific transcript expression.

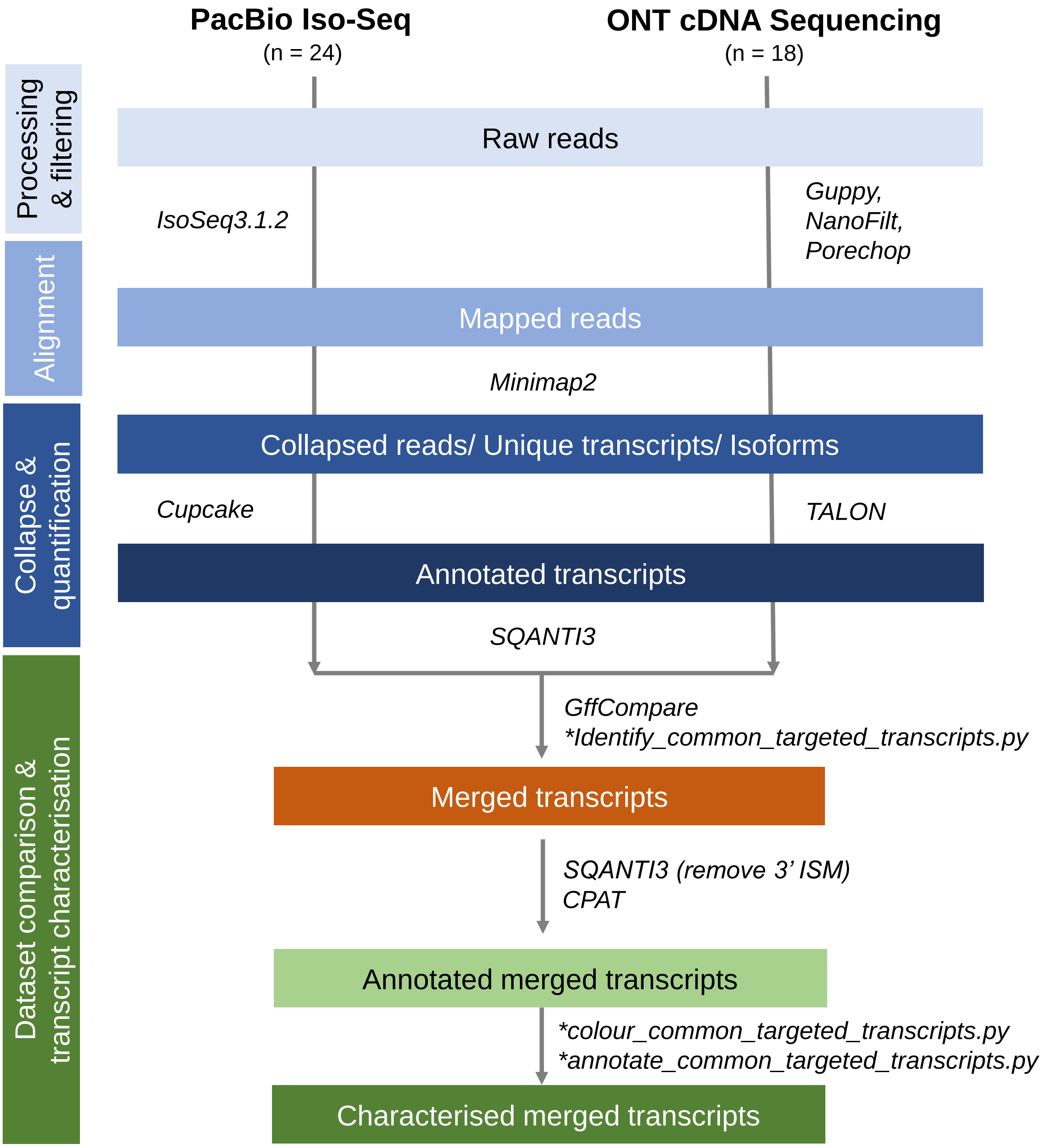

PacBio Iso-Seq
(n = 24)
ONT cDNA Sequencing
(n = 18)
Processing & filtering
Alignment
Collapse & quantification
Dataset comparison & transcript characterisation
Raw reads
IsoSeq3.1.2
Guppy, NanoFilt, Porechop
Mapped reads
Minimap2
Collapsed reads/ Unique transcripts/ Isoforms
Cupcake
TALON
Annotated transcripts
SQANTI3
GffCompare
*Identify_common_targeted_transcripts.py
Merged transcripts
SQANTI3 (remove 3' ISM)
CPAT
Annotated merged transcripts
*colour_common_targeted_transcripts.py
*annotate_common_targeted_transcripts.py
Characterised merged transcripts

Figure 6.7: Bioinformatics pipeline for merging targeted Iso-Seq and ONT datasets. Shown is an outline of the bioinformatics pipeline for processing Iso-Seq and ONT reads from the targeted profiling of the mouse cortex.

6.2.8 Characterisation of AS events and transcript visualisation

Developed for analysing short-read RNA-Seq data, existing tools for assessing alternative splicing events fail to capture the connectivity and complexity of long-read-derived isoforms, particularly those generated from targeted profiling where a deep sequencing coverage is achieved. We therefore developed a custom python script ("annotate_common_targeted_transcripts.py") to accurately assess the occurrence of alternative splicing events by comparing splice sites (exon) coordinates between long-read-derived transcripts and reference transcripts (mm10) (depicted in **Figure 6.7**). Common alternative splicing events such as alternative first exons (AF), alternative last exons (AL), alternative 5' splice sites (A5), alternative 3' splice sites (A3), intron retention (IR) and exon skipping (ES) were assessed (depicted in **Figure 1.9**). Alternative 5' and 3' splice sites were defined as splice sites differing by more than 10bp from the known splice site, and an intron was considered retained if the exon splice site differed by more than 100bp from the known splice site (depicted in **Figure 6.8**). Other regulatory mechanisms such as alternative transcription initiation (defined by alternative TSS) and termination (defined by alternative TSS), and the presence of novel exons, were also evaluated.

Open reading frames were predicted using the Coding-Potential Assessment Tool[367] (CPAT) (v3.0.2) under default parameters, and transcripts with coding potential score ≥ 0.44 (recommended threshold) were predicted as protein-coding. Isoforms were predicted to undergo nonsense-mediated decay (NMD) if the distance between the predicted open reading frame and the last exon-exon junction was > 50bp. Finally, a separate custom python script ("colour_common_targeted_transcripts.py") was applied to colour transcripts by coding potential (green for protein-coding, red for non-protein-coding) and also shade them by abundance. Isoforms were then grouped by splicing patterns and visualised using the UCSC genome browser.

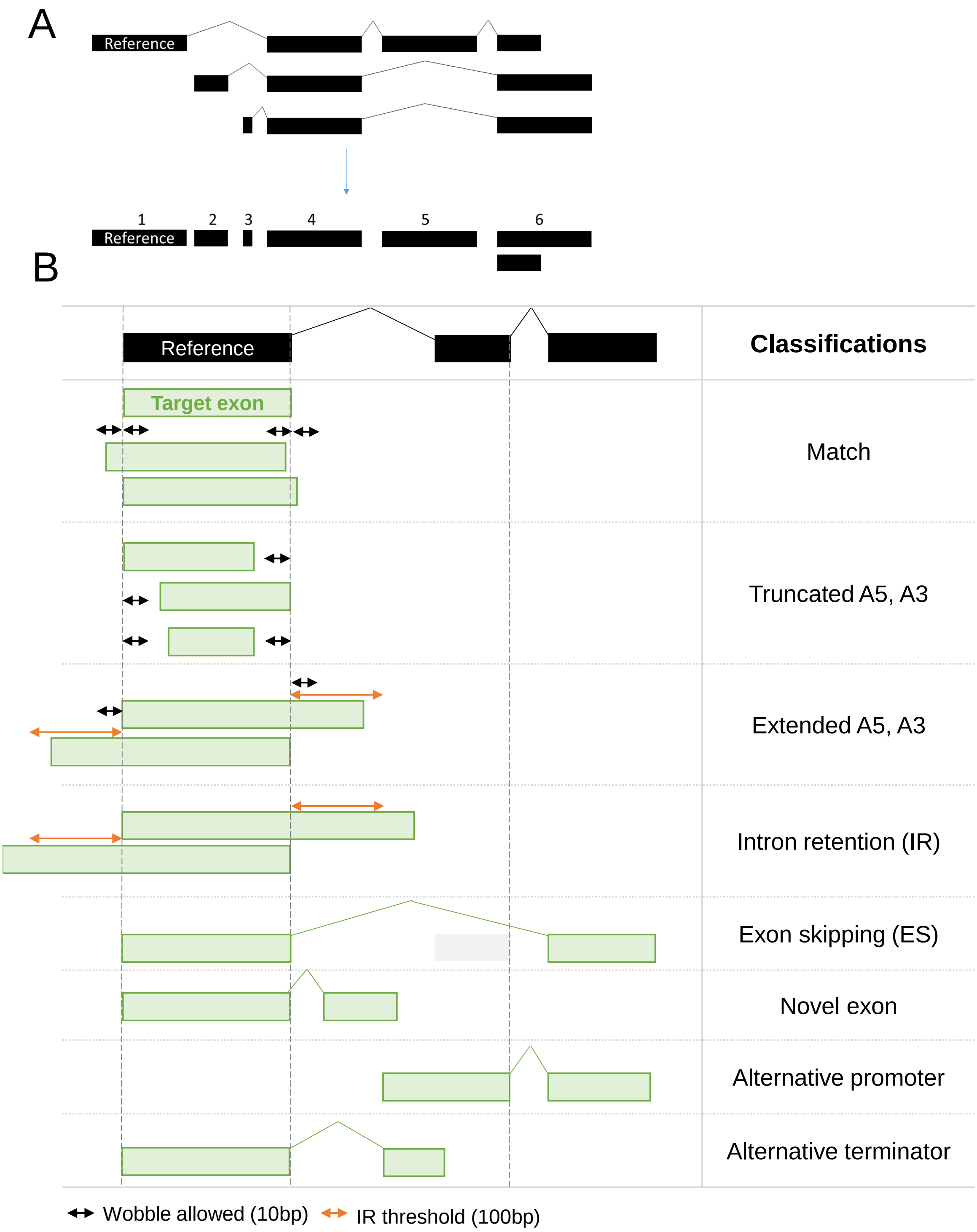


Figure 6.8: Characterisation of isoforms detected in targeted profiling. Shown is a schematic figure of our approach for examining alternative splicing events in the Iso-Seq-defined transcriptome. **(A)** Reference transcripts were "flattened" to obtain splice site coordinates. **(B)** Exon-level comparison of long-read-derived transcripts and reference transcripts was then performed by comparing splice site coordinates to assess occurrence of alternative splicing sites. Splice sites differing < 10bp ("wobble") were considered identical, > 10bp as truncation, > 10bp but < 100bp as extension, and > 100bp as intron retention.

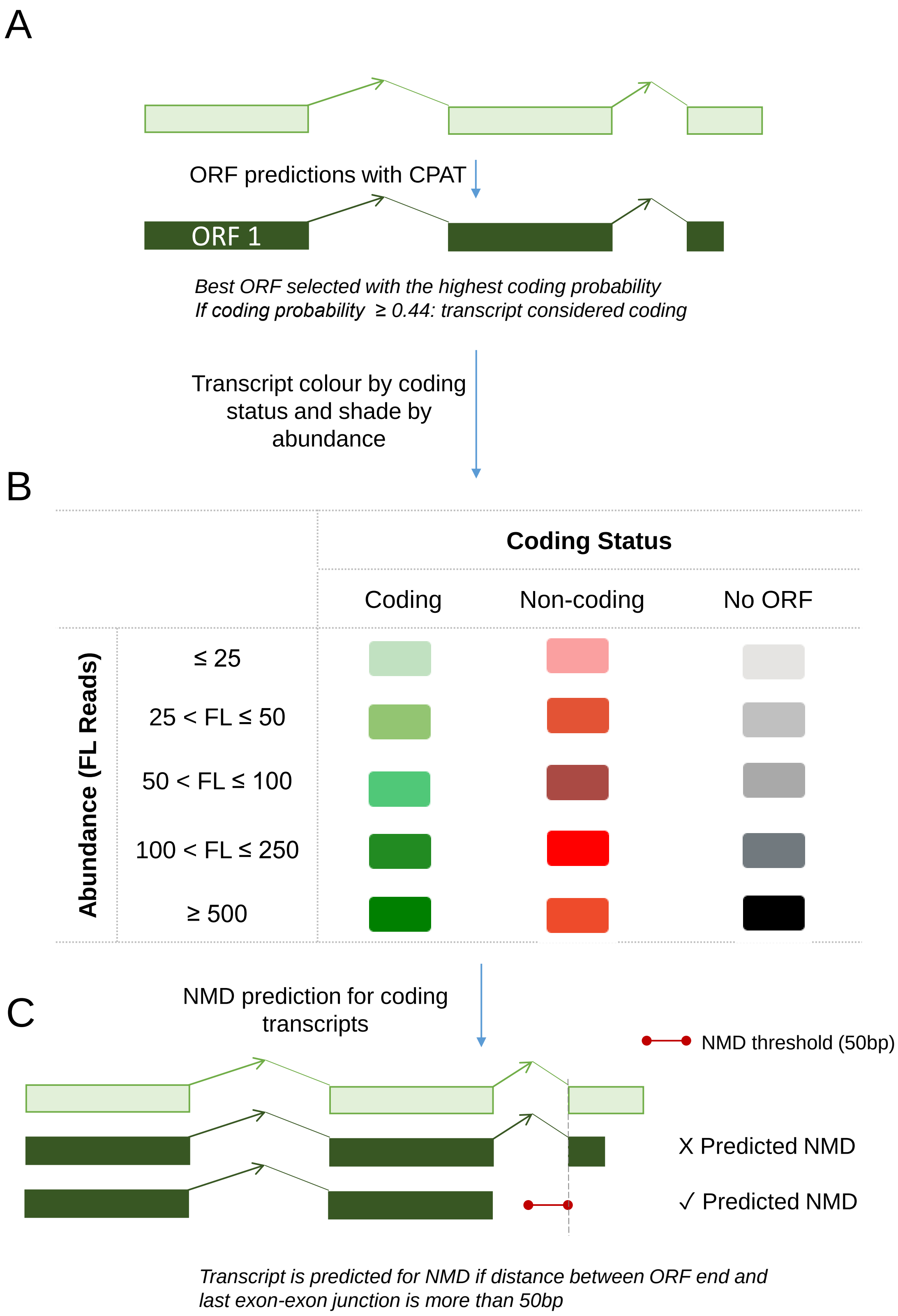

A
ORF predictions with CPAT
ORF 1
Best ORF selected with the highest coding probability
If coding probability ≥ 0.44: transcript considered coding
Transcript colour by coding status and shade by abundance
B
Coding Status
Coding
Non-coding
No ORF
Abundance (FL Reads)
≤ 25
25 < FL ≤ 50
50 < FL ≤ 100
100 < FL ≤ 250
≥ 500
C
NMD prediction for coding transcripts
NMD threshold (50bp)
X Predicted NMD
✓ Predicted NMD
Transcript is predicted for NMD if distance between ORF end and last exon-exon junction is more than 50bp

Figure 6.9: Visualising isoforms by coding potential, abundance and NMD status. Shown is a flow diagram for isoform visualisation. **(A)** Open reading frames were determined using *CPAT* and **(B)** isoforms were subsequently coloured by protein-coding potential (green for protein-coding, red for non-protein-coding, and grey for no open reading frame) and shaded by abundance (described in **Section 6.2.6**). **(C)** A schematic figure illustrating our approach for predicting nonsense-mediated decay (NMD). ORF - Open reading frame.

6.2.9 Differential expression and splicing analyses

Differential expression analysis was performed using *tappAS* with Iso-Seq and ONT full-length read counts as proxies of expression (fully described in **Section 3.3.4**). Briefly, *tappAS* filters out lowly-expressed isoforms, normalises read counts using the TMM approach and implements *maSigPro*[256–258] to elucidate the effects of genotype and age (as shown in the **Equation 5.1**). Normalised counts generated from *tappAS* were used to generate plots illustrating differential expression changes and isoform usage. Gene expression was deduced from the summation of normalised counts from associated isoforms. The distribution of isoform expression (i.e. isoform usage) was determined by dividing the mean expression of each isoform across biological replicates (n = 3) over the total mean expression of all the isoforms. Minor isoforms with low expression (< 0.5 mean normalised expression in WT and TG mice) were removed, and isoforms that constituted < 5% of the total isoform proportions were classified as “Other”.

6.3 Results

6.3.1 Iso-Seq run performance and sequencing metrics

Following Iso-Seq library preparation and SMRT sequencing, we generated a total of 62.8Gb sequencing data (n = 24 samples, mean yield = 20.9Gb, s.d = 2.84Gb, range = 19.25Gb - 24.2Gb, **Table 6.5**). While the sequencing yield was comparable to global transcriptome profiling (**Table 4.2**), the run performance varied across the three Iso-Seq targeted datasets, particularly between batch 1 (n = 6 samples) and batches 2 (n = 9 samples), 3 (n = 9 samples), which were sequenced pre- and post-Covid-19 lockdown respectively. The run performance metrics for batch 1 were optimal (**Table 6.5**). Conversely, batches 2 and 3 had a poor loading rate (P1: batch 1 = 71%, batch 3 = 38.1%) with sequencing yields that were comparable to batch 1, despite containing more samples. We suspect that this reduced run performance was a likely result of sample degradation, given samples were stored in -20°C for > 9 months (due to Covid-19 lockdown) before sequencing.

Following the Iso-Seq bioinformatics pipeline, raw reads were processed and clustered to unique consensus transcripts, which were then mapped to the mouse reference genome. A total of 966K CCS reads (mean = 332K, s.d = 126K, range = 221K - 469K) and 930K FLNC reads (mean = 310K, s.d = 77.7K, range = 2556K - 399K) were successfully generated (**Figure 6.10A**). Where there was evident difference in the number of CCS reads obtained for batch 1 and batches 2, 3 - a reflection of the run performance - batches 2 and 3 had a significantly greater coverage of target genes than batch 1 (**Figure 6.11**). These results indicate that the sub-optimal run performance and subsequent lower sequencing yield of batches 2 and 3 were compensated by a bigger sample size, generating more full-length reads associated to target genes. In contrast, the better run performance but lower sample size of batch 1 resulted in quicker saturation of target genes and the generation of more full-length reads associated to off-target genes.

Finally, we noted that the number of full-length transcripts obtained per sample varied within each batch (**Figure 6.10B**), despite the careful pooling of samples with equal molarity during library preparation. This variability was not associated with RIN (Spearman's rank: corr = 0.147, *P* = 0.492). However, no significant difference in the number of full-length transcripts was observed between WT and TG mice across the runs (Wilcoxon rank sum test: W = 73, *P* = 0.977, **Figure 6.10C**).

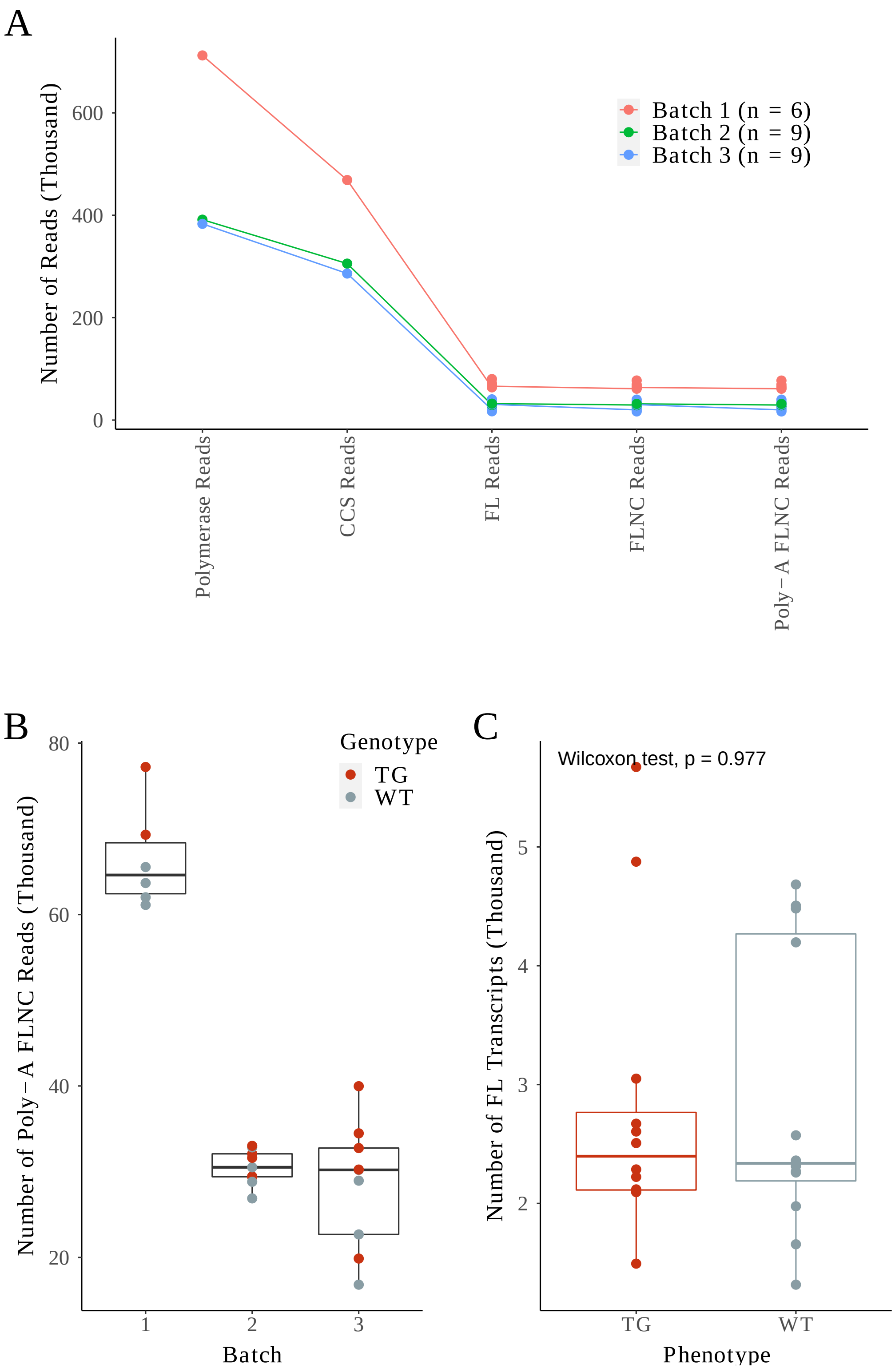


Figure 6.10: Despite batch variability in Iso-Seq targeted datasets, no difference was reported in the number of FL transcripts between WT and TG mice. Shown is **(A)** a scatter plot of the number of reads generated through the Iso-Seq bioinformatics pipeline from the initial generation of CCS reads, FL reads with primer removal to poly(A) FLNC reads after removing artificial concatemers and trimming of poly(A) tails. Samples were multiplexed and sequenced in three runs (batch 1, 2 and 3). Shown are box plots of the **(B)** number of poly(A) FLNC reads by batch and genotype, and the **(C)** final number of FL transcripts by genotype. CCS - Circular consensus sequence, FLNC - Full-length non-chimeric, FL - Full-length, TG - rTg4510 transgenic mice, WT - Wild-type.

Table 6.5: Iso-Seq sequencing yield for targeted profiling of the rTg4510 mouse cortex. The rTg4510 cortex (n = 9 WT, n = 9 TG) was sequenced using the Iso-Seq approach on 3 runs (batch 1, 2 and 3) after multiplexing and enrichment of 20 AD-associated genes. Further details on the evaluation of the Iso-Seq run performance are provided in **Section 3.1.3**. K - Thousand, Pol - Polymerase. N50 is defined as the sequence length of the shortest read at 50% of all reads.

Run	Total bases (Gb)	Polymerase reads (K)	Read length (kb)				Productivity		
			Polymerase		Subread		P0	P1	P2
			Mean	N50	Mean	N50			
Batch 1	24.2	712	34.0	70.5	1.4	1.85	4.62% (46613)	71.58% (722026)	24.76% (249707)
Batch 3	19.3	384	50.5	100	1.6	2.02	18.68% (189549)	38.11% (386743)	43.56% (442054)

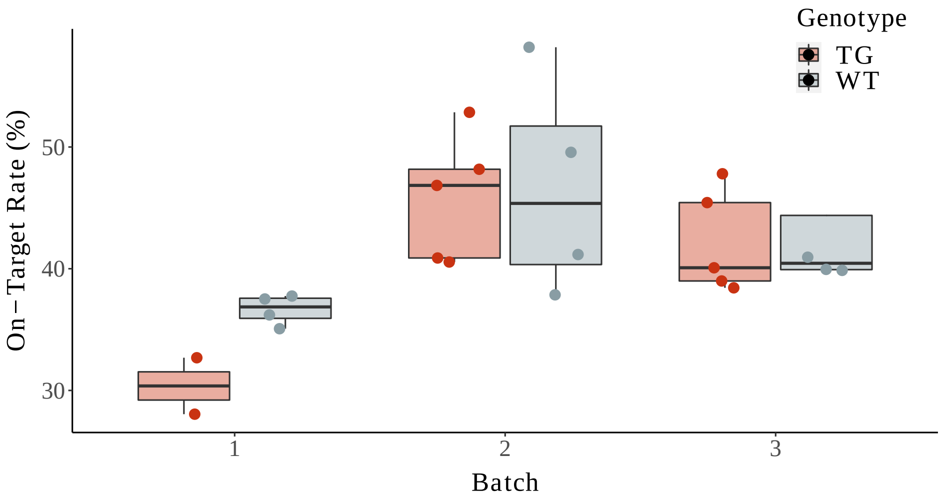


Figure 6.11: Higher coverage of target genes in batches 2 and 3 due to more samples multiplexed and sequenced. Shown is a box plot of the on-target rate, which is defined by the proportion of mapped transcripts with sequences that overlapped with at least one target probe (i.e. transcript is annotated to AD-associated target gene). Of note, a difference in the on-target rate between WT and TG in batch 1 is a likely reflection of the sample variability in sequencing (**Figure 6.10B**). TG - rTg4510 transgenic mice, WT - Wild-type.

6.3.2 Iso-Seq targeted profiling approach detects many novel, rare isoforms annotated to AD-associated genes

Following stringent quality-control and filtering of technical artefacts, we detected 19,659 isoforms in the Iso-Seq targeted dataset. Among these, 2,015 isoforms (10.2%) were annotated to the 20 AD-associated genes enriched in the rTg4510 cortex (n = 12 WT, n = 12 TG). This was in stark contrast to the global transcriptome profiling datasets, where we detected 175 isoforms (0.25% of total number of isoforms detected from global transcriptome profiling) annotated to the same set of AD-associated genes, highlighting the power of targeted profiling for deep coverage of target genes. As expected, targeted sequencing of the same matched samples (n = 6 WT, 6 TG, **Table 4.1**) detected many more AD-associated transcripts than the whole transcriptome profiling approach (**Figure 6.12A**, Iso-Seq global dataset: n = 46 unique isoforms; Iso-Seq targeted dataset: n = 658 unique isoforms; Iso-Seq global and targeted dataset: n = 221 commonly-detected isoforms). The majority of these isoforms unique to the Iso-Seq targeted dataset were novel (n = 525 isoforms, 79.8%, **Figure 6.12B**) as NIC (n = 218 isoforms, 33.1% of total novel isoforms) and NNC (n = 307 isoforms, 46.7% of total novel isoforms), and less abundant (**Figure 6.12D**), highlighting the greater sensitivity of the targeted profiling approach to detect the novel, rarer transcripts. Strikingly, the global transcriptome profiling approach detected all the target genes with the exception of *Trpa1*, the most lowly-expressed target gene in the mouse cortex (**Table 6.1**). This suggests that the gene sensitivity with 5.6 million CCS reads (n = 12 samples) using the Iso-Seq global transcriptome profiling approach was capped between 0.1 TPM (mouse cortex *Trpa1* expression) and 0.5 TPM (mouse cortex *Rhbdf2* expression, the second least expressed target gene). Given that our Iso-Seq global datasets approached saturation, particularly at the gene level (**Figure 4.4A**), it is unlikely that we would have been able to detect *Trpa1* transcripts with more samples using the global transcriptome profiling approach.

Further comparison of the isoform landscape of AD-associated genes in the Iso-Seq global and targeted datasets revealed a similar distribution of isoform length (Iso-Seq global dataset: mean = 2.67kb, s.d = 1.6kb, range = 0.3kb - 10.3kb; Iso-Seq targeted dataset: mean = 2.4kb, s.d = 1.2kb, range = 0.14kb - 10.3kb; Mann-Whitney-Wilcoxon test: W = 1.91 x 10^6, P = 0.068) and exon number (Iso-Seq global dataset: mean = 12.9, s.d = 9.6, range = 1 - 50; Iso-Seq targeted dataset: mean = 11.7, s.d = 7.9, range = 1 - 50; Mann-Whitney-Wilcoxon test: W = 1.86 x 10^6, P = 0.22). Drawing parallels to the isoform landscape of the global transcriptome, approximately half of the isoforms identified in Iso-Seq targeted dataset were novel (n = 919 isoforms, 45.6%) as NIC (n = 485 isoforms, 24.1%) and NNC (n = 434 isoforms, 21.5%) (**Figure 6.13A**), with the remaining half identified as known and predominantly ISM (n = 913 isoforms, 45.3%). However, AD-associated isoforms in Iso-Seq targeted dataset were less enriched near CAGE peaks (median distance from CAGE peak = 335 bp; 646 (32.1%) transcripts located within 50bp of a CAGE peak) than AD-associated isoforms detected in Iso-Seq global dataset (median distance from CAGE peak = 2 bp; 122 transcripts (69.7%) located with 50bp of a CAGE peak), and were located further to annotated transcription start sites (Iso-Seq targeted dataset: median distance = 808bp; Iso-Seq global dataset: median distance = 8bp) and transcription termination sites (Iso-Seq targeted dataset: median distance = 2bp; Iso-Seq global dataset: median distance = 1bp).

Generating highly-parallel RNA-Seq data on the same samples (n = 24 samples, total number of uniquely mapped reads = 360 million), we further found that the vast majority of these isoforms were not supported by RNA-Seq reads at the splice junctions (n = 1,658 isoforms, 82.2%). Given the stringent process of the Iso-Seq bioinformatics pipeline and the deep coverage afforded by targeted sequencing, this is a likely reflection of the relatively low RNA-Seq coverage to comprehensively span these novel junctions and the elevated power of long-read sequencing to deeply profile transcript structure. Nonetheless, a strong correlation was observed across both methods at the gene level (Pearson's correlation: n = 20 genes, corr = 0.86, $P = 1.18 \times 10^{-6}$, **Figure 6.13B**).

Finally, we observed a relatively high off-target rate of Iso-Seq targeted experiments with detection of isoforms (n = 17,644 isoforms, 89.8%) that were not associated with target genes. Comparison of the Iso-Seq targeted datasets using matched samples (n = 6 WT, 6 TG, n = 7,925 off-target isoforms, 89.8%) revealed that the overwhelming majority of these isoforms (n = 7,418, 93.6% off-target isoforms) were also detected in the Iso-Seq global datasets. These commonly-detected off-target isoforms were more abundant than isoforms unique to the Iso-Seq global dataset (**Figure 6.12C**), suggesting that the capture of off-target genes is predominantly driven by abundance rather than sequence homology to target genes (i.e. off-target binding).

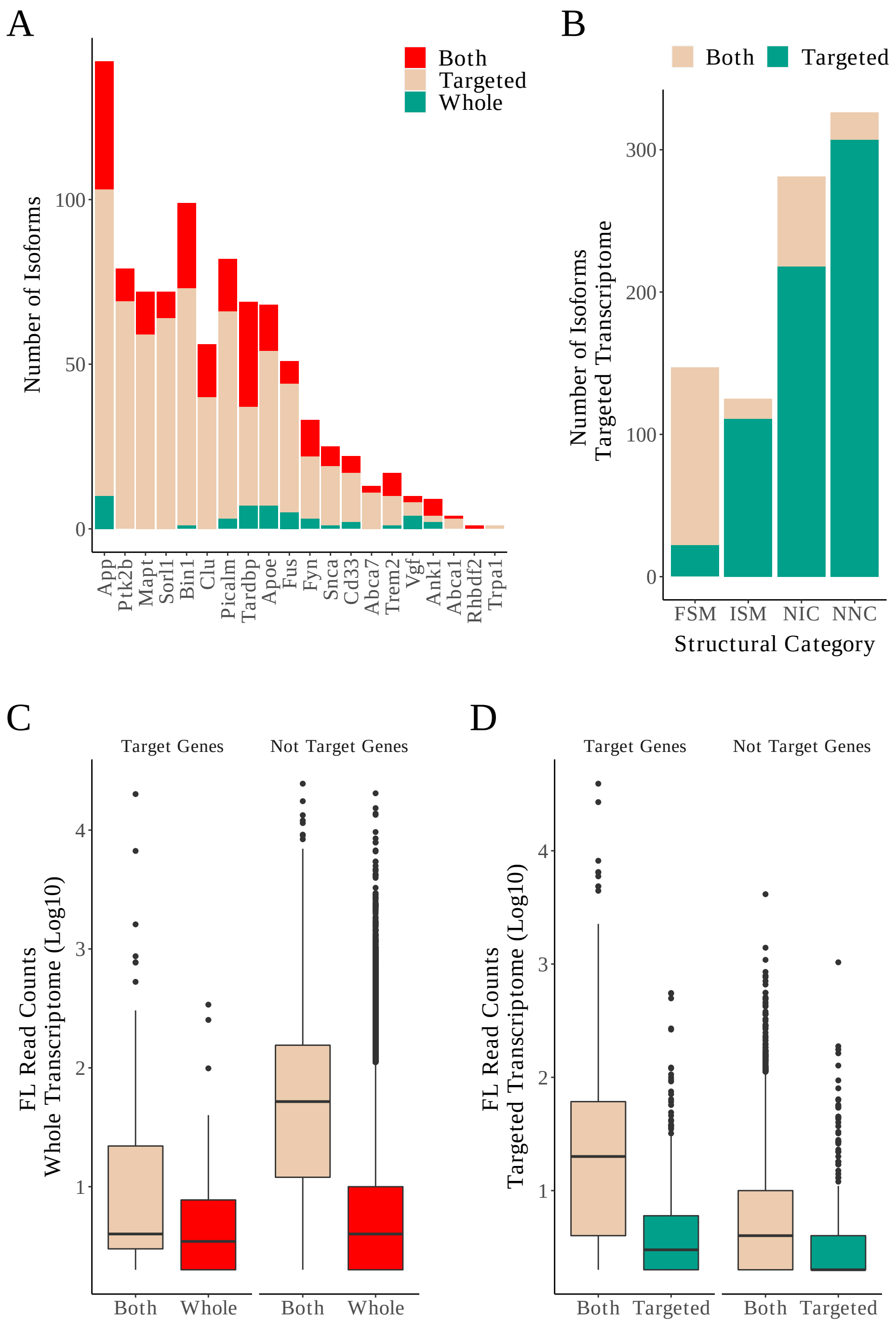


Figure 6.12: Iso-Seq targeted approach detected many more novel and rarer transcripts than global transcriptome profiling of the mouse cortex. Shown are **(A)** bar charts of the number of isoforms per target gene that were uniquely detected using the Iso-Seq targeted profiling approach ("Targeted"), uniquely detected in the Iso-Seq global transcriptome profiling approach ("Whole") and in both datasets ("Both"), and the **(B)** bar charts of the number of isoforms in the Iso-Seq targeted approach stratified by structural category. **(C)** A box plot of the full-length read counts of isoforms associated to target and non-target genes in the Iso-Seq global and **(D)** targeted datasets. Target genes refer to the panel of 20 AD-associated genes that were enriched for targeted sequencing. Only transcripts from matched samples were compared. FSM - Full Splice Match, ISM - Incomplete Splice Match, NIC - Novel In Catalogue, NNC - Novel Not in Catalogue.

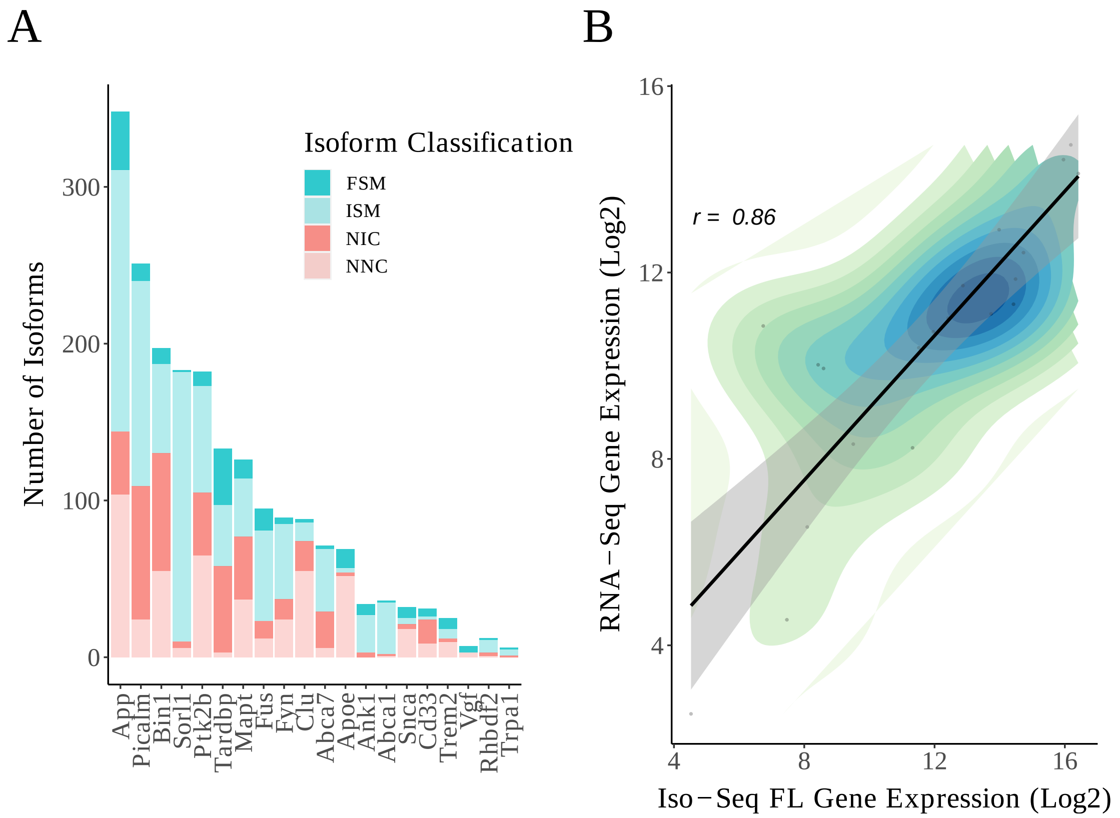


Figure 6.13: Widespread isoform diversity was observed in AD-associated genes with detection of many novel isoforms in the rTg4510 cortex. Shown is **(A)** a bar chart of the number of isoforms detected per AD-associated gene from the Iso-Seq targeted dataset, stratified by *SQANTI* classifications, and **(B)** a density plot of the RNA-Seq and Iso-Seq gene expression. Iso-Seq gene expression was determined from the summation of full-length read counts of associated transcripts, whereas RNA-Seq gene expression was deduced from the normalised *DESeq* counts of aligned RNA-Seq reads to the mouse reference genome.[90]

6.3.3 Confirmation that the *MAPT* transgene is only expressed in rTg4510 TG mice

The mouse *Mapt* gene was one of the 20 target genes enriched for the targeted profiling of the rTg4510 mouse cortex. Given the high homology between the mouse *Mapt* and human *MAPT* coding sequence (as seen in **Section 5.3.2**), we anticipated that the target enrichment approach would also capture the human *MAPT* transgene. As expected, BLAST analysis of the species-specific *Mapt/MAPT* sequences showed only detection of the human-specific *MAPT* sequences in reads from TG mice (**Figure 6.14A,B**), confirming stable activation of the human *MAPT* transgene. There was also an enrichment of full-length reads corresponding to the mouse *Mapt* gene, which was not previously noticeable in the Iso-Seq global dataset (**Figure 5.3A**), as a further testament to the success of the targeted experiments. Notably, there was a striking difference in the number of reads corresponding to the human *MAPT* transgene and the mouse *Mapt* gene in the ONT targeted dataset (**Figure 6.14B**) - a likely reflection of the extra-deep sequencing coverage provided using ONT nanopore sequencing following enrichment of the mouse *Mapt* gene.

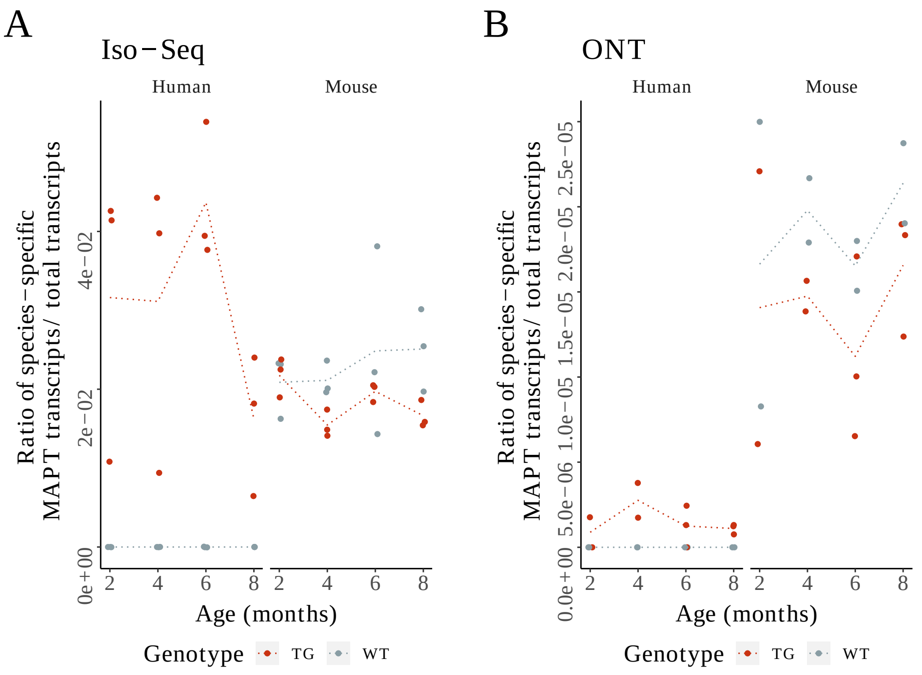


Figure 6.14: Human-specific *MAPT* sequences were only present in transgenic mice, with enrichment of mouse *Mapt* reads. Shown are two scatter plots of the proportion of full-length transcripts that were mapped to human-specific *MAPT* and mouse-specific *Mapt* sequences in the **(A)** Iso-Seq targeted dataset and **(B)** ONT targeted dataset. Red and grey dots refer to TG and WT samples, and dotted lines represent the mean paths across age. Of note, probes were designed to mouse *Mapt* gene and not the human *MAPT* gene.

6.3.4 ONT run performance and sequencing metrics

Following library preparation and nanopore sequencing on the majority of samples (n = 8 WT, n = 10 TG, **Table 6.3**), a total of 28.54M reads (39.68Gb) were generated across two flow cells (batch 2 and batch 3) and a total of 22.8M (80%) reads were successfully basecalled using *Guppy* (**Table 6.6**). Although both flow cells achieved good sequencing yield with similar read lengths, batch 3 had a significantly greater throughput and generated more basecalled reads after filtering (number of pre-filtered reads: batch 2 = 12.3M, batch 3 = 16.3M; number of post-filtered reads: batch 2 = 9.68M (78.8%), batch 3 = 13.13M (80.7%)). Evaluation of the run performance and QC using *PycoQC* revealed that this disparity was the result of lower sequencing channel activity in batch 2 (as shown in **Figure 6.15**). Consequently, the number of bases generated over the course of the run was significantly lower in batch 2 than in batch 3 (**Figure 6.16**).

Table 6.6: Comparable run performance and yield output from targeted nanopore sequencing of Tg4510 mice. Tabulated is a summary of the sequencing yield generated on a subset of mouse samples (n = 18) after sequencing on the ONT MinION using two separate flow cells over 48 hours (batch 2: n = 4 WT, n = 5 TG; batch 3: n = 4 WT, n = 5 TG). Following the ONT bioinformatics pipeline, raw ONT reads were basecalled and filtered by read accuracy (Phred, Q > 7). Active channels refer to the total number of channels that were detected with sequencing activity over the course of the run. N50 refers to the sequence length at which 50% of reads are sized at or over. Gb - Gigabases, kb- kilobases, M - Million.

Run	Active channels	Basecalled reads		Filtered basecalled reads					
		Total bases (Gb)	Number (M)	Total bases (Gb)	Number (M)	Read length (bp)			Mean read quality
						Mean	N50	Longest read (kb)	
Batch 2	479	16.9	12.2	14.2	9.68 (78.8%)	1,478	1,779	19.1	10.2
Batch 3	425	22.8	16.2	19.41	13.1 (80.7%)	1,468	1,813	20.5	9.9

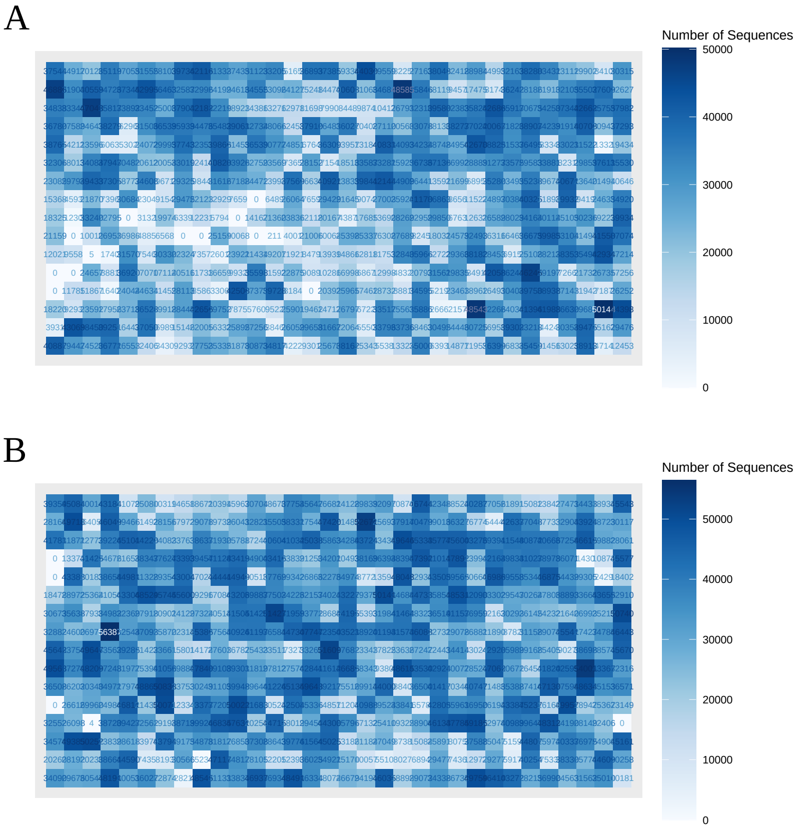


Figure 6.15: High sequencing channel activity from nanopore targeted sequencing of the rTg4510 cortex. Shown is a spatial heat-map representation of channel productivity for **(A)** batch 2 (sequencing yield = 16.9Gb) and **(B)** batch 3 (sequencing yield = 22.8Gb). Each square refers to a channel that contains four nanopores. The channel productivity is determined by the total number of DNA sequences (represented by a colour density from white to dark blue) that translocate through each nanopore in the duration of a given run.

A contrast in activity can be seen across the two runs. Notably, there a concentrated patch of inactive channels (white box, 0 reads translocated through pore of interest) in the batch 2 run, whereas there are significantly more active channels (dark blue, > 50,000 reads translocated through pore of interest) in the batch 3 run. Similar amounts of cDNA templates were loaded onto the flow cells (batch 2: 540ng, batch 3: 500ng).

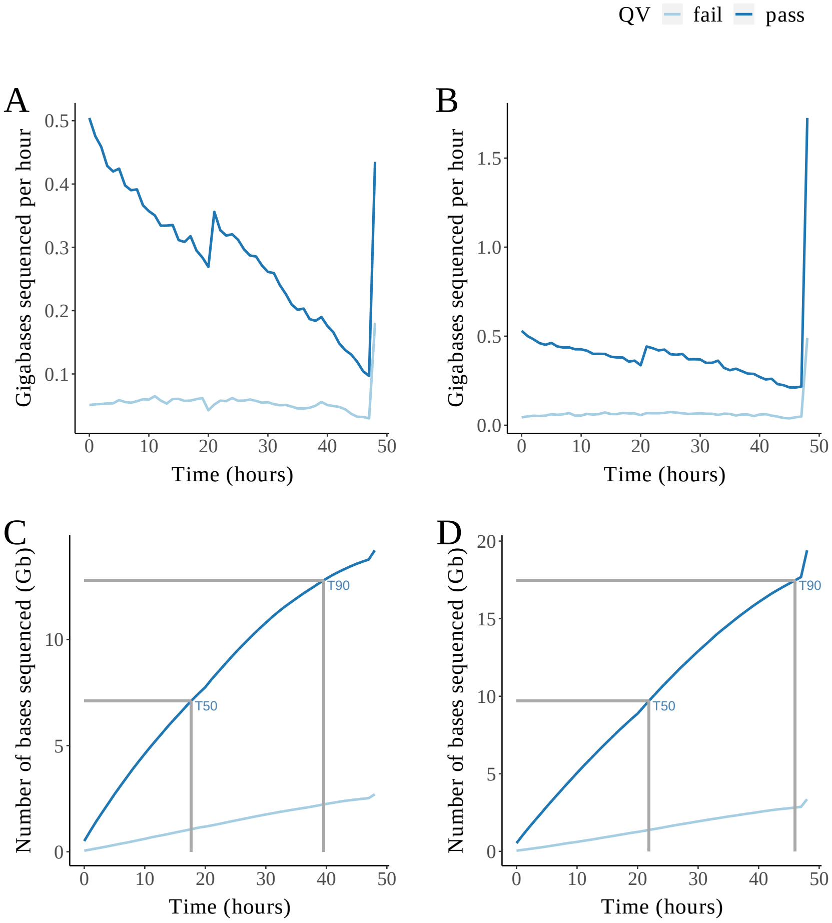


Figure 6.16: ONT temporal run performance for targeted profiling. Shown are time-series plots displaying the **(A)** number of bases generated per hour over the course of the run for batch 2 and **(B)** batch 3, and **(C)** the reads generated cumulatively for batch 2 and **(D)** batch 3. The reads were classified as "pass" (dark blue) if QV $\geq$ 7 and "fail" (light blue) if QV < 7. T50 and T90 refer to the time (hours) at which 50% and 90% of the total number of bases were sequenced. Gb - Gigabases.

Although the sequencing yield from ONT nanopore sequencing was comparable to Iso-Seq after target enrichment (Iso-Seq: range = 19.3Gb - 24.2Gb; ONT: range = 16.9Gb - 22.8Gb), we detected significantly more raw ONT reads (range: 12.3M - 16.3M) than Iso-Seq polymerase reads (range: 0.3M - 0.7M) and subsequently more full-length reads per sample (ONT mean basecalled, filtered reads: n = 918K, range = 667K - 1.32M, **Figure 6.19A**; Iso-Seq mean poly(A) FLNC reads: n = 38.7K, range = 16.8K - 77.2K, **Figure 6.10A**). The on-target rate from ONT nanopore sequencing was also comparable to that seen in Iso-Seq (**Figure 6.11**, **Figure 6.17**), suggesting that the ONT targeted dataset provides a deeper coverage of the target genes. We suspect that this reflects an inherent difference between the two technologies: an insert (cDNA sequence of interest) would be sequenced multiple times from multiple polymerase passes in Iso-Seq (**Figure 3.1A**), whereas the same insert would only be sequenced once following translocation in nanopore sequencing (**Figure 3.13A**). The yield in Iso-Seq was thus limited by the number of wells available for sequencing (1M ZMWs), whilst the yield in ONT nanopore sequencing was constrained by the amount of material and channel activity, which can be easily maximised to ensure a high throughput. However, we also observed that this high ONT sequencing yield was achieved at the expense of read accuracy, given reads were only sequenced once. The average ONT read accuracy was 90% (mean Phred Q = 10; **Table 6.6**, **Figure 6.18A,B**) in comparison to the 99.9% accuracy of Iso-Seq reads. Of note, this relatively low ONT read accuracy was expected and in line with the development of nanopore chemistry at the time of research (summarised in **Figure 3.16**).

Finally, we noted that the number of full-length reads obtained per sample varied within each batch (**Figure 6.19B**), reflecting the data from Iso-Seq targeted profiling. We also detected slightly more reads for TG than WT mice in batch 2 (Wilcoxon rank sum test: W = 18, *P* = 0.063) and batch 3 (Wilcoxon rank sum test: W = 17, *P* = 0.11) (**Figure 6.19C**). However, the difference was not significant at the 5% level after merging both datasets (Wilcoxon rank sum test: W = 59, *P* = 0.10, **Figure 6.19D**). Deeper examination revealed that this variability was not associated with RIN (Spearman's rank: corr = -0.267, *P* = 0.284) or barcode from multiplexing (Spearman's rank: corr = -0.058, *P* = 0.819, **Figure 6.19**), but a reflection of sequencing more TG samples across both runs (batch 2: n = 4 WT, n = 5 TG; batch 3: n = 4 WT, n = 5 TG). Notably, this was not an issue with Iso-Seq targeted profiling where the number of WT and TG mice was overall balanced after including samples from batch 1.

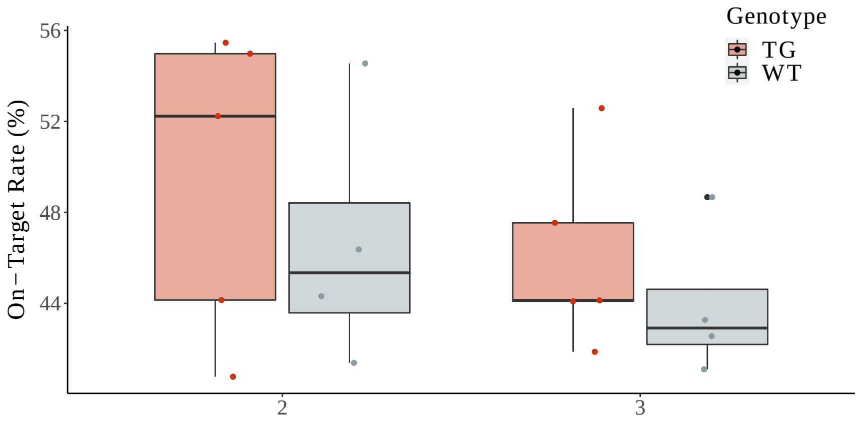


Figure 6.17: On-target rate of ONT nanopore targeted sequencing was comparable to Iso-Seq targeted profiling. Shown is a box plot of the on-target rate observed in ONT nanopore sequencing, which was similar to that observed in Iso-Seq targeted sequencing (**Figure 6.11**). Of note, a difference in the on-target rate between WT and TG in both batches is a likely reflection of the sample variability in sequencing (**Figure 6.19C,D**). TG - rTg4510 transgenic mice, WT - Wild-type.

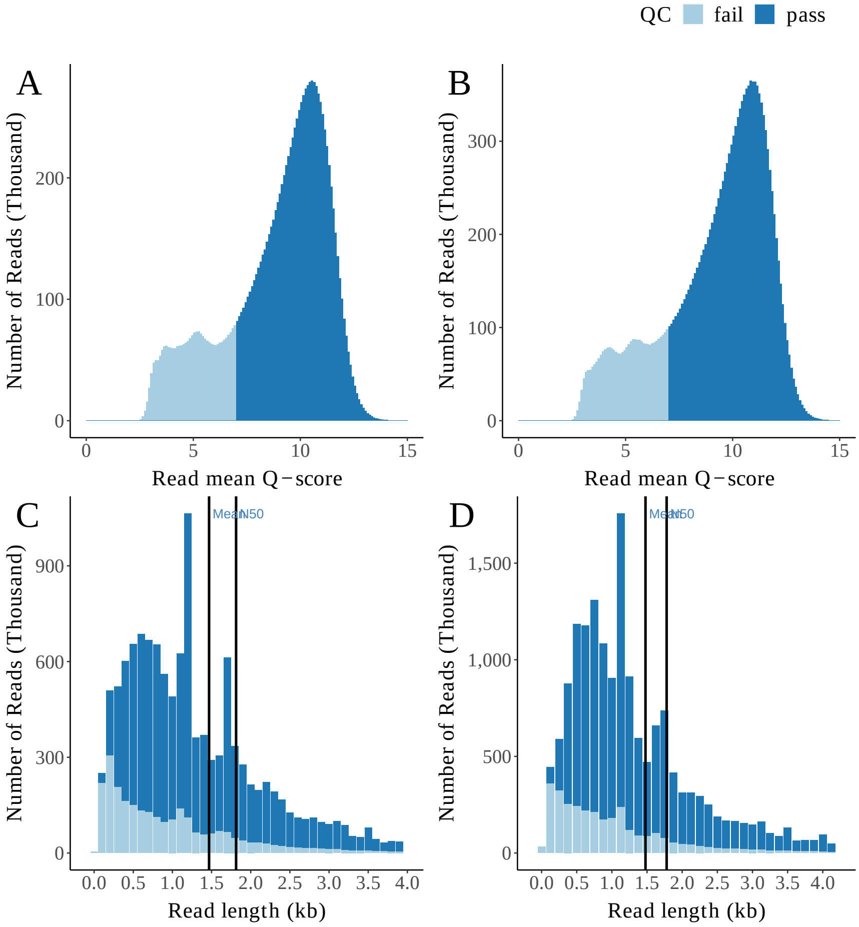


Figure 6.18: Expected length and quality distribution of ONT basecalled reads. Shown are histograms displaying the distribution of **(A)** mean read quality score of batch 2 and **(B)** batch 3, and **(C)** read lengths in batch 2 and **(D)** batch 3. The distribution is shaded by filtering: light blue for failed reads (Q < 7) and dark blue for passed reads (Q ≥ 7).

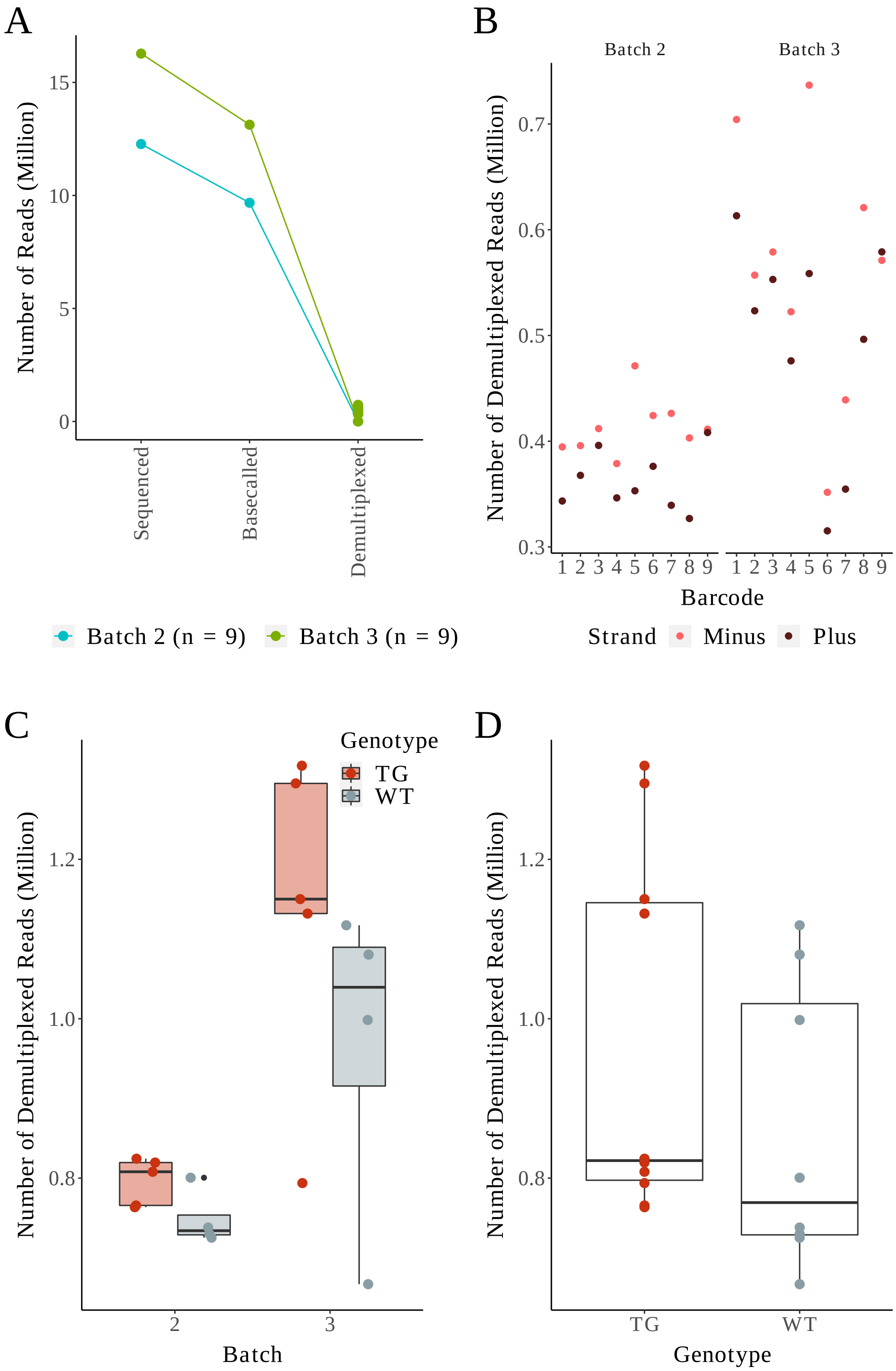


Figure 6.19: The number of reads generated by ONT nanopore targeted sequencing varied by batch and genotype. Shown are scatter plots of the **(A)** number of reads generated from the optimised ONT bioinformatics pipeline after basecalling and demultiplexing, **(B)** number of minus and plus reads for each sample after demultiplexing (see **Figure 3.19** for structure of ONT read template), and box plots of the **(C)** number of filtered basecalled reads by batch and genotype, and **(D)** by genotype after merging data from both runs. TG - rTg4510 transgenic mice, WT - Wild-type.

6.3.5 The vast majority of ONT transcripts were lowly abundant and not detected across biological replicates

Following the ONT bioinformatics pipeline, we detected a total of 1,367,866 isoforms in ONT targeted dataset of which 445,457 isoforms (32.5%) were annotated to 20 AD-associated genes enriched in the rTg4510 cortex (n = 8 WT, n = 10 TG). Filtering these isoforms by expression (minimum 5 reads across 2 samples) using *TALON*, however, drastically reduced the number of isoforms (fold change = -0.988) to 5,947 (1.19%) isoforms annotated to AD-associated genes. This suggests that the vast number of ONT transcripts were lowly abundant and not reproducibly detected across biological replicates. Nonetheless, we detected almost twice as many AD-associated isoforms (fold change = 1.66) using ONT nanopore sequencing (n = 5,331 isoforms) than Iso-Seq (n = 2,015 isoforms), despite sequencing fewer samples (Iso-Seq: n = 24 samples; ONT: n = 18 samples). In line with the sequencing yield generated by the respective technologies, this is again a reflection of the inherent differences in the two technologies (described in **Section 6.3.4**).

In order to compensate the opposing drawbacks of the two long-read targeted sequencing approaches (high accuracy but relatively lower sequencing coverage of Iso-Seq targeted dataset vs relatively lower accuracy but high sequencing coverage of ONT targeted dataset), we merged both targeted datasets using *GffCompare* to comprehensively characterise the AD-associated target genes in the rTg4510 cortex (depicted in **Figure 6.7** and described in **Section 6.2.6**). This strategy further allowed us to perform stringent filtering, while also retaining rare ONT-derived isoforms detected in the Iso-Seq targeted dataset that would have otherwise been filtered.

Comparison of the two datasets using custom scripts revealed that the majority of Iso-Seq-derived transcripts were also detected in ONT nanopore sequencing (n = 617 transcripts, 65.4%), whereas only a relatively small proportion of filtered ONT-transcripts were detected in the Iso-Seq targeted dataset (n = 701 transcripts, 15.1%) (**Figure 6.20**). Examination of the unique filtered ONT-derived transcripts revealed them to be shorter (W = 3.14 x 10^6, *P* = 4.0 x 10^{-125}, **Figure 6.21A**) and with fewer exons (W = 3.03 x 10^6, *P* = 1.37 x 10^{-100}, **Figure 6.21B**) than the commonly-detected ONT-derived transcripts. In contrast, no difference in isoform length (W = 1.76 x 10^5, *P* = 0.72, **Figure 6.21A**) was observed between the unique and commonly-detected Iso-Seq derived transcripts, suggesting that these transcripts might be unique to the remaining samples that were not sequenced with ONT. Finally, we observed that the unique Iso-Seq-derived transcripts and ONT-derived transcripts were more abundant than the commonly-detected transcripts (**Figure 6.21C**), suggesting that transcript expression was not a differentiating factor for whether a transcript was detected in one technology and not the other.

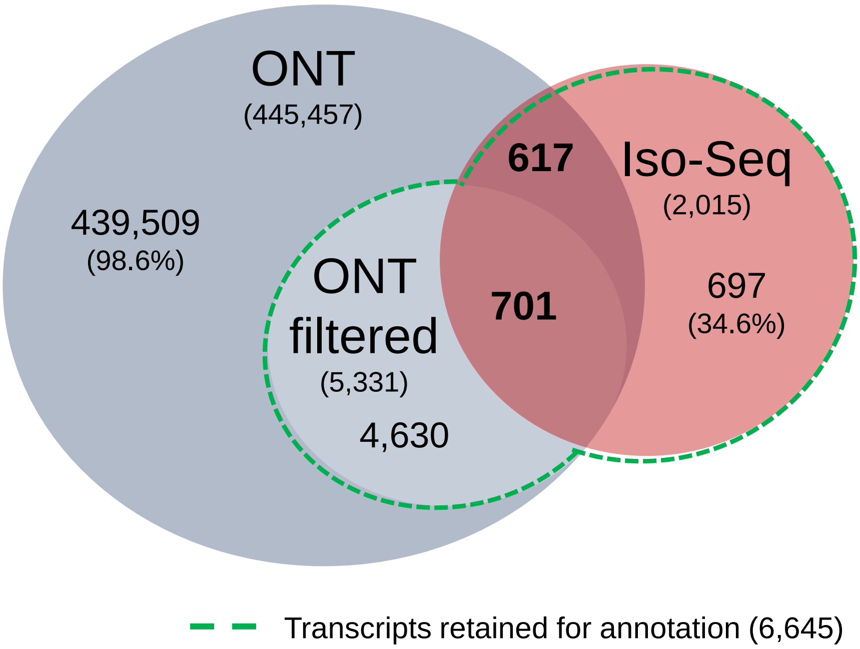


Figure 6.20: Total number of transcripts detected from Iso-Seq and ONT targeted sequencing. Shown is a Venn diagram of the total number of transcripts annotated to the 20 AD-associated target genes detected in Iso-Seq (shaded red) and ONT (shaded blue) targeted datasets. "ONT filtered" transcripts refer to the subset of ONT transcripts that were retained after *TALON* filtering (minimum 5 reads in at least 2 samples). Transcripts in the overlapping sector were defined as complete exact match (class code: "=") using *GffCompare*. The green dash encompasses the subset of transcripts from ONT and Iso-Seq targeted datasets that were taken further for downstream annotation and isoform-based analyses.

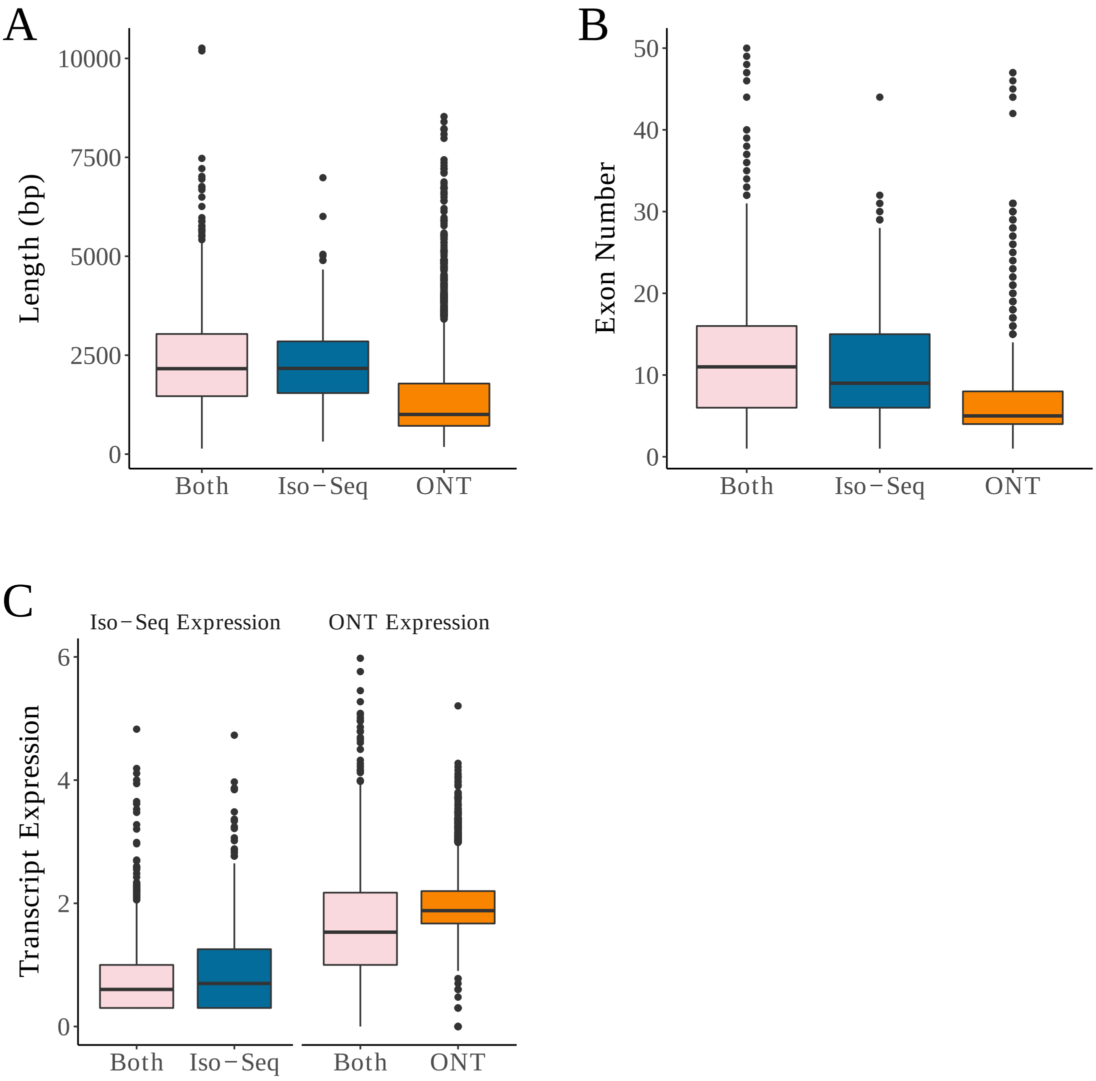


Figure 6.21: Transcripts detected in both Iso-Seq and ONT targeted datasets were more abundant, and longer with more exons than isoforms unique to ONT dataset. Shown are box plots of the **(A)** length, **(B)** exon number and **(C)** expression of transcripts annotated to AD-associated genes (target genes) that were either detected in both Iso-Seq and ONT targeted datasets, or unique to the Iso-Seq (n = 24 samples) and ONT dataset (n = 18 samples). "Both" refer to transcripts that were detected in Iso-Seq targeted and ONT targeted dataset, "Iso-Seq" and "ONT" refer to transcripts that were only detected in Iso-Seq and ONT targeted datasets, respectively. The Iso-Seq and ONT transcript expression refer to the respective full-length read count for the associated isoform.

6.3.6 ONT nanopore sequencing achieves significantly deeper sequencing coverage than Iso-Seq with enrichment of shorter novel transcripts

Following merging of the two targeted datasets, we detected a total of 6,645 isoforms annotated to the 20 AD-associated target genes (**Figure 6.20**). Among these isoforms, the majority were solely derived from ONT nanopore sequencing (n = 4,630 isoforms, 69.7%) with a considerable overlap between the two targeted datasets (n = 1,318 isoforms, 19.8%) (**Figure 6.20**). Landscape evaluation of these retained ONT-derived isoforms (**Figure 6.22A**) revealed a striking contrast to the Iso-Seq targeted dataset (**Figure 6.13A**), with an overwhelming majority of isoforms identified as novel and NNC (n = 4,728 isoforms) with novel combination of splice junctions. Significantly more isoforms were detected across the panel of target genes, particularly *Apoe* with over 2000 isoforms detected. Conversely, *Apoe* was one of the fewer "isoformic" gene in the Iso-Seq targeted dataset with only 69 isoforms detected (**Figure 6.13A**), indicating robust differences between the two targeted datasets.

After further filtering of partial isoforms including 5' degradation products, examination of the AD-associated target isoforms (n = 5,587 isoforms) detected using Iso-Seq and ONT nanopore sequencing revealed a clear difference in isoform length and number of exons. While transcripts in the Iso-Seq targeted dataset were typically sized 2-3kb (mean = 2.38kb, s.d = 1.2kb, range = 0.14kb - 10.3kb, **Figure 6.23A**) with 10 exons (s.d = 7.87, range = 1 - 50, **Figure 6.23B**), we noted a prominent enrichment of short transcripts in the ONT targeted dataset sized 1-2kb (mean = 1.79kb, s.d = 1.45kb, range = 0.153kb - 10.7kb, **Figure 6.23A**) with 5 exons (s.d = 7.06, range = 1 - 50, **Figure 6.23B**). This suggests an over-representation of shorter transcripts in the ONT library, a phenomenon that has been previously reported and may be attributed to premature termination of ONT nanopore sequencing.[187] However notably, we observed a similar distribution of distance to the nearest annotated CAGE peak (**Figure 6.23C**), annotated transcription start sites (**Figure 6.23D**) and termination sites (**Figure 6.23E**) with ONT-derived transcripts more likely to be annotated within 50bp of a CAGE peak, TSS and TTS. The proportion of isoforms with the presence of a known poly(A) site was also similar between Iso-Seq-derived and ONT-derived transcripts, with high support across all isoform categories (**Figure 6.23F**), indicating transcript completeness at both the 5' and 3' end. In summary, these results support the validity of these novel ONT-derived transcripts.

Finally, using RNA-Seq data generated on matched samples, we found that ~80% (n = 712, 77.7%) of transcripts detected in both targeted datasets were supported by RNA-Seq reads. Unsurprisingly, this support was low for transcripts unique to the Iso-Seq targeted dataset (n = 243, 63.9%), and significantly lower for ONT targeted dataset (n = 305, 6.63%) - a reflection of the higher sensitivity of ONT targeted sequencing to detect rarer novel transcripts and the insufficient coverage of RNA-Seq reads to span the junctions of such transcripts. Nonetheless, in comparing the gene-expression level deduced from RNA-Seq data vs ONT targeted data, we observed an even stronger correlation (Pearson's correlation: corr = 0.92, P = 1.27 x 10^{-8}, **Figure 6.22B**) than when the comparison was made with Iso-Seq targeted data (Pearson's correlation: corr = 0.86, P = 1.18 x 10^{-6}, **Figure 6.13B**), a further testament to the deep nanopore sequencing coverage.

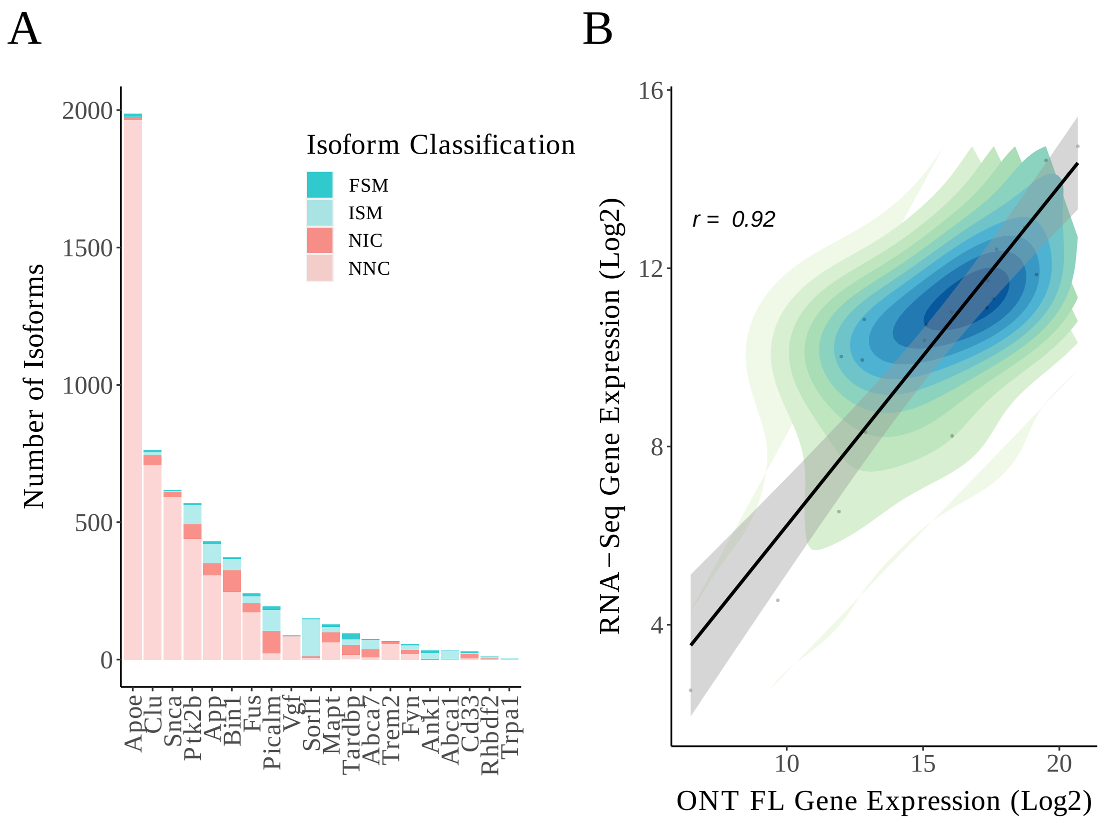


Figure 6.22: ONT is more sensitive than Iso-Seq with greater power to detect novel transcripts. Shown is the **(A)** number of isoforms detected per target gene from the ONT targeted dataset, either classified as known (FSM, ISM) or novel (NIC, NNC). **(B)** A strong correlation was observed between ONT gene expression and RNA-Seq gene expression. ONT gene expression was determined from the summation of full-length read counts of associated transcripts, whereas RNA-Seq gene expression was deduced from the normalised *DESeq* counts of aligned RNA-Seq reads to the mouse reference genome.[90]

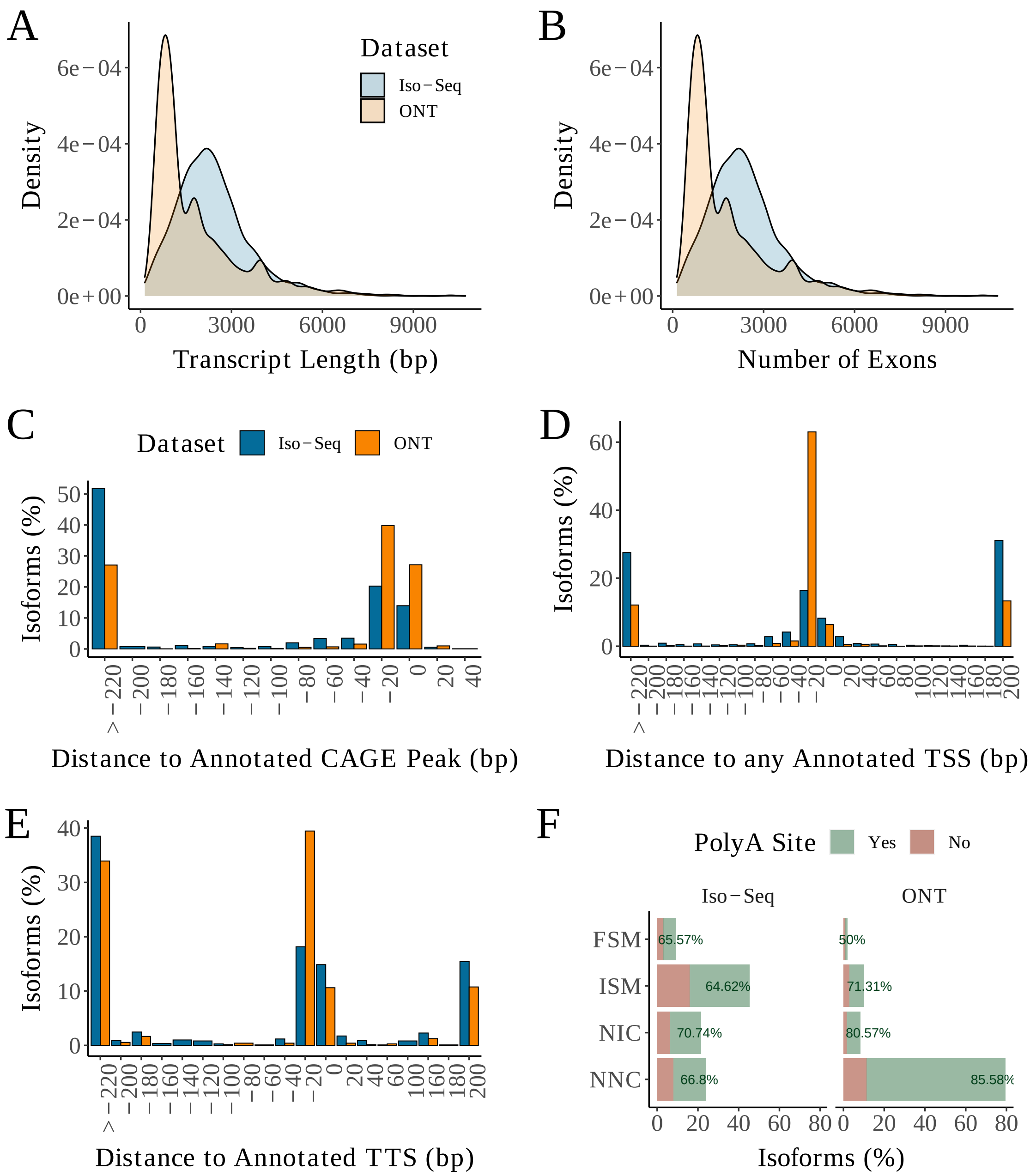


Figure 6.23: While ONT-derived isoforms were generally shorter with fewer exons, the 5' and 3' ends are more within the range of annotated sites and CAGE peaks than Iso-Seq derived isoforms. Shown are density plots of **(A)** the distribution of the transcript lengths, and **(B)** exon number in the targeted Iso-Seq (n = 24 samples) and ONT datasets (n = 16 samples). Distance between **(C)** TSS and closest annotated CAGE peak (a negative value refers to a CAGE peak located upstream of TSS), **(D)** TSS and reference TSS (a negative value refers to a query start downstream of reference), **(E)** TTS and reference TSS (a negative value refers to a query end upstream of reference). **(F)** A bar chart of the proportion of isoforms in the Iso-Seq and ONT targeted dataset with a poly(A) site. Percentages denoted in green refer to the proportion of isoforms within the respective category with a poly(A) site. TSS - Transcription start site, TTS - Transcription termination site. Iso-Seq and ONT refer to isoforms from the Iso-Seq and ONT targeted dataset, respectively.

6.3.7 Characterisation of AS events in AD-associated genes

The depth of sequencing coverage achieved with target gene enrichment, particularly with ONT nanopore sequencing, enabled us to identify hundreds of novel transcripts across the panel of AD-associated genes (**Figure 6.24**, **Table 6.7**). Using custom scripts developed to comprehensively annotate such transcripts (illustrated in **Figure 6.8** and described in **Section 6.2.8**), we identified widespread alternative splicing events (n = 17,826 events, **Table 6.8**) in our panel of AD-associated target genes.

In line with our previous findings from global transcriptome profiling of the mouse cortex (**Section 4.3.8**), we observed widespread usage of alternative 5' and 3' splice site (A5', A3') (n = 8,520 events, 47.8%) followed by exon skipping (n = 6,695 events, 37.6%). Usage of alternative splice sites, defined as a site differing by 10 - 20bp of the reference splice site, was detected for all AD-associated target genes (**Figure 6.26A**). While there were gene-specific variations (**Figure 6.26B**), such as *Vgf* with extensive usage of alternative splice sites of the final exon (alternative last exon) (**Figure 6.29A**), the majority of genes were dominated by alternative 5' and 3' splice sites of internal exons.

Focusing on internal exons, we detected thousands of AD-associated target isoforms with exon skipping (n = 2,431 isoforms, **Figure 6.27A**). Although the vast majority of transcripts were characterised with skipping of several exons, some genes were characterised by significantly more exon skipping events; a number of isoforms annotated to *App* (n = 23 isoforms, **Figure 6.29B**) and *Bin1* (n = 21 isoforms, **Figure 6.29C**) were characterised with skipping of > 10 exons (**Figure 6.27B**). In contrast to the initial theory that exon skipping predominantly occurs with "alternative exons" (i.e. exons that are not present in all reference isoforms), deeper investigation revealed that over a third of the total exons skipped (n = 2,591 exons, 38.5%) were "constitutive" (i.e. exons present in all reference isoforms) (**Figure 6.27C**). Furthermore, several genes were characterised with widespread exon skipping; across the 358 isoforms annotated to *Ptk2b*, 93.5% (n = 29 exons) of *Ptk2b* exons (n = 31 total exons) were found skipped, the overwhelming majority of which were constitutive (n = 28, 96.6% of skipped exons) (**Figure 6.27C**). No correlation was observed between the known number of exons and number of exon skipping events (Pearson's correlation: corr = -0.195, *P* = 0.409).

Conversely, intron retention (IR) was one of the least observed splicing event characterised (n = 747 events, 4.19%), corroborating previous findings from global transcriptome profiling of the mouse cortex (**Section 4.3.8**). Although the majority of IR transcripts were characterised with only one distinct IR event (defined by the presence of an exon with retention of an intronic region spanning more than > 100bp from the reference splice site), several genes were associated with a few novel rare transcripts with multiple IR events (**Figure 6.28A**); this was particularly evident in *Abca7* (4 IR events: n = 1 transcripts; 3 IR events: n = 3 transcripts, **Figure 6.29D**). The majority of IR events were further found to span across at least two exons, with a significant proportion of isoforms characterised with extensive intron retention spanning across 4 exons (**Figure 6.28B**). Finally, we found an association between increased intron retention events and transcript expression (**Figure 6.28C**), corroborating our findings (described in **Section 4.3.9**) and previous studies[368] suggesting that IR is associated with reduced transcript abundance.

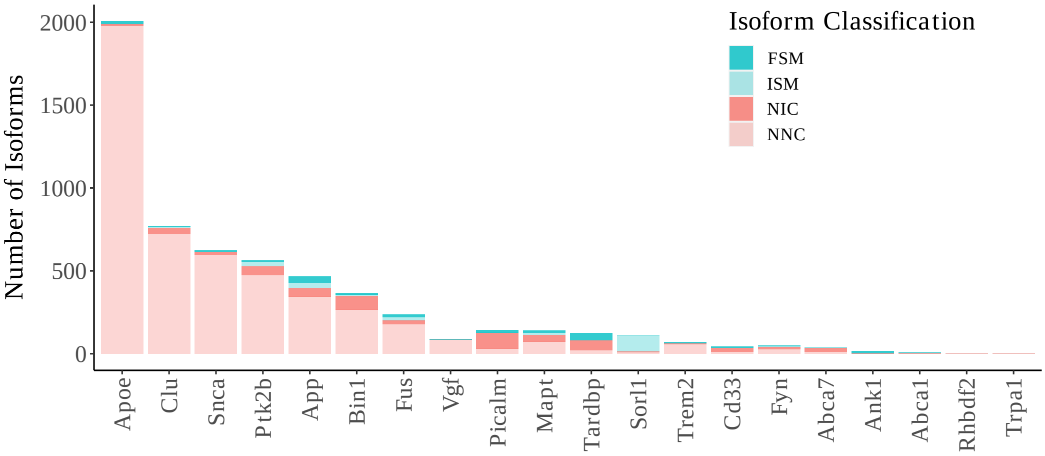


Figure 6.24: Targeted profiling of the rTg4510 cortex identify hundreds of novel transcripts annotated to AD-associated target genes. Shown is a bar chart of the final number of isoforms detected per AD-associated gene after merging Iso-Seq and ONT targeted datasets.

Table 6.7: Overview of the isoform landscape of AD-associated genes. Tabulated is a summary of the isoform landscape per AD-associated gene after targeted profiling of the rTg4510 cortex. Splice junctions were defined by the two pairs of dinucleotides present at the exon-intron boundary, and any other combinations aside from GT-AG, GC-AG and AT-AC pairs were considered non-canonical.

Gene	Number of isoforms						
	Classification			Coding potential		Splice junctions	
	All	Known	Novel	Coding	Non-coding	Canonical	Non-canonical
Abca1	7	1	2	7	0	7	0
Abca7	41	3	34	40	1	41	0
Ank1	17	9	3	15	2	17	0
Apoe	2006	10	1987	1978	28	980	1030
App	466	9	398	451	15	410	56
Bin1	368	6	348	366	2	347	21
Cd33	41	5	34	39	2	43	1
Clu	773	7	756	757	16	457	316
Fus	236	10	200	223	13	219	22
Fyn	50	5	40	48	2	50	0
Mapt	140	9	113	115	25	136	5
Picalm	144	12	126	117	27	144	2
Ptk2b	563	7	528	553	10	413	150
Rhbdf2	5	1	4	5	0	5	0
Snca	622	3	614	220	402	468	154
Sorl1	113	3	12	104	9	113	1
Tardbp	127	21	80	105	22	124	3
Trem2	140	3	61	63	77	66	4
Trpa1	4	1	3	3	1	4	0
Vgf	90	2	86	52	38	19	71

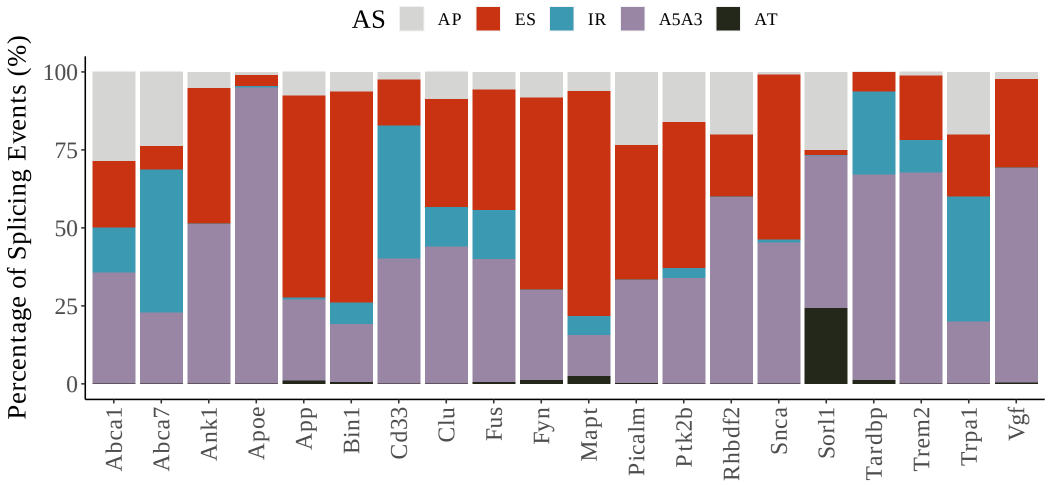


Figure 6.25: Targeted profiling of the rTg4510 cortex identify widespread usage of splicing events. Shown is a bar chart of the proportion of alternative splicing events identified per AD-associated gene after merging Iso-Seq and ONT targeted datasets. AS - Alternative splicing, AT - Alternative termination, ES - Exon skipping, IR - Intron retention.

Table 6.8: Characterisation of alternative splicing events in AD-associated genes. Tabulated is a summary of alternative splicing events per AD-associated gene after targeted profiling of the rTg4510 cortex. AF - Alternative first exon, AP - Alternative promoter, AS - Alternative splicing, AT - Alternative termination, ES - Exon skipping, IR - Intron retention.

Gene	Number of transcripts		Number of events					
	ES	IR	A5A3	AF	AP	AT	ES	IR
Abca1	2	2	5	0	4	0	3	2
Abca7	9	25	27	0	28	0	9	54
Ank1	9	0	20	0	2	0	17	0
Apoe	125	8	4054	0	40	2	156	16
App	406	13	752	0	218	28	1872	17
Bin1	290	52	295	0	101	9	1072	109
Cd33	18	23	49	0	3	0	18	52
Clu	270	119	844	0	169	0	660	241
Fus	103	32	200	0	29	3	196	79
Fyn	44	0	46	0	13	2	98	0
Mapt	115	24	97	0	45	18	527	44
Picalm	83	0	114	0	81	1	148	0
Ptk2b	358	33	632	0	301	0	873	57
Rhbdf2	2	0	6	0	2	0	2	0
Snca	483	14	784	0	16	2	917	14
Sorl1	3	0	193	0	99	96	6	0
Tardbp	8	38	104	0	0	2	10	42
Trem2	28	12	118	0	2	0	36	18
Trpa1	1	2	1	0	1	0	1	2
Vgf	74	0	179	0	6	1	74	0

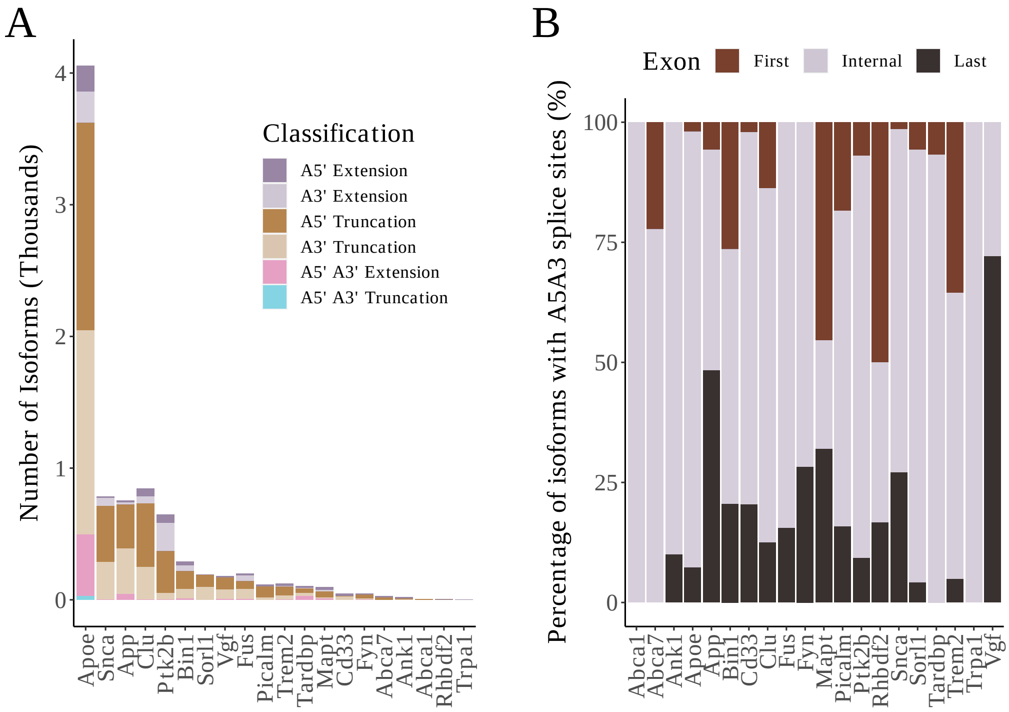


Figure 6.26: Extensive usage of alternative 5' and 3' splice sites in AD-associated genes. Shown are bar charts of the **(A)** number of isoforms annotated to AD-associated genes with alternative 5' (A5') and alternative 3' (A3') splice sites, further classified by the relative location of the novel splice site to the reference splice site (refer to **Figure 6.8** for further details), and of the **(B)** proportion of isoforms with alternative splice sites in the first, internal and last exons.

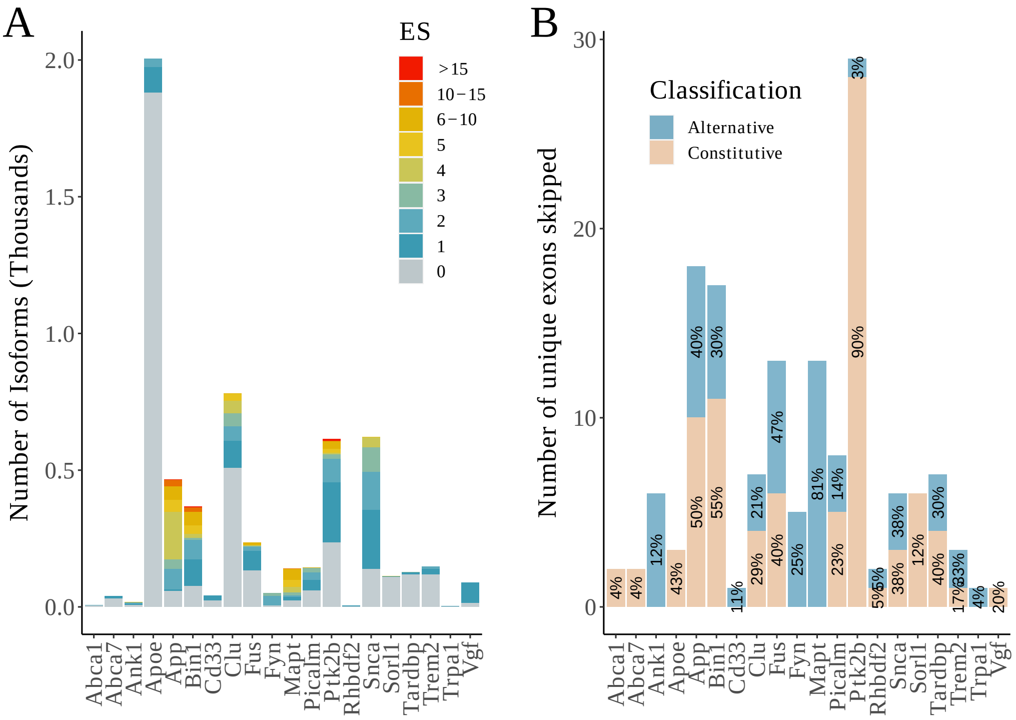


Figure 6.27: Extensive occurrence of exon skipping events in AD-associated genes. Shown are bar charts of the **(A)** number of isoforms annotated to AD-associated genes with exon skipping events, **(B)** number of unique exons skipped. Constitutive exons refer to exons that are present in all known reference isoforms.

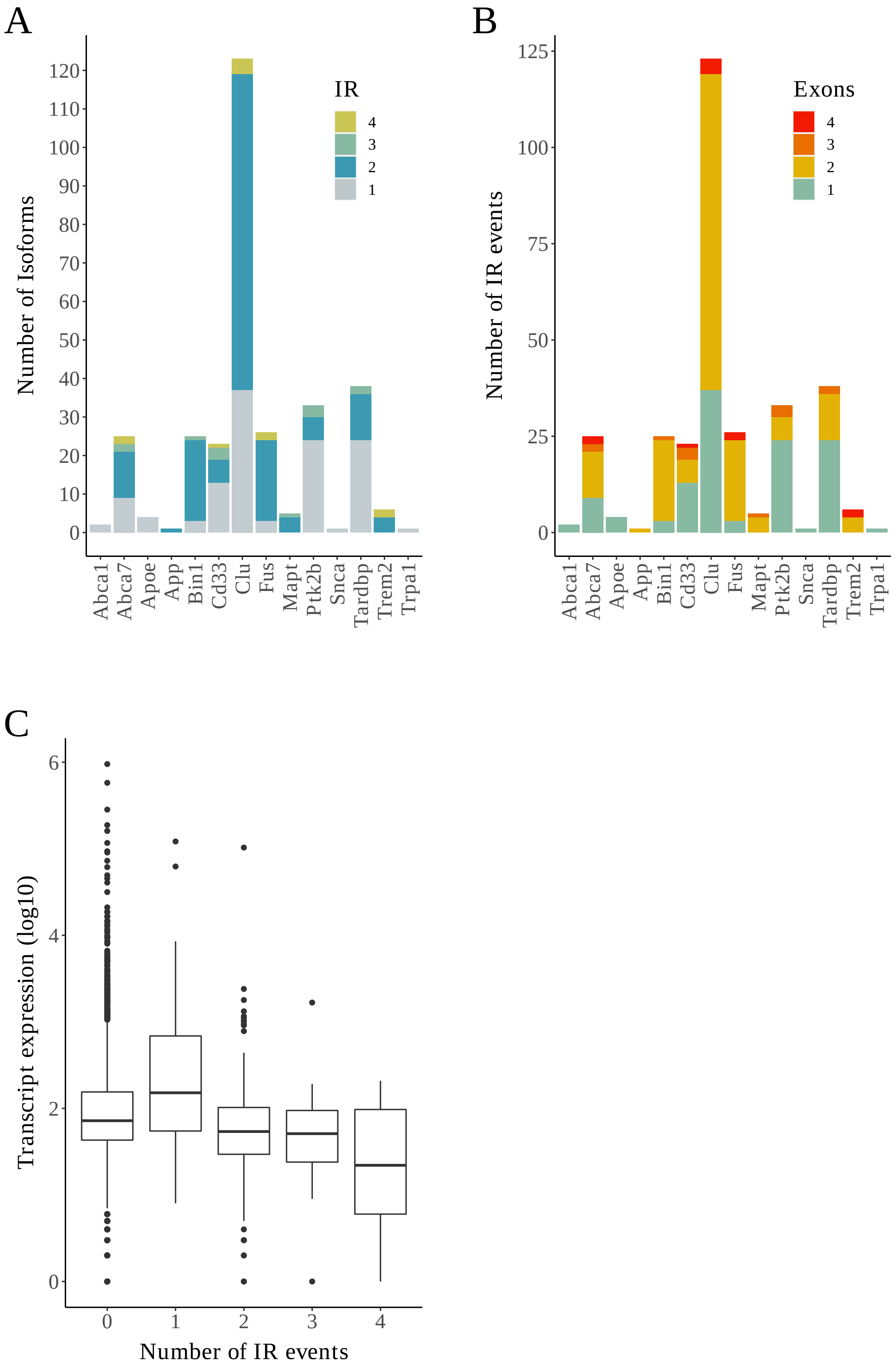


Figure 6.28: Relatively few occurrence of intron retention events, which typically spanned across two exons in lowly-expressed transcripts. Shown is **(A)** a bar chart of the number of isoforms annotated to AD-associated gene with intron retention events, and **(B)** the number of IR events and exons for which intron retention span across, and **(C)** a box plot of the expression of transcripts with multiple intron retention events. Transcript expression is deduced from normalised ONT full-length read counts.

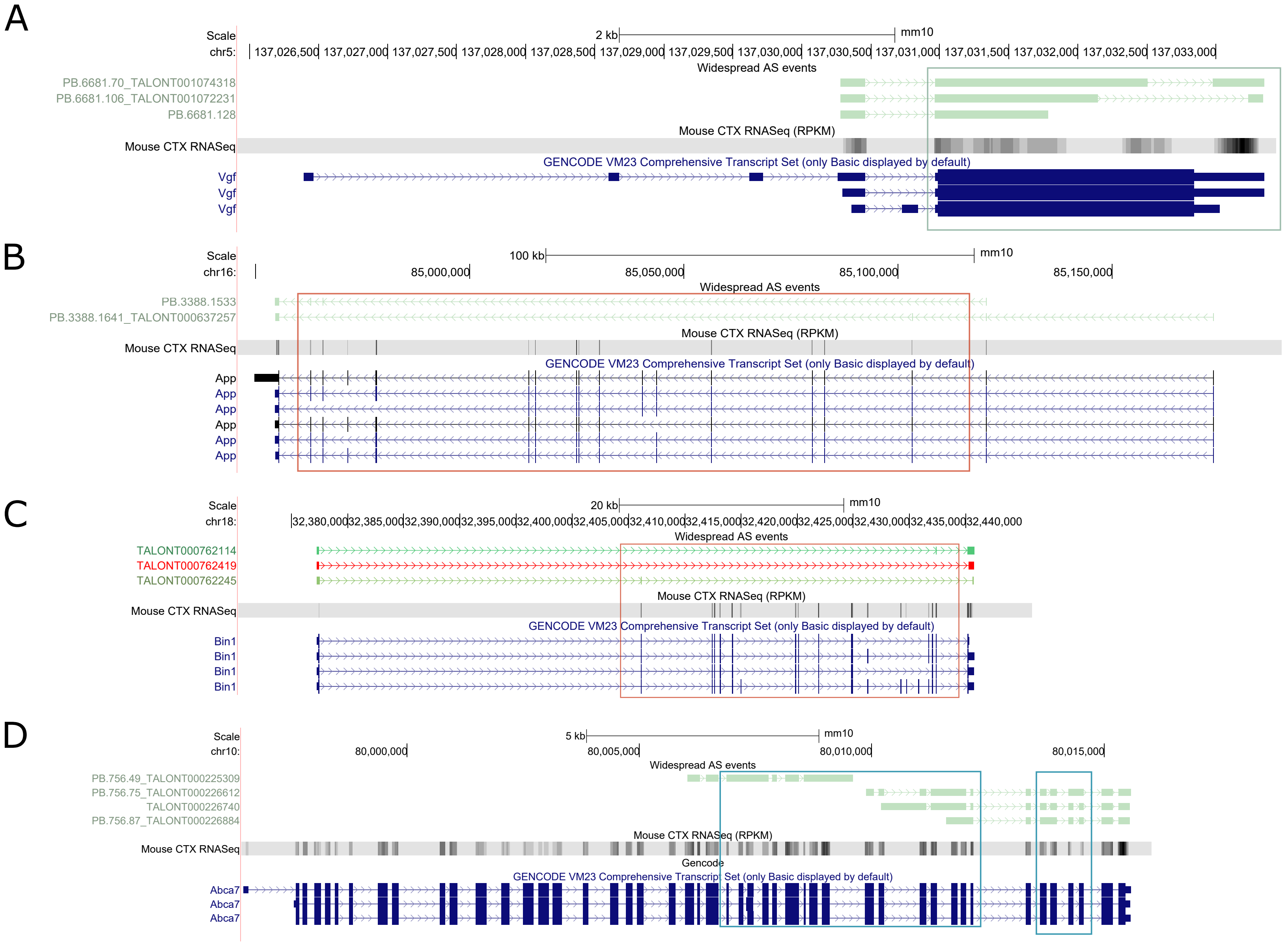


Figure 6.29: Examples of AD-associated genes with extensive usage of splicing events. Shown are UCSC genome browser tracks of **(A)** three *Vgf*-associated isoforms with usage of alternative splice sites in the last exon (green box), **(B)** two *App*-associated isoforms **(C)** three *Bin1*-associated isoforms with multiple exons skipped (> 15 exons, red box), and **(D)** four *Abca7*-associated isoforms with multiple intron retention events (blue box).

6.3.8 Comprehensive characterisation of AD-associated genes

This section details comprehensive transcript annotations from the merged Iso-Seq and ONT targeted datasets, which were generated from the selective profiling of 20 AD-associated genes in the rTg4510 mouse entorhinal cortex. A series of UCSC genome browser tracks and cluster dendrograms were generated for visualisation of each target gene. Examples of these can be found in **Figure 6.30** and **Figure 6.31**, respectively.

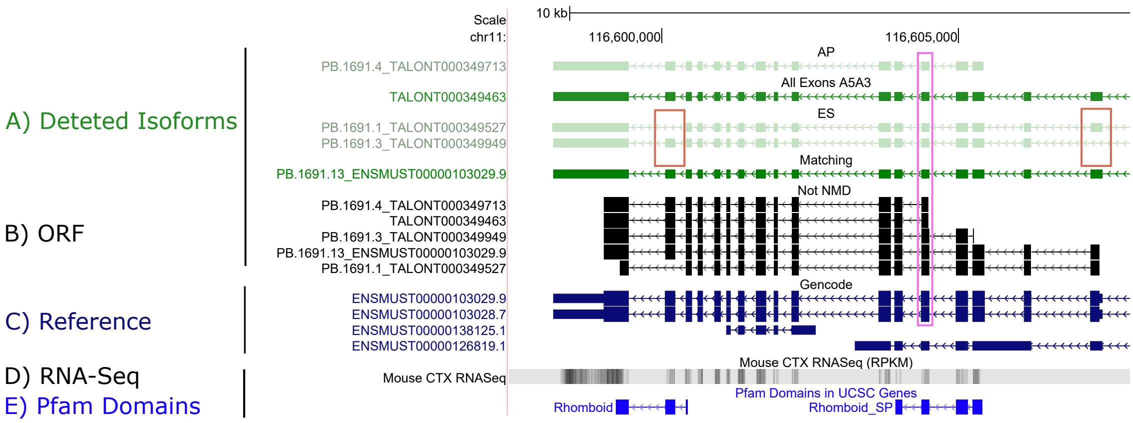


Figure 6.30: Example of a UCSC genome browser track of an AD-associated target gene. Shown is an example of a UCSC genome browser track of isoforms annotated to *Cd33*. Tracks are typically displayed in four panels in the following order:

- **(A)** isoforms detected from the merged targeted dataset of the rTg4510 cortex (n = 12 WT, n = 12 TG) are coloured by protein-coding potential (green for protein-coding, red for non-protein-coding) and shaded by abundance. Isoforms detected using both Iso-Seq and ONT nanopore sequencing are labelled with two IDs separated by an underscore ("_") with the first and second part denoting to the PacBio and ONT isoform IDs, respectively. Conversely, isoforms unique to the Iso-Seq and ONT targeted datasets are labelled as "PB.XXX" and "TALONTXXXX", respectively.
- **(B)** predicted open reading frames (black) using *CPAT*.
- **(C)** known isoforms (blue) from existing mouse reference annotations (mm10, GENCODE, vM22).
- **(D)** RNA-Seq data from global transcriptome profiling of the rTg4510 cortex[90] (n = 30 WT, n = 29 TG).
- **(E)** Pfam domains from the pfam database - a large curated collection of protein families and domains - as part of UCSC genome browser tracks.

Of note, only selected isoforms of interest are displayed given the extremely large number of isoforms detected from targeted sequencing. For genes supplemented with several UCSC tracks, some of the tracks exclude display of RNA-Seq and Pfam domains. Some tracks may also be presented in "squish" mode with compression of intronic regions and exon-only display to ease visualisation.

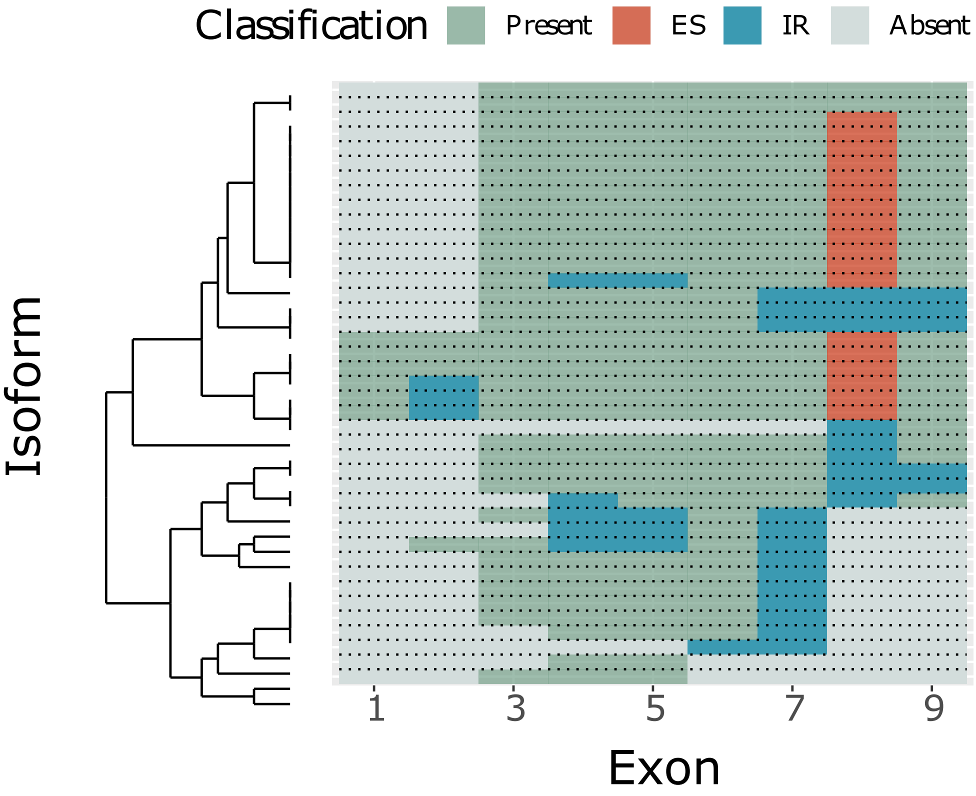


Figure 6.31: Example of a cluster dendrogram of an AD-associated target gene. Shown is an example of a cluster dendrogram of the isoforms annotated to *Cd33*. Of note, only the final, merged set of isoforms from the stringent processing and filtering of Iso-Seq and ONT targeted datasets are displayed. Each row corresponds to an isoform and each column represents an exon. The isoforms are further clustered by exonic structure and two key splicing events - exon skipping (ES) and intron retention (IR), to ease visualisation. Providing an overview of the isoform landscape, we can evidently see occurrence of intron retention events (box blue) in almost all the exons except for exon 1 and 3, and prevalent skipping of exon 8.

6.3.8.1 *Abca1*

The ATP-binding cassette transport A1 gene, *ABCA1*, is hypothesised to be a risk gene for AD as a consequence of its role in cholesterol transport and lipid metabolism.[302] Involved in ApoE lipidation, Abca1 has been shown to facilitate Aβ clearance in mouse models.[304]

Spanning over 129kb on chromosome 4, the mouse *Abca1* gene is characterised with 50 unique exons and two known isoforms. Despite the large number of exons, we only detected 7 isoforms annotated to *Abca1* in the rTg4510 mouse cortex (**Figure 6.32A**), including the known long canonical isoform (Abca1-201, ENSMUST00000030010.3). While most of the isoforms were significantly shorter, we identified a long isoform (PB.5746.2_TALONT000974760) that spanned the length of Abca1-201 (**Figure 6.32A**). Sequenced using both Iso-Seq and ONT technologies, this isoform differed from Abca1-201 with skipping of exons 24 and 35 (**Figure 6.32B**) and the presence of a novel exon located between exons 24 and 25 (**Figure 6.32C**). Skipping of exon 35, which partially encodes the transmembrane domain (ABC2 membrane 3), was also observed in one of the shorter isoforms (PB.5746.23_TALONT000972642). Embedded in the membrane bilayer, the *Abca1* transmembrane domain is involved in substrate transport across the membrane. While it is less conserved than the ATP-binding domain, the structure of the transmembrane domain determines the specificity and binding affinity for substrates. ORF prediction showed that skipping of exon 35 shortened but maintained the open reading frame, and inclusion of the novel exon did not translate to additional protein domains. In contrast to this long isoform, we identified two shorter isoforms (PB.5746.81, PB.5746.85_TALONT000975469) with intron retention in the last exon (**Figure 6.32A,B**). ORF prediction of such transcripts showed a shortened but similar reading frame with no prediction of nonsense-mediated decay.

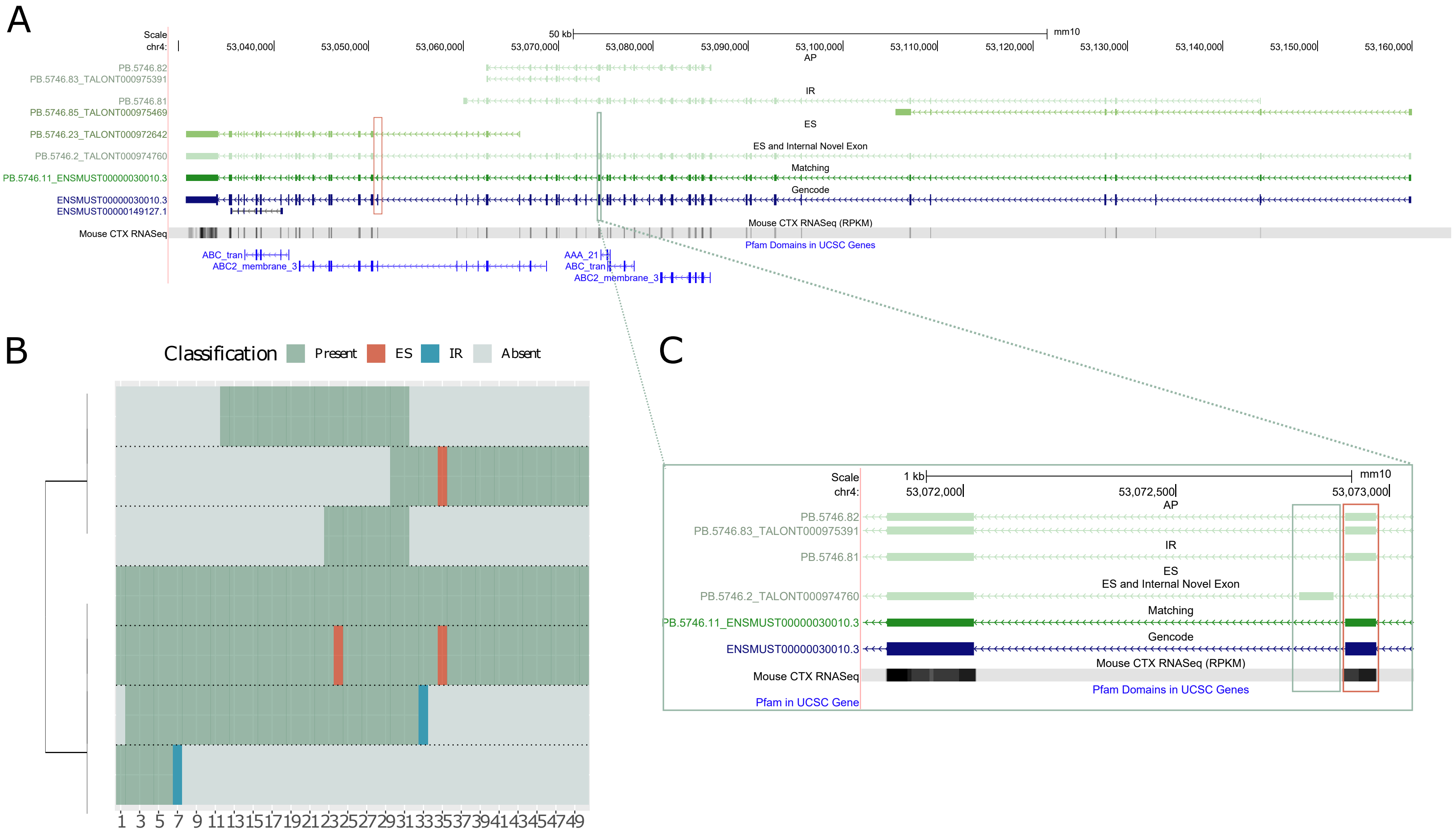


Figure 6.32: Characterisation of *Abca1* isoforms in the rTg4510 cortex. Shown are **(A)** a UCSC genome browser track of isoforms annotated to *Abca1*, **(B)** a cluster dendrogram for an overview of the *Abca1* isoform landscape, and **(C)** a zoomed-in track showing skipping of exon 24 (boxed red) and inclusion of a novel exon (boxed green).

6.3.8.2 *Abca7*

Another member of the ATP-binding cassette transporters, the ATP-binding cassette transport A7 gene, *ABCA7*, is also a risk gene for AD with the identification of both common and rare variants associated with the disease.[305–307] Analogous to ABCA1, ABCA7 is involved in regulating lipid transport and metabolism, including ApoE lipidation.[369]

Spanning ~20kb on chromosome 10, the mouse *Abca7* gene is characterised with 47 unique exons and three known isoforms. In contrast to *Abca1*, we detected 41 isoforms associated with *Abca7*. A reflection of the almost-identical structure of the three known isoforms (**Figure 6.33D**), we identified four novel isoforms that spanned the full-length of the gene, shared similar exonic structure, and only slightly differed at the splice sites ("wobble", < 10bp) (**Figure 6.33D**). In contrast to these long isoforms, we also detected significantly shorter isoforms that broadly fell into 2 categories: i) isoforms that preserved the exonic structure at the 5' end with an alternative last exon characterised with intron retention (n = 5 isoforms, 12.2%), and ii) isoforms with an alternative first exon and a 3' end characterised with exon skipping and intron retention events (n = 23 isoforms, 56%) (**Figure 6.33A**). Exon skipping was further found exclusive to exon 31 (n = 1 isoform, 2.4%) and exon 32 (n = 8 isoforms, 19.5%) (**Figure 6.33B**), both of which partially encode the ABC2 membrane 3 (*Abca7* transmembrane domain).[369] Intron retention was also particularly enriched between exons 37 and exon 39 (**Figure 6.33C**), both of which also encode the ABC2 membrane 3. Analogous in structure to Abca1, the transmembrane domain is also involved in substrate transport across the membrane lipid bilayer by undergoing a conformational change.

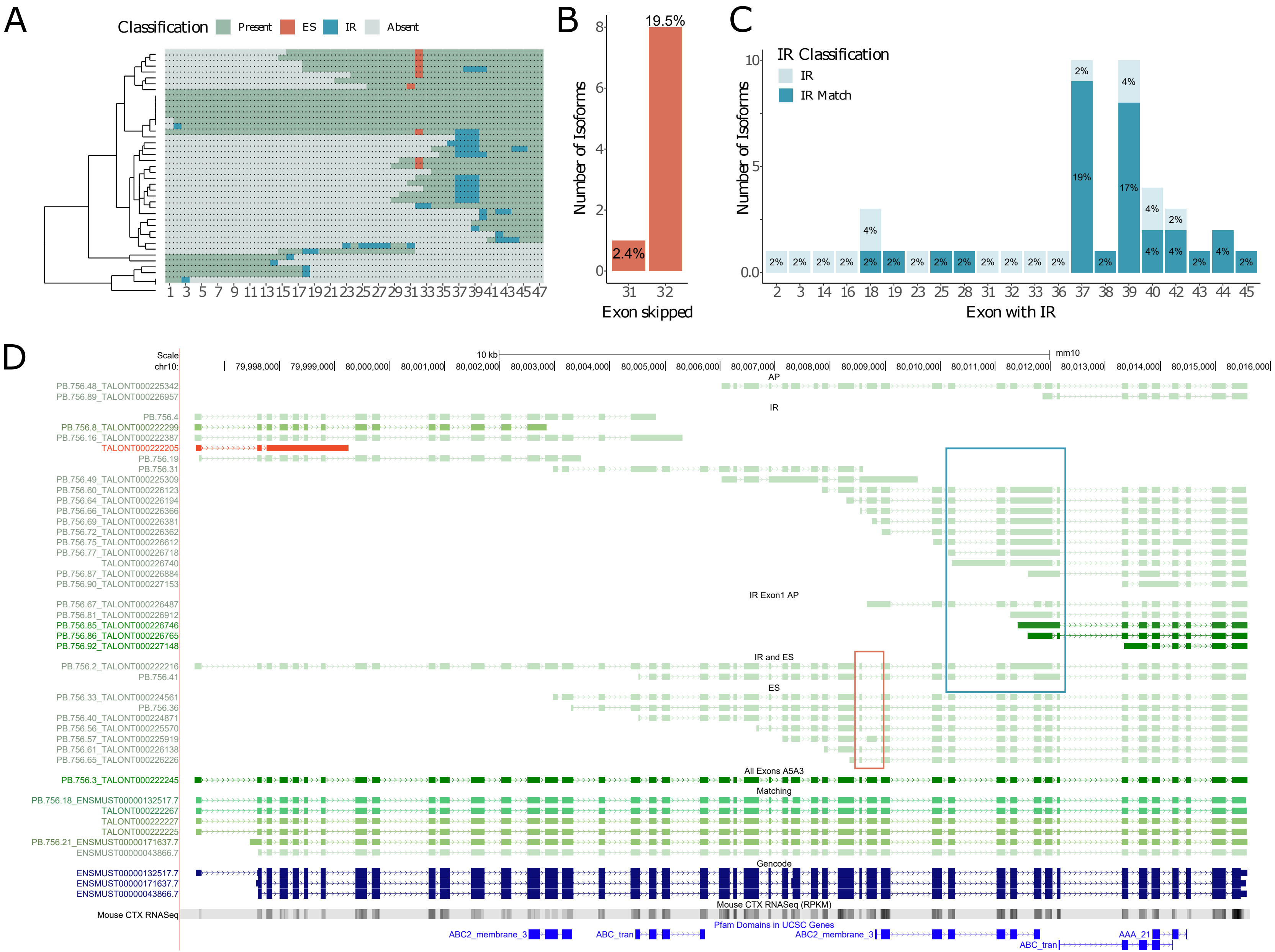


Figure 6.33: Characterisation of *Abca7* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Abca7* isoform landscape, and bar charts of the **(B)** number of isoforms with exon skipping, and the **(C)** number of isoforms with intron retention (IR) events, and **(D)** a UCSC genome browser track of the isoforms annotated to *Abca7*. Intron retention events are further classified as “IR” if it occurs at one exon or “IR Match” if it spans across multiple exons.

6.3.8.3 *Ank1*

Recent epigenome-wide association studies (EWAS) have identified a number of genetic loci at which variable DNA methylation is associated with increased risk of AD pathology.[310,311] One locus that is consistently hypermethylated in AD post-mortem brain tissue resides in *ANK1*, a gene that encodes an integral membrane involved in mediating attachment of membrane proteins to the underlying cytoskeleton.[310,311] Important for key activities such as cell mobility and proliferation, ANK1 is associated with multiple isoforms with varying lengths and affinity for target membrane proteins.

Spanning over 175kb on chromosome 8, the mouse *Ank1* gene is characterised with unique 48 exons and 17 known isoforms. In our dataset, *Ank1* stood out from the panel of target genes for its enrichment of the significantly shorter known isoforms (**Figure 6.34A, B**). While we detected several of the longer known isoforms (n = 4 isoforms, 23.5%, length = 6.5kb - 8.2kb), the majority of the 17 *Ank1*-associated isoforms were less than ~2kb (n = 11 isoforms, 64.7%) in alignment with the shorter known *Ank1*-associated isoforms (sAnk1) that span the 3' end. The two most abundant isoforms also corresponded to one of the short known isoforms (Ank1-208, ENSMUST00000121075.7) and the short non-coding isoform (Ank1-211, ENSMUST00000130311.1). The expression of such short, truncated isoforms have been previously found to be specific to striated muscles and driven by the activity of a second alternative promoter.[370] Notably, recent studies have identified a type 2 diabetes risk allele that increases promoter activity and sAnk1 expression.[371]

Aside from the length disparities among the isoforms detected, we found that the splicing pattern was generally conserved across the gene with consistent skipping of certain exons, notably exons 44 and 47 (**Figure 6.34C**). Both exons were present in the majority of known isoforms (n = 11 isoforms, 64.7%), and did not encode for any ankyrin repeat domains. Strikingly, ORF predictions indicated that while exon 44 skipping maintained the reading frame, inclusion of this exon resulted in a stop codon (**Figure 6.34D**). Isoforms without exon 44 skipping were subsequently predicted for nonsense-mediated decay (**Figure 6.34D**).

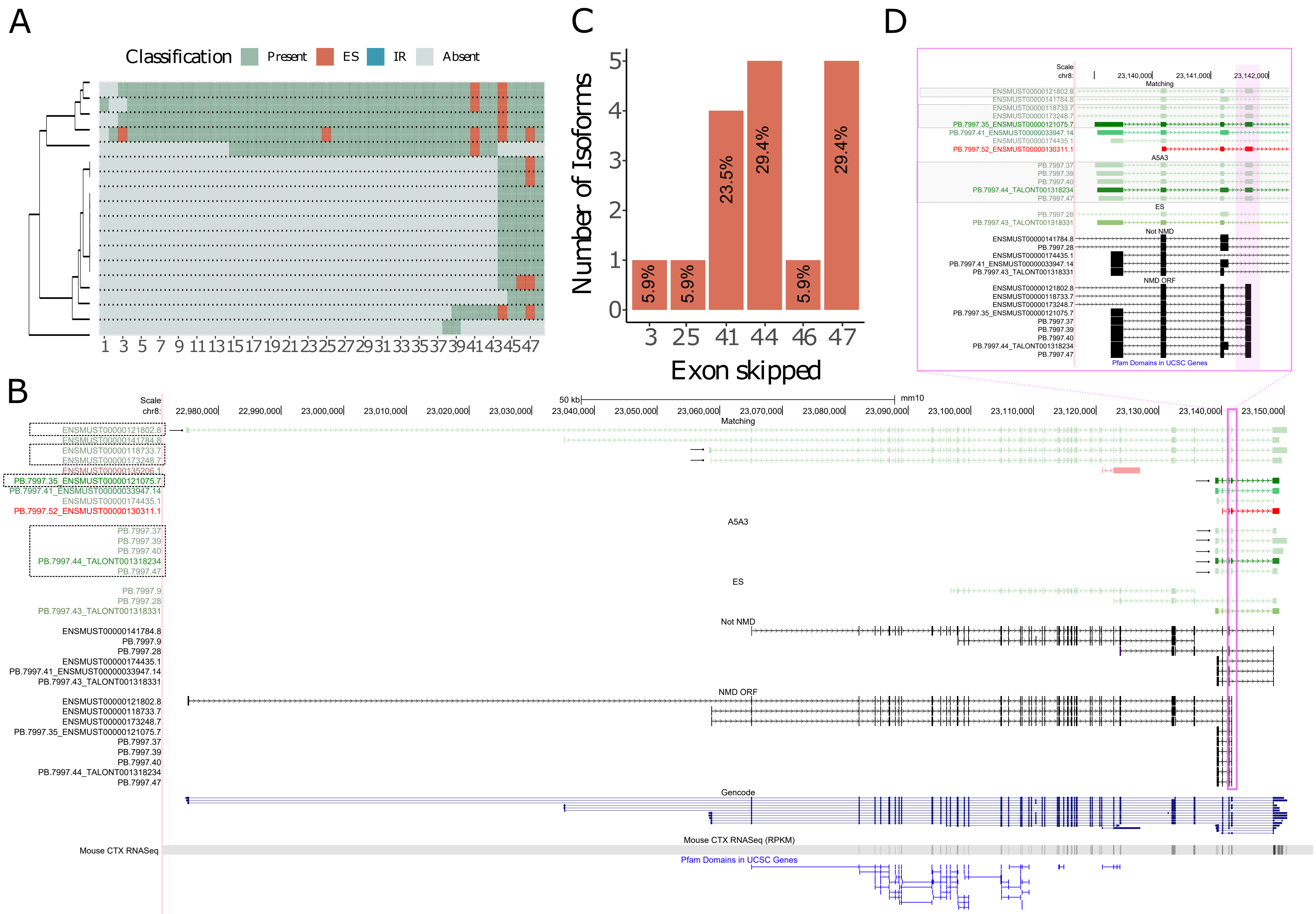


Figure 6.34: Characterisation of *Ank1* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Ank1* isoform landscape, **(B)** a UCSC genome browser track of the isoforms annotated to *Ank1*. Isoforms indicated with a black box and black arrow contain exon 44 and are subsequently not predicted for nonsense-mediated decay (NMD). **(C)** A bar chart of the number of isoforms with exon skipping, and **(D)** a zoomed-in track showing isoforms with exon 44 inclusion, generating truncated reading frames and products predicted for nonsense-mediated decay.

6.3.8.4 *Apoe*

To date, inheritance of the ϵ4-allele of *APOE*, which encodes the apolipoprotein E, is the strongest risk factor for late-onset AD (detailed in **Section 1.1.3**). Involved in regulating lipid homeostasis, ApoE facilitates lipid transport essential for CNS development and maintenance. Characterised with three well-known human isoforms, ApoE exhibit isoform-specific activity and binding affinity for substrates, including β-amyloid peptides.[313]

Despite only containing 4 exons (7 if the 3 alternative exons are included) and spanning just over 3kb on chromosome 7, the mouse gene *Apoe* was the most "isoformic" gene among our panel of AD-associated target genes; we detected an overwhelming total of 2,006 isoforms. While the majority of isoforms contained all 4 exons (n = 1,390 isoforms, 69.2%) (**Figure 6.35A**), a deeper examination of *Apoe* revealed complex variations of exon 6 (last/penultimate exon depending on the reference isoform of interest) and the 3' UTR (**Figure 6.35B**). Notably, this 831bp exon encodes the apolipoprotein domain that binds to lipids. Supported by RNA-Seq data from matched samples, these variations included usage of alternative 5' and 3' splice sites of this exon, or of matched 5' and 3' end sites but skipping within the exon resulted in two enclosed exons (**Figure 6.35B**). Noteworthy, one of the known isoforms, Apoe-202 (ENSMUST00000167646.8) was also characterised with this "internal exon skipping" phenomenon.

In contrast to exon 6, the other exons were relatively conserved with significantly fewer variations of 5' and 3' splice sites (**Figure 6.35D**). However, unlike exon 6 which was present in nearly all isoforms, we detected skipping events of such exons (**Figure 6.35E**): exon 4 (n = 61 isoforms, 48.8% of isoforms with ES) and exon 5 (n = 89 isoforms, 71.2% of isoforms with ES). Despite this widespread isoform diversity, we only detected four isoforms that contained the known alternative first exon present in one of the known isoforms, Apoe-204 (ENSMUST00000172983.7). Notably, one of these isoforms (TALONT00166063) was characterised with both the first canonical exon and the alternative first exon, indicating that these exons are not mutually exclusive (**Figure 6.35F**).

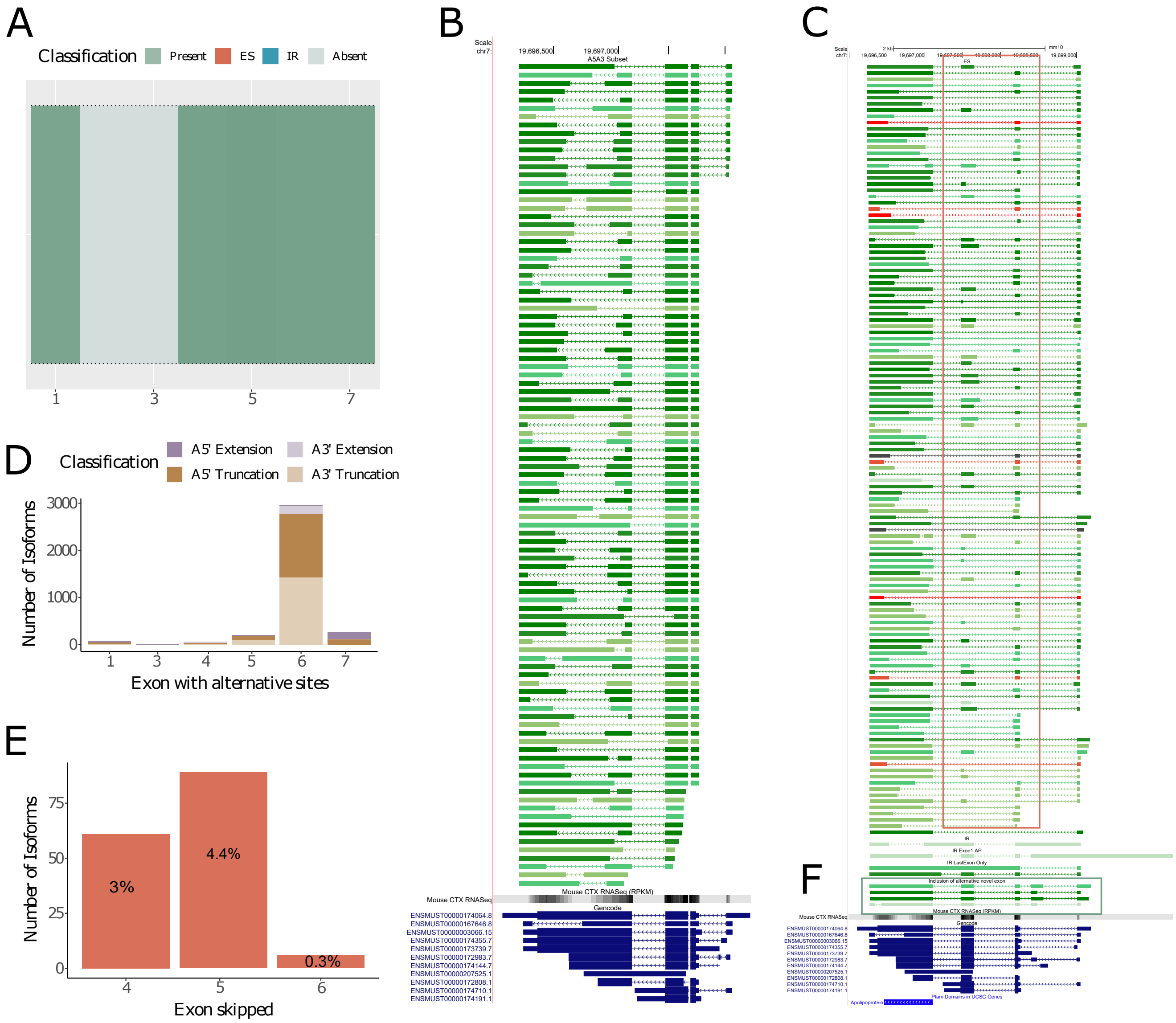


Figure 6.35: Characterisation of *Apoe* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Apoe* isoform landscape (exons 2, 3 and 4 are alternative first exons from other known reference isoforms), **(B)** UCSC genome browser tracks of the isoforms annotated to *Ank1* with usage of alternative 5' and 3' splice sites, and **(C)** exon skipping events (box red), and **(D)** a bar chart of the number of isoforms with alternative 5' and 3' splice sites, **(E)** exon skipping events. **(F)** UCSC tracks of isoforms with novel alternative first exons.

6.3.8.5 *App*

The amyloid precursor protein gene, *APP*, is well-established in playing a key role in AD pathogenesis. Central to the amyloid cascade hypothesis (described in **Section 1.1.3**), *APP* encodes the integral membrane protein that is sequentially cleaved to generate Aβ peptides of varying lengths. Over 50 mutations are identified in *APP* and are known to cause Familial Alzheimer's disease.[41–43]

Spanning over 224kb on chromosome 16, the mouse *App* gene is characterised with 20 unique exons and 11 known isoforms. In contrast to the mouse reference annotations and in spite of the relatively few exons associated with *App*, we detected 466 isoforms annotated to *App*. We identified isoforms with varying lengths with less than half spanning the full length of the gene (n = 183 isoforms, 39.9%) (**Figure 6.36A**). A number of the shorter isoforms shared similar exonic structure to the two known short isoforms: App-205 (ENSMUST00000227654.1) and App-210 (ENSMUST00000228375.1) (**Figure 6.36H**). However, the majority of the shorter isoforms (n = 428 isoforms, 91.8%) were characterised with alternative first exons while preserving the 3' end of the exonic structure (**Figure 6.36A**), which contained the uncleaved β-amyloid peptide.

Despite this variation in isoform length, we observed a consistent splicing pattern with exon skipping (n = 406 isoforms, 87.1%) enriched in two regions of the gene (**Figure 6.36C**): i) exon 7 (n = 289 isoforms) and exon 8 (n = 304 isoforms) (**Figure 6.36B, F**), which encode the Kunitz protease inhibitor (KPI) domain, and ii) exon 14 (n = 392 isoforms) and exon 15 (n = 390 isoforms) (**Figure 6.36B, E**), which are alternative last exons present in only two of the known *App* isoforms (App-208, ENSMUST00000227753.1 and App-209, ENSMUST00000227990.1). With over 60% of isoforms characterised by four or more skipping events (n = 291 isoforms, 62.4% of ES-isoforms) (**Figure 6.36D**), we also detected isoforms with skipping of exon 18 (n = 39 isoforms) and exon 19 (n = 54 isoforms) (**Figure 6.36B, G**), which encode the uncleaved β-amyloid peptide. While the majority of known isoforms did not contain exons 7 and 8, most reference isoforms contained the KPI (KPI+) domain, a 57-amino-acid insert with homology to the Kunitz family of serine protease inhibitors. Recent studies investigating differential transcript expression in AD human post-mortem brains have revealed significant down-regulation of two isoforms lacking the KPI domain.[118] Noteworthy, increased mRNA and protein expression of KPI(+) transcripts were associated with increased β-amyloid accumulation.[314]

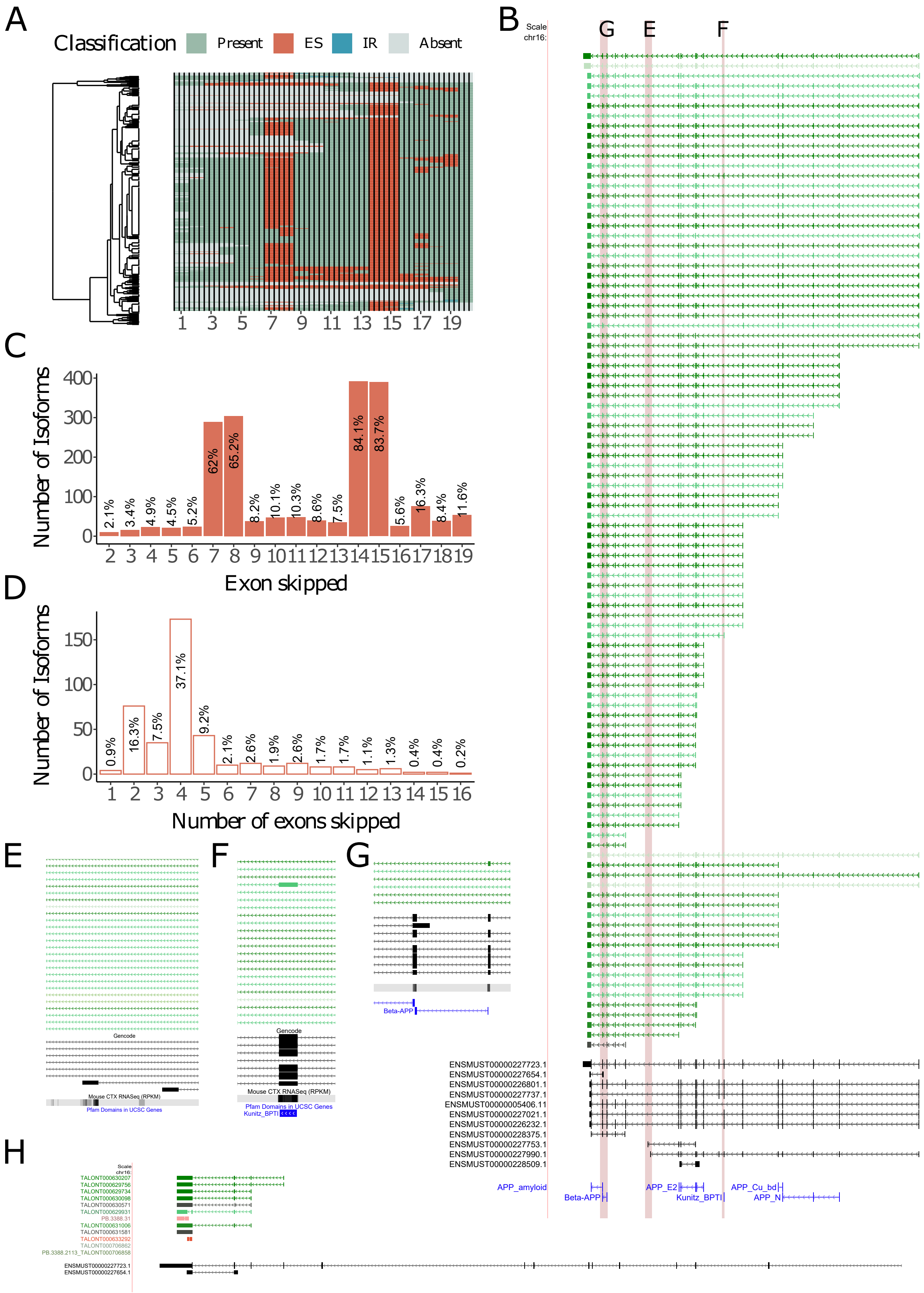


Figure 6.36: Characterisation of *App* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *App* isoform landscape, **(B)** a UCSC genome browser track of a subset of *App*-associated isoforms with exon skipping (boxed in red), **(C)** a bar chart of the number of isoforms with exon skipping, and **(D)** isoforms with the total number of exons skipped, **(E)** zoomed-in UCSC genome browser tracks showing skipping of exons 14 and 15, **(F)** exon 7 encoding the KPI domain, **(G)** exons 18 and 19 encoding the uncleaved β-amyloid peptide, and **(H)** a UCSC genome browser track of the short isoforms that aligned with App-205 (ENSMUST00000227654.1).

6.3.8.6 *Bin1*

Bridging integrator 1 gene, *BIN1*, is a well-established AD risk gene with multiple AD-associated SNPs identified from GWAS, and remains second only after *APOE* in genome-wide significance.[25] Although the mechanisms underlying the role of *BIN1* in AD pathogenesis is not fully understood, recent transcriptomic profiling studies on human post-mortem AD tissue have revealed differential cell-specific *BIN1* transcript expression associated with tau accumulation and AD-related traits.[319]

Spanning over 58kb on chromosome 18, the mouse *Bin1* gene is characterised with 20 unique exon and 6 known isoforms. In our dataset, we detected 368 isoforms annotated to *Bin1* and identified widespread occurrence of exon skipping, particularly within certain regions of the gene (**Figure 6.37A**). Drawing parallels to the human-equivalent *BIN1* isoform landscape,[319] inclusion of exons 14 - 16, which encode the CLAP domain involved in endocytosis, was highly variable among isoforms (**Figure 6.37B,C**) (Exon 14 skipping: n = 104 isoforms; Exon 15 skipping: n = 174 isoforms; Exon 16 skipping: n = 249 isoforms). Strikingly, the most abundant isoform detected in both our Iso-Seq (8,769 Iso-Seq full-length reads) and ONT targeted datasets (40,622 ONT full-length reads) was a novel isoform (PB.3915.2_TALONT000761829) that shared a similar exonic structure to the known canonical isoform (ENSMUST00000025239.8, which was the second most abundant isoform with 4,120 Iso-Seq full-length reads and 18,451 ONT full-length reads) with the exclusion of exon 15. ORF prediction of this isoform showed that skipping of this exon maintained the reading frame. This region, encoding the CLAP domain, was also notably enriched with occurrence of intron retention events (n = 48 events, 25 isoforms) (**Figure 6.37D,E**). In contrast, the first 10 exons, which encode the N-BAR domain involved in membrane curvature, were relatively more conserved in these long isoforms with fewer splicing events.

Despite the relatively conserved nature of the N-BAR domain, we detected a number of shorter isoforms that were characterised with an alternative first exon and subsequent exclusion of this domain (**Figure 6.37F**). Sharing similar exonic structure to Bin1-204 (ENSMUST00000234373.1), some of these isoforms were further characterised with a long alternative first exon that spanned across the CLAP domain.

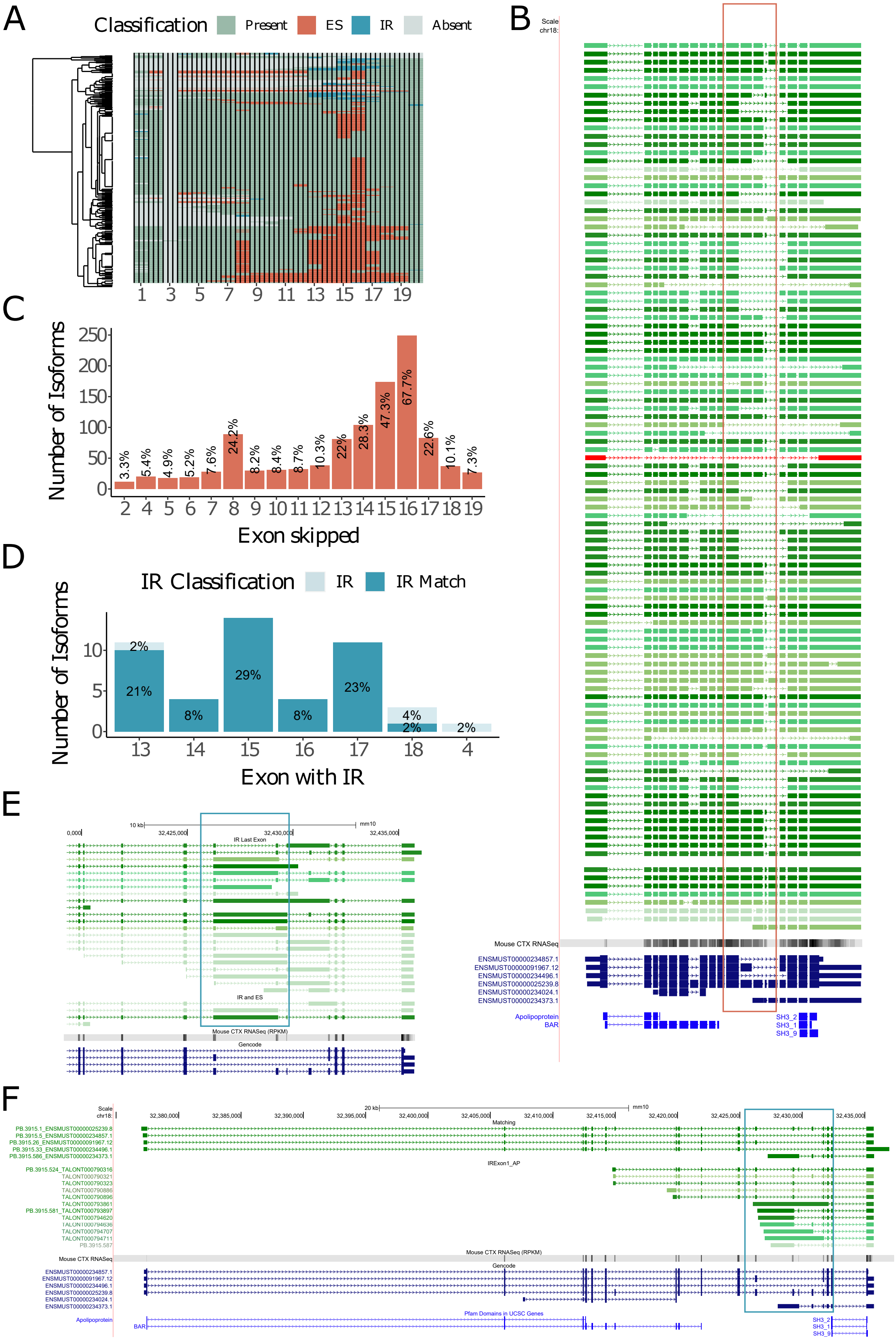


Figure 6.37: Characterisation of *Bin1* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Bin1* isoform landscape, **(B)** a UCSC genome browser track of the subset of *Bin1*-associated with exon skipping (boxed in red), **(C)** a bar chart of the number of isoforms with exon skipping, and **(D)** intron retention events, **(E)** zoomed-in UCSC genome browser tracks showing intron retention events spanning across exons 16 and 17, **(F)** UCSC genome browser track of the short isoforms that aligned with Bin1-204 (ENSMUST00000234373.1) with the alternative long exon boxed in blue.

6.3.8.7 *Cd33*

Sialic acid-binding immunoglobulin-like lectin 3 gene, known as *Cd33*, is implicated in AD by the identification of various AD-associated variants from GWAS. Encoding a myeloid-specific transmembrane receptor involved in key cell-signalling pathways, *Cd33* is implicated in cell adhesion, immune cell growth and cytokine release.[323] Notably, the protective *CD33* AD-associated variant has been correlated with decreased levels of Aβ peptides in AD brains as a consequence of enhanced phagocytosis.[324]

Spanning across 16kb on chromosome 7, the mouse *Cd33* gene is characterised with 9 unique exons and 6 known isoforms. In our dataset, we detected 41 isoforms annotated to *Cd33*, including 5 of the known isoforms (**Figure 6.38**). Reflecting the isoform landscape of the mouse reference annotations, the majority of isoforms detected were similarly short, lacked the first two exons and skipped exon 8 (which was only present in one of the known isoforms Cd33-201, ENSMUST00000004728.11) (**Figure 6.38A**). ORF predictions showed that skipping of this exon slightly reduced but maintained the reading frame (**Figure 6.39A**). In contrast, deeper evaluation revealed that truncation (alternative 3' splice site) of exon 4 shifted the reading frame as a result of generating a stop codon (**Figure 6.39A**). Consequently, the reading frame of such isoforms appeared to be driven by a downstream initiator codon in exon 5, resulting in exon 4 exclusion. Notably, exons 4 and 5 encode the two immunoglobulin domains.

Although the majority of isoforms were relatively short, we detected a few novel isoforms that spanned the length of the gene and incorporated features from different known isoforms. One example included a novel isoform that shared the 5' exonic structure of Cd33-202 (ENSMUST00000039861.6), while simultaneously harbouring the longer 3' UTR only present in Cd33-203 (ENSMUST00000205503.1) (**Figure 6.38D**). We further detected two novel isoforms with a novel exon present between exon 1 and exon 2 (66bp, Chr7:43529866-43529932, **Figure 6.39B**). ORF predictions showed that inclusion of the novel exon did not alter the reading frame.

While exon skipping was localised to exon 8, we observed widespread occurrence of intron retention (IR) events across *Cd33*. Over a third of the isoforms detected (n = 14 isoforms, 34.21%) contained at least one IR event (**Figure 6.38B**), with one isoform containing an IR event that spanned across 4 exons (TALONT001237573, **Figure 6.39C**). Deeper characterisation revealed that intron retention primarily occurred around exon 7 (**Figure 6.38C**) and extended to the final two exons, exon 8 and exon 9, with varying lengths (**Figure 6.39C**). ORF predictions showed that an extended intron retention at exon 7 revealed a stop codon resulting in a shortened reading frame, whereas an intact exon 7 resulted in a slightly longer reading frame.

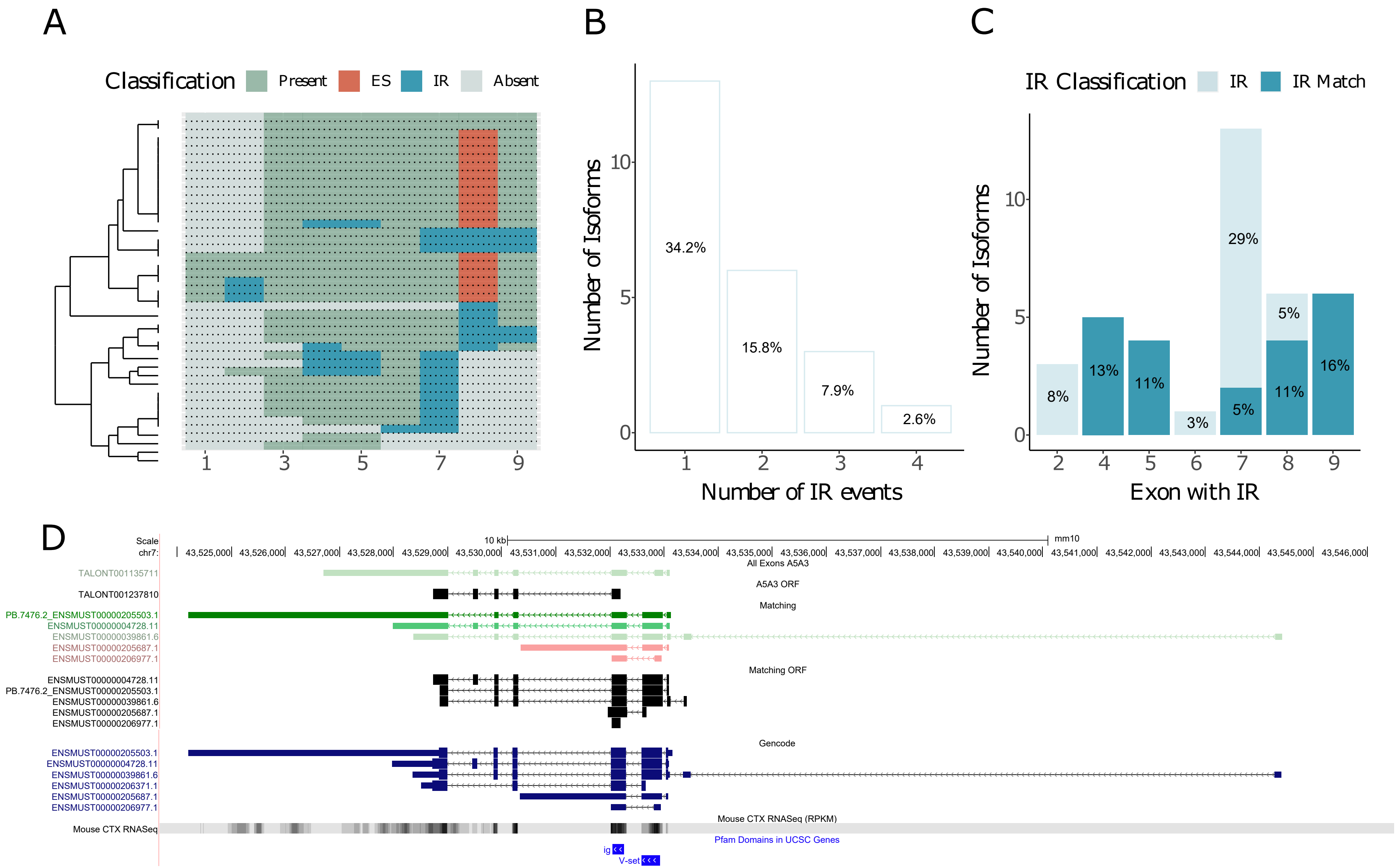


Figure 6.38: Characterisation of *Cd33* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Cd33* isoform landscape, **(B)** a bar chart of the number of isoforms with intron retention events, **(C)** the number of exons characterised with IR, and **(D)** a UCSC genome browser track of the detected isoforms that aligned with known *Cd33* isoforms. Intron retention events were further classified as “IR” if it occurs at one exon or “IR Match” if it spans across multiple exons.

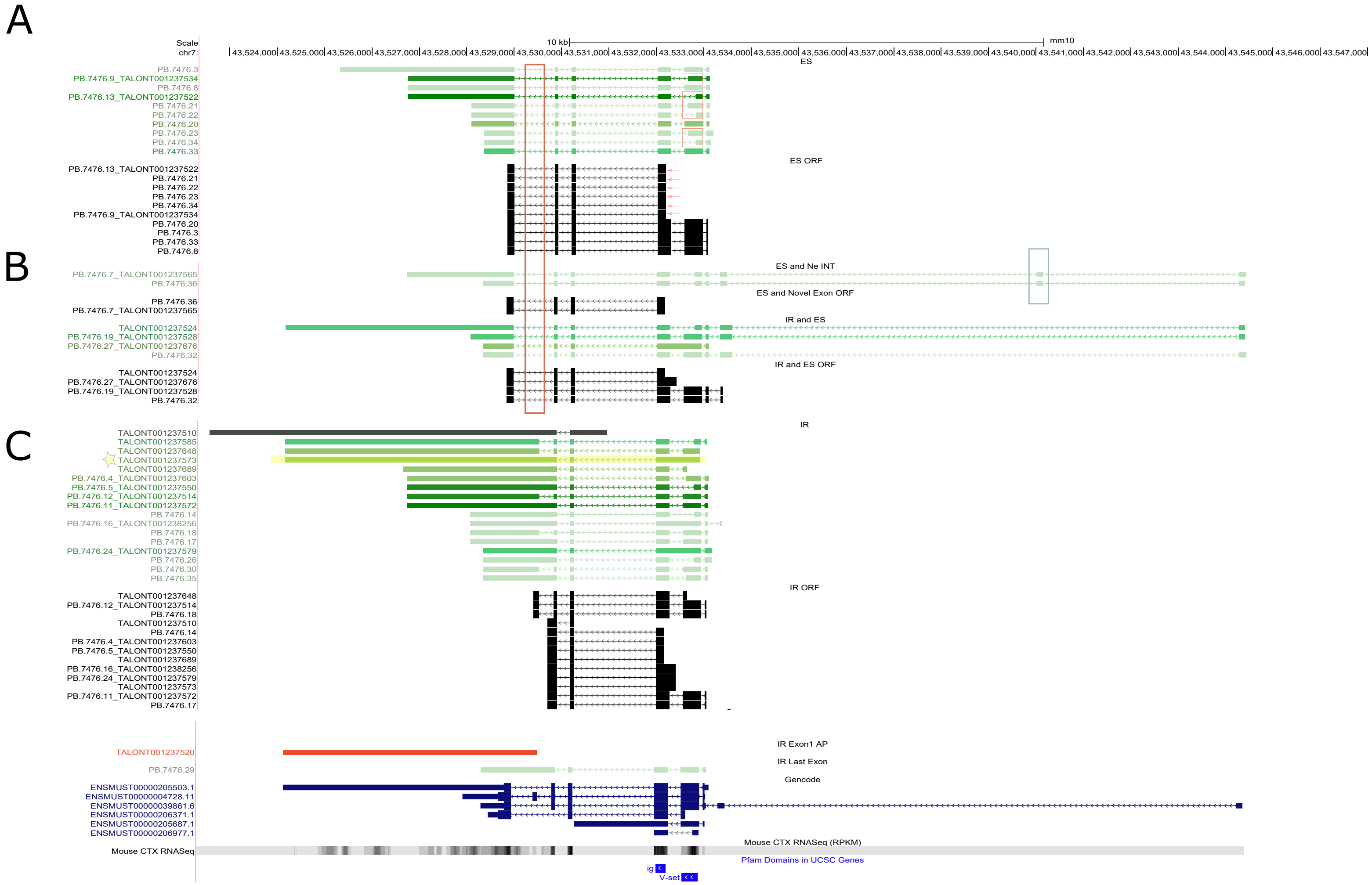


Figure 6.39: Characterisation of *Cd33* splicing events in the rTg4510 cortex. Shown are UCSC genome browser tracks of **(A)** *Cd33* isoforms with exon skipping and their respective predicted open reading frames - the exon truncation and the shift in reading frame are denoted by the pink box and pink arrow, respectively, **(B)** two novel long isoforms with the presence of a novel exon located between exon 1 and exon 2 (boxed), and **(C)** isoforms characterised with intron retention (IR) events - the isoform with IR spanning across 4 exons is highlighted in yellow. Skipping of exon 8 is denoted by the red box.

6.3.8.8 *Clu*

The clusterin gene, *CLU*, is strongly associated with late-onset AD along with *APOE* and *BIN1*.[27] Encoding a glycoprotein with chaperone function, clusterin has been implicated in various pathways including immune regulation, lipid homeostasis and apoptosis.[372] While the role of clusterin in AD pathology is not fully understood, recent studies have reported up-regulation of clusterin in AD brains with a suggestive role in altering Aβ aggregation.[327]

Spanning over 13kb on chromosome 14, the mouse *Clu* gene is characterised with 14 unique exons and 9 known isoforms. In our dataset, we identified 733 isoforms annotated to *Clu* in the mouse rTg4510 cortex. Although the isoform diversity in the mouse reference genome annotations was predominantly driven by alternative first exons (n = 6 alternative first exons across 9 known isoforms), the majority of isoforms detected (n = 402 isoforms, 55%) were derived from the first upstream exon from Clu-201 (ENSMUST00000022616.13) and spanned the full-length of the clusterin domain (**Figure 6.40A,B**). Notably, we identified a handful of novel isoforms with novel alternative first exons (**Figure 6.40B**), highlighting the extensive usage of alternative first exons as a transcriptional mechanism of *Clu* expression. ORF predictions, however, showed that translation was initiated at the start site located in exon 7, bypassing all the alternative first exons. Drawing parallels to the human *CLU* gene,[372] alternative start codons were also identified in the upstream exons, though their functional importance is unknown. We further identified the N-terminus endoplasmic reticulum (ER) signal peptide with a cleavage site (position 21 - 22) within exon 7, allowing generation of secreted CLU.[372]

Deeper characterisation of our dataset revealed that *Clu* isoform diversity was primarily driven by extensive usage of alternative 5' and 3' splice sites, and occurrence of exon skipping and intron retention events localised to certain regions of the gene (**Figure 6.40A,H**), notably: skipping of exon 9 (n = 143 isoforms, 19.5%), exon 10 (n = 129 isoforms, 17.6%) and exon 11 (n = 165 isoforms, 22.5%) (**Figure 6.40C,D**). ORF predictions showed that skipping of these exons reduced but maintained the reading frame. Conversely, intron retention events (n = 217 events) predominantly occurred at the 5' end of the gene with intron retention events spanning across exons 6, 7 and 8 (**Figure 6.40E,G**). Noteworthy, exon 6 refers to an alternative first exon from Clu-206 (ENSMUST00000144619.1). ORF prediction of these isoforms showed that there was a shift in open reading frame with initiation of translation at exon

10. On the other hand, the effect of IR spanning across exons 7 and 8 on the reading frame appeared to be dictated by the upstream first exon; usage of the first exon from the canonical isoform, Clu-201 (ENSMUST00000022616.13), resulted in a truncated protein from translation at exon 10, whereas usage of the first exon from Clu-207 (ENSMUST00000146990.1) resulted in a truncated protein destined for nonsense-mediated decay (**Figure 6.40F**).

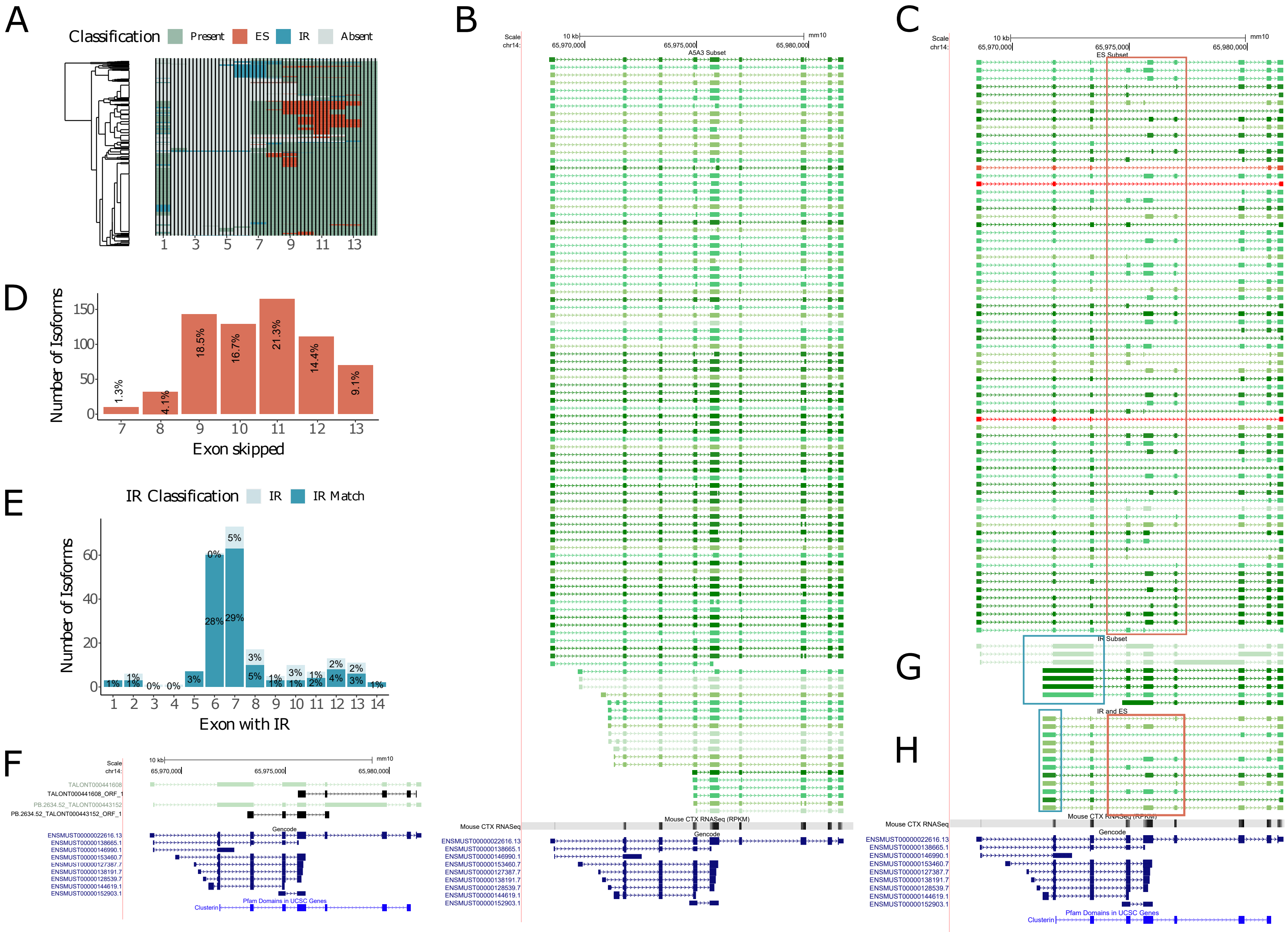


Figure 6.40: Characterisation of *Clu* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Clu* isoform landscape, **(B)** UCSC genome browser tracks of isoforms annotated to *Clu* that shared exonic structure with the long *Clu* isoform (Clu-201), and **(C)** isoforms characterised with exon skipping events, **(D)** bar charts of the number of isoforms with exon skipping, and **(E)** intron retention, and **(F)** UCSC genome browser tracks of two *Clu* isoforms with intron retention spanning across exon 7 and exon 8, resulting in two different open reading frames, **(G)** isoforms with intron retention and **(H)** exon skipping events together.

6.3.8.9 *Fus*

The fused in sarcoma gene, *Fus*, is a well-established causative gene for a number of neurodegenerative diseases, including ALS and FTD. Encoding the subunit of the heterogeneous nuclear ribonucleoprotein complex, a multi-functional DNA/RNA-binding protein, *Fus* is implicated in key cellular processes including transcriptional regulation, DNA repair and alternative splicing.[373] Increasing evidence suggests that aggregates of the FUS proteins followed by formation of intracellular inclusion bodies are key initiator events in disease onset and pathology.[373] Notably, *Fus* neuronal aggregates have become the characteristic pathological hallmark for the subset of sporadic FTD cases that lack the more established inclusion markers of TDP-43 and tau.[374]

Spanning across 18kb on chromosome 7, the mouse *Fus* gene is a complex gene characterised with multiple known isoforms of varying lengths. However, in contrast to the *Fus* isoform landscape in the mouse reference genome annotations, the *Fus* splicing pattern in our dataset was relatively simpler (**Figure 6.41A**): a quarter (n = 53 isoforms, 25.7%) of the isoforms detected (n = 236 isoforms) largely shared the exonic structure of the long known isoform that spanned the full-length of the gene, differing only by minor variations ("wobble", < 10bp) at the splice site (**Figure 6.41B**).

Despite this relatively simple isoform landscape, we detected widespread occurrence of exon skipping (**Figure 6.41C,D**) and intron retention events (**Figure 6.41E,F**) that were not present in the reference mouse genome. Over 40% of isoforms detected (n = 103 isoforms, 43.6%) were characterised with at least one exon skipping event, with exon 8 skipped in half of the these isoforms (n = 50 isoforms, n = 48.5% of isoforms with exon skipping, **Figure 6.41D**). Intron retention events were further localised to this region of the *Fus* gene: 17 isoforms were identified with intron retention events spanning across exons 6, 7 and 8. While such IR events were also present in the known isoforms - Fus-206 (ENSMUST00000128851.7) - the majority of the detected IR events spanning across this region belonged to the first exon rather than an internal exon (**Figure 6.41F**). Consequently, the isoforms detected contained an alternative first exon. Finally, we also detected a number of intron retention events at the 3' end of the *Fus* gene between exons 12 and 14 (**Figure 6.41F**), which encode the zinc-finger-containing RGG-Znf-RGG domain (zF-RanBP) essential for RNA recognition and binding.[375]

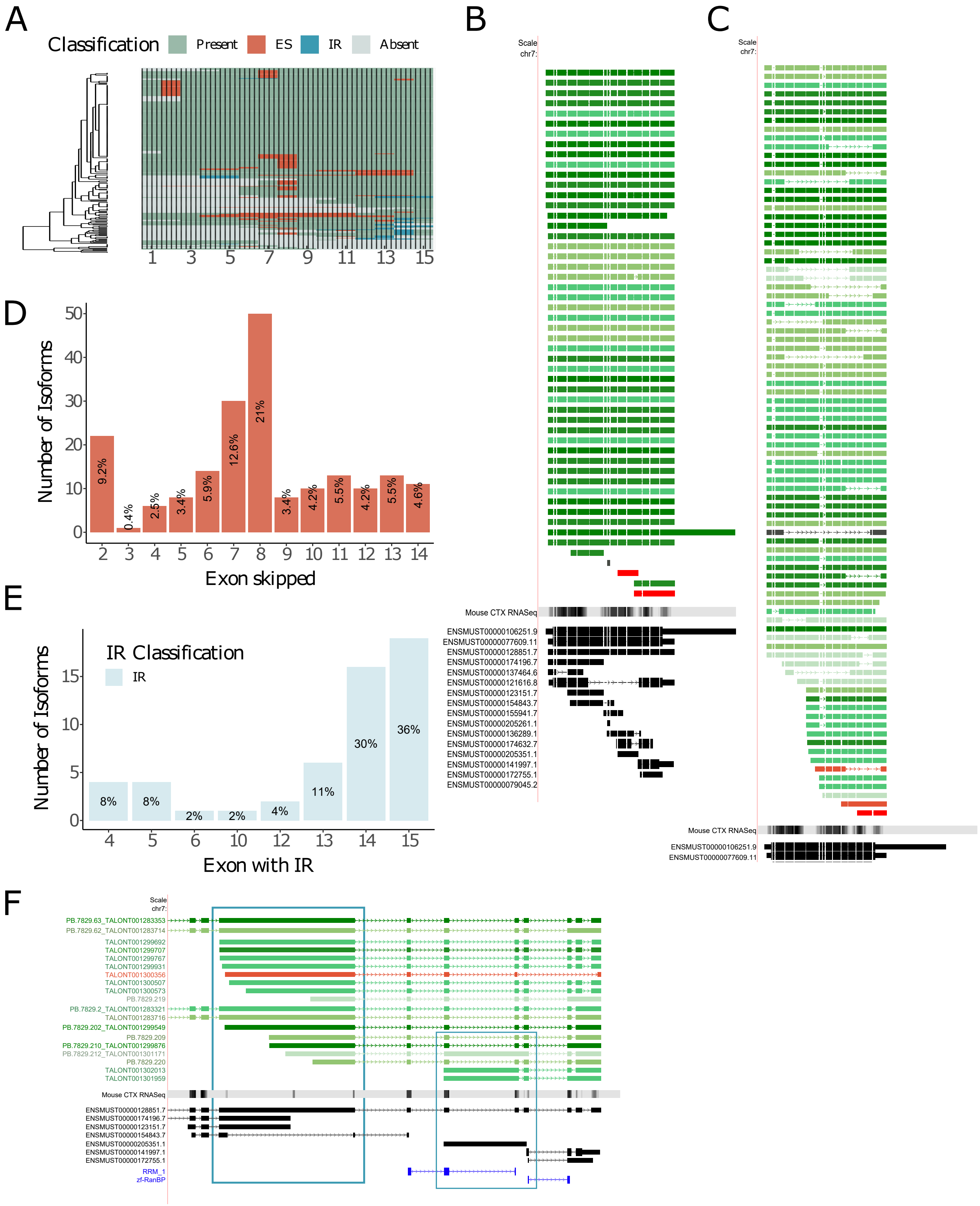


Figure 6.41: Characterisation of *Fus* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Fus* isoform landscape, **(B)** UCSC genome browser tracks of the subset of isoforms that spanned the full-length of the *Fus* gene with minor splice site variation, **(C)** isoforms characterised with exon skipping events, **(D)** bar charts of the number of isoforms with exon skipping and **(E)** intron retention events, and **(F)** a zoomed-in UCSC genome browser track of isoforms with intron retention spanning across exons 6, 7 and 8 and between exons 12 and 14 (boxed in blue).

6.3.8.10 *Fyn*

The non-receptor tyrosine kinase, FYN, is implicated in AD pathogenesis as a consequence of its interactions with tau.[333] Overexpression of FYN, which is known to directly phosphorylate tau,[333] was shown to accelerate synaptic and cognitive impairments in a mouse model of AD.[334] Recent studies have further identified an isoform-specific role of *FYN* in modulating neurofibrillary degeneration with evidence of isoform switching in the AD neocortex.[335]

Spanning over 196 kb on chromosome 10, the mouse *Fyn* gene is associated with 20 unique exons and 9 known isoforms that primarily differ by an alternative first exon. Detecting 50 isoforms annotated to *Fyn*, we found these first exons (exons 1 - 6) were mutually exclusive with isoforms containing either exons 2 and 3 or exon 4 (**Figure 6.42A**). Conversely, exons 1, 5 and 6 - the other three alternative first exons - were not solely featured in any of the isoforms except in an intron retention event (**Figure 6.42A**). Notably, we detected a number of novel isoforms with novel alternative first exons located at three distinct regions (**Figure 6.42C**): i) between exons 2 and 3, ii) exons 4 and 5 and ii) near exon 8. Despite the widespread usage of alternative first exons and promoter, ORF prediction showed that inclusion of these novel first exons did not alter the reading frame, which was still predicted to start from exon 7.

In contrast to the complexity at the 5' end of the *Fyn* gene, which does not encode for any protein domains, the exonic structure downstream of exon 7 was relatively conserved (**Figure 6.42A**). Exons that encode the SH3 domain (exons 7 to 12), which is known to interact with tau, were present in all the isoforms detected (**Figure 6.42**). Notably, exons 13 and 14, which encode the protein kinase domain, were found to be mutually exclusive (**Figure 6.42A,B,C**); the majority of detected isoforms contained exon 13 (n = 36 isoforms) and skipped exon 14 (**Figure 6.42**). Finally, we noted a complete exclusion of exon 16, which was only present in Fyn-203 as an alternative first exon (**Figure 6.42A,C**).

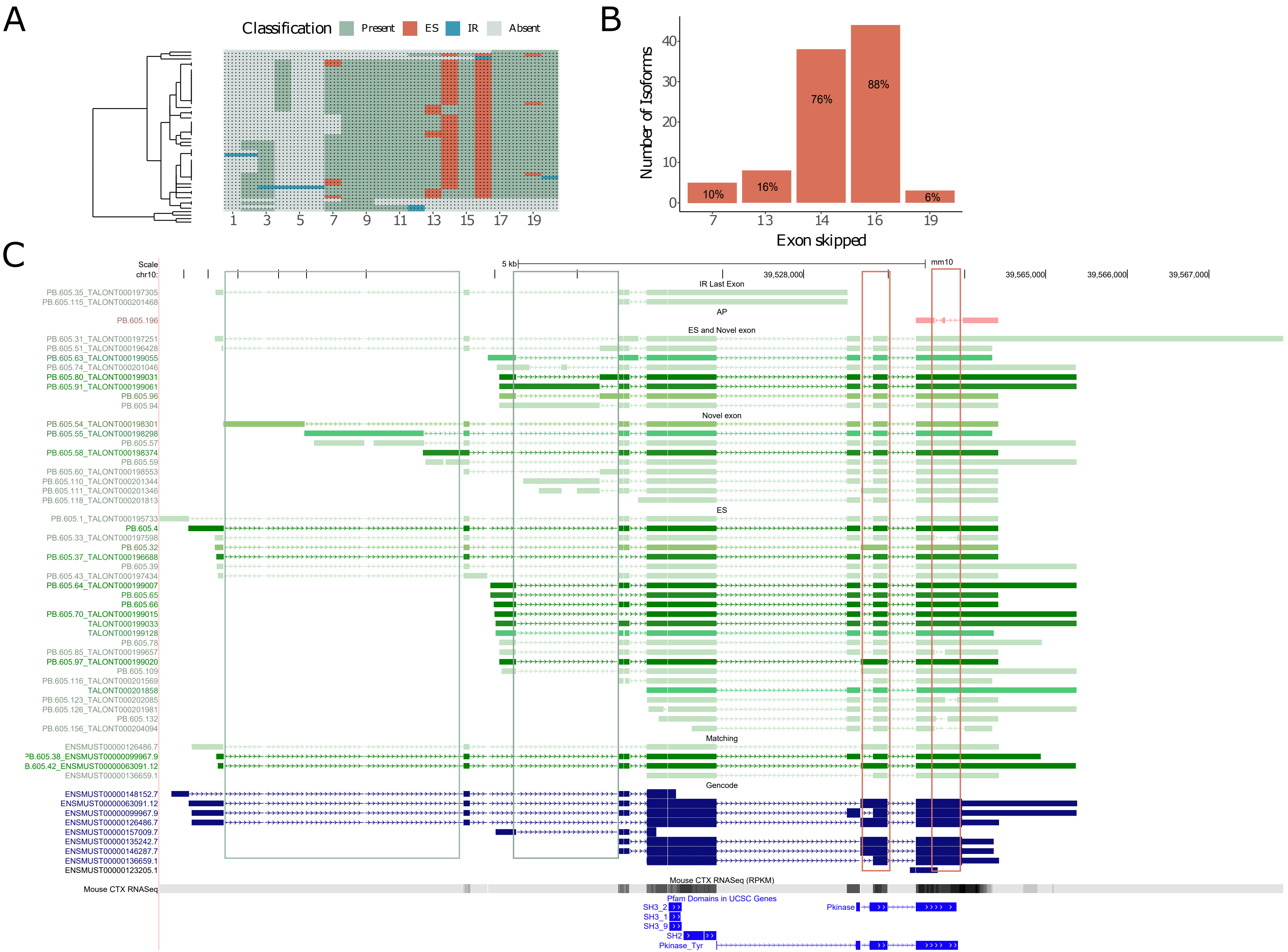


Figure 6.42: Characterisation of *Fyn* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Fyn* isoform landscape, **(B)** a bar chart of the number of isoforms with exon skipping, and **(C)** a UCSC genome browser track of the *Fyn*-associated isoforms characterised by exon skipping (boxed red) events and the presence of novel exons (boxed green).

6.3.8.11 *Mapt*

The microtubule-associated protein tau gene, *MAPT*, is a well-established gene in AD pathogenesis. Central to the tau hypothesis (described in **Section 1.1.3**), *MAPT* encodes the tau protein essential for microtubule stability and maintenance. Tau mutations associated with frontotemporal dementia and parkinsonism (FTDP), which result in altered ratio of tau isoforms, are known to induce tau phosphorylation and aggregation.[337] While no causative *MAPT* mutations have been identified in AD, tau aggregation and subsequent formation of neurofibrillary tangles are one of the key hallmarks of AD pathology (described in **Section 1.1.1**).

Spanning over 100kb on chromosome 11, the mouse *Mapt* gene is characterised with 16 unique exons and 12 known isoforms. In our dataset, we detected 140 isoforms annotated to *Mapt*, including 7 of the known isoforms. Deeper examination of the *Mapt* isoform landscape revealed consistent exon skipping events that alternated across the gene (**Figure 6.43A,B,C**); exons 4, 6, 8, 10 and 13 were typically skipped, whereas exons 2, 7, 11, 12 and 14 were typically included. The majority of isoforms were characterised with at least one exon skipping event (n = 115 isoforms, 82.1%) with most isoforms skipping 3 or more exons (n = 98 isoforms, 70%). ORF predictions showed that all these exon skipping events reduced but maintained the reading frame.

Finally, using Iso-Seq and ONT nanopore sequencing, we detected some significantly shorter isoforms that were distinguished by the presence of an alternative first exon characterised with intron retention (**Figure 6.43D**). These broadly appeared between exons 9 and 11, which encode the tubulin-binding repeat domain. ORF predictions of these isoforms revealed that intron retention that significantly spanned across exons 7, 8 and 9 generated a truncated product destined for nonsense-mediated decay, whereas intron retention events that spanned across exons 10 and 11 maintained the reading frame (**Figure 6.43D**).

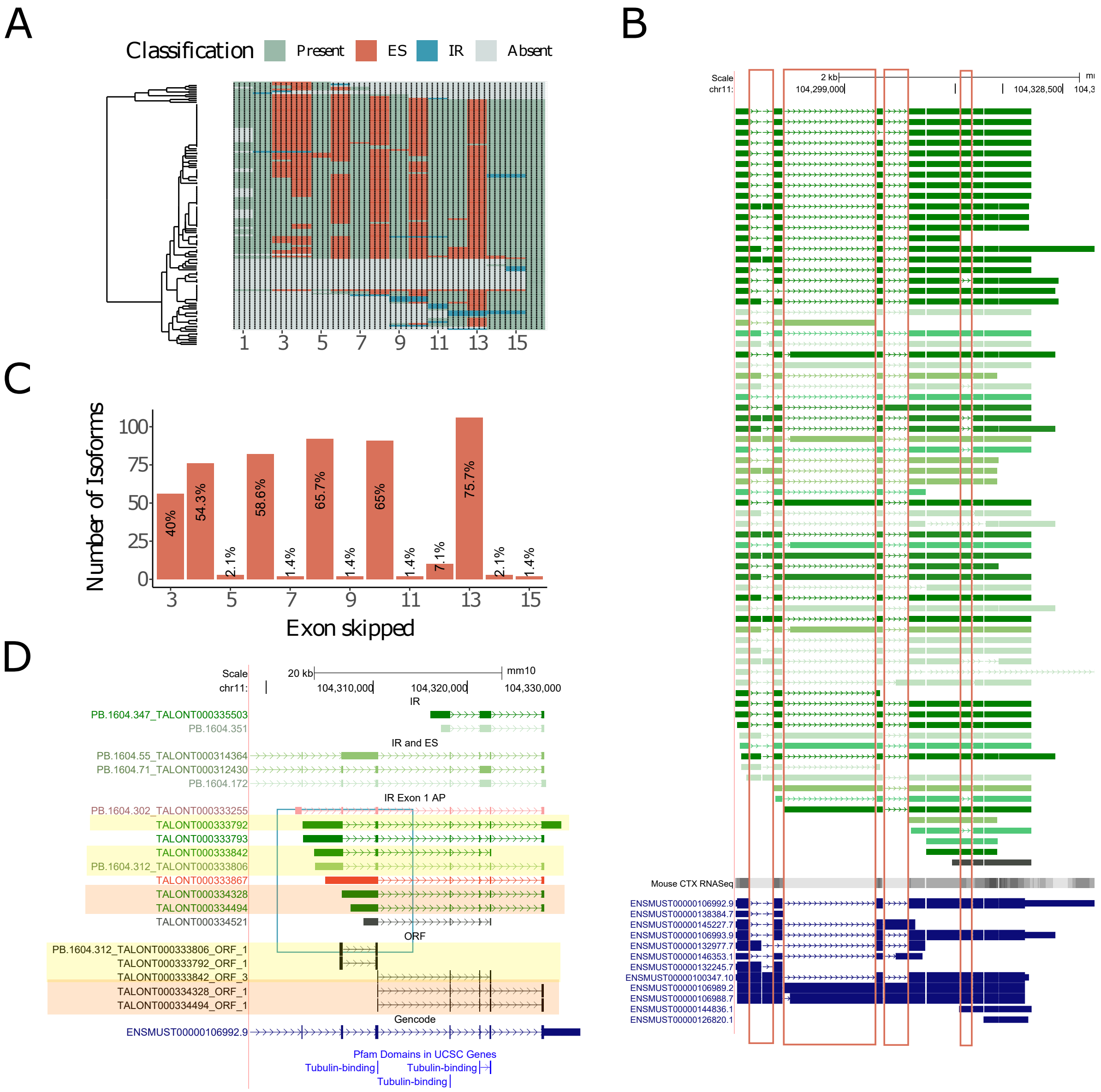


Figure 6.43: Characterisation of *Mapt* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Mapt* isoform landscape, **(B)** a UCSC genome browser track of a subset of isoforms with exon skipping (boxed in red) events, **(C)** a bar chart of the number of isoforms with exon skipping, and **(D)** a zoomed-in track of the shorter isoforms with an alternative first exon characterised with intron retention. The isoforms with intron retention spanning across exons 7 and 9 generated a truncated reading frame predicted for nonsense-mediated decay (highlighted in yellow), whereas the isoforms with intron retention spanning across exons 10 and 11 maintained the reading frame (highlighted orange).

6.3.8.12 *Picalm*

The phosphatidylinositol-binding clathrin assembly protein gene, *PICALM*, is another reproducible AD-associated gene identified by GWAS. Emerging evidence suggest that *PICALM*, which encodes an adaptor protein involved in clathrin-mediated endocytosis, mediates AD-associated by modulating production, trafficking and clearance of Aβ peptide.[341] Recent studies further showed that increased expression of *PICALM* rescued endocytic effects associated with *APOE*4.[340]

Spanning over 83kb on chromosome 7, the mouse *Picalm* gene is associated with 22 unique exons and 15 known isoforms. In our dataset, we detected 144 isoforms annotated to *Picalm*, including all the known isoforms except the non-coding isoform, Picalm-208 (ENSMUST00000207949.1). Approximately a fifth of the isoforms detected (n = 31 isoforms, 21.5%) spanned the full-length of the *Picalm* gene, while the remaining isoforms were significantly shorter and characterised with an alternative first exon (**Figure 6.44A,D**). Notably, we identified a novel isoform (PB.7635.2_TALONT001254093) that contained a novel exon 59kb upstream of exon 1 (**Figure 6.44F**). ORF prediction of this isoform, which was detected using both PacBio Iso-Seq and ONT nanopore sequencing, predicted a reading frame initiated from this novel first exon.

Although the exonic structure was broadly conserved across the isoforms (**Figure 6.44B,C**), we observed consistent exon skipping of exons 13, 18, and 21 (**Figure 6.44C,E**), which neither encode for the ANTH nor the ENTH protein domains; the ANTH domain is a membrane binding domain implicated in the formation of clathrin-coated pits, while the ENTH domain mediates membrane curvature and subsequent endocytosis. ORF predictions showed that skipping of such exons shortened but maintained the reading frame.

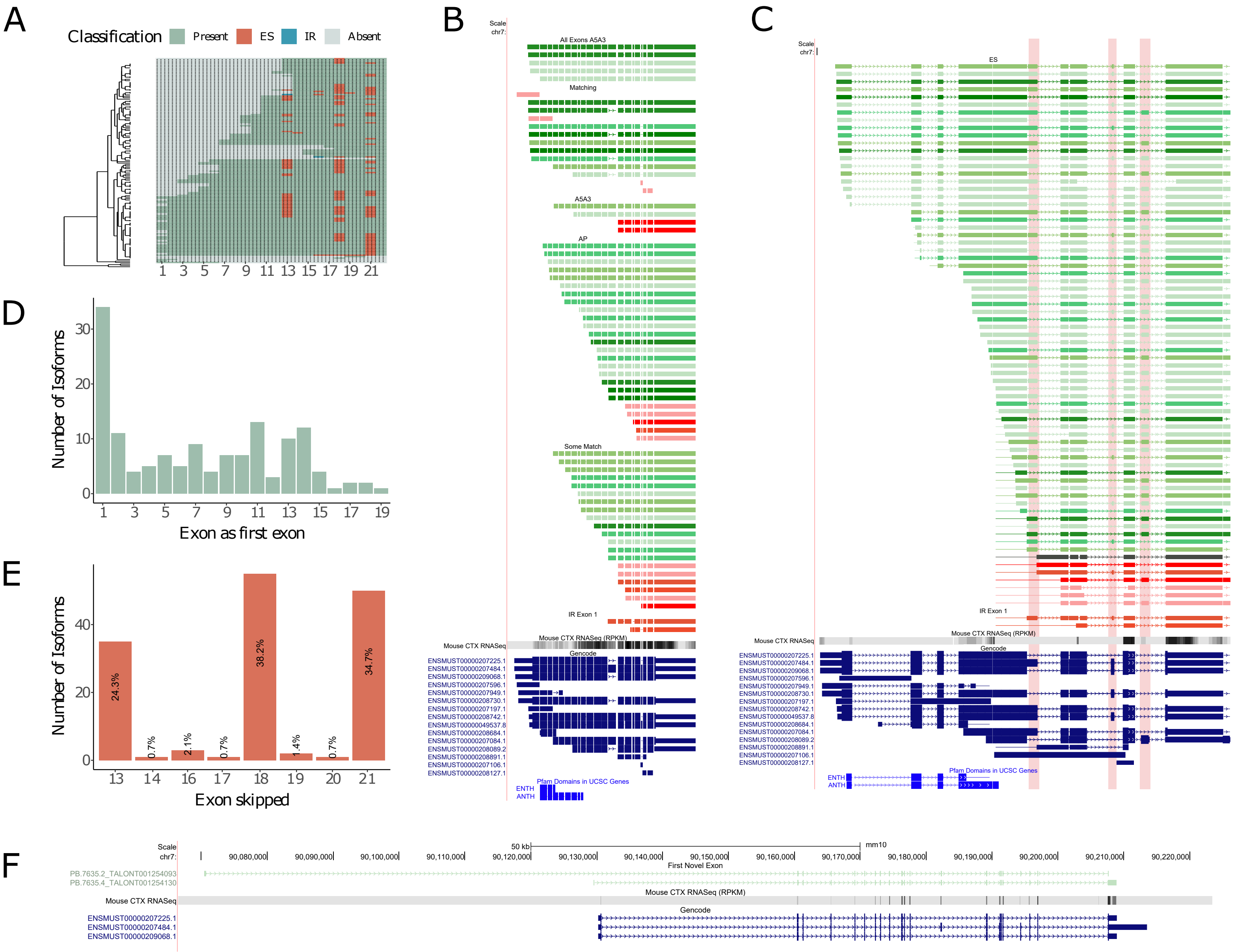


Figure 6.44: Characterisation of *Picalm* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Picalm* isoform landscape, **(B)** UCSC genome browser tracks of the isoforms that aligned to the reference isoforms, **(C)** isoforms with exon skipping events (highlighted in red), **(D)** bar charts of the number of isoforms with alternative first exons, and **(E)** exon skipping, and **(F)** a UCSC genome browser track of the novel isoform containing a novel exon upstream of the gene.

6.3.8.13 *Ptk2b*

Compelling evidence implicates the protein-tyrosine kinase 2-beta gene, *PTK2B*, in AD pathology with consistent identification of genetic variants associated with AD risk.[25,27] Encoding a protein tyrosine kinase (Pyk2) involved in key signalling pathways, *PTK2B* is involved in synaptic plasticity and activity. Of note, increased *Ptk2b* expression in a mouse model corrected deficits in synaptic proteins and improved the behaviour phenotype of transgenic mice.[342] A recent study further revealed a direct association of AD-associated *PTK2B* genetic variant with altered splicing,[92] though the mechanism underlying the role of *PTK2B* in AD pathology remains unclear.

Spanning over 127kb on chromosome 14, the mouse *Pt2kb* gene is associated with 31 unique exons and 8 known isoforms. Drawing parallels to the human-equivalent *PTK2B*, the mouse *Pt2kb* gene is similarly characterised with a tyrosine kinase domain flanked by a N-terminus FERM domain and a C-terminus FAT (focal adhesion targeting) domain.[376] In our dataset, we detected 563 isoforms annotated to *Ptk2b*. Deeper examination of *Pt2kb*-annotated isoforms revealed isoforms with varying lengths, containing exons that encode the kinase and FAT domain, but not the N-terminus FERM domain (**Figure 6.45A**). Previous studies have similarly identified human-equivalent isoforms missing the FERM and kinase domains,[376] and these isoforms were predicted to be transcribed from an alternative promoter as an endogenous regulator of Pyk2 activity.[376]

Finally, we noticed that while the exonic structure was largely conserved across the gene (**Figure 6.45B**), we detected high occurrence of exon skipping events localised to exon 27 (n = 296 isoforms, 52.5%) (**Figure 6.45C,D**). ORF predictions showed that skipping of this 24bp exon, which was present across all the known isoforms (i.e. "constitutive"), shortened but maintained the reading frame. We also detected a few isoforms with intron retention events (n = 6 isoforms, 1.1%). Examination of these IR events found them to be localised to the first exon of the shorter isoforms (**Figure 6.45E**), which shared a similar exonic structure to Ptk2b-205 (ENSMUST00000136216.7) in spanning across the 3' end and containing exons that encode the FAT domain. ORF predictions revealed that IR events spanning across exons 29 and 30 resulted in a reading frame shift with a truncated isoform predicted for nonsense-mediated decay (**Figure 6.45E**). Conversely, intron retention events that spanned across exons 25 to 28, thereby keeping exons 29 and 30 intact, maintained the reading frame with potential translation of a 125-amino-acid-peptide containing the FAT domain (**Figure 6.45E**).

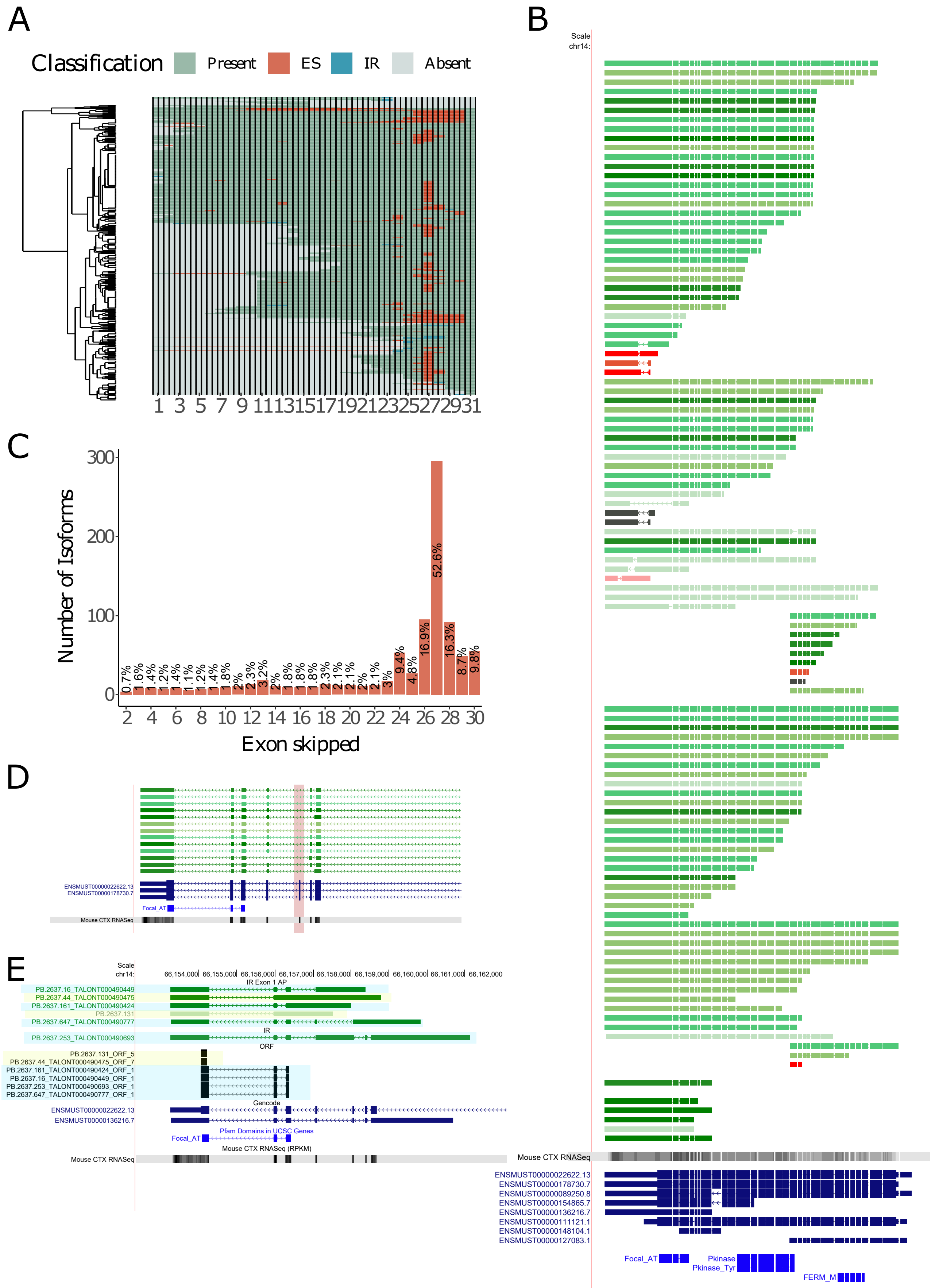
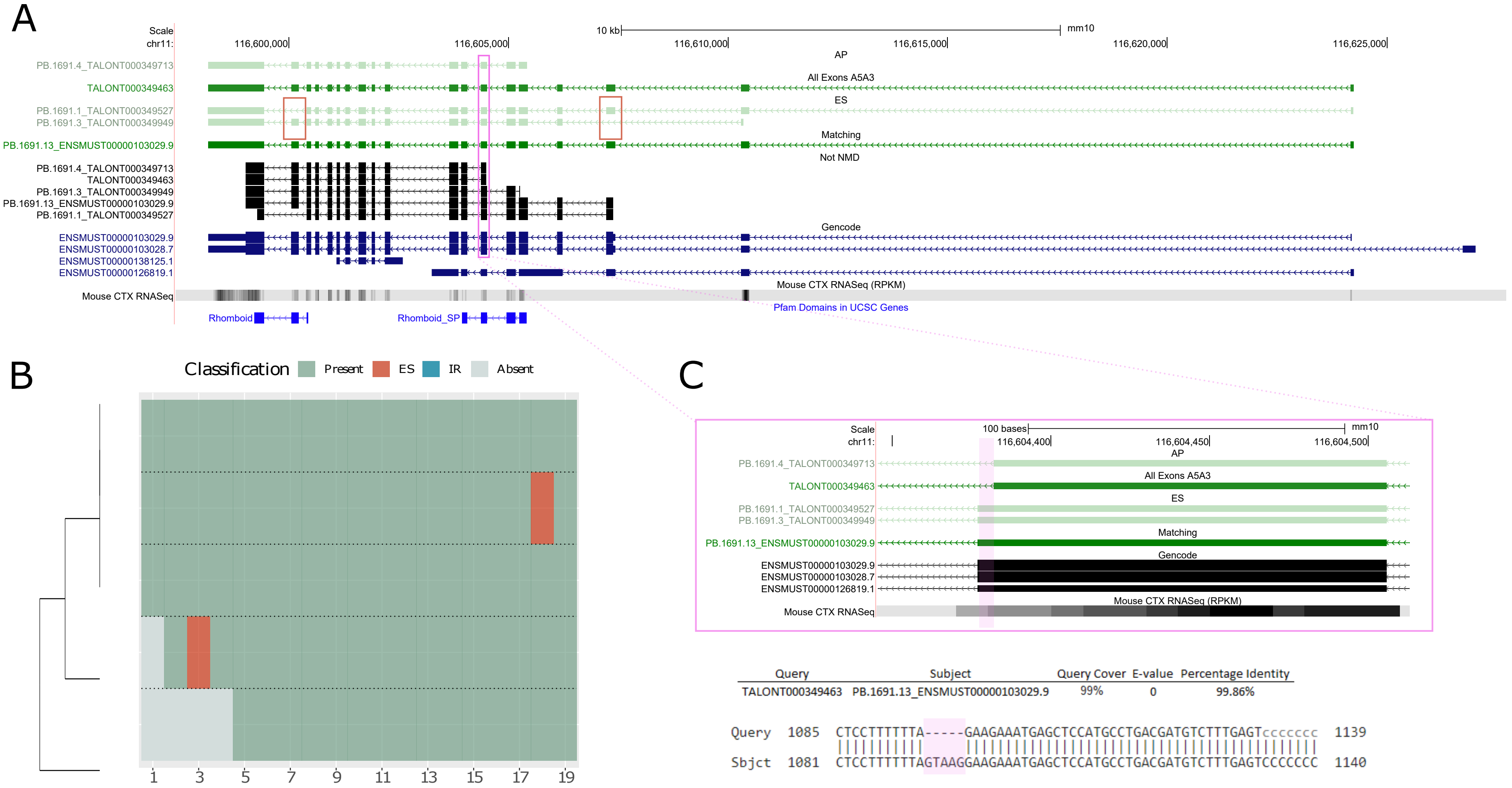


Figure 6.45: Characterisation of *Ptk2b* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Ptk2b* isoform landscape, **(B)** a UCSC genome browser track of a subset of isoforms that shared a similar internal exonic structure but varied at the 5' end, **(C)** bar charts of the number of isoforms with exon skipping, **(D)** zoomed-in tracks of isoforms with skipping of exon 27 and of **(E)** isoforms with intron retention events. Isoforms characterised with an intron retention event spanning across exons 29 and 30 were predicted for NMD (boxed yellow, isoforms not predicted for NMD are boxed in blue)

6.3.8.14 *Rhbdf2*

Aside from *Ank1* (described in **Section 6.3.8.3**), AD EWAS have consistently identified a differentially-methylated region residing in *RHBDF2*, a gene that encodes a rhomboid serine protease essential for TNFα secretion. The role of *RHBDF2* in AD pathogenies, however, remain poorly understood.

Spanning across 29kb on chromosome 11 with known 19 unique exons, the mouse *Rhbdf2* gene is the second least expressed gene from our panel of AD-associated target genes. In our dataset, we detected 5 isoforms, including the canonical isoform, Rhbdf2-202 (ENSMUST00000103029.9), and 2 exon skipping events (**Figure 6.46A,B**): exon 3 and exon 18, the latter which encodes the rhomboid domain. While *Rhbdf2* exhibited the least complex splicing pattern, ORF predictions of these isoforms revealed reading frame predictions of varying lengths (**Figure 6.46B**): i) the known canonical isoform appeared to be translated from an alternative start codon present in exon 3, ii) skipping of exon 3 was associated with a reading frame initiated at exon 5, iii) skipping of exon 18 retained the reading frame from exon 3 but reduced the frame at the 3' end, and iv) a 5bp deletion at exon 7 (**Figure 6.46C**), which encodes the protease domain, shifted the reading frame and subsequently resulted in usage of an alternative start codon at exon 7 rather than exon 3.



Query	Subject	Query Cover	E-value	Percentage Identity
TALONT000349463	PB.1691.13_ENSMUST00000103029.9	99%	0	99.86%

```
Query  1085  CTCCTTTTTTA-----GAAGAAATGAGCTCCATGCCTGACGATGTCTTTGAGTccccccc  1139
             |||||||||||     |||||||||||||||||||||||||||||||||||||
Sbjct  1081  CTCCTTTTTTAGTAAGGAAGAAATGAGCTCCATGCCTGACGATGTCTTTGAGTCCCCCCC  1140
```

Figure 6.46: Characterisation of *Rhbdf2* isoforms in the rTg4510 cortex. Shown are **(A)** a UCSC genome browser track of the isoforms annotated to *Rhbdf2*, **(B)** cluster dendrogram for an overview of the *Rhbdf2* isoform landscape, and **(C)** a zoomed-in track showing the 5bp deletion at exon 7, which resulted in a reading frame shift and the subsequent usage of the downstream start codon at exon 7.

6.3.8.15 *Snca*

Aggregates of the α-Synuclein protein, encoded by the *SNCA* gene, are one of the defining hallmarks for a number of neurodegenerative diseases, collectively known as synucleinopathies. Notably, up to 50% of AD patients are presented with co-morbid αSyn pathology.[377] While the precise mechanisms driving synucleinopathies pathogenesis are yet to be determined, there is increasing evidence for the role of altered *SNCA* splicing as a key mechanism for disease development.[347,348] Of note, recent studies have shown that skipping of exon 6 result in a truncated SNCA protein that exhibit increased propensity to aggregate.[347,348]

Spanning across 98kb on chromosome 6, the mouse *Snca* gene is characterised with 8 unique exons and 4 known isoforms. Despite having relatively few exons, *Snca* was the third most "isoformic" gene in our dataset with 622 isoforms. In line with findings from studying the human *SNCA* gene,[148] the mouse-equivalent full-length *Snca* isoform (Snca-201, ENSMUST00000114268.4), was the most abundant transcript sequenced using Iso-Seq (15,469 Iso-Seq full-length reads) and nanopore sequencing (283,239 ONT full-length reads), making up 60% of the total *Snca* mRNA transcripts.

Comprehensive characterisation revealed widespread usage of alternative splicing events, including exon skipping (**Figure 6.47A**), and alternative 5' and 3' splice sites (**Figure 6.47B**). Over 75% of isoforms (n = 483 isoforms, 77.7%) were identified with at least one exon skipping event. Exon 4 and 5, which partially encode the synuclein domain, were skipped the most (exon 4: n = 249 isoforms; exon 5: n = 206 isoforms) (**Figure 6.47C,D**), despite being present in all the known isoforms (i.e constitutive). Extensive usage of alternative 5' and 3' sites, particularly A5' truncation, was also observed across the gene with the exception of exon 4 (**Figure 6.47E**). Strikingly, approximately two-thirds of isoforms detected were either non-protein-coding (n = 391 isoforms, 62.8%) or not characterised by an open reading frame (n = 11 isoforms, 1.7%) (**Figure 6.47A,B**), a likely consequence of the widespread usage of alternative splice sites combined with exon skipping events.

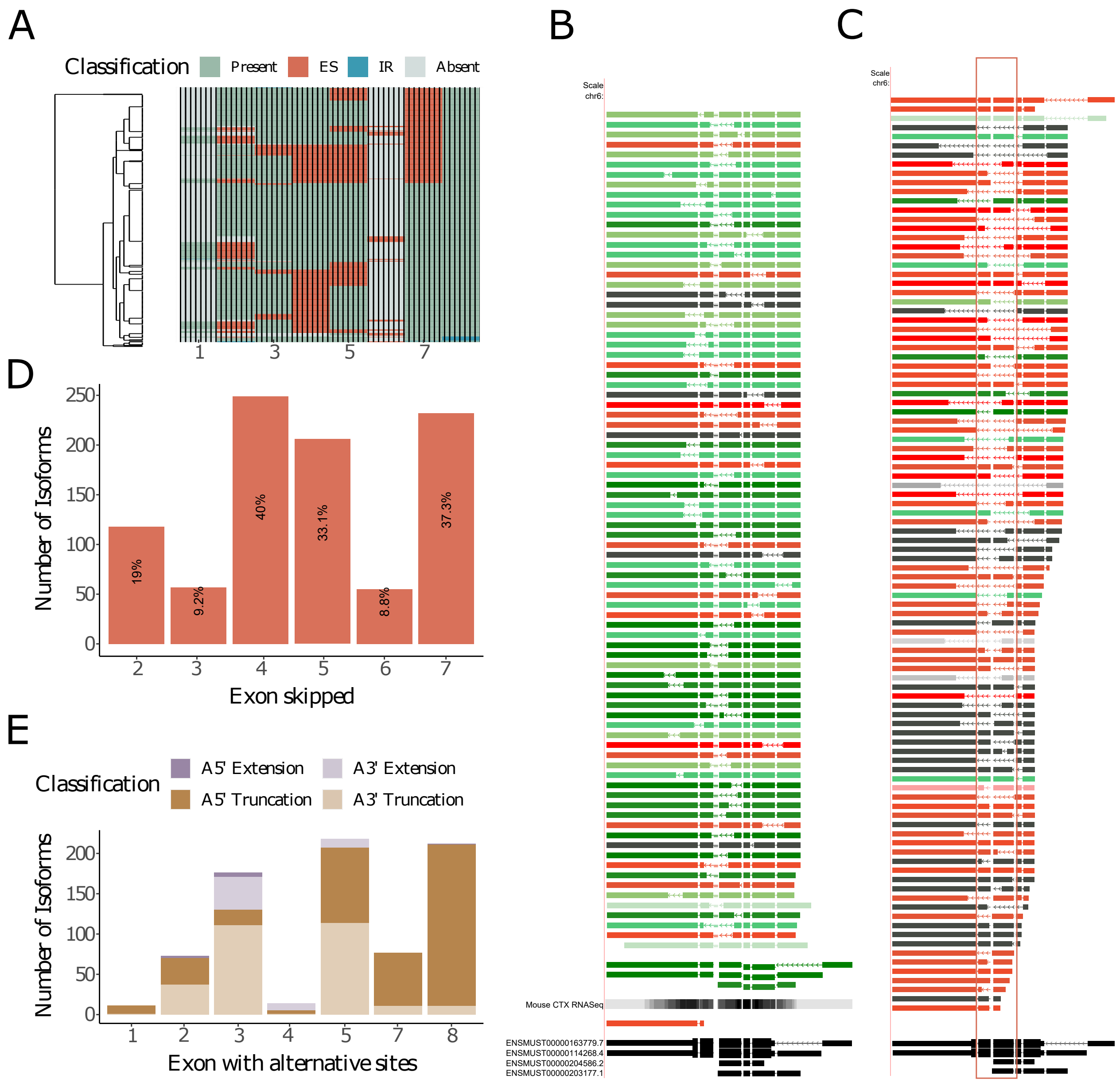


Figure 6.47: Characterisation of *Snca* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Snca* isoform landscape, **(B)** UCSC genome browser tracks of the subset of isoforms with alternative 5' and 3' splice sites, and **(C)** subset of isoforms characterised with exon skipping events, particularly skipping of exons 4 and 5 (boxed in red). Isoforms are coloured by protein-coding potential (green for protein-coding and red for non-protein-coding) and shaded by abundance. **(D)** Bar charts of the number isoforms with exon skipping, and **(E)** alternative 5' and 3' splice sites.

6.3.8.16 *Sorl1*

Genetic variation annotated to the sortilin-related receptor gene, *SORL1*, has been repeatedly identified as associated with AD risk.[349] Encoding an endocytic sorting receptor, *SORL1* is involved in trafficking of APP and regulation of Aβ production.[350] Of note, recent studies have found that deletion of *SORL1* in hiPSCs (human induced pluripotent stem cells) resulted in cell-type specific endosome enlargement with altered APP localisation.[350] Studies on post-mortem AD brain tissue have further revealed altered splicing of *SORL1* with decreased expression of the full-length SORL1 isoform, but consistent expression of the isoform lacking exon 2.[308]

Spanning over 160kb on chromosome 9, the mouse *Sorl1* gene is characterised with 49 unique exons and 4 known isoforms. This was in stark contrast to our dataset where we detected 113 isoforms annotated to *Sorl1*. Deeper examination of these isoforms, however, revealed that the majority largely shared the same internal exonic structure with few occurrences of exon skipping and intron retention events (**Figure 6.48A, F**). In contrast, over 75% (n = 88 isoforms, 77.9%) of the isoforms were characterised with an alternative first exon (**Figure 6.48C**), and can be broadly classified into three distinct groups by their alternative last exons: i) exon 37, ii) exon 38 (**Figure 6.48D**) and iii) exon 41 (**Figure 6.48E**). While the alternative last exons of such isoforms were in perfect alignment with one another (**Figure 6.48D,E**), they did not fully match the exon of interest, resulting in extensive variation of exons 37, 38 and 41 (**Figure 6.48B**). Notably, exons 37 and 38 encode the fibronectin type III (fn3) domain, where most of the rare AD-associated variants were found located.[378]

Sorl1 isoform variation in our dataset was thus driven by usage of alternative promoter and termination, generating isoforms of varying lengths that either contained the C-terminus Vps10-domain that binds to neurotrophic factors or the N-terminus fn3 domain. ORF prediction of these isoforms show a shortened, but otherwise intact reading frame with no prediction for nonsense-mediated decay.

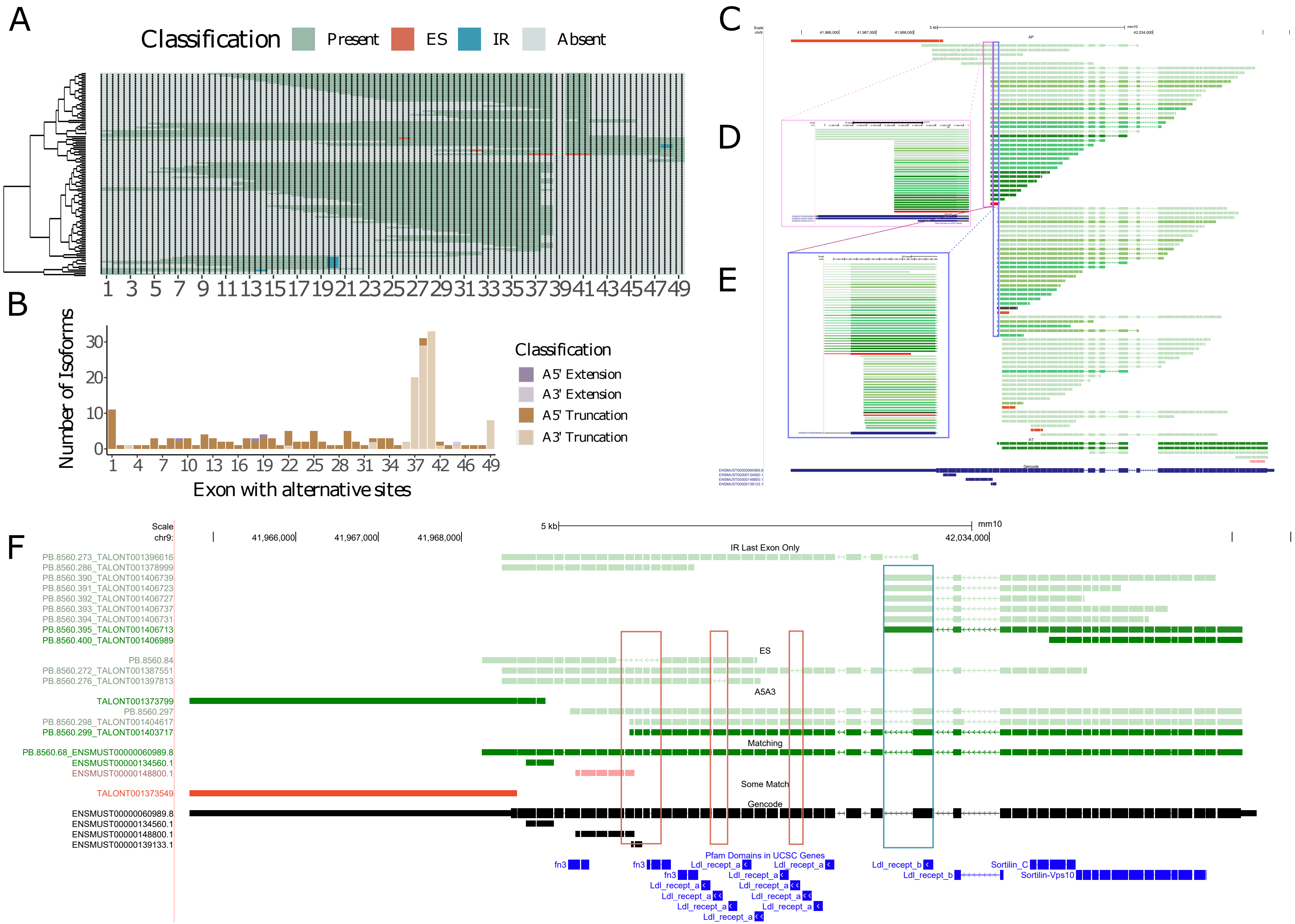


Figure 6.48: Characterisation of *Sorl1* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Sorl1* isoform landscape, **(B)** a bar chart of the number of isoforms with alternative splice sites, **(C)** a UCSC genome browser track of the isoforms that shared an identical internal exonic structure but contained an alternative first exon, **(D)** zoomed-in figure of these isoforms with alternative first exon overlapping exon 41, **(E)** isoforms with alternative first exon overlapping exon 38, and **(F)** a UCSC genome browser track displaying isoforms with exon skipping (boxed red) and intron retention (boxed blue) events.

6.3.8.17 *Tardbp*

Aggregates of the transactive response DNA binding protein (TDP-43), encoded by *TARDBP*, has long been established as a hallmark for ALS and Frontotemporal lobar degeneration (FTLD).[379] Notably, deposition of TDP-43 has been associated with the development of severe AD pathology with up to 60% of AD patients also characterised by TDP-43 pathology.[352] Furthermore, inheritance of APOE4 is associated with increased frequency of TDP-43 pathology, further implicating the role of TDP-43 in AD pathology.[379]

Spanning over 14kb on chromosome 4, the mouse *Tardbp* gene is characterised with 10 unique exons and 30 known isoforms. Despite only containing 10 exons, *Tardbp* is one of the most complex gene from our panel of AD-associated target genes; multiple isoforms are characterised with multiple exon overlap across the 3' end of the gene. Detecting 127 isoforms annotated to *Tardbp* in our dataset, we observed a similarly complex isoform landscape, capturing the full complement of known non-protein-coding and protein-coding isoforms (**Figure 6.49A,B**). Supplementing the complexity of the 3' end of the *Tardbp* gene, we further observed an enrichment of intron retention (IR) events between exons 6, 7 and 8 (**Figure 6.49B,D**). Deeper examination of IR events revealed them to belong to the final exon of some of the shorter detected isoforms (**Figure 6.49C**), resulting in the generation of a novel alternative last exon (**Figure 6.49F**). ORF prediction of these IR isoforms, which were further characterised with alternative 5' and 3' splice sites of the upstream exons (**Figure 6.49E**), revealed a truncated reading frame with such isoforms predicted for nonsense-mediated decay.

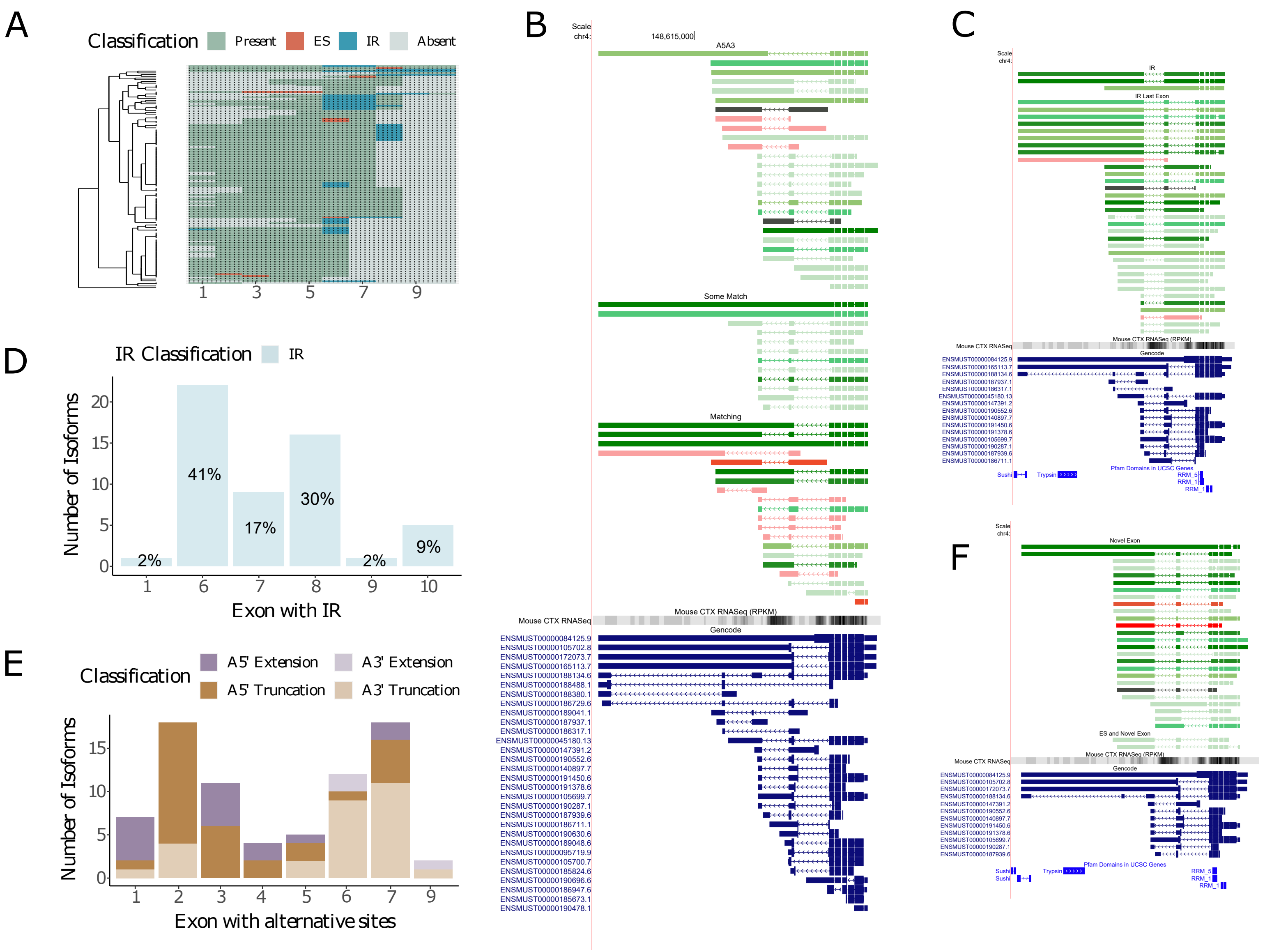


Figure 6.49: Characterisation of *Tardbp* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Tardbp* isoform landscape, **(B)** UCSC genome browser tracks of the isoforms that aligned to known *Tardbp* isoforms, **(C)** isoforms with intron retention events, **(D)** bar charts of the number of isoforms with intron retention events, and **(E)** usage of alternative 5' and 3' splice sites. **(F)** UCSC tracks of isoforms characterised by intron retention in the final exon.

6.3.8.18 *Trem2*

The triggering receptor expressed on myeloid cells 2 gene, *Trem2*, is a AD risk gene nominated by GWAS. Encoding a microglial-specific receptor of the innate immune response, *TREM2* is implicated in a range of microglial functions including inflammation, phagocytosis and proliferation. Notably, *TREM2* AD-associated variants have been found to induce partial loss-of-function of the TREM2 protein and modulate TREM2 signalling in microglia, impacting their response to Aβ plaques.[356,357] This is supported by recent studies, which show reduced microglia recruitment and phagocytosis of amyloid plaques in mouse models lacking *Trem2*.[355]

Spanning across 7kb on chromosome 4, the mouse *Trem2* gene is associated with 6 unique exons and 4 known isoforms. In our dataset, we detected 70 isoforms associated with *Trem2*, including the 3 known isoforms. While we did not detect the fourth known isoform, *Trem2-204* (ENMUST00000148545.1) - a non-protein-coding transcript - we identified a novel isoform that incorporated the unique exon (exon 3) associated with this isoform (**Figure 6.50A**), thereby containing a total of six exons.

Nonetheless, the vast majority of isoforms (n = 64 isoforms, 91.4%) were characterised with five exons or less (**Figure 6.50A, B**). These isoforms primarily differed in the usage of alternative 5' and 3' splice sites (**Figure 6.51B**), particularly in exon 2 which encodes the Ig-like V-type domain. This variability in exon 2 was further supported by RNA-Seq data from matched samples (**Figure 6.51B**). Strikingly, exon 2 was observed with the fewest exon skipping events (n = 2 isoforms, 2.9%) with such isoforms predicted to be non-coding (**Figure 6.50D**), highlighting the importance of exon 2. Noteworthy, the majority of AD-associated *TREM2* variants were located to the Ig-like V-type domain in the human-equivalent *TREM2* isoforms. In contrast, exon 4 was characterised with the fewest A5' and A3' splices sites (n = 5 isoforms, **Figure 6.50C,D**), but the most skipping events (n = 11 isoforms, 15.7%).

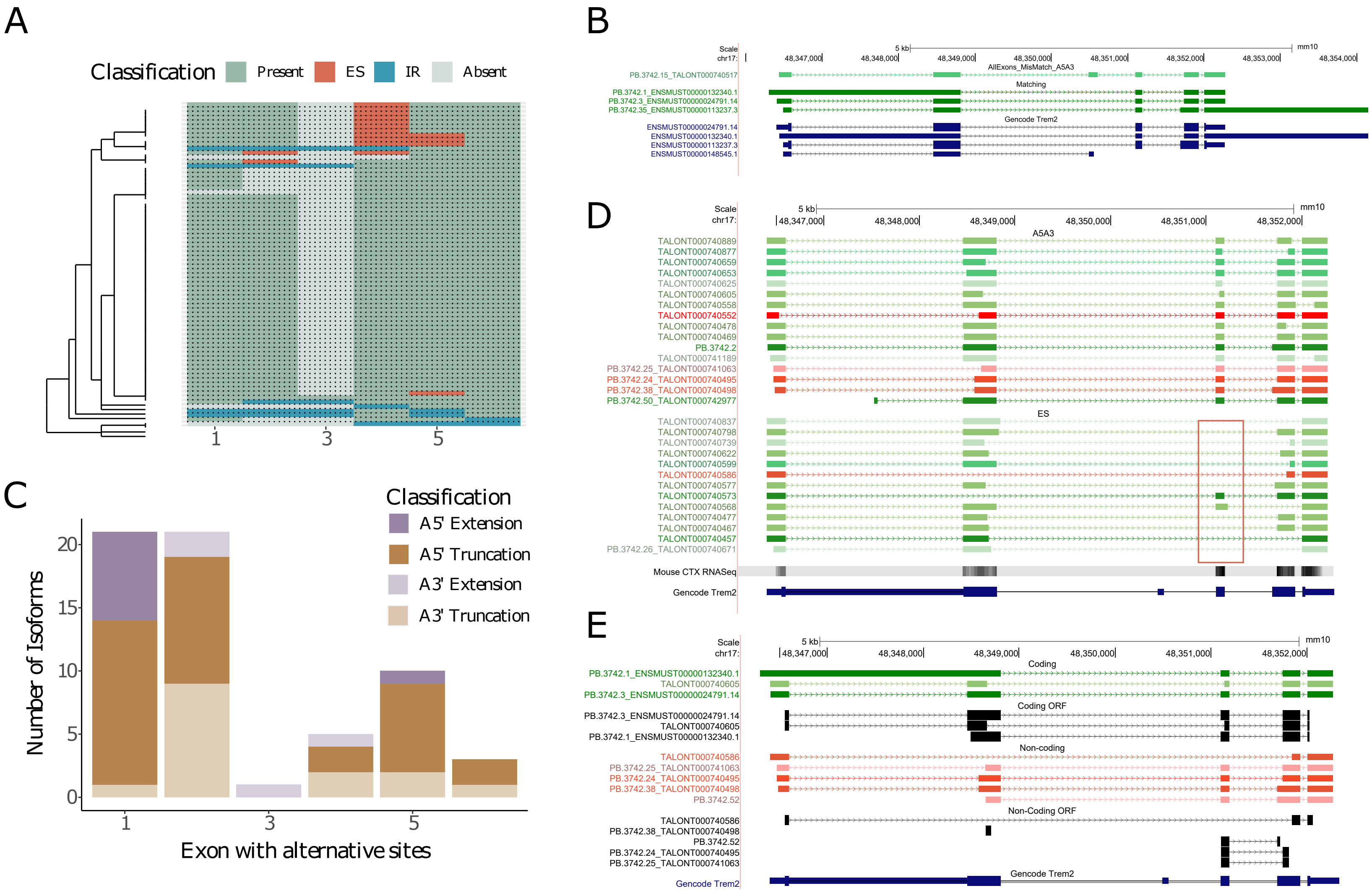


Figure 6.50: Characterisation of *Trem2* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Trem2* isoform landscape, **(B)** a UCSC genome browser track of the isoforms that aligned to known *Trem2* isoforms, **(C)** a bar chart of the number of isoforms with alternative splice sites, **(D)** a UCSC genome browser track of the *Trem2*-associated isoforms with alternative 5' and 3' splice sites and exon skipping events (boxed red), and of **(E)** protein-coding and non-protein-coding isoforms with their respective open reading frames (ORF).

Finally, we also detected 12 novel isoforms characterised with novel exons (**Figure 6.51**), which were confined to the 5' end of the *Trem2* gene: i) upstream of the first known exon (n = 3 isoforms), and ii) located between exon 1 and exon 2 with two varying lengths (~49bp = 4 isoforms, ~96 - 109bp = 5 isoforms). ORF predictions showed that the upstream novel exons did not encode for a start codon with the open reading frame still being initiated from the known first exon. In contrast, the internal novel exons were retained within the reading frame (**Figure 6.51**).

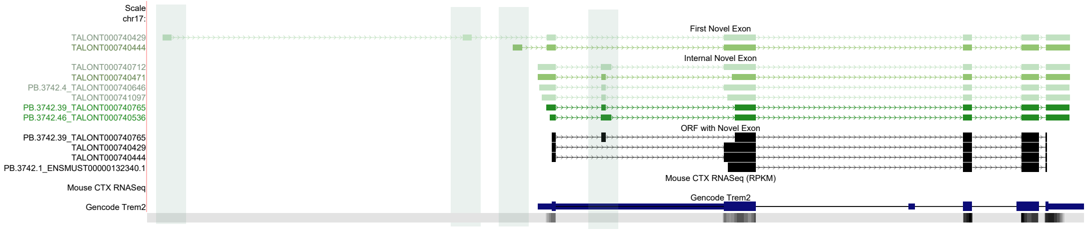
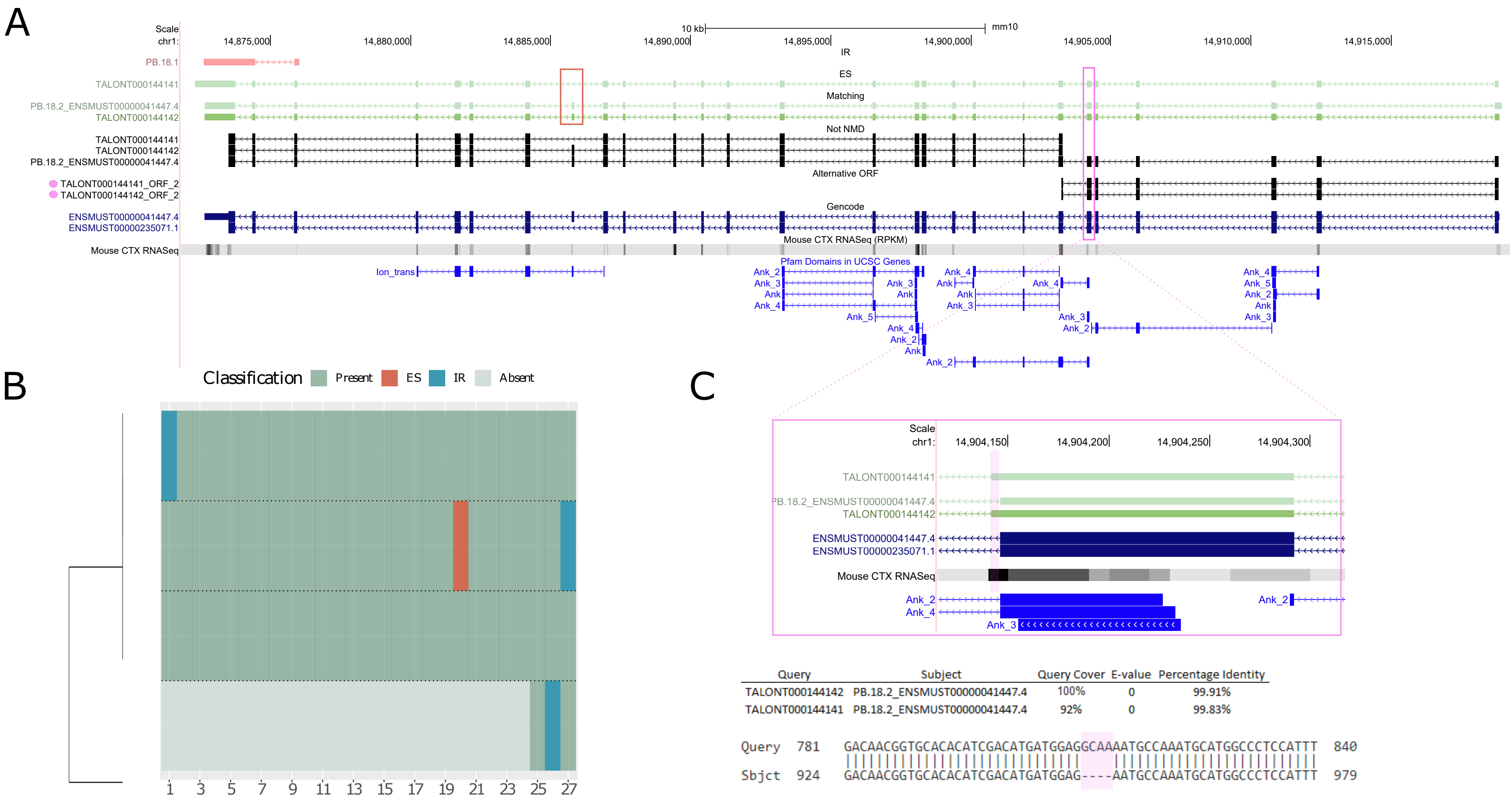


Figure 6.51: Characterisation of *Trem2* splicing events in the rTg4510 cortex. Shown is a UCSC genome browser track of a subset of isoforms annotated to *Trem2*, which contain novel exons located upstream of the gene and between exon 1 and 2. The predicted open reading frames from these isoforms are also shown (black tracks).

6.3.8.19 *Trpa1*

The transient receptor potential ankyrin 1 gene, *TRPA1*, encodes a non-selective calcium channel that is implicated in astrocytic hyperactivity at AD onset.[361] Supporting evidence showed that inhibition of TRPA1 normalised astrocyte activity and subsequently preserved synaptic integrity.[360] Deletion of TRPA1 in mouse models further reduced morphological damage and memory loss after Aβ injection, implicating a detrimental role of *TRPA1* in the early stages of AD pathology.[361]

Spanning over 46kb on chromosome 1, the mouse *Trpa1* gene is the least expressed gene amongst our panel of AD-associated target genes. Despite containing 27 exons, *Trpa1* is only associated with only 2 known isoforms. Unsurprisingly, *Trpa1* was thus characterised with the fewest isoforms in our dataset (n = 4 isoforms) (**Figure 6.52A**). Apart from detecting one of the two known canonical isoform, Trpa1-201 (ENSMUST00000041447.4), we detected a short novel non-coding isoform and two novel isoforms that spanned the length of the *Trpa1* gene (**Figure 6.52A,B**). Blast analysis of these isoforms revealed that the two novel long isoforms generally shared the exonic structure of Trpa1-201, with the exception of exon 20 skipping and a 4-nucleotide addition at the end of exon 6 (extension of the 3' splice site) (**Figure 6.52C**). ORF predictions showed that skipping of exon 20, which partially encodes the ion channel domain, shortened but maintained the reading frame (**Figure 6.52A**). In contrast, ORF predictions revealed that the 4-nucleotide addition at exon 6, which was validated by RNA-Seq data (**Figure 6.52C**), generated a short product destined for nonsense-mediated decay as a consequence of an in-frame stop codon (**Figure 6.52A**). Any translation of these two novel isoforms were thus predicted from the alternative start codon at exon 7 (**Figure 6.52A**), bypassing upstream exons that encode a subset of the ankyrin domains.



Query	Subject	Query Cover	E-value	Percentage Identity
TALONT000144142	PB.18.2_ENSMUST00000041447.4	100%	0	99.91%
TALONT000144141	PB.18.2_ENSMUST00000041447.4	92%	0	99.83%

```
Query  781   GACAACGGTGCACACATCGACATGATGGAGGCAAATGCCAAATGCATGGCCCTCCATTT  840
             |||||||||||||||||||||||||||||    ||||||||||||||||||||||||||
Sbjct  924   GACAACGGTGCACACATCGACATGATGGAG----AATGCCAAATGCATGGCCCTCCATTT  979
```

Figure 6.52: Characterisation of *Trpa1* isoforms in the rTg4510 cortex. Shown are **(A)** a UCSC genome browser track of isoforms annotated to *Trpa1* with exon 20 skipping and extension of exon 6 denoted in red and pink box respectively, **(B)** a cluster dendrogram for an overview of the *Trpa1* isoform landscape, and **(C)** a zoomed-in track showing the 4bp addition at the end of exon 6, which resulted in a reading frame shift and the generation of a stop codon (marked with a pink circle).

6.3.8.20 *Vgf*

The VGF nerve growth factor inducible gene, *VGF*, was first implicated in AD pathology after the repeated detection of decreased VGF-derived peptide levels in AD samples.[362] As a neurosecretory protein, VGF undergoes proteolytic processing to generate at least 12 VGF-peptides, which are essential for neurogenesis and synaptogenesis.[362] Administration of these peptides in AD mouse models reduced plaque burden, microglial activation and formation of defective dendrites.[380] Recent studies further showed that overexpression of VGF partially rescued memory impairment and neuropathology, suggesting a causal role for VGF in protecting against AD development.[364]

Spanning over 7kb on chromosome 5, the mouse *Vgf* gene is characterised with 6 unique exons and 4 known isoforms. Despite containing relatively few exons, we detected 90 isoforms associated with *Vgf* in our targeted dataset. Initial examination of this gene remarkably suggested a relatively simple splicing pattern (**Figure 6.53A**): i) sole usage of the alternative first exon from Vgf-201 (ENSMUST00000041543.8) with no detection of the first three exons from the long isoform, Vgf-204 (ENSMUST00000190827.6), ii) skipping of exon 5, which was only present in Vgf-202 (ENSMUST00000186451.1), in the majority of isoforms, and iii) a few intron retention events.

However, deeper examination revealed complex variations of the final exon and 3' UTR, supported by matched RNA-Seq data (**Figure 6.53B,C**). The vast majority of isoforms detected (n = 87 isoforms, 96.7%) were characterised with either usage of i) an alternative 5' splice site of the last exon (n = 35 isoforms, 38.9%), ii) an alternative 3' splice site of the last exon (n = 4 isoforms, 0.04%), or iii) matched 5' and 3' splice site but skipping within the last exon resulting in two enclosed exons (**Figure 6.53B**). This phenomenon was observed in isoforms that were detected using both Iso-Seq and ONT nanopore sequencing, and has been previously observed in *Apoe* (**Section 6.3.8.4**). ORF predictions of these isoforms showed that while this internal skipping phenomenon did not result in nonsense-mediated decay, it generated significant variations of the reading frame particularly at the 3' end. Conversely, isoforms with an alternative 5' start site of the last exon were predicted as either non-protein-coding or missing a reading frame (**Figure 6.53B**). We anticipate that this widespread isoform diversity, driven by an alternative last exon, would result in the generation of multiple VGF protein isoforms with differing cleavage sites.

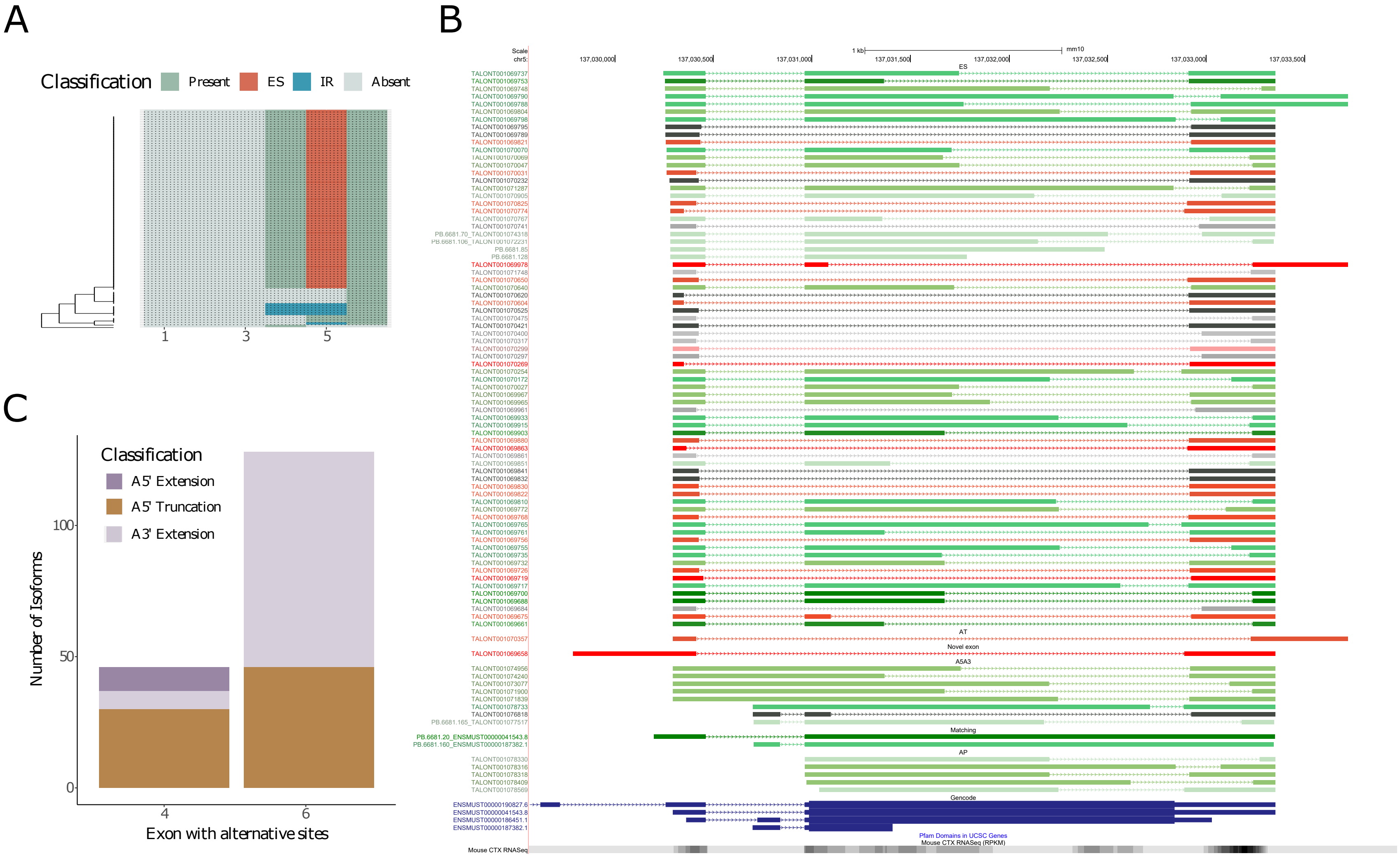


Figure 6.53: Characterisation of *Vgf* isoforms in the rTg4510 cortex. Shown are **(A)** a cluster dendrogram for an overview of the *Vgf* isoform landscape, **(B)** a UCSC genome browser track of a subset of isoforms illustrating the complex variation of the last exon, and **(C)** a bar chart of the number of exons with alternative 5' and 3' splice sites.

6.3.9 Improved sensitivity from targeted sequencing detects up-regulation of AD-associated genes in rTg4510 TG mice

Despite the success of capturing full-length transcripts from long-read sequencing of the global transcriptome, we have shown that this approach fails to robustly detect less abundant genes and isoforms (**Section 6.3.2**). In contrast, we have illustrated that target enrichment achieves deep sequencing of 20 AD-associated target genes, revealing unprecedented diversity of alternatively-spliced isoforms including hundreds of novel transcripts not previously described in existing reference annotations or in the global Iso-Seq dataset (**Section 6.3.8**). We anticipated that this deep sequencing coverage would further allow more accurate quantification of gene and isoform expression using normalised full-length read counts (**Figure 3.21C**), forgoing the need of short-read RNA-Seq data for quantitative analyses (**Figure 3.21B**). Subsequently, we sought to characterise transcriptional differences of these well-known AD-associated genes between rTg4510 WT and TG mice.

Using targeted Iso-Seq reads for both annotation and quantification, we identified three genes that were up-regulated with progressive tau pathology in rTg4510 TG mice (summarised in **Table 6.9**): *Trem2* ($\log_2$ fold change between TG and WT mice aged 8 months ($\log_2FC_g$) = 2.26, R^2 = 0.788, P = 2.45 x 10^{-16}), *Cd33* ($\log_2FC_g$ = 1.79, R^2 = 0.588, P = 4.5 x 10^{-9}) and *Rhbdf2* ($\log_2FC_g$ = 1.39, R^2 = 0.5, P = 3.06 x 10^{-6}). Up-regulation of these genes was also observed using normalised ONT read counts, with significantly greater expression differences between WT and TG mice, due to the greater sequencing depth achieved with ONT nanopore sequencing (*Trem2*: $\log_2FC_g$ = 2.47, R^2 = 0.938, P = 1.91 x 10^{-40}; *Cd33*: $\log_2FC_g$ = 2.25, R^2 = 0.823, P = 2.97 x 10^{-20}; *Rhbdf2*: $\log_2FC_g$ = 1.31, R^2 = 0.564, P = 1.36 x 10^{-6}). In each of these validated genotype-associations, the direction of effect was the same (**Table 6.9**).

Finally, we detected a significant increase in *Apoe* ($\log_2FC_g$ = 1.44, R^2 = 0.76, P = 1.45 x 10^{-8}), *Clu* ($\log_2FC_g$ = 1.36, R^2 = 0.80, P = 1.39 x 10^{-16}) and *Abca1* ($\log_2FC_g$ = 1.56, R^2 = 0.7, P = 5.66 x 10^{-5}) gene expression using normalised ONT but not Iso-Seq FL read counts; the direction of effect however was the same, suggesting that the Iso-Seq-derived analyses were constrained by power. Our findings corroborated with a recent RNA-Seq study that reported *Trem2* and *Apoe* up-regulation in isolated-microglia from rTg4510 mice.[351] Notably, these gene expression differences were not recapitulated using counts from the Iso-Seq global dataset, highlighting the higher sensitivity of targeted sequencing for gene expression analyses.

Table 6.9: Differential gene and transcript expression analysis from targeted profiling of the rTg4510 cortex. Tabulated is a summary of the differential expression analyses performed using full-length counts derived from Iso-Seq and ONT nanopore sequencing. Grey blocks refer to no significant difference in expression.

Target gene	Differential gene expression[a]		Differential transcript expression[b]			
			Iso-Seq		ONT	
	Iso-Seq	ONT	Known	Novel	Known	Novel
Abca1		1.56 (0.7, 5.66 x 10^{-5})			1	
Abca7						2
Ank1						
Apoe		1.44 (0.76, 1.45 x 10^{-8})		1	4	134
App			3	5	1	18
Bin1					3	48
Cd33	1.79 (0.588, 4.5 x 10^{-9})	2.25 (0.823, 2.97 x 10^{-20})			2	7
Clu		1.36 (0.802, 1.39 x 10^{-16})	1		1	165
Fus						21
Fyn						2
Mapt					2	16
Picalm					2	3
Ptk2b				1	2	12
Rhbdf2	1.39 (0.499, 3.06 x 10^{-6})	1.31 (0.564, 1.36 x 10^{-6})		1		1
Snca						40
Sorl1					1	3
Tardbp						2
Trem2	2.26 (0.788, 2.45 x 10^{-16})	2.47 (0.938, 1.91 x 10^{-40})	1	3	3	41
Trpa1						
Vgf						12

[a] Statistics reported as $\log_2$ fold change at 8 months TG vs WT (R^2, *P*)

[b] Total number of known and novel transcripts identified as differentially expressed

6.3.10 Gene-specific differential transcript expression and usage in TG mice

Given the improved sensitivity of target enrichment and the extensive mapping of isoform landscape of AD-associated genes (detailed in **Section 6.3.8**), we next sought to identify differences in transcript expression between rTg4510 TG and WT mice. In total, we detected 553 differentially expressed transcripts using normalised ONT reads counts (summarised in **Table 6.9**), the majority (n = 485, 87.7%) of which were associated with progressive tau pathology and the remaining associated with either age or genotype alone. Among these, 448 transcripts (81%) were novel with the greatest number of differentially expressed transcripts annotated to *Clu* (n = 151, 31.1%). Of note, differential transcript expression was detected for all 20 AD-associated target genes with the exception of *Ank1* and *Trpa1*.

Despite this unprecedented detection of novel transcripts whose expression altered with increased tau pathology, we found that the majority (n = 366, 75.4%) of these isoforms were lowly-expressed ($<$ 20 normalised full-length read counts) and accounted for less than 5% of the respective isoform fraction. Further examination revealed that the isoform landscape across the 20 AD-associated genes were characterised by a few dominant isoforms (**Figure 6.54**). This corroborates with recent findings from the VastDB[381] (the largest resource documenting genome-wide AS events in vertebrates, to date) that reported simultaneous expression of multiple major isoforms in more than 18% of genes.[381]

Reflecting the relatively lower sequencing depth of the Iso-Seq targeted dataset, we only detected 16 differentially expressed transcripts using Iso-Seq normalised read counts. Comparison of the ONT and Iso-Seq targeted dataset revealed 6 (37.5%) that were commonly identified as differentially expressed (**Table 6.10**): three transcripts were annotated to *Trem2* (**Figure 6.55**), one novel transcript to *Clu* (**Figure 6.56A,B,C**), one known transcript to *Ptk2b* (**Figure 6.56D,E,F**) and one novel transcript to *Apoe* (**Figure 6.56G,H,I**). All 6 transcripts were up-regulated with progressive tau pathology in the rTg4510 mice with similar effect size in the Iso-Seq and ONT targeted datasets (**Table 6.10**).

The following sections describe the transcriptional profiles of *Trem2* (**Section 6.3.10.1**), *Cd33* (**Section 6.3.10.2**) and *Bin1* (**Section 6.3.10.3**) in detail, which exhibited significant variation associated with progression of tau pathology. Profiles of the remaining 17 AD-associated genes can be found in **Appendix E**.

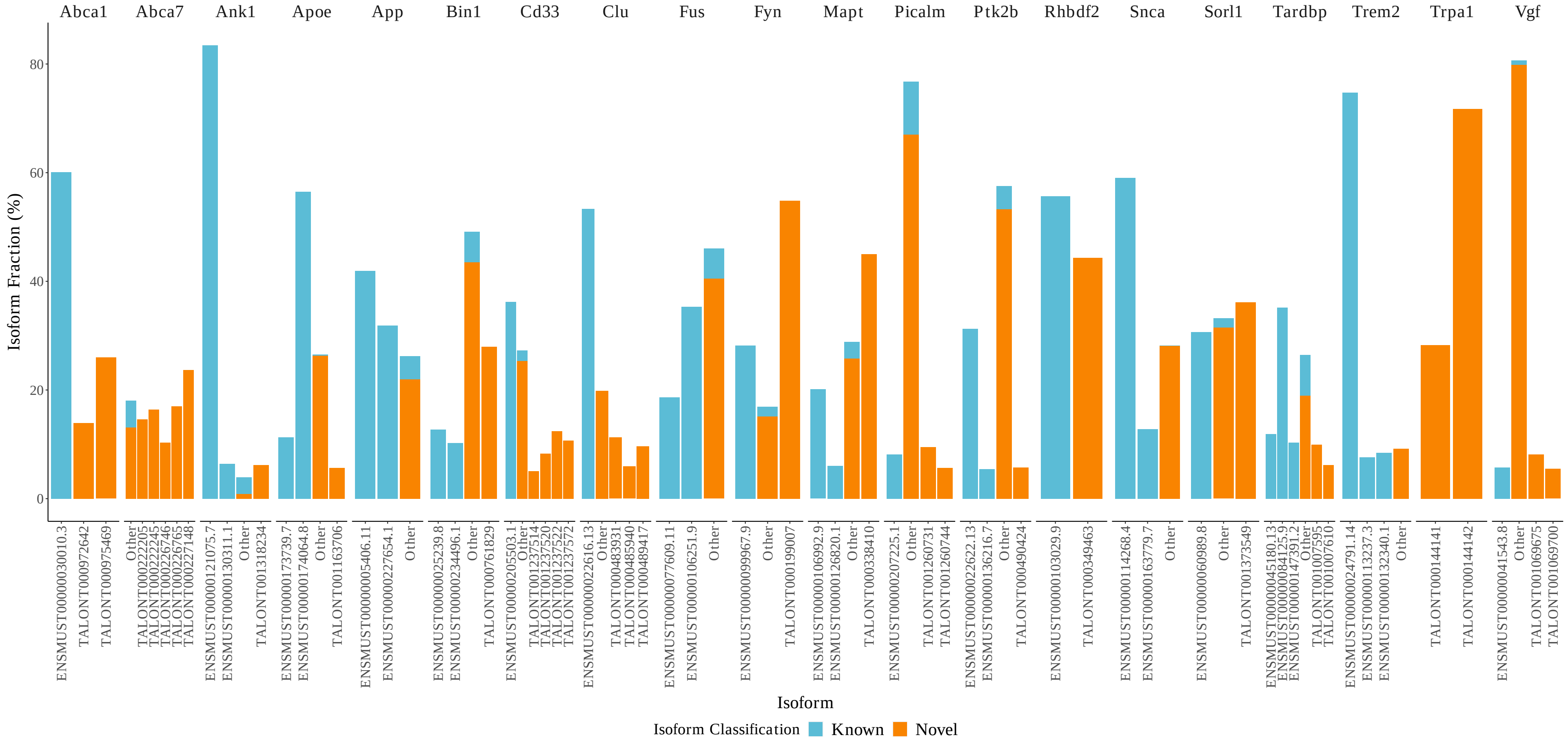


Figure 6.54: The isoform landscape for the majority of AD-associated genes were dominated by a few major isoforms. Shown is a bar chart of the proportion of major isoforms annotated to the 20 AD-associated target genes that were enriched for targeted sequencing of the rTg4510 cortex. Isoforms that constitute < 5% of total counts are clustered as "Other". Known isoforms refer to isoforms existing in mouse reference annotations (mm10, GENCODE, vM22). The proportion of each isoform (isoform fraction) is calculated by dividing the mean expression (ONT full-length normalised count) of the respective isoform across biological replicates over the total mean expression of all the isoforms across all of the samples in the ONT dataset (n = 8 WT, n = 10 TG).

Table 6.10: Common differentially spliced transcripts identified from Iso-Seq & ONT targeted profiling. Tabulated is a summary of the 6 transcripts commonly identified as differentially expressed using Iso-Seq and ONT targeted profiling of rTg4510 mice.

Transcript		Gene	Iso-Seq summary statistics				ONT summary statistics			
PacBio ID	ONT ID[a]		FDR[b]	R^{2}[c]	$\log_2FC_g$[d]	$\log_2FC_a$[e]	FDR[b]	R^{2}[c]	$\log_2FC_g$[d]	$\log_2FC_a$[e]
PB.3742.3	ENSMUST00000024791.14	*Trem2*	1.75 x 10^{-14}	0.79	2.08	2.27	1.5 x10^{-41}	0.939	2.42	2.46
PB.3742.1	ENSMUST00000132340.1	*Trem2*	1 x 10^{-6}	0.561	Inf	Inf	6.47 x 10^{-22}	0.873	1.93	2.18
PB.2634.256	TALONT000465283	*Clu*	4.63 x 10^{-7}	0.593	1.12	1.3	1.47 x 10^{-13}	0.827	1.4	1.57
PB.2637.336	ENSMUST00000089250.8	*Ptk2b*	2.8 x 10^{-13}	0.74	2.05	1.65	2.62 x 10^{-6}	0.609	1.46	1.02
PB.3742.12	TALONT000740634	*Trem2*	1.98 x 10^{-18}	0.773	2.66	2.93	3.91 x 10^{-6}	0.577	2.34	2.64
PB.7333.32	TALONT001163706	*Apoe*	1.49 x 10^{-4}	0.549	0.615	0.891	1.07 x 10^{-3}	0.546	0.976	1.17

[a] The isoform classification can be inferred from the ONT ID whereby the prefix “ENSMUST” and “TALONT” refers to known and novel isoform, respectively

[b] False discovery rate

[c] R^2 is a statistical measure that represents the amount of variance explained by the model

[d] Log_2 fold change of TG aged 8 months vs WT aged 8 months

[e] Log_2 fold change of TG aged 8 months vs TG aged 2 months

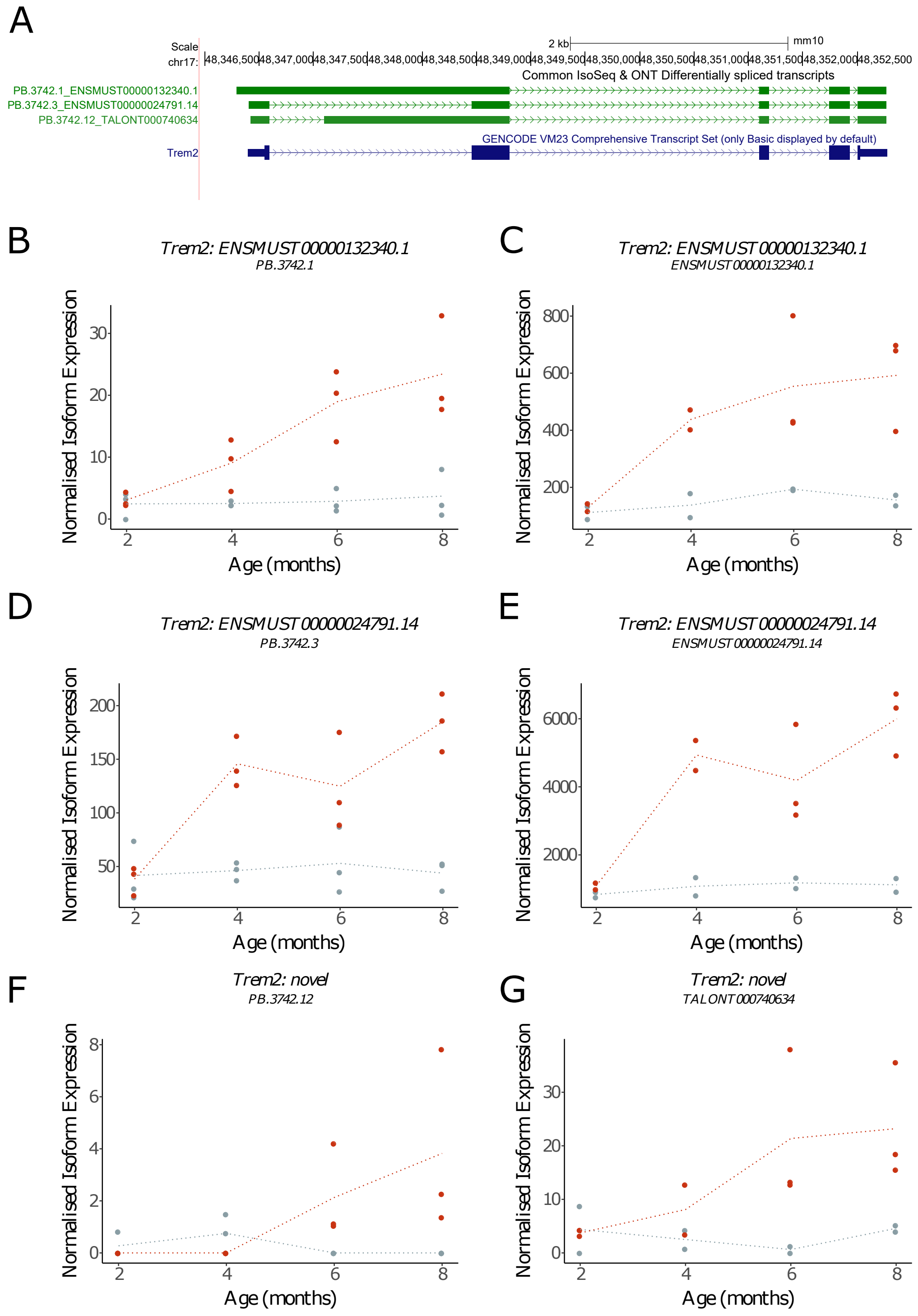


Figure 6.55: Three common differentially expressed transcripts annotated to *Trem2*. Shown is **(A)** a UCSC genome browser of the three *Trem2* transcripts that were commonly identified as differentially expressed in Iso-Seq and ONT targeted dataset, and their respective transcript expression in WT (grey) and rTg4510 TG mice (red) measured using **(B, D, F)** normalised Iso-Seq full-length read counts and **(C, E, G)** normalised ONT full-length read counts. Dotted lines represent the mean paths across age.

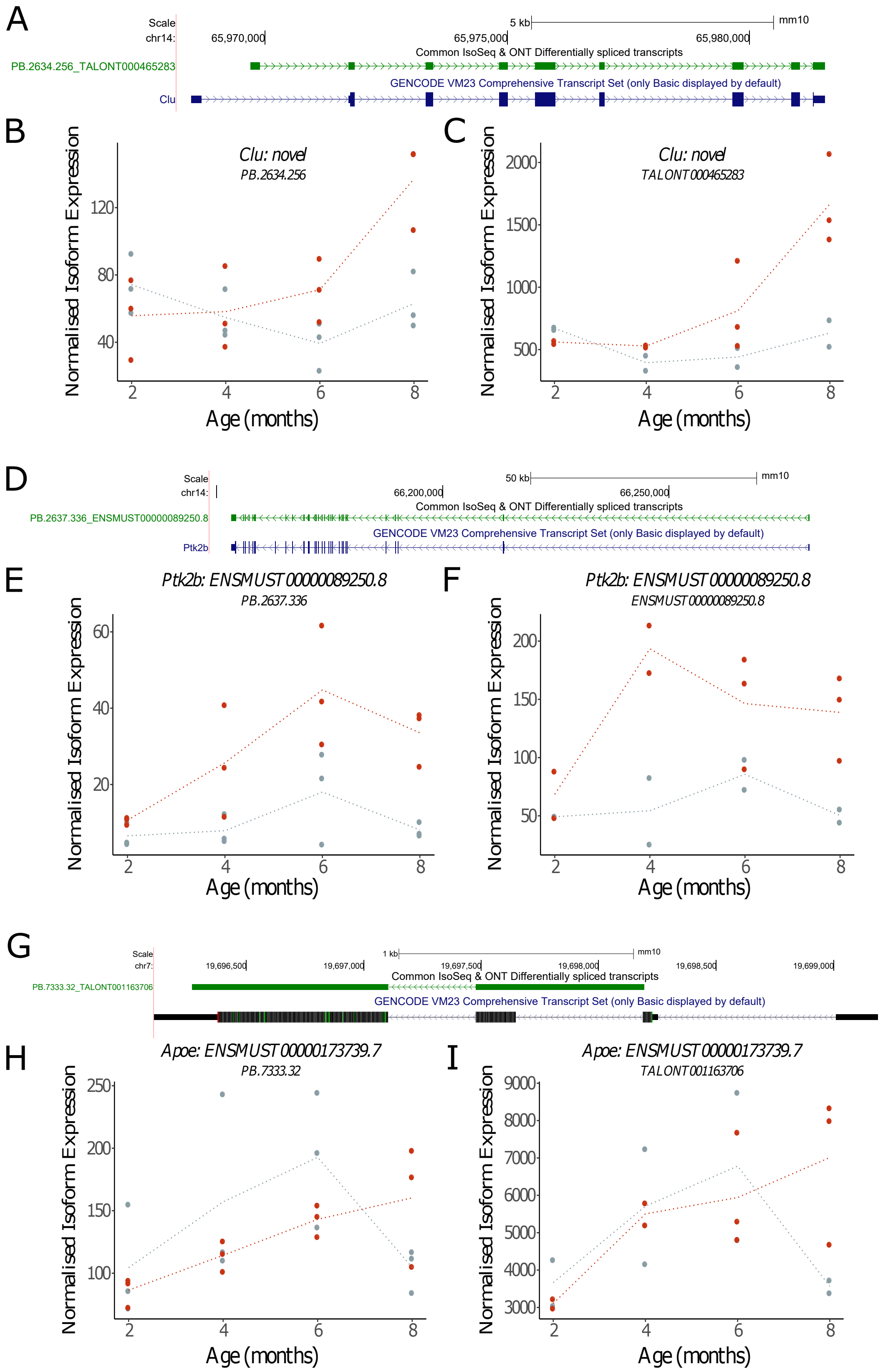


Figure 6.56: Common differentially expressed transcripts annotated to *Clu, Ptk2b* and *Apoe*. Shown are **(A, D, G)** UCSC genome browser tracks of the common (i.e. identified in both ONT and Iso-Seq targeted dataset) differentially expressed transcripts annotated to *Clu*, *Ptk2b* and *Apoe*, and their respective transcript expression in WT (grey) and rTg4510 TG mice (red) measured using **(B, E, H)** normalised Iso-Seq full-length read counts and **(C, F, I)** normalised ONT full-length read counts. Dotted lines represent the mean paths across age.

6.3.10.1 Global increase of *Trem2*-associated isoforms, particularly Trem2-201

The top ranked differentially expressed transcript between WT and TG mice was a known isoform (Trem2-201, ENSMUST00000024791.14, **Figure 6.57A**) annotated to *Trem2* (detailed annotations of *Trem2* are provided in **Section 6.3.8.18**). Expression of this known isoform significantly dominated that of the novel isoforms (**Figure 6.57C,G**), and was strongly up-regulated with progressive tau pathology (**Figure 6.57D,H**); this was evident in both the ONT and Iso-Seq targeted dataset adding confidence to this finding. Drawing parallels to *Gfap* (**Figure 5.9**) and *C4b* (**Figure 5.10**), up-regulation of Trem2-201 mirrored that of *Trem2* gene expression (**Figure 6.57B,F**), indicating that the increased *Trem2* gene expression in aged rTg4510 TG mice was primarily driven by one dominant isoform. Despite this up-regulation of Trem2-201, we found that there was no change in isoform usage across genotype or with age (**Figure 6.57E,I**); Trem2-201 occupied over 75% of the isoform proportion in rTg4510 irrespective of genotype and age. The remaining isoform proportions were equally divided between the two other known isoforms (**Figure 6.57I**) - Trem2-202 (ENSMUST00000113237.3) and Trem2-203 (ENSMUST00000113237.3) - and all of the other novel isoforms combined. These findings suggest that the drastic up-regulation of the dominant isoform (Trem2-201) in aged rTg4510 TG mice was also accompanied with an increased expression of all minor isoforms by the same magnitude, resulting in a consistent isoform proportion. Hierarchical clustering of individual samples based on *Trem2* isoform expression level confirmed the robust differences between TG and WT groups by age, reflecting the global increase of *Trem2*-associated transcript expression with progressive tau pathology.

6.3.10.2 Drastic up-regulation of Cd33-203 accompanied with reduced usage of other isoforms

Aside from *Trem2*, we identified differences in expression of the canonical isoform annotated to *Cd33* with accumulation of tau in rTg4510 TG mice (detailed isoform annotations of *Cd33* are provided in **Section 6.3.8.7**). Using normalised ONT read counts, we noted a significant increase in gene expression of *Cd33* (**Figure 6.58B,E**) and its known isoform - Cd33-203 (ENSMUST00000205503.1) - in aged rTg4510 TG mice (**Figure 6.58A,G**). However, unlike *Trem2*, the *Cd33* isoform landscape was relatively more complex (**Figure 6.58C,F**), suggesting that the increased *Cd33* gene expression was not solely driven by this one isoform. Several novel isoforms were found to be abundantly expressed, occupying ~50% of *Cd33* isoform proportion (**Figure 6.58H**). Among these, two isoforms (TALONT001237522, TALONT001237572) were also significantly up-regulated with progressive tau pathology (**Figure 6.58G**) (TALONT001237572: $\log_2$FC = 2.65, R^2 = 0.79, *P* = 2.37 x 10^{-14}; TALONT001237522: $\log_2$FC = 1.91, R^2 = 0.90, *P* = 7.62 x 10^{-25}). Both isoforms differed from the known isoform by an alternative last exon characterised with an intron retention event spanning across exon 7 and exon 8 (**Figure 6.58A**). Finally, we observed a notable change in isoform usage between rTg4510 TG and WT mice aged 8 months (**Figure 6.58H**) with up-regulation of Cd33-203 coupled with the down-regulation of a novel mono-exonic isoform (TALONT001237520, R^2 = 0.50, *P* = 2.38 x 10^{-5}) that spanned the 3' UTR.

6.3.10.3 Gradual isoform switch in expression of *Bin1* known isoforms

Drawing parallels to *Cd33*, we similarly identified expression differences in known isoforms annotated to *Bin1* (comprehensive characterisation of the *Bin1* isoform landscape in the rTg4510 cortex is provided in **Section 6.3.8.6**). Using normalised ONT read counts, we observed a significant increase in expression of the two known isoforms - Bin1-205 (ENSMUST00000234496.1) and Bin1-206 (ENSMUST00000234857.1) - associated with progressive tau pathology (**Figure 6.59A,G**). Both these isoforms spanned the full-length of the *Bin1* gene, and differed by an additional exon skipping event (**Figure 6.59A**); Bin1-205 and Bin1-206 were both characterised by skipping of exon 7, which partially encoded the N-BAR domain and exons 12 - 15, with exon 16 also skipped in Bin1-205. The significant up-regulation of both isoforms were marked with a notable change in isoform usage between rTg4510 TG and WT mice aged 8 months (**Figure 6.59H**) with down-regulation of the two other major isoforms: i) Bin1-201, the know isoform which contains all the exons, (ENSMUST00000025239.8) and ii) a novel isoform (PB.3915.2_TALONT000761829), which shares a similar exonic structure to Bin1-201 bar the skipping of exon 16. Notably, these findings corroborate with a recent study that showed differential isoform expression in human AD post-mortem brain tissue;[319] the down-regulated mouse Bin1-201 isoform encodes for the human *BIN1* isoform 1 (ENST00000316724.10, 87.2% homology) similarly down-regulated in AD brain, whereas the up-regulated mouse Bin1-205 isoform corresponds to the up-regulated human *BIN1* isoform 9 (ENST00000409400.1, 88% homology). Finally, despite these striking expression alterations at the isoform level, there was no significant gene expression difference between WT and TG mice, highlighting the importance of performing isoform-based analyses (**Figure 6.59B,E**).

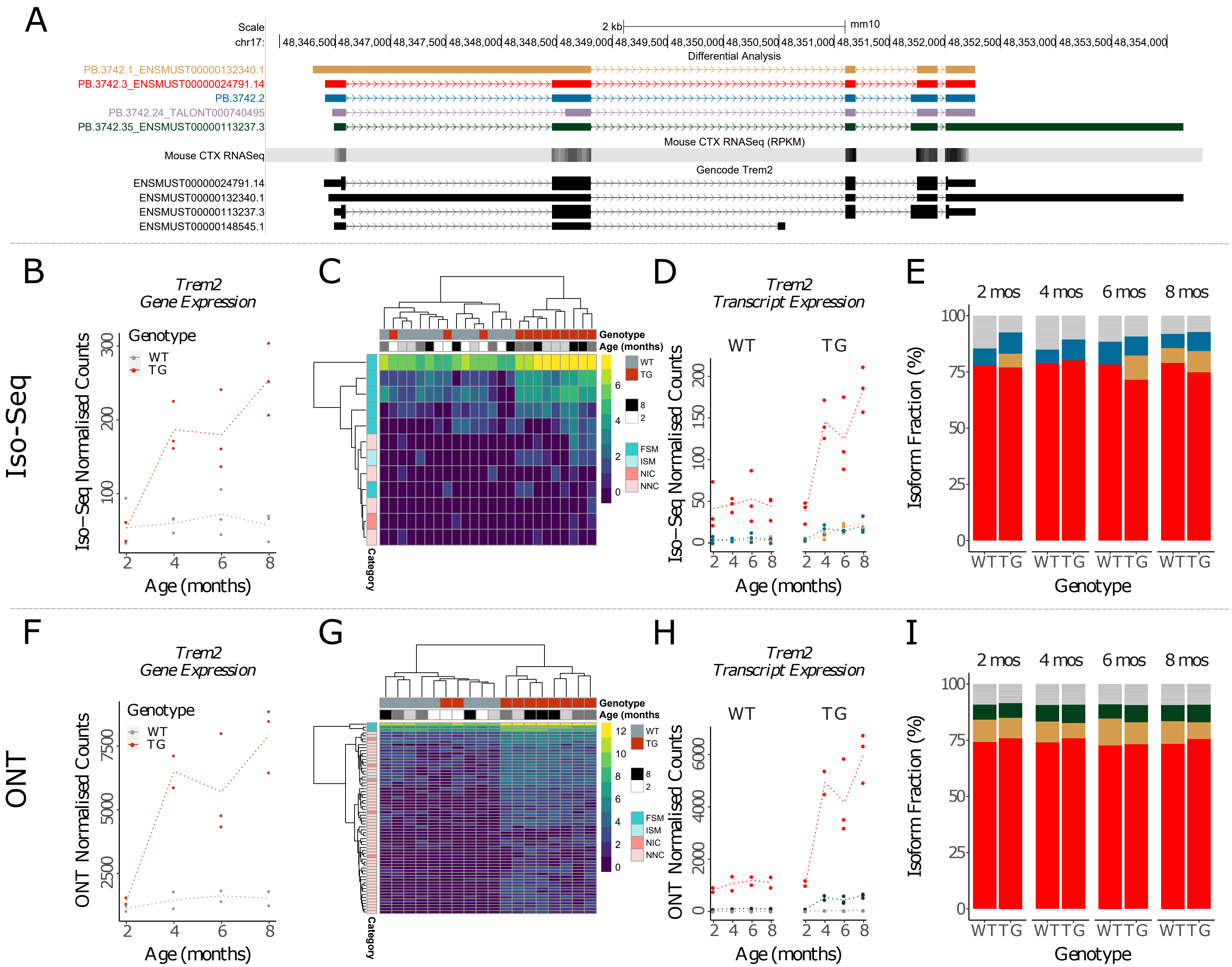


Figure 6.57: Global increase of *Trem2*-associated isoforms, particularly Trem2-201: Shown are plots generated from the differential expression and splicing analyses of *Trem2* in the rTg4510 cortex. *Legend continues on the following page.*

Figure 6.57: Shown are three panels relating to **(A)** UCSC genome browser track of the *Trem2*-associated isoforms of interest with the reference mouse annotations (mm10, GENCODE, vM22) and RNA-Seq data from matched samples, **(B - E)** differential expression analyses using normalised Iso-Seq full-length counts, and **(F - I)** normalised ONT full-length counts for quantification.

In detail, **(B)** and **(F)** are scatter plots of *Trem2* gene expression determined using normalised Iso-Seq and ONT counts, respectively. Red and grey dots refer to TG and WT samples, and dotted lines represent the mean paths across age.

(C) and **(G)** are heat-maps representing expression of all the *Trem2*-associated isoforms detected in the Iso-Seq and ONT targeted dataset, respectively. Each row refers to an isoform, labelled using *SQANTI* classification, and each column refers to a sample with the genotype and age provided.

(D) and **(H)** are scatter plots of the top three ranked differentially-expressed *Trem2* isoform using normalised Iso-Seq and ONT counts, respectively. The coloured isoforms correspond to those displayed on the UCSC genome browser track (Figure A).

(E) and **(I)** show the isoform proportion of *Trem2* by age and genotype using normalised counts from Iso-Seq and ONT counts, respectively. The coloured isoforms correspond to those displayed on the UCSC genome browser track (Figure A). Light grey bars refer to the fraction of novel isoforms that individually account < 5% of the total count.

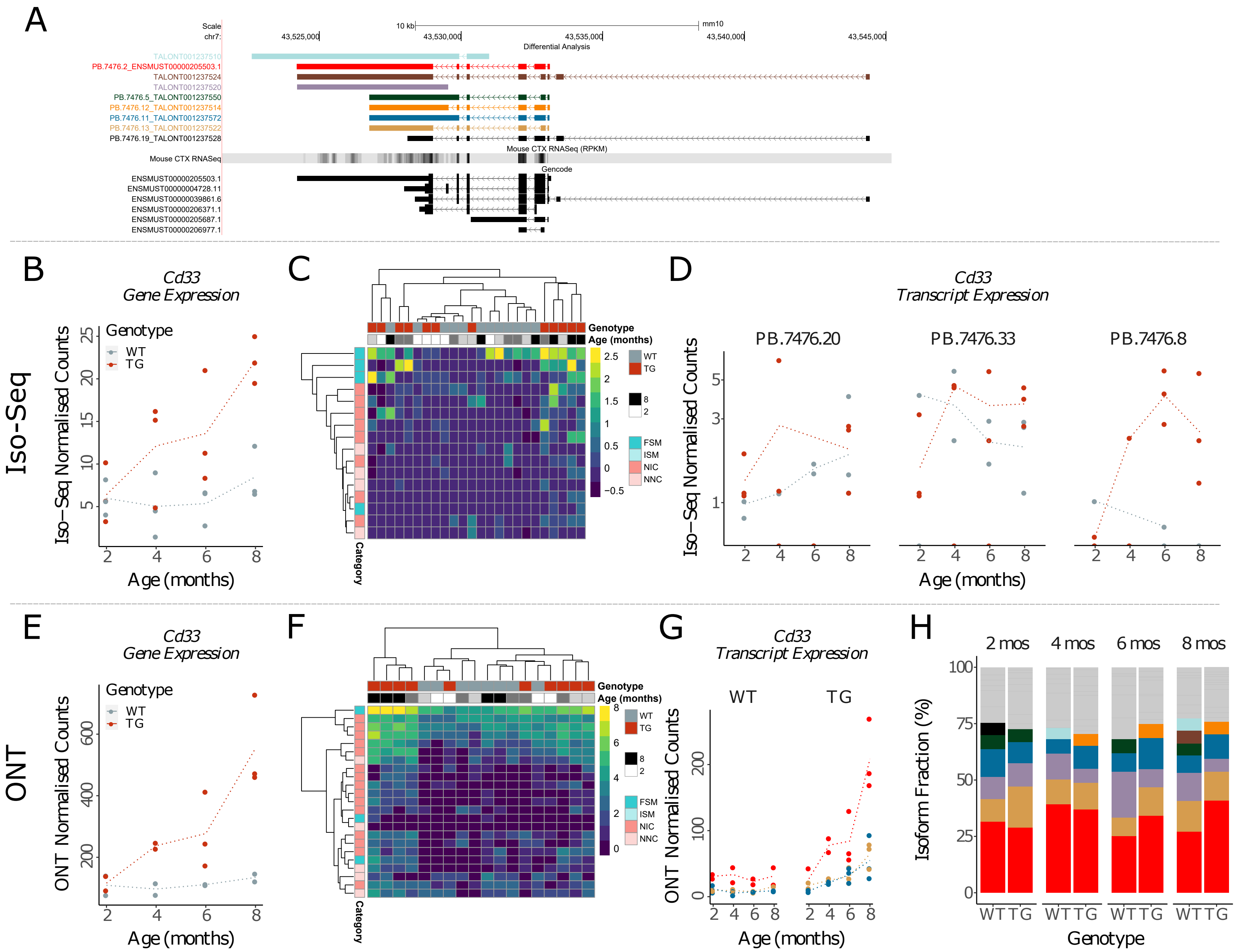


Figure 6.58: Drastic up-regulation of Cd33-203 accompanied with reduced usage of other isoforms: Shown are plots generated from the differential expression and splicing analyses of *Cd33* in the rTg4510 cortex. *Legend continues on the following page.*

Figure 6.58: Shown are three panels relating to **(A)** UCSC genome browser track of the *Cd33*-associated isoforms of interest with the reference mouse annotations (mm10, GENCODE, vM22) and RNA-Seq data from matched samples, **(B - D)** differential expression analyses using normalised Iso-Seq full-length counts, and **(E - H)** normalised ONT full-length counts for quantification.

In detail, **(B)** and **(E)** are scatter plots of gene expression determined using normalised Iso-Seq and ONT counts, respectively. Red and grey dots refer to TG and WT samples, and dotted lines represent the mean paths across age.

(C) and **(F)** are heat-maps representing expression of all the isoforms detected in the Iso-Seq and ONT targeted dataset, respectively. Each row refers to an isoform, labelled using *SQANTI* classification, and each column refers to a sample with the genotype and age provided.

Shown are scatter plots of **(D)** the top three most abundant isoforms detected using Iso-Seq, and **(G)** the top three most differentially-expressed isoform using normalised ONT counts, respectively. The coloured isoforms correspond to those displayed on the UCSC genome browser track (Figure A). No change in differential isoform expression was detected using Iso-Seq counts (Figure D), likely due to the relatively lower sequencing depth.

(H) show the isoform proportion of *Cd33* by age and genotype using normalised ONT counts. The coloured isoforms correspond to those displayed on the UCSC genome browser track (Figure A). Light grey bars refer to the fraction of novel isoforms that individually account < 5% of the total count. The respective plot for normalised Iso-Seq counts is not shown here, given Iso-Seq did not detect differential isoform expression changes.

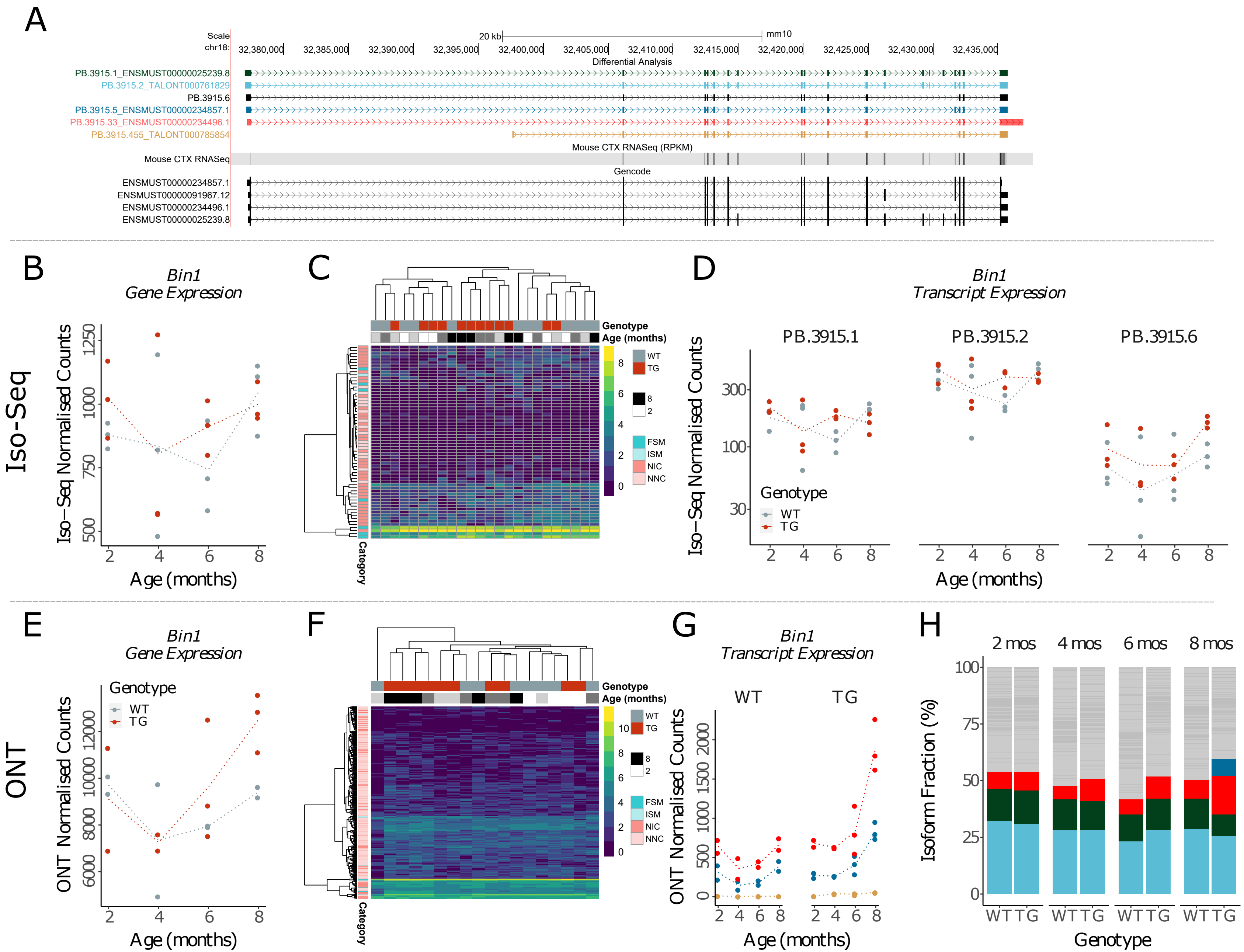


Figure 6.59: Gradual isoform switch in expression of *Bin1* known isoforms: Shown are plots generated from the differential expression and splicing analyses of *Bin1* in the rTg4510 cortex. *Refer to* ***Figure 6.58*** *for the same caption but with reference to Bin1.*

6.4 Discussion

In this chapter, we combined the advantages of long-read sequencing and target capture to map the transcriptional landscape of 20 AD-associated genes in the mouse cortex. To our knowledge, this is the first study to apply this approach at such scale, and the first to profile a transgenic mouse model, rTg4510, enabling us to comprehensively characterise the transcriptional variation of these AD-associated genes as a consequence of tau accumulation.

6.4.1 Overview of results

Using custom-designed biotinylated probes, we successfully enriched and sequenced full-length transcripts using PacBio Iso-Seq and ONT nanopore sequencing. We revealed unprecedented diversity of alternatively-spliced isoforms numbering in the thousands, detecting many more novel isoforms annotated to AD-associated genes than previously detected using global transcriptome profiling of the mouse cortex, highlighting the power of Capture-Seq for targeted sequencing. We subsequently developed an analysis pipeline with various custom scripts (available on GitHub) to handle and accurately document the complexity of these long-read-derived isoforms and extensive usage of alternative splicing events, which were used to ease visualisation of isoforms on the UCSC genome browser.

Comparison of the datasets generated using Iso-Seq and nanopore sequencing revealed striking differences inherent in the technology of the two long-read sequencing platforms; notably, PacBio Iso-Seq generated fewer but more accurate raw reads, whereas ONT generated significantly more reads but with lower accuracy. We observed that 85% of ONT reads were only detected once in one sample (shown in **Figure D.3**), suggesting that many of these are technical artefacts. Merging of these two datasets from sequencing the same samples across multiple long-read platforms thus improved our confidence of the isoform annotations generated in these experiments.

By comprehensively characterising the merged isoform landscape, our findings shed light on the complexity of transcriptional regulation and highlight the significant extent to which alternative splicing events contribute to isoform diversity in the cortex. Although we identified widespread usage of alternative 5' and 3' splice sites, there were notable gene-specific variation in the splicing events that dominated the isoform landscape. These included: i) extensive variation of the 3' UTR in isoforms annotated to *Apoe* and *Vgf*, ii) the consistent skipping of certain exons, which were often found constitutively expressed in mouse reference annotations, as illustrated in the *Mapt, Fyn, Snca* isoform landscape, iii) the occurrence of intron retention events localised to certain regions of the gene, which may encode a protein domain, and iv) the presence of various alternative first exons, as seen in *Clu, Fyn, Sorl1*, despite transcription being predicted to initiate from the canonical first exon downstream. Finally, our annotations provide further insights into the extreme precision of splicing, whereby a subtle change at a specific splice site generated a reading frame shift and the prediction of a truncated product destined for nonsense-mediated decay (particularly showcased in *Rhbdf2, Trpa1*).

The deep sequencing depth achieved using target enrichment allowed us to reliably determine transcript expression using normalised full-length read counts derived from long-read sequencing, forgoing the need of RNA-Seq data. As such, we identified widespread transcriptional variation associated with progressive tau pathology across the vast majority of AD-associated genes with evidence of altered splicing and transcript expression. Among these alterations, we identified robust tau-associated up-regulation of transcripts annotated to microglial-specific genes, *Trem2* and *Cd33.* Of note, a recent study has revealed crosstalk between CD33 and TREM2, suggesting that TREM2 acts downstream of CD33 in modulating microglial cell response to Aβ plaques.[323] A working model of this crosstalk has been proposed, implicating the role of altered *TREM2* and *CD33* splicing in the dysregulation of intracellular signalling pathways essential for phagocytosis.[65]

Finally, we detected robust differential expression changes in *Bin1* with evidence of differential isoform usage in aged rTg4510 transgenic mice. This is in agreement with a recent RNA-Seq study in human AD post-mortem brain tissue.[118] We show that the differentially expressed isoforms primarily differed by the presence/absence of exons encoding the clathrin-binding domain (CLAP) domain, which is involved in endocytosis and is also highly variable among human-equivalent *BIN1* isoforms.[319] Our results thus reflect altered exon splicing as a potential mechanism contributing to the role of *Bin1* in tau pathology, as observed in other studies of human AD brains.[118,319]

6.4.2 Limitations

Our results should be interpreted in the context of several limitations. Firstly, in following the official Iso-Seq protocol, we did not perform 5’-cap selection. Although our cDNA synthesis kit preferentially enriched for full-length cDNA sequences and stringent filtering was performed as part of our bioinformatics pipeline, we cannot exclude the possibility that some of the shorter isoforms may be a reflection of 5' degradation. While sequencing the library using two separate long-read sequencing platforms allowed us to validate our isoform annotations and reduce the number of these artefacts, this caveat becomes particularly apparent when we still detect isoforms that only differ at the transcription start site. Of note, we did trial a protocol (detailed in **Appendix C**) borrowed from the Wellcome Trust Advanced Course that I attended during my PhD; however, we were unable to generate sufficient material for library preparation. Moving forward for new long-read sequencing studies, we should optimise and integrate some of the 5'-cap protocols recently released[382, 383] into the lab workflow to guarantee the generation of full-length transcripts.

Secondly, PCR amplification was required following target enrichment. Our approach, as with most target enrichment methods, was thus constrained by the length of cDNA inserts used for library preparation; it becomes more challenging to amplify longer transcripts, particularly those sized above 5 - 6kb.[212] Evaluation of both Iso-Seq and ONT targeted datasets, however, showed that we detected transcripts up to 10kb, with detection of many long isoforms (> 8kb) known to span the full-length of the gene. Nonetheless, we acknowledge that there is an inherent length bias in preferentially sequencing the shorter transcripts. These challenges, however, could be addressed by a novel method recently introduced by Oxford Nanopore Technologies, "ReadUntil". This method allows nanopore devices to selectively eject reads from nanopores in real time (based on the identity of the initial set of sequenced bases), allowing rapid enrichment of targeted regions through a purely computational approach and eliminating the need for customised target-specific sample preparation.[384] Without constraints to the number and size of regions that can be simultaneously targeted, recent studies have illustrated the capacity of this method to enrich thousands of disease-specific genes with accurate detection of single-nucleotide polymorphisms and methylation.[385, 386]

Finally, we observed a relatively high off-target rate and inter- and intra-batch variability, despite increasing the number of samples sequenced per run and conducting sample randomisation to ensure equal representation. The assessment of off-target reads indicated that we had reached saturation of our target genes with off-target sequencing of other abundantly-expressed transcripts. We could have therefore included significantly more samples per run, thereby reducing the number of batches and batch variability. Moving forward for new long-read targeted sequencing studies, a power calculation should be performed, with consideration of the number and expression of target genes, in order to maximise throughput of relevant reads per sequencing run.

6.4.3 Conclusion

In summary, our study revealed unprecedented diversity of alternatively-spliced isoforms annotated to AD-associated genes. We identified robust transcriptional and splicing differences in these AD-risk genes paralleling the development of tau pathology. Among these changes, we found global up-regulation of *Trem2*-associated isoforms and isoform switches in *Bin1* and *Cd33*, further supporting a role for the dysregulation of the immune response in the development of AD pathology. Altogether, our findings demonstrate the utility of performing targeted long-read sequencing to enable comprehensive characterisation of the AD transcriptomic landscape and accelerate the discovery of meaningful alterations in the AD brain.

Chapter 7

General Discussion

This chapter concludes my thesis by summarising the key findings and implications of our results in light of the existing literature on transcriptional variation in AD and the current status of long-read sequencing approaches for transcriptome profiling. This discussion will also summarise some of the key limitations and caveats that should be considered when interpreting our results. Finally, this chapter concludes by giving an outline of the future directions of the research presented in this thesis.

7.1 Key findings

We hypothesised that transcriptional regulation, particularly alternative splicing, is dysregulated in the development of AD pathology. However, previous studies investigating splicing and transcript-level expression variations have been constrained by inherent limitations of short-read RNA-Seq approaches. The primary aim of this thesis was to address these limitations and leverage the power of novel long-read sequencing technologies to accurately characterise the transcriptomic landscape in a mouse model of AD, with a particular focus on identifying splicing patterns associated with the progression of tau pathology.

In meeting the objectives set out in **Chapter 1**, we:

- optimised a novel laboratory workflow and bioinformatics pipeline in **Chapter 3** to profile full-length transcripts using two state-of-the-art, long-read sequencing approaches: PacBio isoform sequencing (Iso-Seq) and ONT nanopore cDNA sequencing.

- characterised the global transcriptome landscape of the mouse cortex using Iso-Seq in **Chapter 4**, revealing widespread cortical isoform diversity and extensive usage of alternative splicing events.
- identified transcriptional and splicing alterations associated with the progression of tau pathology in rTg4510 mice in **Chapter 5**, finding evidence for profound isoform switching events in genes previously implicated in AD.
- performed long-read targeted sequencing of 20 AD-risk genes in **Chapter 6**, identifying robust transcript expression differences in microglial-specific genes associated with tau accumulation in rTg4510 mice.

7.2 Implications and limitations

7.2.0.1 Incomplete reference annotations with detection of novel isoforms

One of the common overarching themes in our long-read sequencing analyses - and those recently published by others - is the extensive detection of novel isoforms not present in current reference annotations, highlighting the constraints of previous transcriptome studies to accurately study the regulation of alternative splicing. Global transcriptome profiling of the mouse cortex (**Chapter 4**) detected over 20,000 novel isoforms, with evidence of novel splice junctions and splicing events. The power to discover more novel and rare isoforms with long-read sequencing is highlighted in our target enrichment of AD-associated genes (**Chapter 6**), with over 2,000 novel isoforms annotated to *Apoe* alone. Given that genetic and transcriptomic studies are fundamentally reliant on accurate gene annotations, this "incompleteness" of existing annotations limits our understanding of the role of transcriptional variation in complex diseases. Of note, a recent study found that genes associated with neurodegenerative disorders, *SNCA, APOE* and *CLU* (also included in the target gene list), were significantly under-represented in the human reference annotations after leveraging transcriptome data from the GTEx consortium.[387]

I envisage that the number of novel isoforms detected will continue to grow with advances in sequencing technologies. However, one of the major challenges in long-read sequencing is how to best assess the validity and quality of these isoforms. Despite the capacity to enrich for full-length transcripts, the standard long-read sequencing protocol is still reliant on cDNA synthesis and PCR amplification, creating artefacts that can be misinterpreted as novel isoforms generated from non-canonical splicing. Although we have optimised the number of PCR cycles, random amplification errors (i.e. PCR errors) and unequal amplification (PCR bias, due to intrinsic differences in amplification efficiency of templates with challenges from amplifying > 10kb transcripts) are likely to still exist. Future experiments should consider the addition of unique molecular identifiers (UMIs) to the sequencing library before amplification to provide error correction and enable the accurate identification of PCR duplicates.[388]

RNA degradation can also result in isoforms with incomplete 5' ends resulting in misinterpretation of these artefacts as novel isoforms with novel alternative initiation start sites.[382] Of note, we found that more than 98% of our ONT novel transcript were covered with less than five reads across two samples. While we are confident about the validity of our transcriptome annotations - since we undertook stringent filtering and cross-validated isoforms using two independent long-read sequencing approaches - it will be important to undertake further experimental validation and integrate with other functional data.

7.2.0.2 Functional importance of RNA isoforms on proteome diversity

As we (and others) have shown, the high-throughput sequencing of full-length transcripts highlights the widespread isoform diversification through alternative splicing of the transcriptome. However, large-scale proteomic studies have failed to recapitulate this isoform diversity at the protein level, sparking fierce debate about the impact of alternative splicing on protein production and function.[389–391] This disparity is likely to be driven by the poor sensitivity and incompleteness of current protein reference databases, limiting its utility for isoform-based analyses.[392] Notably, recent studies using "long-read proteogenomics" - an integrative approach that incorporates ORF annotations derived from long-read sequencing data with peptides detected using mass-spectrometry (MS) - revealed hundreds of protein isoforms undetectable using traditional MS.[393,394] This approach further allowed identification of novel peptides that corresponded to novel exons and splicing events derived from improved long-read annotations.[?,393]

We show that the isoform landscape for the majority of AD-risk genes characterised in this thesis were dominated by a few major isoforms (**Chapter 6**). This raises the question of how functionally important these alternatively-spliced novel isoforms are, particularly when they only differ slightly from the major isoforms. However, we know that minor changes in the open reading frame, while not inducing large protein conformational changes, can disrupt post-translational modifications and impact protein function.[392] This is supported by identification of *TREM2* variants known to impact ligand binding and modulate downstream signalling pathways, while broadly maintaining protein structure and stability.[356] Furthermore changes in the 5' and 3' UTR, while resulting in no observable effect on the protein product, can affect transcript stability, export, localisation and translation efficiency.[392] Finally, it is possible that the additive effects resulting from the expression of multiple low-abundant, but mis-spliced isoforms could have deleterious impacts globally, by: i) altering the ratio of canonical isoforms, ii) overwhelming the ubiquitin-proteasome pathway with accumulation of aberrant polypeptides, and iii) generating insoluble protein aggregates, which largely underlie the neuropathology of AD and other neurodegenerative diseases.[395]

7.2.0.3 Understanding the development of tau pathology through transcriptome profiling

In providing a "snapshot" of the cellular state, transcriptome profiling can provide key insights into the mechanisms underlying disease changes, enabling the elucidation of key pathogenic pathways. My experiments also included the profiling of the cortex across multiple time points (2, 4, 6 and 8 months) spanning the development of tau pathology in the rTg4510 mouse model (**Chapter 5, 6**), further allowing us to evaluate temporal transcriptional changes.

In line with findings from RNA-Seq studies using other mouse models, we found widespread transcriptional differences associated with tau accumulation (**Chapter 5**). Furthermore, our long-read sequencing datasets were powered to identify robust differences in isoform expression, which drove gene expression alterations that were well established in the development of human AD (**Chapter 5, 6**). Highlighting the utility of isoform-level analyses, we also identified a number of genes that presented differential transcript usage and isoform switches without overall gene-level expression differences, particularly in the later stages of tau pathology. Characterisation of these alternatively-spliced isoforms suggest these alterations could have functional biological consequences, although more work would be needed to confirm this.

However, such studies are inherently limited in its ability to dissect the cause-and-effect between transcriptional alterations and pathology. While rTg4510 transgenic mice develop tau pathology as a consequence of the overexpression of the human *MAPT* transgene, the expression and splicing changes identified could be a consequence rather than a cause of tau pathology. Additional functional cellular studies will be required to establish the cause-and-effect relationship and the role of splicing in driving pathological changes.

7.2.0.4 Cellular heterogeneity in isoform diversity in the brain

Alternative splicing is known to define tissue-specificity, with more than a third of human genes found to express tissue-dominant isoforms characterised by tissue-specific splicing events.[93, 95] Recent studies using single-cell sequencing further support the evidence for cell-type-specific splicing, particularly in the brain where it plays an important role in neuronal development and maintenance.

However, the vast majority of AD transcriptome profiling studies, including the studies presented in this thesis, have been limited to analyses on "bulk" tissue comprised of a complex mix of different cell-types. We were therefore unable to draw conclusions about cell-specific splicing variations, despite recent studies reporting differential expression of cell-specific *BIN1* isoforms in the AD brain.[319] Furthermore, given AD pathogenesis is characterised by progressive changes in cell composition that vary across different brain regions, it becomes more challenging to disentangle tissue and cell-specific AD-associated variants. Having not accounted for cellular heterogeneity in our studies, we acknowledge that our results may be a partial reflection of microgliosis and neuronal loss. While this challenge can be addressed using a combined approach of single-cell and long-read sequencing, as shown in recent studies (reviewed in **Table 1.6**), this strategy is currently limited to achieve the depth required to detect reliable disease- and cell-specific splicing variations.

7.2.0.5 Translational relevance of AD mouse models

One of the key challenges of studying transcriptomic variation in human post-mortem brain tissue is that they typically represent the end-stage of the disease. This subsequently limits the power to detect progressive disease-associated variations and the ability to infer causality. Although mouse models act as valuable reductionist tools to dissect the mechanisms that drive the onset and progression of AD pathology, there are some concerns about how representative they are of sporadic LOAD in human. Future work will be essential to perform cross-species analyses and translate our findings to human.

Notably, the rTg4510 model does not develop amyloid pathology which is a key hallmark of AD, implicating that this model is better suited to study disorders characterised solely by tau pathology, e.g. FTD. However, it is also important to note that there is currently no AD mouse model that encapsulates all the defining features of AD. To date, no causative *MAPT* mutations have been identified in AD despite the fact that tau pathology correlates better with disease progression and cognitive decline than amyloid plaques.[51–53] Consequently, the rTg4510 model remains a good model to investigate the mechanisms associated with the spread of NFTs, which closely parallels the progression of the Braak stages in human AD brains. Furthermore, by profiling this mouse model at multiple ages selected to encompass the development of pathology, my studies were well powered to identify transcriptomic variation associated with both genotype and the progression of tau pathology. Of note, we identified transcriptional differences that broadly overlapped with previous studies of the human AD brain transcriptome - including *Gfap* upregulation and *Bin1* differential isoform expression

However, one potential limitation of the rTg4510 model is the integration of the calcium-calmodulin kinase IIa promoter (CaMKIIα-tTA) and the human *MAPT* transgene in the rTg4510 mouse model. This has been found to disrupt the expression of five endogenous mouse genes (*Vipr2*, *Wdr60*, *Esyt2*, *Ncapg2* and *Ptprn2*), which may contribute to the neurodegenerative phenotype observed in these mice.[85,90] In line with previous work from our group,[90] I also used wild-type mice with no *MAPT* transgene insertion as controls. While this allows us to ascertain the genotype of the mice sequenced, some of our findings may be a reflection of the transgene insertion rather than tau pathology. We could have alternatively used doxycycline(DOX)-treated rTg4510 mice as controls, which carry but do not express the *MAPT* transgene, thereby eliminating any potential confounders inherent in transgenic mice.

7.3 Future directions

In this thesis, I have presented an optimised laboratory and bioinformatics pipeline, a set of empirical findings, and a comprehensive resource, which serve to deepen our understanding of the cortical isoform diversity and its role in the development of AD pathology. However, as discussed above, the findings presented are not without limitations inherent to the research question and methods chosen. As such, the following section details a number of future directions proposed to address these limitations and foster additional research in this area.

7.3.0.1 Integration with other datasets

As described in **Section 1.2.1**, alternative splicing is highly regulated in a concerted manner and require multiple mechanistic components, including epigenetic modifications (such as DNA methylation and histone modifications). Epigenetics refers to the heritable, but reversible, alterations to the chromatin structure that have been shown to influence splicing through the recruitment of splicing factors.[396,397,397–399] A number of studies, for example, have shown that gene body DNA methylation can influence polymerase processivity and elongation rate, which can in turn determine the splicing of alternative exons with weak splice sites.[396,397] Integrating our transcriptomic dataset with epigenetic data (DNA methylation) generated on the same mouse samples (Castanho et al., unpublished data, 2022) would be important to further understand the mechanisms driving transcriptional regulation in AD development and provide insights into the cross-talk between splicing and epigenetics. Notably, preliminary analyses revealed a differentially-methylated cytosine that coincided with an exon skipping event in a *Bin1*-associated isoform that was upregulated in rTg4510 TG mice, demonstrating the power of multiomics.

One of the key limitations of our analyses is that they were performed on "bulk" tissue, limiting the detection of cell-specific isoforms and meaning that some of our results could be confounded by cellular heterogeneity. Moving forward, mouse single-cell data derived from isolating specific cell populations - using methods such as fluorescence-activated cell sorting (FACS) (Policicchio et al., unpublished data, 2022) - or publicly available mouse single cell RNA-Seq datasets could be used to infer cell populations in our dataset with the potential to assess cell-type specific splicing events.

7.3.0.2 Experimental and functional validation

Although we used ONT nanopore sequencing and short-read RNA-Seq to validate and complement our Iso-Seq data with notable success (**Chapter 6**), further experimental validation of the novel isoforms and empirical findings presented in this thesis is important. This could be achieved using a number of molecular biology techniques, such as i) RT-qPCR with primers that flank the alternatively-spliced region, ii) western blot and enzyme-linked immunosorbent assays (ELISA) to identify protein isoforms, and iii) fluorescence *in-situ* hybridisation (FISH) and immunohistochemistry techniques for isoform visualisation at the RNA and protein level, respectively. Of note, all these methods require specific primers and antibodies that are unique to the isoform of interest, which may be challenging.

An alternative method is to perform mass-spectrometry on the same samples, and utilise an integrative proteogenomics approach to validate novel splicing events (as described in **Section 7.2.0.2**). It is important to note, however, that current mass-spectrometry methods are limited in discriminating different isoforms given that current protocols first require the proteins to be digested and fragmented.

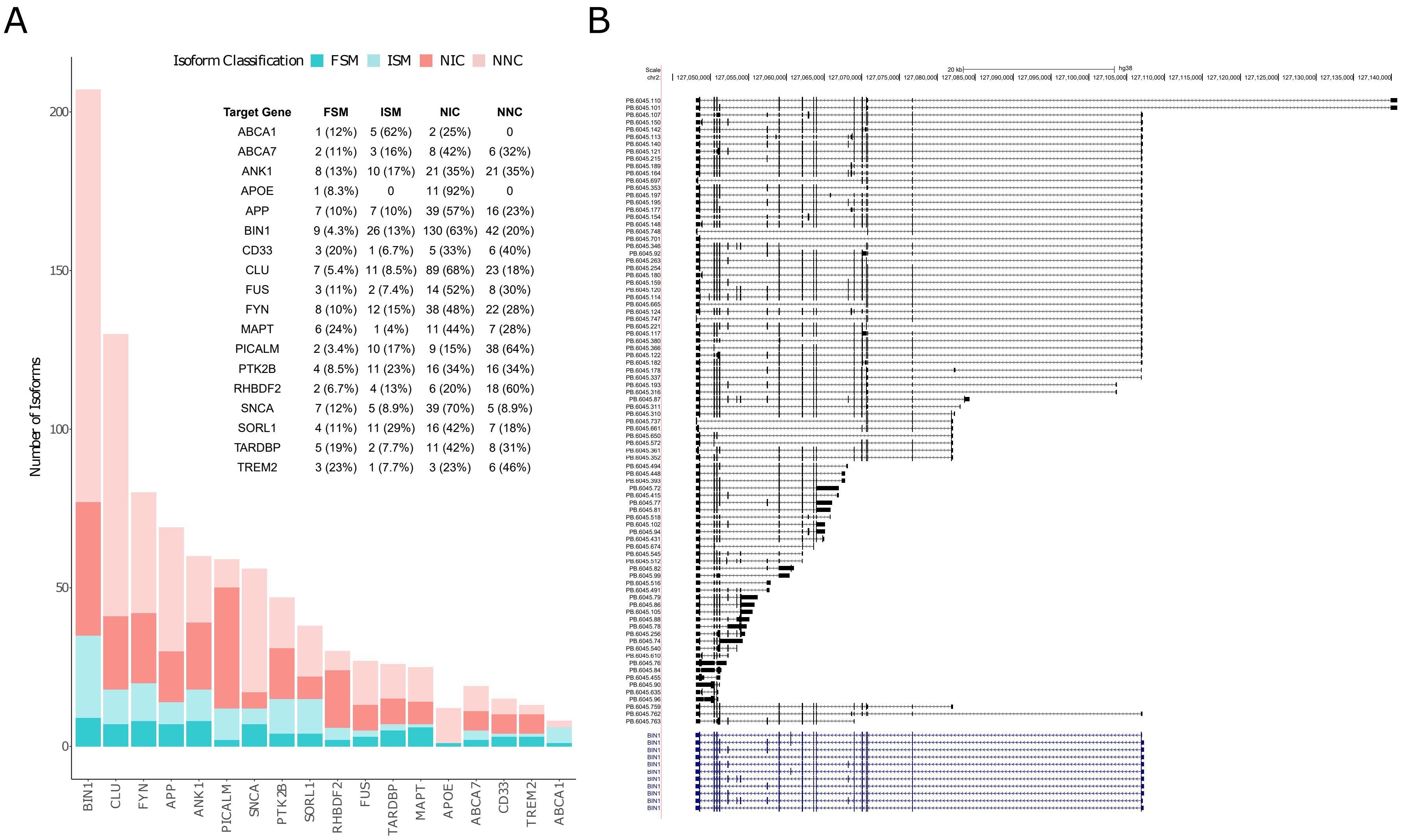
Finally, the functional consequences of alternative splicing events can be explored using various assays, such as the CRISPR-Cas9 base-editing system and minigene splicing reporters among others, to accurately and rapidly recapitulate splicing effects in cultured cells.

7.3.0.3 Cross validation with other mouse models

Given the limitations of the rTg4510 model (as discussed in **Section 7.2.0.5**), future work should consider the transcriptome profiling of other mouse models with amyloid or tau pathology, or cross-validating our findings. Of note, recent advances have been made using CRISPR-Cas9 to develop new AD mouse models that exhibit a more accurate disease phenotype. Examples of these models include: i) App knock-in mouse models that carry a humanized Aβ sequence and clinical AD mutations,[400,401] and ii) a tau knockout strain generated by CRISPR-Cas9-mediated genome editing of *Mapt*, resulting in a model with no overt phenotypes but is resistant to excitotoxicity.[402] By introducing subtle disease-causing mutations, these models develop more accurate pathology without potential artefacts introduced from transgene insertion and/or overexpression.

7.3.0.4 Transcriptome profiling of human post-mortem brain tissues

The ultimate goal of animal models is to translate and recapitulate findings to humans. Consequently, the findings presented in this thesis provide the foundation for a broader follow-up study in human post-mortem brain tissue. While this was beyond the scope of my written thesis, as part of this research, we performed targeted Iso-Seq of 20 AD risk-genes in a subset of prefrontal cortex samples from the Brains for Dementia Research (BDR) cohort (n = 15 controls, 15 AD cases). In applying the same sequencing approach and bioinformatics pipeline described in **Chapter 6**, preliminary analyses show widespread detection of novel isoforms annotated to AD-risk genes (**Figure 7.1A**) characterised by extensive usage of alternative splicing events (**Figure 7.1B**). Following a more detailed mapping of this isoform landscape, future work will involve a comprehensive case-control analyses at the transcript-level and a comparison of these findings with those identified in the tau mouse model. Finally, we have applied other profiling approaches - such as genomic profiling (ATAC sequencing), epigenetics (DNA and histone modifications), transcriptomics (RNA-Seq, miRNA) and proteomics (mass-spectrometry) - on the same samples. This will allow us to perform a comprehensive integrated multiomics approach to gain deeper mechanistic insights into the development of AD pathogenesis.



Target Gene	FSM	ISM	NIC	NNC
ABCA1	1 (12%)	5 (62%)	2 (25%)	0
ABCA7	2 (11%)	3 (16%)	8 (42%)	6 (32%)
ANK1	8 (13%)	10 (17%)	21 (35%)	21 (35%)
APOE	1 (8.3%)	0	11 (92%)	0
APP	7 (10%)	7 (10%)	39 (57%)	16 (23%)
BIN1	9 (4.3%)	26 (13%)	130 (63%)	42 (20%)
CD33	3 (20%)	1 (6.7%)	5 (33%)	6 (40%)
CLU	7 (5.4%)	11 (8.5%)	89 (68%)	23 (18%)
FUS	3 (11%)	2 (7.4%)	14 (52%)	8 (30%)
FYN	8 (10%)	12 (15%)	38 (48%)	22 (28%)
MAPT	6 (24%)	1 (4%)	11 (44%)	7 (28%)
PICALM	2 (3.4%)	10 (17%)	9 (15%)	38 (64%)
PTK2B	4 (8.5%)	11 (23%)	16 (34%)	16 (34%)
RHBDF2	2 (6.7%)	4 (13%)	6 (20%)	18 (60%)
SNCA	7 (12%)	5 (8.9%)	39 (70%)	5 (8.9%)
SORL1	4 (11%)	11 (29%)	16 (42%)	7 (18%)
TARDBP	5 (19%)	2 (7.7%)	11 (42%)	8 (31%)
TREM2	3 (23%)	1 (7.7%)	3 (23%)	6 (46%)

Figure 7.1: Widespread detection of novel isoforms annotated to AD-risk genes from targeted profiling of AD human post-mortem brain tissue. Shown are some preliminary results, showing **(A)** the number of known (FSM, ISM) and novel isoforms (NIC, NNC) annotated to AD-risk genes, classified using *SQANTI* (detailed in **Section 3.1.4.4**), and **(B)** a UCSC genome browser track of a subset of isoforms annotated to *BIN1* after Iso-Seq targeted profiling of AD human post-mortem brain tissue. We applied the same Iso-Seq targeted laboratory workflow and bioinformatics pipeline that was optimised for the mouse cortex in **Chapter 6**. Iso-Seq-derived isoform annotations and human reference annotations (hg38, GENCODE) are in black and blue, respectively.

7.4 Conclusion

Taken altogether, my thesis harnessed the power of long-read sequencing to extend our understanding of the mouse cortical transcriptome and assess splicing variation associated with the development of tau pathology in a transgenic mouse model. To our knowledge, it is the first study to profile a mouse model of AD pathology, rTg4510, using both PacBio Iso-Seq and ONT nanopore sequencing. In providing a framework for the comprehensive characterisation of the AD isoform landscape at a global and targeted level, we revealed unprecedented diversity of alternatively-spliced isoforms annotated to AD-risk genes. We identified widespread transcriptional and splicing variations paralleling the development of tau pathology, with evidence of differential isoform expression and usage. Our findings corroborated data from previous studies implicating a role for altered splicing and immune response in the development of AD pathology. The data presented in this thesis provides a strong foundation for characterising the transcriptomic landscape in the human AD brain and represents a valuable resource to the research community.

Appendices

Appendix A

Integrated Iso-Seq protocol

A.1 Requirement of sample quality

The following sample conditions are important to ensure high quality sequencing library:

- Double-stranded DNA generated from cDNA synthesis of extracted RNA with preferable RIN > 8
- Minimum freeze-thaw cycles
- No exposure to high temperature (> 65°C) or pH extremes (< 6, > 9),
- 1.8 - 2 OD 260/280, and 2.0 - 2.2 OD 260/230
- No insoluble material
- No RNA contamination or carry over contamination (e.g polysaccharides)
- No exposure to UV or intercalating fluorescent dyes
- No chelating agents, divalent metal cations, denaturants or detergents

A.2 General

The following sections are general steps that are applicable throughout the entire protocol.

A.2.0.1 AMPure bead purification

Throughout the protocol, DNA is purified using AMPure PacBio (PB) beads. Exact relative concentration of AMPure beads, sufficient amount of freshly-prepared ethanol, and not over-drying of beads are critical to remove adapters and dimers, and for high DNA recovery.

1. Prepare AMPure PB beads for use by allowing them to equilibrate to room temperature for a minimum of 15 minutes. Resuspend by vortexing.
2. After adding specified ratio of AMPure PB beads (ratio differs pending on the part of protocol), mix the bead/DNA solution thoroughly.
   - Ensure exact concentration is used particularly for 0.4X AMPure beads - higher concentration would result in retainment of undesired short inserts, whereas lower concentration would result in significant yield loss.
3. Quickly spin down the tubes (1 second) to collect beads.
4. Allow the DNA to bind to beads by shaking in a VWR vortex mixer at 2000rpm for 10 minutes at room temperature.
5. Spin down both tubes (1 second) to collect beads.
6. Place tubes in a magnetic bead rack, and wait until the beads collect to the side of tubes and solution appears clear (2 minutes).
   - The actual time required to collect the beads to the side depends on the volume of beads added.
7. With the tubes still on the magnetic bead rack, slowly pipette off cleared supernatant and save in other tubes. Avoid disturbing the bead pellet.
   - If the DNA is not recovered at the end of this procedure, equal volumes of AMPure PB beads can be added to the saved supernatant and repeat the AMPure PB bead purification steps to recover the DNA.
8. With the tubes still on the magnetic bead rack, wash beads with 1.5ml freshly prepared 70% ethanol by slowly dispensing it against the side of the tubes opposite the beads. Avoid disturbing the bead pellet.
   - Freshly-prepared 70% ethanol should be used for efficient washing, and should be stored in a tightly capped polypropylene tube for no more than 3 days.
   - Wash beads thoroughly by adding 70% ethanol to the rim of the tube, otherwise it can result in retention of short and adapter dimers.
9. Repeat step 3.
10. Remove residual 70% ethanol by taking tubes from magnetic bead rack and spin to

pellet beads. Place the tubes back on magnetic bead rack and pipette off any remaining 70% ethanol.

11. Repeat step 5 if there are remaining droplets in tubes.
12. Remove tubes from magnetic bead rack and allow beads to air-dry (with tube caps open) for 30 seconds.
    - Important to not over-dry pellet (over 60 seconds), as would otherwise result in low yield due to challenges to perform efficient sample elution.
13. Elute with specified amount of PacBio Elution Buffer (amount depends on which part of the protocol).
14. Tap tubes until beads are uniformly re-suspended. Do not pipette to mix.
15. Elute DNA by letting the mix stand at room temperature for 2 minutes.
16. Spin the tube down to pellet beads, then place the tube back on the magnetic bead rack. Let beads separate fully and transfer supernatant to a new 1.5ml Lo-Bind tube. Avoid disturbing beads.

A.2.0.2 Assessment of DNA quantity using Qubit

Accurate quantification of DNA using Qubit where stated is essential for accurate binding reaction conditions, and to avoid overloading/under-loading, which would otherwise result in high P2 (off polymerase-to-template ratio) and low sequencing yield.

As part of QC across the various stages of library preparation, quantify DNA using Qubit dsDNA High Sensitivity Assay Kit (ThermoFisher Scientific), following manufacturer's instructions.

1. Set up and label the required number of Qubit assay tubes (0.5mL) for samples and 2 standards.
    - Do not label the side of the tubes as this can interfere with sample readout.
2. Prepare the Qubit working solution by diluting Qubit dsDNA HS Reagent in Qubit dsDNA HS Buffer of a ratio 1:200, and mix well.
3. Add 190μL of Qubit working solution to tubes designated for standards, and 10μL of Qubit working solution to tubes designated for samples.
4. Add 10μL of each standard and 190μL of respective samples to the appropriate labelled tubes, totalling to a final volume of 200μL per tube.
5. Mix all Qubit assay tubes well by vortexing for 2 - 3 seconds, and incubate at room

temperature for 2 minutes.

6. Run the standards and samples on the Qubit 3.0 Fluorometer, using the dsDNA High Sensitivity option, and account for dilution factor to determine final concentration.

A.2.0.3 Assessment of DNA library size using TapeStation or Bioanalyzer

Also as part of QC across the various stages of library preparation in parallel with performing the Qubit assay, run DNA using D5000 ScreenTape or DNA 12000 Assay (Agilent), following manufacturer's instructions.

D5000 ScreenTape on 2200 TapeStation

1. Allow reagents to equilibrate at room temperature for minimum 30 minutes, and vortex.
2. Prepare samples by mixing 5μL of D5000 Sample Buffer and 1μL of respective sample.
3. Prepare ladder by mixing 1μL of D5000 Sample Buffer and 1μL of D5000 ladder.
    - Note: While electronic ladder is not available on the D5000 assay, it is not absolute necessary to run the ladder, particularly if only checking for intact library distribution size.
4. Vortex at 2000rpm for 1 minute and briefly spin down.
5. Load and run samples on D5000 ScreenTape using 2200 TapeStation instrument.

DNA 12000 Assay on 2100 Bioanalyzer

1. Set up the chip priming station and the Bioanalyzer 2100, decontaminating the electrodes with water.
2. Allow reagents to equilibrate at room temperature for minimum 30 minutes.
3. Prepare and load the gel-dye matrix into the appropriate wells of the chip.
4. Pipette 5μL of marker into the ladder and 12 sample wells.
5. Pipette 1μL of ladder into the appropriate well, and 1μL of sample or water in respective 12 sample wells.
6. Vortex chip for 60 seconds at 2400rpm and insert into the 2100 Bioanalyzer.

A.3 First-strand cDNA synthesis

1. For each sample, add 200ng of RNA with 1μL of barcoded/non-barcoded poly(T) primer in a micro centrifuge on ice (**Table A.1**), mix and spin briefly.
2. Incubate tubes at 72°C in a 105°C hot-lid thermal cycler for 3 minutes, slowly ramp to 42°C at 0.1°C/sec, then let sit for 2 minutes.
3. During incubation, prepare PCR reaction mix by combining the following reagents in **Table A.1** in the order shown. Scale reagent volumes accordingly to the number of samples prepared.
   - Important: Only add reverse transcriptase to the master mix just prior to step 4, and go immediately into step 5.
4. Within the last 1 minute of RNA reaction tubes sitting at 42°C, incubate PCR reaction mix at 42°C for 1 minute and proceed immediately to step 5.
5. Aliquot 5.5μL of PCR reaction mix into each RNA reaction tube. Mix tubes by tapping and spin briefly
6. Incubate tubes at 42°C for 90 minutes, followed by 70°C for 10 minutes.
7. Add 90μL of PacBio Elution Buffer (EB) to each RNA reaction tubes.

Table A.1: Reagent composition for SMARTer cDNA synthesis

Reagents	Volume (μL)
Total RNA (200ng)	X
3'SMART CDS Primer IIA (12μM)	1
Nuclease-free water	10 - X
5X First-Strand Buffer	2
DTT (100mM)	0.25
dNTP (10mM)	1
SMARTer IIA Oligonucleotide (10mM)	1
RNase Inhibitor	0.25
SMARTScribe RT (100 U)	1
Total volume per sample	*10*

A.4 PCR cycle optimisation

1. Prepare a PCR reaction mix (**Table A.2**), scale up accordingly by the number of samples.
2. Aliquot 45μL of PCR reaction mix to a micro centrifuge for each sample.
3. Add 5μL of respective diluted cDNA from first-strand synthesis, mix and spin down.
4. Cycle the reaction with the conditions outlined in **Table A.3** using 105°C heated lid.
   - At cycles 10, 12, 14, 16 and 18, take 5μL from reaction tubes and transfer to new microcentrifuge tube.
   - Flick and spin down reaction tubes, before returning them back to thermocycler to continue for incubation.
5. Run 5μL of cDNA from each sample and cycle on a 1% agarose gel (**Section A.5**) at 110V for 20 minutes with 1μL 100bp ladder.
   - Note: input of 5μL of cDNA rather than 10μL, as stated in protocol, otherwise insufficient amount of diluted cDNA to proceed with both PCR cycle optimisation and PCR large scale amplification.
6. Determine the number of optimum PCR cycles to generate a sufficient amount of ds-cDNA without the risk of over-amplification (**Section 3.1.2.1**)

Table A.2: Reagent composition for PCR cycle optimisation

Reagents	Volume (μL)
5X PrimeSTAR GXL buffer	10
dNTP Mix (2.5mM each)	4
5’ PCR Primer IIA (12/μM)	1
Nuclease-free water	29
PrimeSTAR GXL DNA Polymerase (1.25U/μL)	1
Total volume per sample	*45*

A.5 Running an agarose gel

1. Weigh 1.5mg of agarose and place into a beaker containing 100ml 1X TBE buffer.
2. Microwave beaker for 10 - 20 seconds until the solution appears clear and allow to cool for 2 -3 minutes.
3. Prepare the casket with insert of comb (ensure the number of wells > number of samples).

Table A.3: PCR conditions for PCR cycle optimisation

Segments	Temperature (°C)	Time	Cycles
1	98	30 seconds	1
2	98	10 seconds	10
	65	15 seconds	
	68	10 minutes	
	68	5 minutes	1
3	98	10 seconds	2
	65	15 seconds	
	68	10 minutes	
	68	5 minutes	1
4	Take 5μL, and repeat step 3 for a total of 20 cycles		

4. Add 1.75μL of ethidium bromide into the beaker, and pour agarose solution into the casket.
5. Cool gel for ~20 minutes.

A.6 Large-scale PCR

1. Set up and label 16 microcentrifuge tubes for each sample.
2. Prepare a PCR reaction mix for each sample in 1.5mL LoBind Eppendorf (**Table A.4**).
3. Add 50μL of respective diluted cDNA to each PCR reaction mix.
    - Note: input of 50μL of cDNA rather than 100μL, as stated in protocol, otherwise insufficient amount of diluted cDNA to proceed.
4. Mix and briefly spin down.
5. Aliquot 50μL of PCR reaction mix (now 800μL) into 16 micro-centrifuge tubes.
6. Cycle the reaction with the conditions outlined in **Table A.5**.

Table A.4: Reagent composition for large-scale PCR

Reagents	Volume (μL)
5X PrimeSTAR GXL buffer	160
dNTP Mix (2.5mM each)	64
5' PCR Primer IIA (12μM)	16
Nuclease-free water	464
PrimeSTAR GXL DNA Polymerase (1.25U/μL)	16
Total volume per sample for 16 PCR reactions	*750*

Table A.5: PCR conditions for large-scale PCR

Segments	Temperature(°C)	Time	Cycles
1	98	30 seconds	1
2	98	10 seconds	
	65	15 seconds	N cycles
	68	10 minutes	
3	68	5 minutes	1

A.7 Bead purification of large-scale PCR products

A.7.1 Fraction 1 and 2: 1st purification

1. Pool 500μL PCR reactions (10 x 50μL PCR reactions) and add 0.40X volume of AMPure PB (200μL) magnetic beads. This is Fraction 2.
   - Important: Pipette exactly 500μL of PCR reactions and 200μL of AMPure PB magnetic beads as otherwise risk of significant DNA loss.
2. Pool remaining PCR reactions and add 1X volume of AMPure PB magnetic beads. This is Fraction 1.
   - Note: There will be inevitable sample loss through evaporation (~20μL), therefore do not expect to recover 800μL of cDNA.
3. Proceed with AMPure PB bead purification (**Section A.2.0.1**), with 100μL of EB to Fraction 1 and 22μL EB to Fraction 2.
4. Fraction 1 requires a second round of AMPure PB bead purification. Proceed directly to the next section ("Second Purification"). Fraction 2 does not require a second AMPure PB bead purification. Set this tube aside on ice and measure DNA concentration along with Fraction 1 after the second 1X AMPure PB bead purification for Fraction 1.

A.7.2 Fraction 1: 2nd purification

1. Perform a second round of AMPure PB bead purification for Fraction 1 (now in 100μL of EB) using 1X volume of AMPure PB magnetic beads.
2. Proceed with AMPure PB bead purification (**Section A.2.0.1**), with 22μL of EB to Fraction 1.
3. Quantify DNA amount and concentration of Fraction 1 and Fraction 2 using Qubit dsDNA High Sensitivity assay (**Section A.2.0.2**).
4. Determine the library size using the Bioanalyzer with DNA 12000 Kit (**Section A.2.0.3**).

A.7.3 Pooling Fraction 1 (1X) and 2 (0.40X)

Based on sample information from the Qubit and Bioanalyzer, determine the molarity of the two fractions using the following equation:

$$\frac{concentration(\frac{ng}{\mu L}) \times 10^6}{660(\frac{g}{mol}) \times average\ library\ size\ in\ bp^*} = concentration\ in\ nM \tag{A.1}$$

> * the average library size was determined by the start and end point of the cDNA smear on the Bioanalyzer

A minimum 200ng of pooled cDNA is necessary for library construction, despite the minimum recommended 1μg in protocol. If performing target capture, proceed to "Target Capture with IDT Probes" (**Section A.8**) below, otherwise skip to "SMRTbell template preparation" (**Section A.9**).

A.8 Target capture using IDT probes

A.8.0.1 Prepare hybridisation

The probes for all the target genes should be delivered and resuspended in one pooled tube as equimolar amounts.

1. Add 1 – 1.5μg cDNA to a 0.2mL PCR tube.
2. Add 1μL of SMARTer PCR oligo and 1μL poly(T) blocker (both at 1000μM) to the tube containing the cDNA.
3. Close the tube's lid and puncture a hole in the cap.
4. Dry the cDNA Sample Library/ SMARTer PCR oligo/ poly(T) blocker completely in a LoBind tube using a DNA vacuum concentrator (speed vacuum).
   - Place the 0.2mL PCR Tube in a 1.5mL Eppendorf tube. Do not leave tubes in the speed vacuum once they have dried. This will result in over drying the tube contents.
   - Be sure to seal sample tube! (From experience, evaporation with 20μL takes 30 minutes)
5. To the dried-down sample, add reagents listed in **Table A.6**.
6. Cut off the punctured lid and replace with new PCR lid. Ensure fully sealed.

Table A.6: Reagent composition for hybridisation

Reagents	Volume (μL)
2X Hybridisation Buffer	8.5
Hybridisation Buffer Enhancer	2.7
Nuclease-free water	1.8

7. Mix the reaction by tapping the tube, followed by a quick spin.
8. Incubate at 95°C for 10 minutes, lid set at 100°C, to denature the cDNA.
9. Brief spin. Leave the PCR tube at room temperature for ~2 minutes. Probes should never be added while at 95°C.
10. Add 4μL of xGen Lockdown Panel/Probe for a total volume of 17μL. Mix and quick spin.
11. Leave the PCR tube at room temperature for 5 minutes.
12. Incubate in a thermocycler at 65°C for 4 hours, lid set at 100°C.

A.8.0.2 Prepare beads for target capture

1. Allow the Dynabeads M-270 Streptavidin to warm to room temperature for 30 minutes prior to use.
2. Prepare Wash Buffers as tabulated in Table **Table A.7**.
3. Aliquot 200μL of 1X Wash Buffer (Tube 1) to new 1.5ml Eppendorf tube.
4. Mix the Dynabeads M-270 beads thoroughly by vortexing for 15 seconds. Check the bottom of the container to ensure proper reconstituting.
5. For a single sample, aliquot 100μL beads into a 1.5mL LoBind tube.
6. Place the LoBind tube in a magnetic rack until the beads collect to the side of the tube and the solution appears clear.
7. With the tube still on the magnetic rack, slowly pipette off cleared supernatant and save in another tube.
   - Note: Avoid disturbing pellet, not necessary to remove all liquid as will be removed with subsequent wash steps. Allow the Dynabeads to settle for at least 1 - 2 minutes before removing the supernatant. The Dynabeads are “filmy” and slow to collect to the side of the tube.
8. Wash beads with 200μL of 1X Bead Wash Buffer with the tube still on the rack.
9. Remove the tube from the magnetic rack. Vortex/tap tube until the beads are in solution. Quickly spin and place the tube in the magnetic rack until the beads collect to the

side of the tube (2 minutes). Once clear, carefully remove and discard supernatant.

10. Repeat steps 8 – 9.
11. Wash beads with 100μL of 1X Bead Wash Buffer.
12. Remove the tube from magnetic rack. Vortex/tap tube until the beads are in solution. Quickly spin and place the tube in the magnetic rack until the beads collect to the side of the tube (2 minutes). Do not remove the supernatant until ready to add hybridization sample.
13. Once clear, carefully remove and discard supernatant.
14. Proceed immediately to the "Binding cDNA to captured beads". The washed beads are now ready to bind the captured DNA. Do not allow the capture beads to dry. Small amounts of residual Bead Wash Buffer will not interfere with binding of DNA to the capture beads.

Table A.7: Preparation of wash buffers

Reagents	Buffer Volume (μL)	Water Volume (μL)
Wash Buffer I (Tube 1)	40	360
Wash Buffer II (Tube 2)	20	180
Wash Buffer III (Tube 3)	20	180
Stringent Wash Buffer (Tube S)	50	450
Bead Wash Buffer	250	250

A.8.0.3 Binding cDNA to beads

Steps 1 - 4 should be completed one tube at a time, working quickly to prevent the temperature of the hybridized sample from dropping significantly below 65°C.

1. Transfer 17μL hybridized probe/sample mixture prepared in the "Preparing hybridization section" (**Section A.8.0.1**) to the washed capture beads.
2. Mix by tapping the tube until the sample is homogeneous.
3. Aliquot 17μL of resuspended beads into a new 0.2mL PCR tube.
4. Incubate at 65°C for 45 minutes, lid set at 70°C.
    - Every 10 - 12 minutes, remove the tube and gently tap the tube to keep the beads in suspension. Do not spin down.
    - Prepare labelled and pre-heat 1.5μL Eppendorf LoBind tube at 65°C for later transfer of sample.
5. Preheat the following wash buffers at 65°C in water bath: 200μL of 1X Wash Buffer (Tube 1), 500μL of 1X Stringent Wash Buffer (Tube S).

6. Proceed immediately to heated washes (**Section A.8.0.4**).

A.8.0.4 Perform heated washes

Steps 1 - 4 need to be completed at 65°C to minimize non-specific binding of the off-target DNA sequences to the capture probes.

1. Add 100μL of pre-heated 1X Wash Buffer (Tube 1 at 65°C) to bead hybridised sample.
2. Mix thoroughly by tapping the tube until the sample is homogeneous. Be careful to minimise bubble formation.
3. Transfer sample (117μL) from PCR tube to 1.5mL Eppendorf LoBind tube.
4. Place the LoBind tube in a magnetic rack until the beads collect to the side of the tube and the solution appears clear (1 minute).
   - Bead separation should be immediate. To prevent temperature from dropping below 65°C, quickly remove the clear supernatant.
   - With the tube still on the magnetic rack, slowly pipette off cleared supernatant and save in another tube: "supernatant post-binding". Be careful not to disturb the pellet.
5. Remove the tube from the magnetic rack and quickly wash beads with 200μL of pre-heated 1X Stringent Wash Buffer (Tube S).
6. Tap the tube until the sample is homogeneous. Be careful not to introduce bubble formation. Work quickly so that the temperature does not drop below 65°C.
7. Incubate at 65°C for 5 minutes.
8. Place the LoBind tube in a magnetic rack until the beads collect to the side of the tube and the solution appears clear (almost immediate).
9. Repeat steps 5 – 8.
10. Proceed immediately to room temperature washes (**Section A.8.0.5**).

A.8.0.5 Perform room temperature washes

1. Wash beads with 200μL of room temperature 1X Wash Buffer I (Tube 1).
2. Remove the tube from the magnetic rack. Mix tube thoroughly by tapping the tube until sample is homogeneous, important to ensure beads fully resuspended!
3. Incubate for 2 minutes, while alternating between tapping for 30 seconds and resting for 30 seconds, to ensure mixture remains homogenous.
4. Quickly spin and place the tube in the magnetic rack until the beads collect to the side

of the tube (1 minute). When clear, remove and discard supernatant.

5. Wash beads with 200μL of room temperature 1X Wash Buffer II (Tube 2).
6. Repeat steps 2 - 4.
7. Wash beads with 200μL of room temperature 1X Wash Buffer III (Tube 3).
8. Repeat steps 2 - 4.
9. Remove residual Wash Buffer III with a fresh pipette, with the sample tube still on the magnet.
   - Important to ensure all residual wash buffer III removed.
10. Remove tube from the magnetic bead rack and add 50μL of Elution Buffer. This is required enough for two PCR reactions. Stored the beads plus captured samples at -15°C to -25°C or proceed to the next step. It is not necessary to separate the beads from the eluted DNA, as bead/sample mix can be added directly to PCR.

A.8.0.6 Amplification of captured cDNA

1. Prepare PCR reaction mix in a 1.5ml Eppendorf tube (**Table A.8**).
2. Split the PCR reaction mix into two tubes, 100μL each.
3. Cycle with the conditions outlined in **Table A.9**.
4. Pool the 100μL reactions and proceed to AMPure bead purification.

Table A.8: Reagent composition for amplification of captured cDNA

Reagents	Volume (μL)
Nuclease-free water	104.5
10x LA PCR buffer	20
2.5mM each dNTPs	16
SMARTer PCR Oligo (12μM)	8.3
Takara LA Taq DNA Polymerase	1.2
Captured Library	50
Total volume per sample	*200*

Table A.9: PCR conditions for amplification of captured cDNA

Segment	Temperature (°C)	Time
1	95°C	2 minutes
2	95°C	20 seconds
3	68°C	10 minutes
4	Repeat steps 2-3, for a total of 11 cycles	
5	72°C	10 minutes
6	4°C	Hold

A.9 SMRTbell template preparation

A.9.0.1 DNA damage and end repair

1. Prepare a PCR reaction mix in a 1.5mL Eppendorf LoBind tube (**Table A.10**).
2. Mix the reaction well by flicking tube and briefly spin down.
3. Incubate tubes at 37°C for 20 minutes, then return reaction to 4°C.
4. Add 2.5μL End Repair Mix to incubated cDNA.
5. Mix the reaction well by flicking tube and briefly spin down.
6. Incubate at 25°C for 5 minutes, then return reaction to 4°C.

Table A.10: Reagent composition for DNA damage and end repair

Reagents	Volume (μL)
Pooled cDNA (Fraction 1 & 2)	X (200ng - 5ug)
DNA Damage Repair Buffer	5
NAD+	0.5
ATP high	5
dNTP	0.5
DNA Damage Repair Mix	2
Nuclease-Free water	X to adjust to 50
Total volume per sample	*50*

A.9.0.2 DNA purification

1. Proceed with AMPure PB bead purification. (**Section A.2.0.1**), with 1X volume of AMPure beads (50μL) and eluting with 32μL of Elution Buffer.
2. The End-Repaired DNA can be stored overnight at 4°C (or -20°C for longer).

A.9.0.3 Prepare blunt ligation reaction

1. Add the following reagents in **Table A.11** in the order shown to each sample.
2. Mix the reaction well by flicking the tube and briefly spin down.
3. Incubate at 25°C for up to 24 hours, returning reaction to 4°C (for storage up to 24 hours).
4. Incubate at 65°C for 10 minutes to inactivate the ligase, returning reaction to 4°C. Proceed with adding exonuclease.

Table A.11: Reagent composition for blunt ligation reaction

Reagents	Volume (μL)
Pooled cDNA (End Repaired)	31
Blunt Adapter (20μM)	2
Mix before proceeding	
Template Prep Buffer	4
ATP low	2
Mix before proceeding	
Ligase	1
Nuclease-Free water	X to adjust to 40
Total volume per sample	*40*

A.9.0.4 Add exonuclease to remove failed ligation products

1. Add 1μL of Exonuclease III to pooled cDNA.
2. Add 1μL of Exonuclease VII to pooled cDNA.
3. Mix reaction well by flicking the tube and briefly spin down.
4. Incubate at 37°C for 1 hour, returning reaction to 4°C. Proceed with purification.

A.9.0.5 First purification of SMRTbell templates

1. Proceed with AMPure PB bead purification (**Section A.2.0.1**), with 1X volume of AMPure beads (42μL) and eluting with 50μL of Elution Buffer.

A.9.0.6 Second purification of SMRTbell templates

1. Proceed with AMPure PB bead purification (**Section A.2.0.1**), with 1X volume of AMPure beads (50μL) and eluting with 10μL of Elution Buffer.

2. Quantify DNA amount and concentration of Fraction 1 and Fraction 2 using Qubit dsDNA High Sensitivity assay (**Section A.2.0.2**).
3. Determine the library size using the Bioanalyzer with DNA 12000 Kit (**Section A.2.0.3**).

Appendix B

Customised ONT nanopore sequencing protocol

This protocol was adapted from three sources: i) “Wellcome Trust Advanced Course: RNA Transcriptomics (2018)” that I attended as part of my PhD, and provided by J. Ragoussis (hereby referred to as “WTAC”), ii) the official ONT protocol “1D amplicon/cDNA by Ligation (SQK-LSK109)”, and iii) directed under the guidance of Dr Karen Moore, University of Exeter Sequencing Service. In brief, this protocol aimed to complement the Iso-Seq protocol (**Appendix A**) as a direct comparison of the two sequencing technologies. It was therefore important to ensure that all the steps, except the ONT library preparation step, were consistent (**Figure 3.14**). Consequently, cDNA synthesis and amplification was performed, followed by target capture. The lab workflow then branched out upon the respective library preparation.

B.1 cDNA Synthesis and amplification

For a direct comparison of ONT nanopore sequencing and PacBio Iso-Seq approach, the same methods for cDNA synthesis and amplification in the Iso-Seq protocol were used (**Section A.3** - **Section A.6**). There were attempts to perform cDNA synthesis and amplification from WTAC protocol, particularly as it used the capped-dependent TeloPrime Full-Length cDNA Amplification kit (**Appendix C**). However, there were difficulties in achieving sufficient yield for downstream library preparation, in addition to complicating downstream comparative analyses.

B.2 Bead purification of large-scale PCR products

1. Pool 800μL PCR reactions (16 x 50μL PCR reactions) and add 0.9X volume of AMPure PB (200μL) magnetic beads.
    - 20 - 30μL loss is expected from evaporation, therefore do not be able to recover 800μL of cDNA.
    - Note: Only prepare 1 Fraction for downstream library preparation rather than the 2 Fractions in Iso-Seq.
2. Proceed with AMPure PB bead purification (**Section A.2.0.1**), with 51μL of TE Buffer.
3. Quantify DNA amount and concentration using Qubit dsDNA High Sensitivity assay (**Section A.2.0.2**).
4. Determine the library size using the Bioanalyzer with DNA 12000 Kit (**Section A.2.0.3**).

B.3 ONT MinION library preparation

B.3.1 DNA damage and end repair

1. Thaw DNA CS (DCS) at room temperature, mix, spin down, and place on ice.
2. Prepare the NEBNext FFPE DNA Repair Mix and NEBNext End repair/ dA-tailing Module reagents in accordance with manufacturer's instructions, and place on ice.
3. Prepare a PCR reaction mix for each sample in microcentrifuge tube (**Table B.1**).
4. Mix gently by flicking tube and spin down.
5. Incubate in thermal-cycler at 20°C for 5 minutes and 65°C for 5 minutes.

B.3.1.1 Bead purification of cDNA end-repaired products

1. Proceed with AMPure PB bead purification (**Section A.2.0.1**), with 1X of AMPure beads and elute with 61μL of nuclease-free water.

B.3.1.2 Prepare Ligation Reaction

1. Prepare the following reagents:

Table B.1: Reagent composition for DNA and end repairs

Reagents	Volume (μL)
cDNA (1.5μg)	X
DNA CS	1
NEBNext FFPE DNA Repair Buffer	3.5
NEBNext FFPE DNA Repair Mix	2
Ultra II End-prep reaction buffer	3.5
Ultra II End-prep reaction mix	3
Nuclease-free water	Up to 60
Total volume	*60*

- Spin down Adapter Mix (AMX) and T4 Ligase from the NEBNext Quick Ligation Module, and place on ice.
- Thaw Ligation Buffer (LNB) at room temperature, spin down and mix by pipetting. Due to viscosity, vortexing this buffer is ineffective. Place on ice immediately after thawing and mixing.
- Thaw Elution Buffer (EB) and S Fragment Buffer (SFB) at room temperature, mix by vortexing, spin down and place on ice.

2. Prepare PCR reaction mix in a 1.5mL Eppendorf LoBind tube.
3. Mix gently by flicking the tube, and spin down.
4. Incubate the reaction for 10 minutes at room temperature (up to 4 hours).

B.3.1.3 Bead purification of ligated cDNA

1. Prepare the AMPure beads for use by allowing to equilibrate to room temperature for a minimum of 15 minutes. Resuspend by vortexing.
2. Add 40μL of resuspended AMPure XP beads to the reaction and mix the bead/DNA solution thoroughly.
3. Incubate on a Hula mixer (rotator mixer) for 5 minutes at room temperature.
4. Spin down both tubes (1 second) to collect beads.
5. Place tubes in a magnetic bead rack, and wait until the beads collect to the side of the tubes and the solution appears clear (2 minutes).
6. With the tubes still on the magnetic bead rack, slowly pipette off cleared supernatant and save in other tubes. Avoid disturbing the bead pellet.
7. With the tubes still on the magnetic bead rack, wash the beads by adding either 250μL S Fragment Buffer (SFB). Flick the beads to resuspend, then return the tube to magnetic

rack and allow the beads to pellet. Remove the supernatant using a pipette and discard.

8. Repeat the previous step.
9. Remove residual supernatant by taking tubes from magnetic bead rack and spin to pellet beads. Place the tubes back on magnetic bead rack and pipette off any remaining supernatant.
10. Remove tubes from magnetic bead rack and allow beads to air-dry (with tube caps open) for 30 seconds.
11. Elute with 15μL Elution Buffer (EB). Tap tubes until beads are uniformly re-suspended. Do not pipette to mix.
12. Elute DNA by letting the mix stand at room temperature for 10 minutes.
13. Spin the tube down to pellet beads, then place the tube back on the magnetic bead rack. Let beads separate fully and transfer supernatant to a new 1.5mL Eppendorf LoBind tube. Avoid disturbing beads.
14. Quantify DNA amount and concentration of Fraction 1 and Fraction 2 using Qubit dsDNA High Sensitivity assay (**Section A.2.0.2**). Determine library size using the Bioanalyzer with DNA 12000 Kit (**Section A.2.0.3**).

B.4 Priming the Flow Cell

1. Prepare the following reagents:
   - Thaw the Sequencing Buffer (SQB), Loading Beads (LB), Flush Tether (FLT) and one tube of Flush Buffer (FLB) at room temperature before placing the tubes on ice as soon as thawing is complete.
   - Mix the Sequencing Buffer (SQB) and Flush Buffer (FLB) tubes by vortexing, spin down and return to ice.
   - Spin down the Flush Tether (FLT) tube, mix by pipetting, and return to ice.
2. Open the lid of the nanopore sequencing device and slide the flow cell's priming port cover clockwise so that the priming port is visible.
3. Priming and loading the SpotON Flow Cell.
   - Take care to avoid introducing any air during pipetting.
   - Care must be taken when drawing back buffer from the flow cell. The array of pores must be covered by buffer at all times. Removing more than 20 - 30μL risks damaging the pores in the array.

4. After opening the priming port, check for small bubble under the cover. Draw back a small volume to remove any bubble (a few μLs):
   - Set a P1000 pipette to 200μL.
   - Insert the tip into the priming port.
   - Turn the wheel until the dial shows 220 - 230μL, or until you can see a small volume of buffer entering the pipette tip.
   - Visually check that there is continuous buffer from the priming port across the sensor array.
5. Prepare the flow cell priming mix: add 30μL of thawed and mixed Flush Tether (FLT) directly to the tube of thawed and mixed Flush Buffer (FLB), and mix by pipetting up and down.
6. Load 800μL of the priming mix into the flow cell via the priming port, avoiding the introduction of air bubbles. Wait for 5 minutes.
7. Thoroughly mix the contents of the LB tube by pipetting. The Loading Beads (LB) tube contains a suspension of beads. These beads settle very quickly. It is vital that they are mixed immediately before use.

B.4.1 Library loading into the Flow Cell

1. Prepare sample for loading (**Table B.2**).
2. Gently lift the SpotON sample port cover.
3. Load 200μL of the priming mix into the flow cell via the priming port (not the SpotON sample port), avoiding the introduction of air bubbles.
4. Mix the prepared library gently by pipetting up and down just prior to loading.
5. Add 75μL of sample to the flow cell via the SpotON sample port in a drop-wise fashion. Ensure each drop flows into the port before adding the next.
6. Gently replace the SpotON sample port cover, making sure the bung enters the SpotON port, close the priming port and replace the MinION lid.

Table B.2: Reagent composition for loading ONT Flow Cell

Reagents	Volume (μL)
Sequencing Buffer (SQB)	37.5
Loading Buffer (LB), mixed immediately before use	25.5
DNA library	12
Total volume	*75*

Appendix C

Trialling of alternative cDNA synthesis approaches

C.1 Alternative cDNA synthesis approaches

Despite the power of the SMARTer cDNA synthesis kit (Clontech) to enrich for high-quality and full-length cDNA templates, this kit is costly and fails to differentiate between intact and degraded RNA (as described in **Section 2.2.1**). Two alternative methods were therefore trialled and optimised on a mouse cortical sample:

1. An adaptation of the cDNA synthesis protocol (provided by J. Ragoussis) taken from the "Wellcome Trust Advanced Course: RNA Transcriptomics (2018)" (hereby referred to as "WTAC") which I attended during my PhD. This protocol is based on the Smart-seq2 protocol,[403] which formed the basis for the SMARTer cDNA synthesis kit.
2. Full-Length cDNA Amplification (TeloPrime) kit, which ensured the amplification of fully-intact cDNA containing the 5' cap.

Unfortunately, I was unable to generate sufficient cDNA material from the Full-Length cDNA Amplification kit required for downstream library preparation. The following section thus only documents the optimised protocol and results from the adaptation of the WTAC protocol, based on Smart-seq2.

C.1.1 Adaptation of Smart-seq2

In the WTAC protocol, SuperScript IV enzyme (SSIV) was used for reverse transcription of RNA into cDNA due to its high thermal-stability, subsequently allowing it to be used in high temperatures to resolve complex RNA secondary structures. Reverse transcription step was also split into two steps, pre-RT and RT, to first hybridise poly(T) primer to poly(A) tract of mRNA before first-strand synthesis (**Table C.2** - **Table C.5**). cDNA templates were then amplified using Advantage 2 high fidelity polymerase.

One of the challenges in applying the WTAC protocol to the mouse experiments was the high amount of RNA needed for cDNA synthesis: 300ng/μL was specified as the starting amount, but majority of the samples only had an average concentration of 70ng/μL. The volume of pre-RT and RT PCR mix was therefore upscaled to account for additional sample volume, while maintaining the concentration of other reagents.

C.1.1.1 Protocol

Table C.1: List of reagents for Smart-seq2 protocol

Reagents	Supplier (Catalogue)	Step
ERCC RNA Spike-In Mix	Thermo Fisher Scientific (4456740)	Pre-RT
RNAse inhibitor (40 U/μL)	Clontech (2313A)	Pre-RT
Advantage UltraPure PCR dNTP Mix (10 mM each dNTP)	Clontech (639125)	Pre-RT
• SuperScript IV (200U/μL)	Thermo Fischer Scientific (18090010)	RT
• 5X RT Buffer		Pre-RT
• DTT (0.1M)		RT
Betaine (5M)	Sigma-Aldrich (B0300-1VL)	RT
MgCl2 (1M)	Thermo Fisher Scientific (AM9530G)	RT
10X Advantage 2 PCR Buffer	Clontech (639207)	PCR
50X Advantage 2 Polymerase Mix		
50X dNTP Mix		

1. Assuming a stock ERCC RNA spike-in concentration of 30.3ng/μL and a final concentration of 1.8ng/μL, 1μL of stock ERCC RNA spike-in was originally diluted with 15.83 TE buffer (dilution of 1:16.83).
2. Pre-reverse transcription PCR mix was prepared with RNA (**Table C.2**).
3. The sample was mixed by tapping the tube, spun down and incubated in the conditions outlined in **Table C.3** to prime the RNA with poly(T).
4. Reverse transcription PCR mix was prepared with primed-RNA (**Table C.4**).

5. The sample was mixed by tapping the tube, spun down and incubated in the conditions outlined in **Table C.5** for first-strand cDNA synthesis.
6. cDNA from RT reaction was split into 7 PCR tubes, each with 1μL of cDNA and reagents tabulated in **Table C.6**.
   - This step was necessary to test the optimal number of PCR cycles to prevent over-amplification, with the PCR tubes incubated for different cycle durations.
7. All 7 PCR tubes were incubated simultaneously (**Table C.7**), with one PCR tube removed at each cycle from cycle 12 to cycle 18 in segment 3 (as suggested by WTAC protocol) and placed in ice.
8. At the end of segment 3 in **Table C.7**, all 6 PCR tubes were placed back into the thermal cycler for enzyme termination (**Table C.7**, segment 4).
9. PCR tubes were analysed on a 1% agarose gel (described in **Section A.5**) or on a D5000 genomic ScreenTape (described in **Section A.2.0.3**).

Table C.2: Pre-reverse transcription PCR mix. Tabulated is a list of the reagent volume from the WTAC protocol, and of that optimised to use a lower initial RNA concentration. While the *total volume per Sample* was different due to amount of RNA used (200ng), the final concentrations of all the other remaining reagents were maintained.

Reagents	Final concentration	Volume (μL/sample)	
		WTAC protocol	Optimised protocol
Diluted ERCC RNA		0.1	0.1
Rnase Inhibitor	0.64U/μL	0.05	0.08
Poly(T) primer	2.76μM	0.7	1.15
SuperScript IV Buffer	0.17	0.1	0.17
Nuclease-Free Water		0	0.55
dNTP Mix	1.9mM	0.56	0.95
Total RNA		1.5 (300ng)	2 (200ng)
Total volume per Sample		*3*	*5*

Table C.3: PCR conditions for pre-reverse transcription. Tabulated are the PCR conditions for amplification of cDNA templates: initial 72°C to prompt unfolding of RNA secondary structures, 4°C for poly(T) binding and 25°C to encourage more specific binding.

Segments	Temperature (°C)	Time (minutes)
1	72	3
2	4	10
3	25	1

Table C.4: Reagent composition for reverse transcription PCR mix.

Reagents	Final concentration	Volume (μL/sample)	
		WTAC protocol	Optimised protocol
Sample from pre-RT		3	5
dH20		0.85	0.55
Superscript IV Buffer	1.00 U/μL	0.8	1.1
DTT	0.57 mM	0.175	0.25
TSO	2.50 mM	0.7	1
RNAse inhibitor	1.20 uM	0.175	0.25
SuperScript IV RT	10.0 U/μL	0.35	0.5
Betaine	0.50 M	0.7	1
MgCl2	3.57 mM	0.25	0.35
Total volume per sample		*7*	*10*

Table C.5: PCR conditions for reverse transcription. Tabulated is a list of the PCR conditions as taken from the WTAC protocol. Segments 2, 3, 5, 7, 9 allow the unfolding of RNA secondary structures and completion or continuation of RT with successive higher temperatures, whereas segments 1, 4, 6 and 8 allow template switching. Segment 11 is for enzyme termination.

Segments	Temperature (°C)	Time	Cycles
1	50	10 minutes	1
2	55	30 seconds	10
	50	30 seconds	
3	60	30 seconds	5
	55	30 seconds	
4	50	30 seconds	1
5	60	30 seconds	5
	60	30 seconds	
6	50	30 seconds	1
7	70	30 seconds	5
	65	30 seconds	
8	50	30 seconds	1
9	75	30 seconds	5
	70	30 seconds	
10	50	1 minute	1
11	80	10 minutes	1

Table C.6: PCR Mix. Tabulated is a list of the reagent volumes as taken from the WTAC protocol. The volume tabulated is sufficient for only 1μL of cDNA from total RT reaction - upscale accordingly.

Reagents	Volume (μL/sample)
cDNA from RT Reaction	1
Nuclease-free water	6.8
10X Advantage 2 PCR Buffer	1
50X dNTP Mix	0.4
PCR primer	0.4
50X Advantage 2 Polymerase Mix	0.4
Total volume for sample	*10*

Table C.7: PCR conditions for cDNA amplification. Tabulated is a list of PCR conditions as taken from WTAC protocol. All 7 prepared PCR tubes were individually removed and placed on ice in segment 3 (between cycles 12 to 18) to determine the optimum number of PCR cycles for amplification. After cycle 18, all removed PCR tubes were placed back into the thermal cycler for enzyme termination (segment 4).

Segments	Temperature (°C)	Time	Cycles
1	95	1 minute	1
2	95	20 seconds	
	58	4 minutes	5
	68	6 minutes	
3	95	20 seconds	
	64	30 seconds	12 - 18 cycles
	68	6 minutes	
4	72	10 minutes	1

C.1.1.2 Results

Following the optimised WTAC protocol, we similarly determined the optimal number of PCR cycles necessary for cDNA amplification (akin to **Section 3.1.2.1**) by assessing the quality of cDNA templates from cycles 11 to 18 using agarose gel electrophoresis. No difference in bands was observed between the original WTAC and optimised protocol (**Figure C.1A**), indicating that I had successfully optimised the WTAC protocol for smaller initial input of RNA. However, both protocols showed a consistent intensity of DNA bands across the range of PCR cycles rather than an expected gradual increase, suggesting there was already over-amplification by cycle 12 (**Figure C.1A**). The two stark bands at ~600bp and ~1000bp against the smear of cDNA further suggests over-usage of ERCC, possibly due to the overestimation of assumed proportion of mRNA. The optimised protocol was thus repeated with a wider range of cycles from cycle 2 to 20, and with a lower ERCC RNA spike-in concentration to reduce unnecessary sequencing of ERCC (final concentration of 0.6ng/μL and a dilution factor of 1:50.5) (**Figure C.1B**).

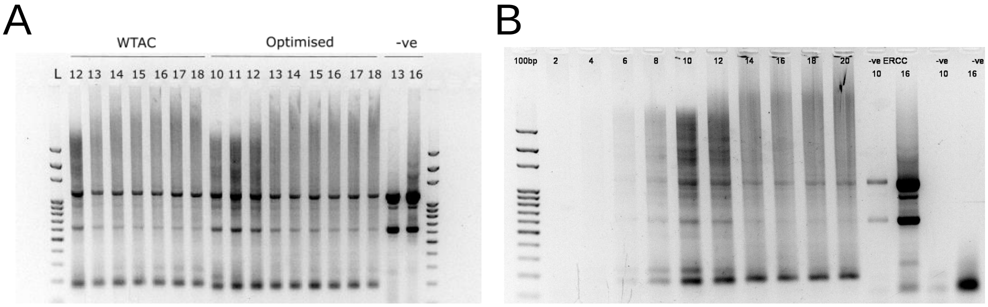

A
WTAC
Optimised
-ve
L 12 13 14 15 16 17 18 10 11 12 13 14 15 16 17 18 13 16
B
100bp 2 4 6 8 10 12 14 16 18 20 -ve ERCC -ve -ve
10 16 10 16

Figure C.1: Successful optimisation of Smart-seq2 protocol. Shown are agarose gel images of amplified cDNA from the **(A)** WTAC and optimised protocol following amplification from cycles 12 to 18, and from **(B)** repeating the optimised protocol with a greater range of PCR cycles and lower ERCC RNA spike-in concentration. Numbers represent the number of PCR cycles, L refers to 100bp ladder, -ve refers to negative control with water.

Despite successfully optimising the WTAC protocol for lower input of RNA and for determining the optimum number of PCR cycles for amplification, it was technically challenging to upscale the final amount of cDNA (10μL) synthesised for large scale PCR amplification (which required 160μL for Iso-Seq). I therefore decided to switch back to the SMARTer cDNA synthesis kit (Clontech) for the processing of mouse cortical samples used for Iso-Seq and ONT profiling in **Chapters 4 and 6**. Nonetheless, this optimised protocol provided insights into the appropriate amount of ERCC to be used, which was subsequently integrated into the final laboratory workflow (as described in **Section 2.3**).

Appendix D

Optimisation of ONT bioinformatics pipeline

This section documents the results generated from processing ONT raw reads using various community-based tools, and provides a rationale for the final ONT bioinformatics pipeline (depicted in **Figure 3.18**) applied following targeted profiling of AD-risk genes in the rTg4510 cortex (**Chapter 6**).

D.0.1 Introduction

As mentioned in **Section 3.2.4**, the bioinformatics pipeline for processing ONT raw reads was less defined than the Iso-Seq bioinformatics pipeline (as depicted in **Figure 3.9**) and relied heavily on community-based tools particularly for the collapse and annotations of mapped ONT reads. These tools included:

1. *TAMA*, which allows collapsing of long reads using a user-defined threshold for "wobble" (the maximum amount of bp difference between exon start and end sites across transcripts to be collapsed together) and comparison of annotated isoforms generated from different pipelines.
2. *FLAIR*, which allows usage of genome annotations and/or short-read RNA-Seq splice junctions for correction of misaligned splice sites prior to collapse of long reads.
3. *TALON*, which also allows reference-based error correction to remove microindels, mismatches and non-canonical splice junctions to generate a reference database. Each

transcript from each sample is then compared to existing transcript models in this database, which is continually updated with incorporation of new transcript models where appropriate. *TALON* further requires novel transcripts to be reproducibly detected across biological replicate samples.

To efficiently process mapped ONT transcripts derived from ONT targeted profiling, I trialled and benchmarked all three tools for isoform annotation and quantification of the rTg4510 cortex.

D.0.2 Processing of ONT reads by *TAMA* revealed necessity for splice-site error correction

Usage of *TAMA* for individual collapse and merging of each sample successfully curated thousands of full-length transcripts annotated to AD-associated genes, the vast majority of which were novel with novel splice-site junctions.

The curated transcriptome was then further filtered with RNA-Seq data and using other parameters to remove artefacts from intra-priming, and reverse transcription template switching (as described in **Section 3.1.4.4**). However, upon using RNA-Seq reads as junction support of isoforms with non-canonical junctions, the number of isoforms that were retained drastically fell from a total 40,498 isoforms (n = 1,332 known isoforms; n = 39,176 novel isoforms, **Figure D.1A**) annotated to AD target genes to 431 isoforms (**Figure D.1B**), with removal of all detected novel isoforms due to low short-read coverage (< 3 reads). Given the relatively high error rate of ONT reads (our ONT reads had an average 90% accuracy, mean Phred Q = 10), we suspected that many of the ONT reads had incorrectly sequenced splice sites, resulting in the generation of non-canonical splice junctions that were not supported by RNA-Seq reads.

D.0.3 Comparison of ONT and Iso-Seq datasets revealed insufficient coverage of RNA-Seq dataset

We therefore performed a splice-site correction step using *FLAIR*,[166] which uses RNA-Seq reads to assess the validity of splice site boundaries: junctions supported by three uniquely mapping short-reads were considered valid, whereas incorrect splice sites were replaced with the nearest valid splice sites within a 10-nucleotide window - the final set of corrected reads only consist of reads with valid splice sites.[166] Including this splice-site correction step after mapping with *Minimap2* and before isoform collapse with *TAMA*, we successfully recovered some novel transcripts (n = 4,945 total isoforms; n = 1,589 known isoforms; n = 3,356 novel isoforms, **Figure D.1C**). Notably, the usage of RNA-Seq reads for further junction filtering in *SQANTI* did not make any difference on the final number of transcripts (**Figure D.1D**).

Next, we compared the isoforms associated with target genes from all three ONT datasets generated either using RNA-Seq reads as support with *SQANTI* or splice site correction with *FLAIR* and Iso-Seq dataset using *TAMA Merge* under default parameters (**Figure D.2**). The vast majority of isoforms (99.8%) detected using PacBio Iso-Seq were detected across all three ONT datasets. However, a third of the Iso-Seq-derived isoforms overlapped with ONT-derived isoforms that were neither corrected with *FLAIR* nor supported by RNA-Seq reads. This suggests that coverage of the RNA-Seq dataset at 360 million reads (n = 24 samples) was insufficient to detect rare novel splice junctions and filtering of ONT reads with RNA-Seq dataset would be over-stringent with removal of true, novel, rare transcripts.

D.0.4 Usage of *TALON* for correction and filtering of ONT reads

TALON was used to address the challenges of processing noisy ONT reads. The aim was to achieve a fine balance between preserving rare novel transcripts and discarding technical artefacts. Rather than using RNA-Seq reads for filtering, novel isoforms were removed if they were lowly-expressed and were not detected across biological replicates. Through this approach (**Figure D.3**), we found that i) the majority of transcripts had only one read and were only detected in one sample, ii) the number of reads in each sample (expression) rather than the number of samples (reproducibility) primarily determined whether a transcript was retained or filtered, and ii) the default parameter (minimum 5 reads across all samples) was too stringent as it removed the vast majority of reads - a lower threshold (minimum 5 reads across minimum 2 samples) was therefore proposed, given that the proportion of reads retained starts to plateau beyond this point (**Figure D.3**).

Notably, aside from the added features of reference-based error correction and quantification-led filtering, *TALON* also superseded *TAMA* and *FLAIR* for simultaneous transcript discovery and quantification (the abundance output from *TAMA*'s tama_read_support_levels.py is computationally intensive to recover the read counts for each sample).

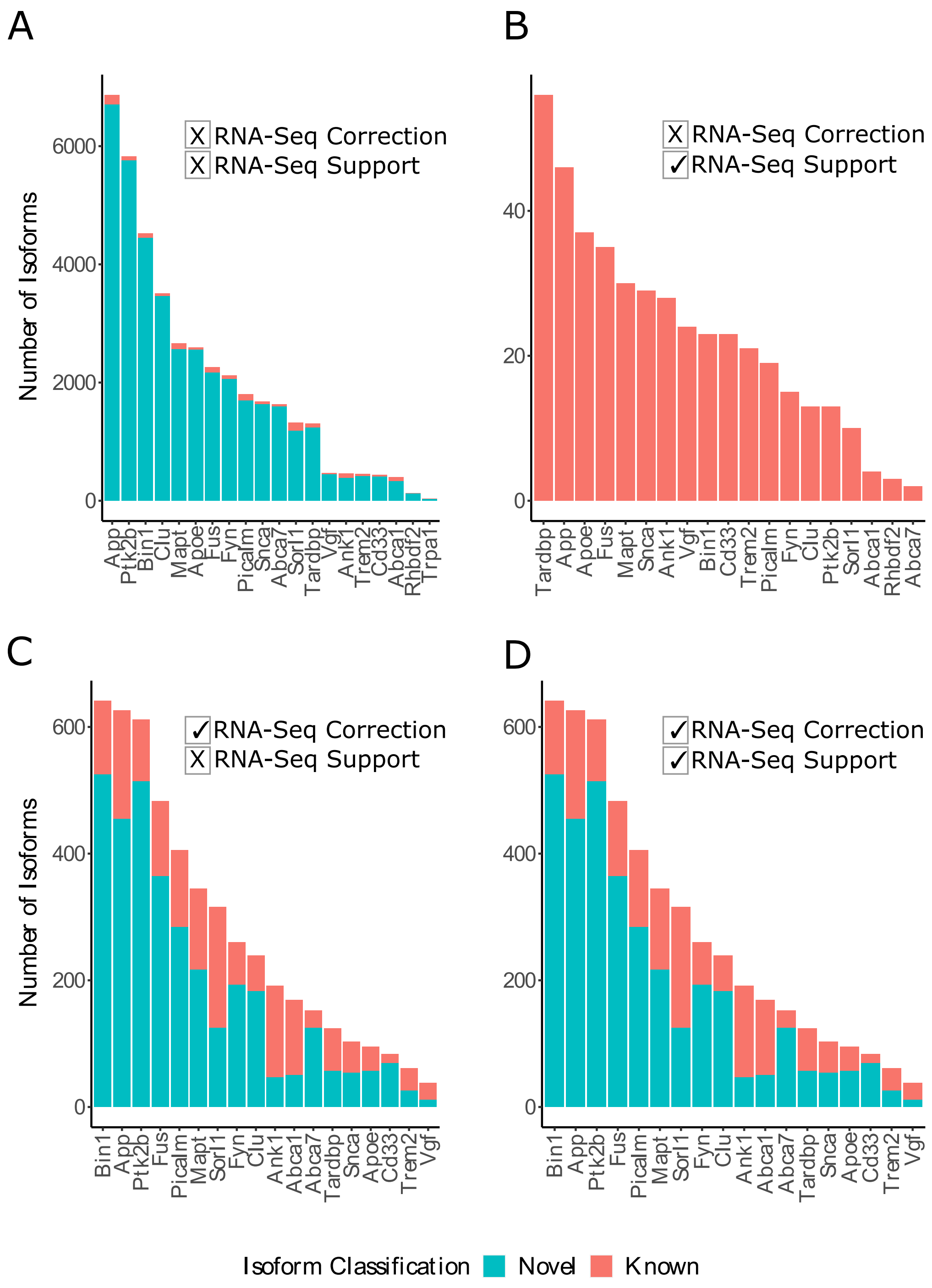


Figure D.1: Number of ONT-derived isoforms associated with AD target genes using different bioinformatics approaches. Shown are bar plots of the final number of isoforms associated with AD target genes from ONT nanopore sequencing after *SQANTI* annotation **(A)** without RNA-Seq reads for filtering, and **(B)** with RNA-Seq reads for filtering as junction support (*SQANTI* filter) (RNA-Seq Support), and **(C)** after splice junction correction with *FLAIR* (RNA-Seq Correction), and **(D)** with RNA-Seq splice junction correction and filtering.

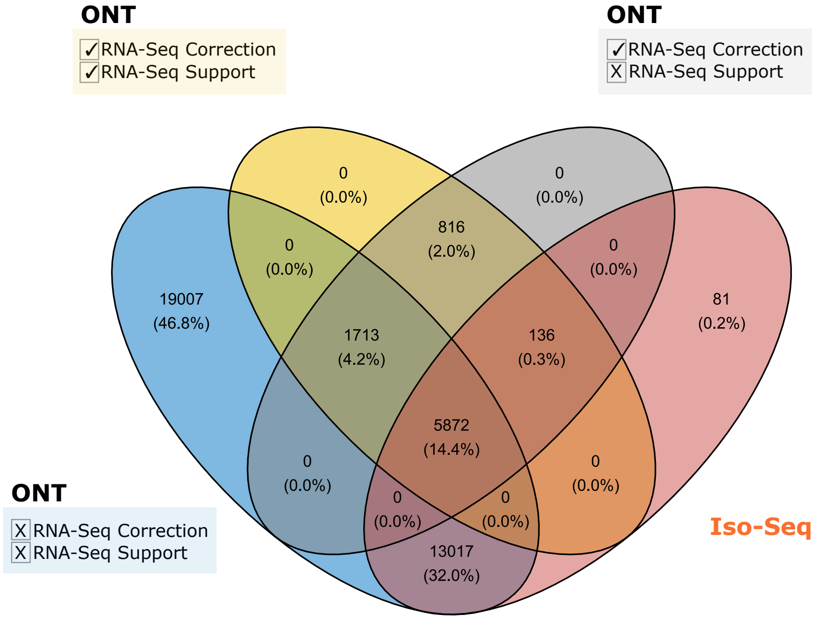


Figure D.2: Overlap of isoforms associated with target genes detected using Iso-Seq and ONT sequencing. Shown is a Venn diagram of the overlap of isoforms associated with AD target genes detected from ONT nanopore sequencing after *SQANTI* annotation. ONT transcripts were corrected with RNA-Seq with *FLAIR* (RNA-Seq Correction) or filtered out with RNA-Seq reads as junction support (*SQANTI* filter) (RNA-Seq Support) or both. Of note, more samples were sequenced with the PacBio Iso-Seq approach (n = 24) than the ONT nanopore sequencing approach (n = 18).

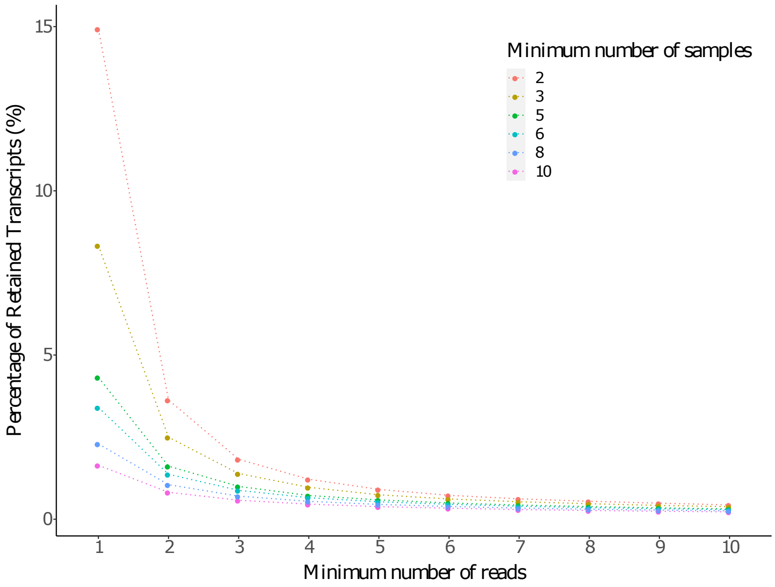


Figure D.3: Sensitivity curve for *TALON* filtering of ONT reads. Shown is a sensitivity curve for the proportion of ONT reads retained after applying different *TALON* parameters for filtering (i.e. the number of samples and reads).

Appendix E

Differential expression analyses of AD-risk genes

This section documents the results from the differential expression and splicing analyses of the other 17 AD-risk genes profiled in the rTg4510 cortex:

1. *Abca1* (**Figure E.1**)
2. *Abca7* (**Figure E.2**)
3. *Ank1* (**Figure E.3**)
4. *Apoe* (**Figure E.4**)
5. *App* (**Figure E.5**)
6. *Clu* (**Figure E.6**)
7. *Fus* (**Figure E.7**)
8. *Fyn* (**Figure E.8**)
9. *Mapt* (**Figure E.9**)
10. *Picalm* (**Figure E.10**)
11. *Ptk2b* (**Figure E.11**)
12. *Rhbdf2* (**Figure E.12**)
13. *Snca* (**Figure E.13**)
14. *Sorl1* (**Figure E.14**)
15. *Tardbp* (**Figure E.15**)
16. *Trpa1* (**Figure E.16**)
17. *Vgf* (**Figure E.17**)

Equivalent plots for *Trem2*, *Cd33* and *Bin1* are discussed in detail in **Chapter 6**, under **Section 6.3.10.1**, **Section 6.3.10.2** and **Section 6.3.10.3**, respectively.

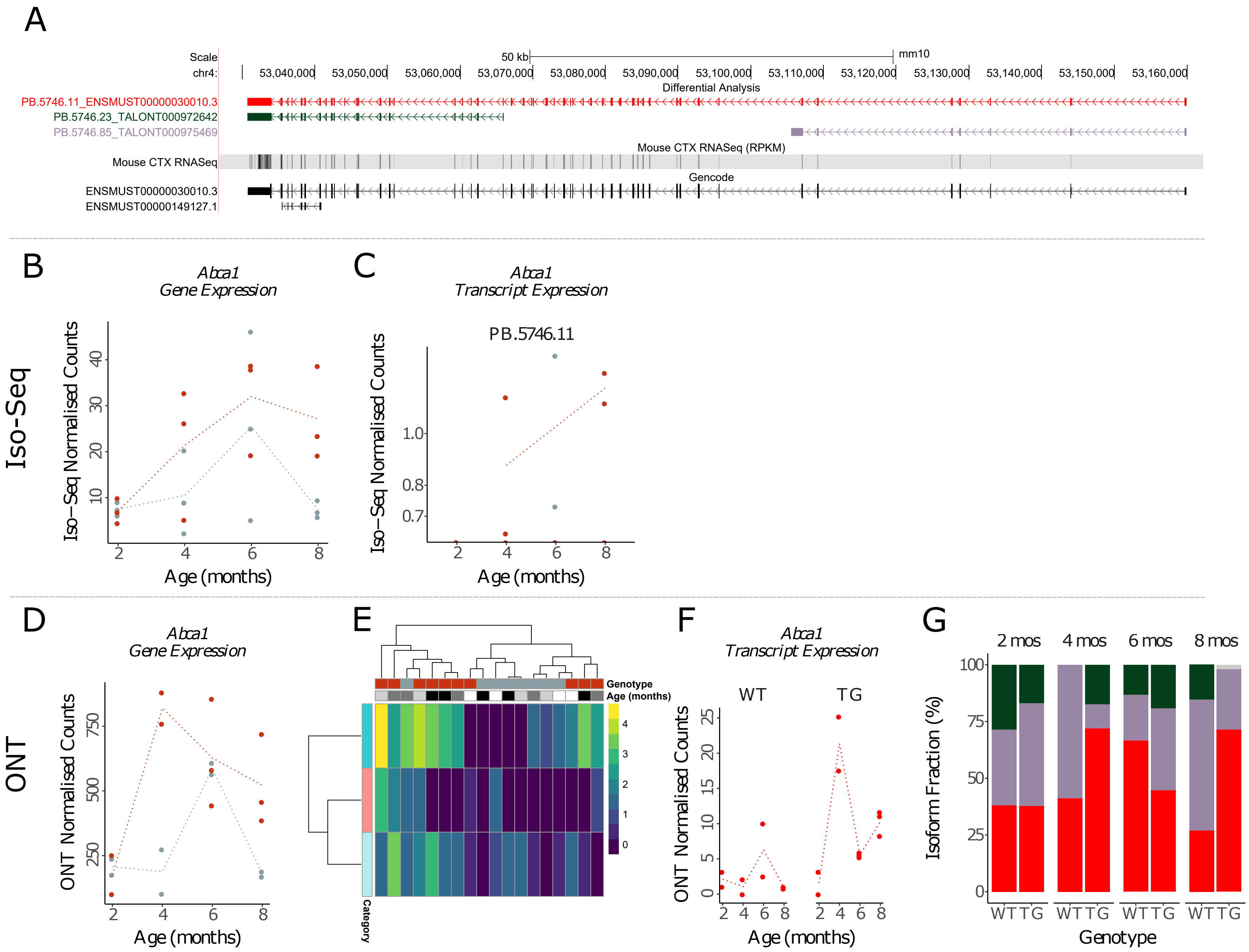


Figure E.1: Differential *Abca1* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Abca1* in the rTg4510 cortex. *Refer to **Figure 6.58** for the same legend.*

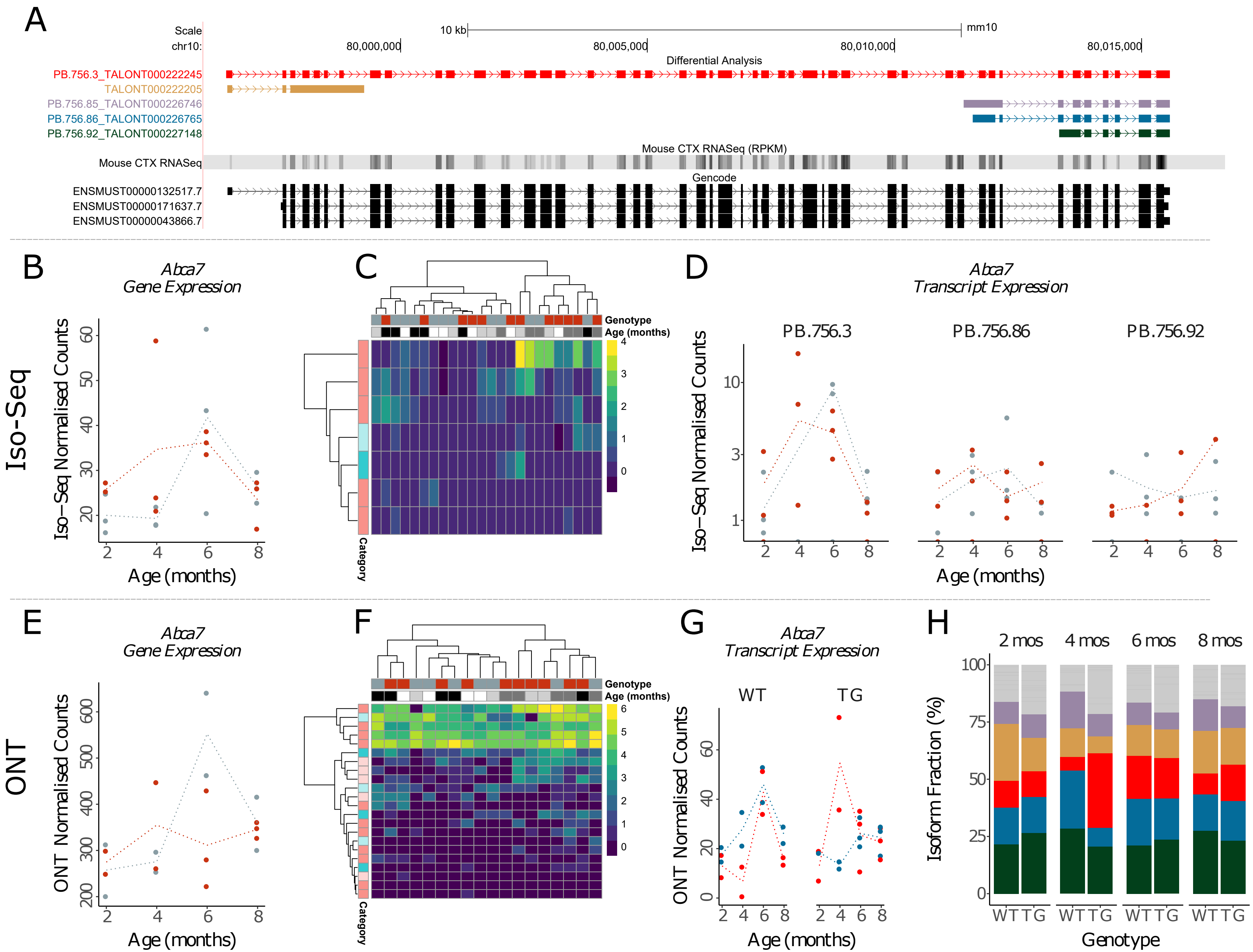


Figure E.2: Differential *Abca7* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Abca7* in the rTg4510 cortex. *Refer to **Figure 6.58** for the same legend.*

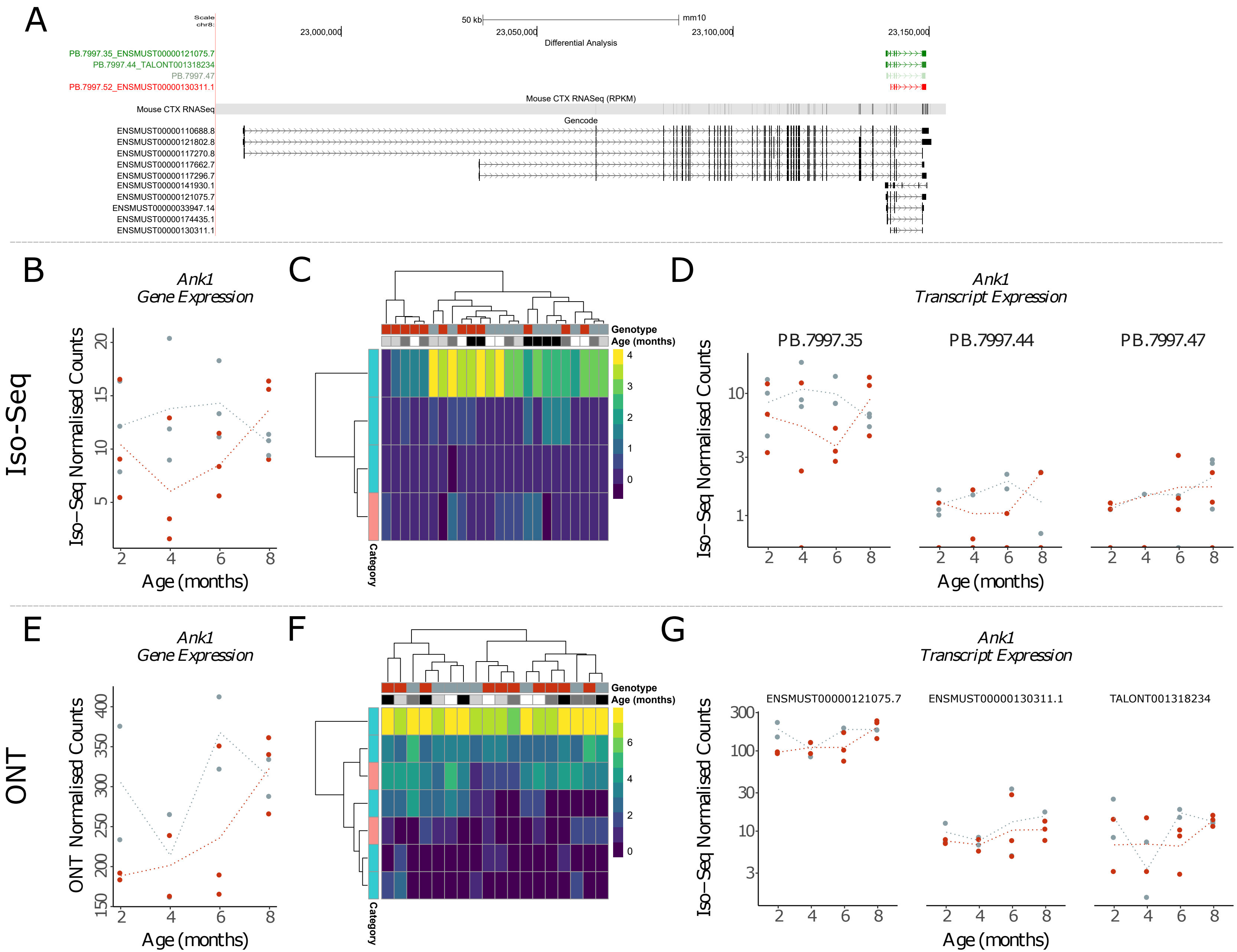


Figure E.3: Differential *Ank1* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Ank1* in the rTg4510 cortex. *Refer to **Figure 6.58** for the same legend.*

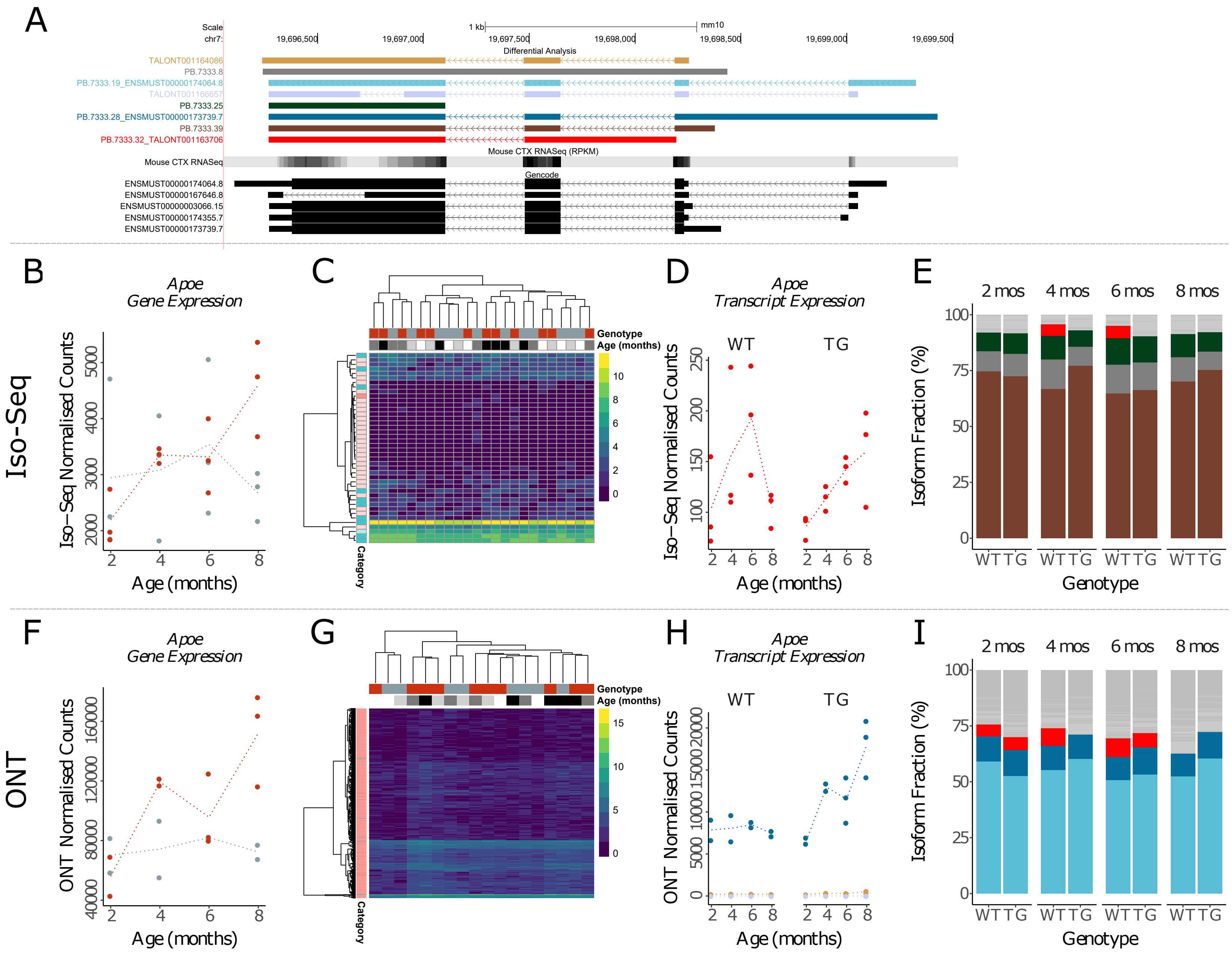


Figure E.4: Differential *Apoe* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Apoe* in the rTg4510 cortex. *Refer to* ***Figure 6.58*** *for the same legend.*

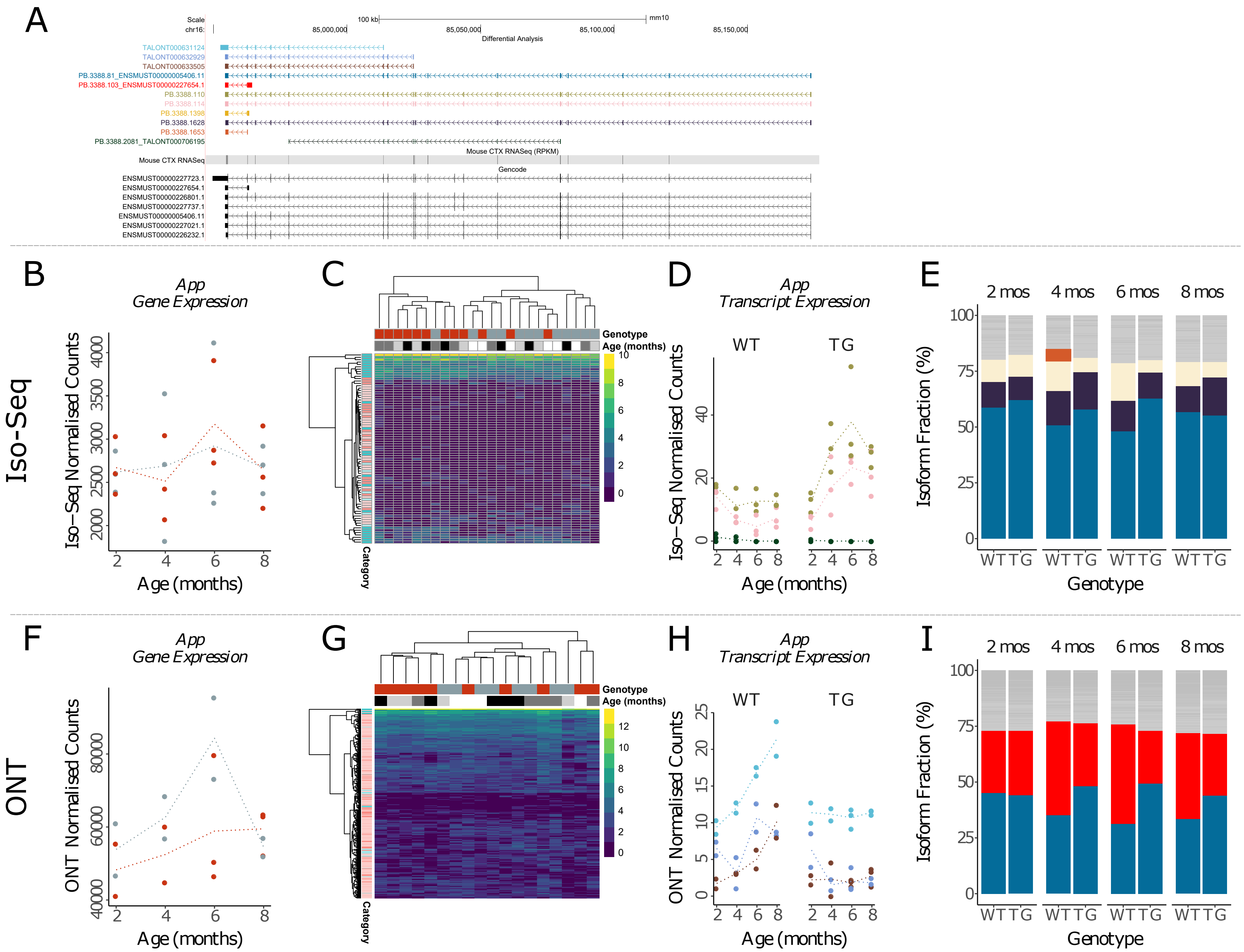


Figure E.5: Differential *App* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *App* in the rTg4510 cortex. *Refer to **Figure 6.58** for the same legend.*

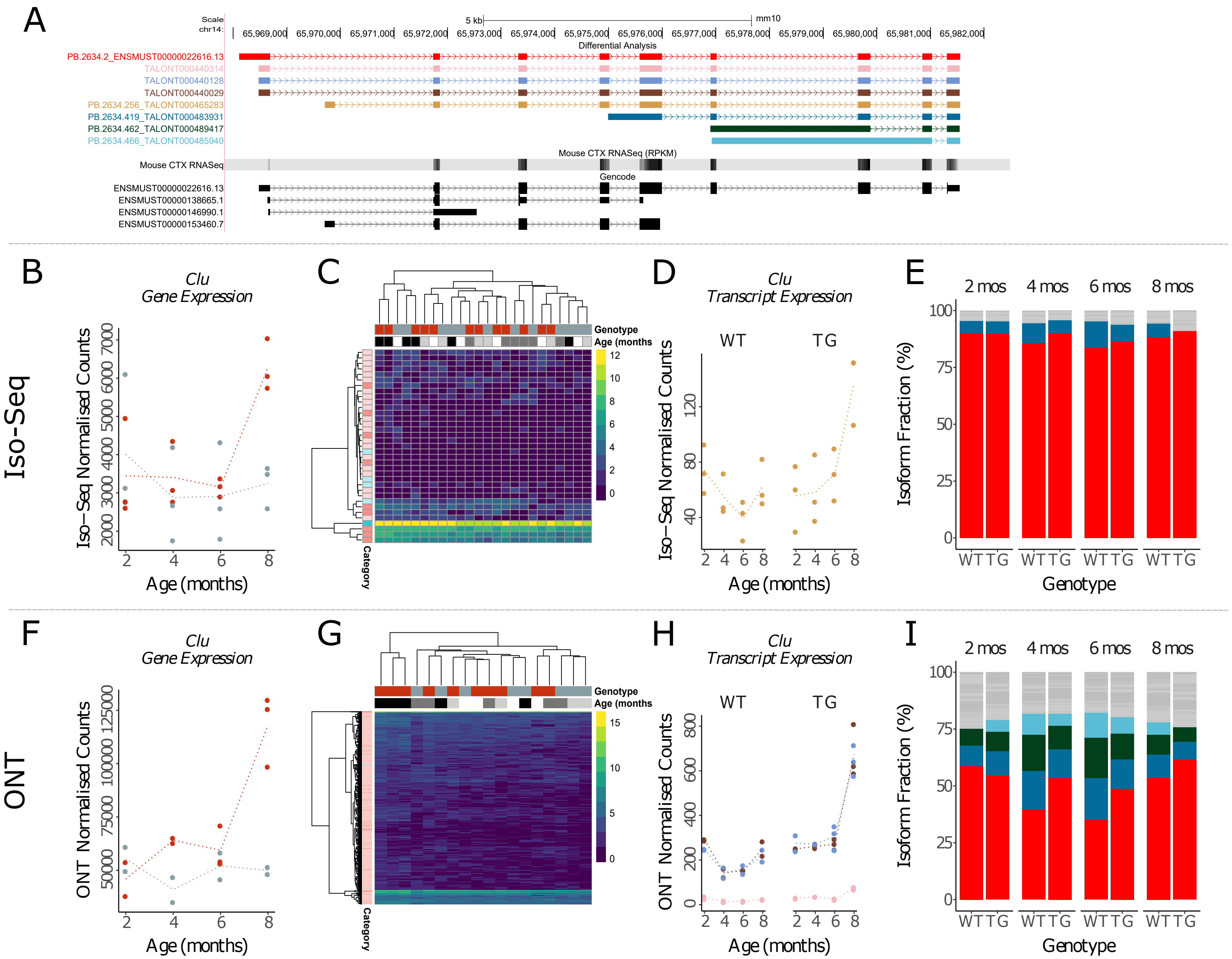


Figure E.6: Differential *Clu* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Clu* in the rTg4510 cortex. *Refer to **Figure 6.58** for the same legend.*

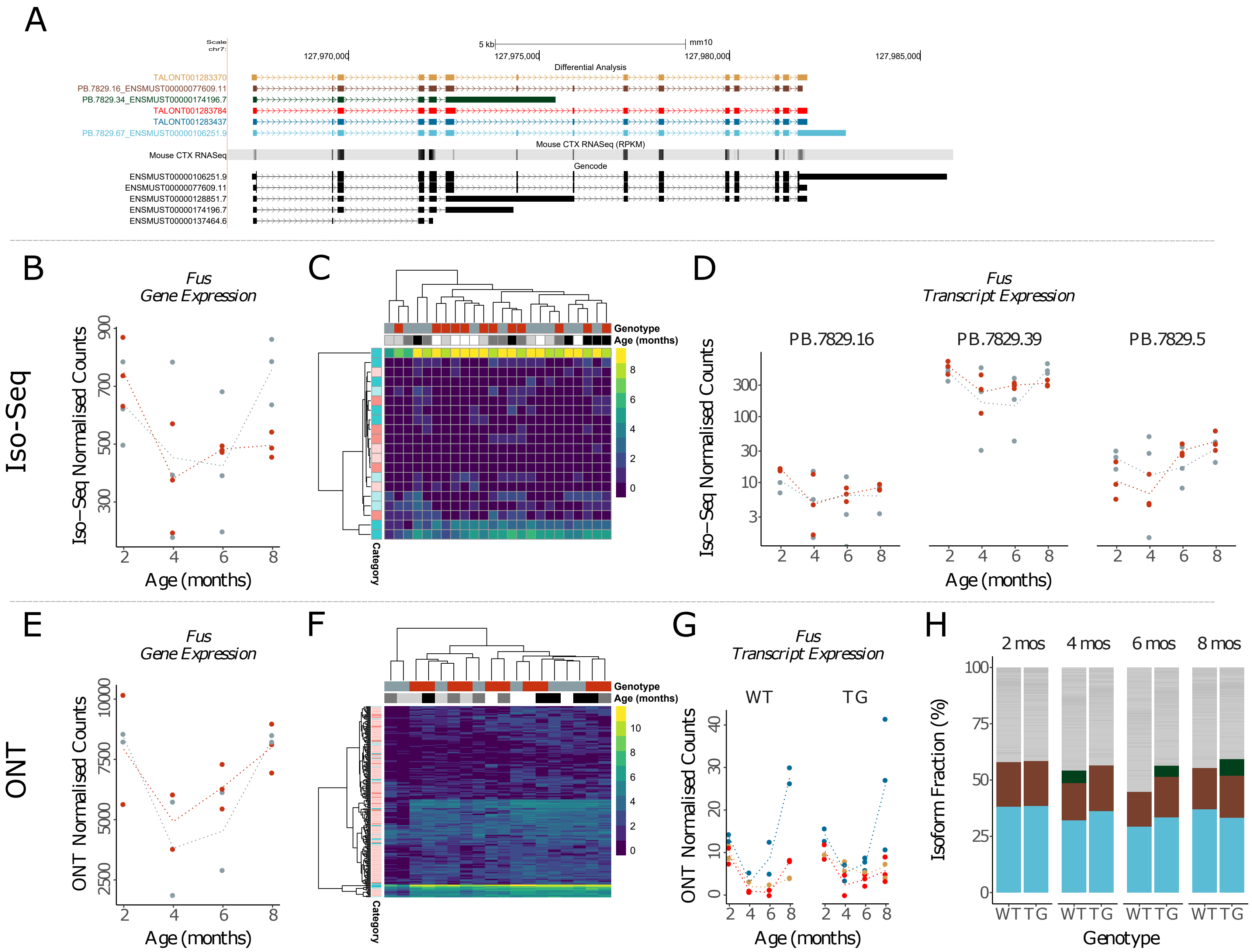


Figure E.7: Differential *Fus* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Fus* in the rTg4510 cortex. *Refer to **Figure 6.58** for the same legend.*

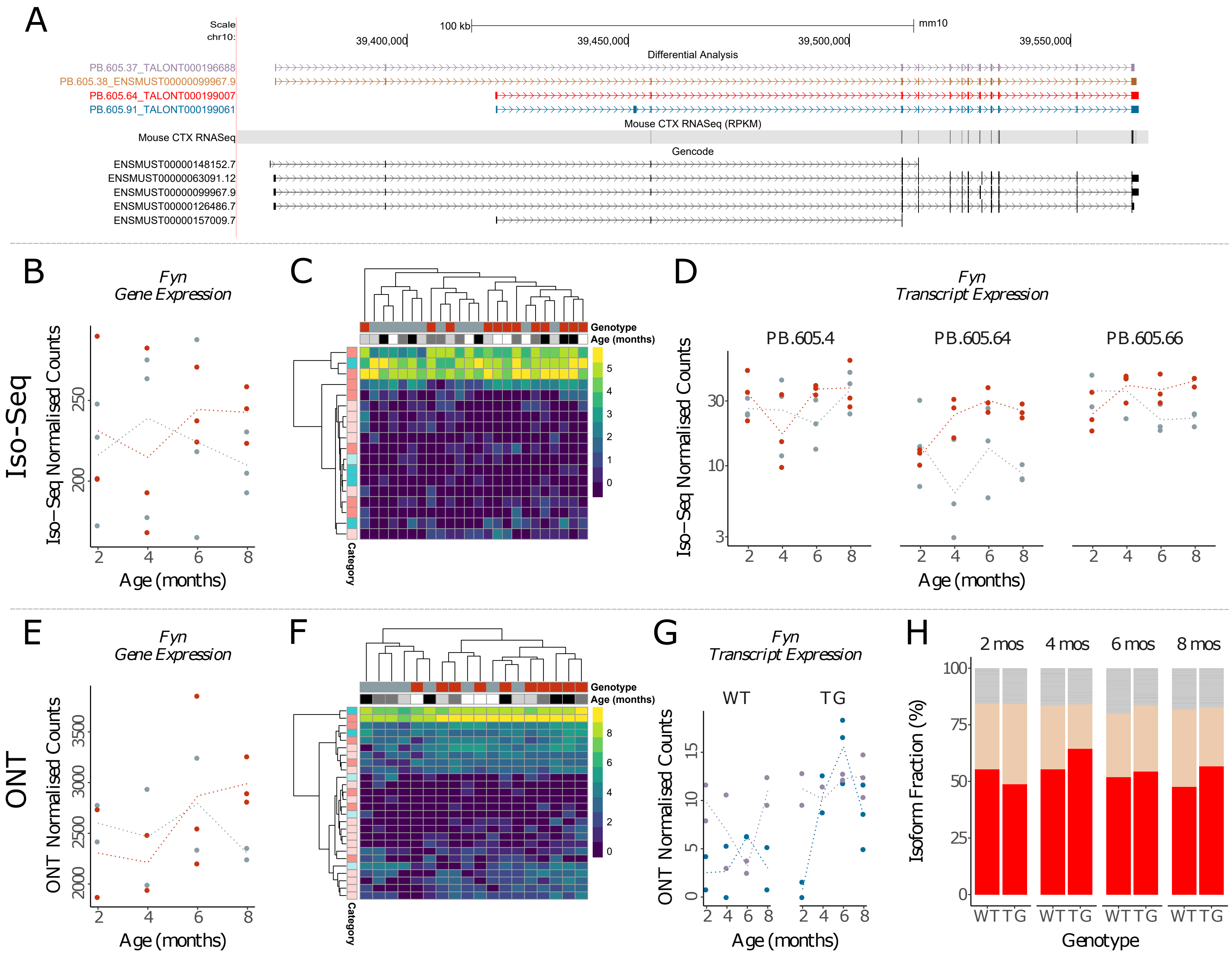


Figure E.8: Differential *Fyn* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Fyn* in the rTg4510 cortex. *Refer to* ***Figure 6.58*** *for the same legend.*

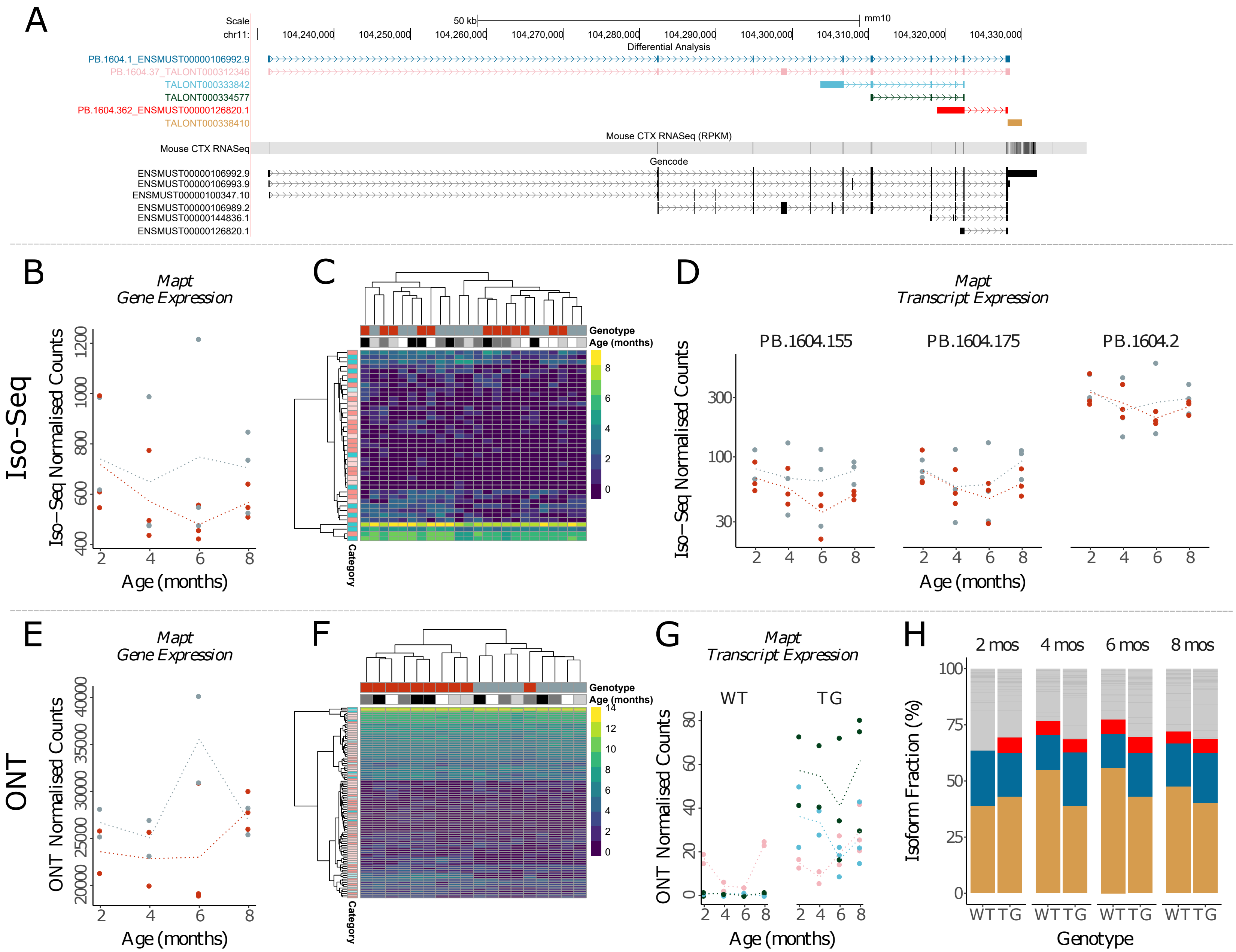


Figure E.9: Differential *Mapt* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Mapt* in the rTg4510 cortex. *Refer to **Figure 6.58** for the same legend.*

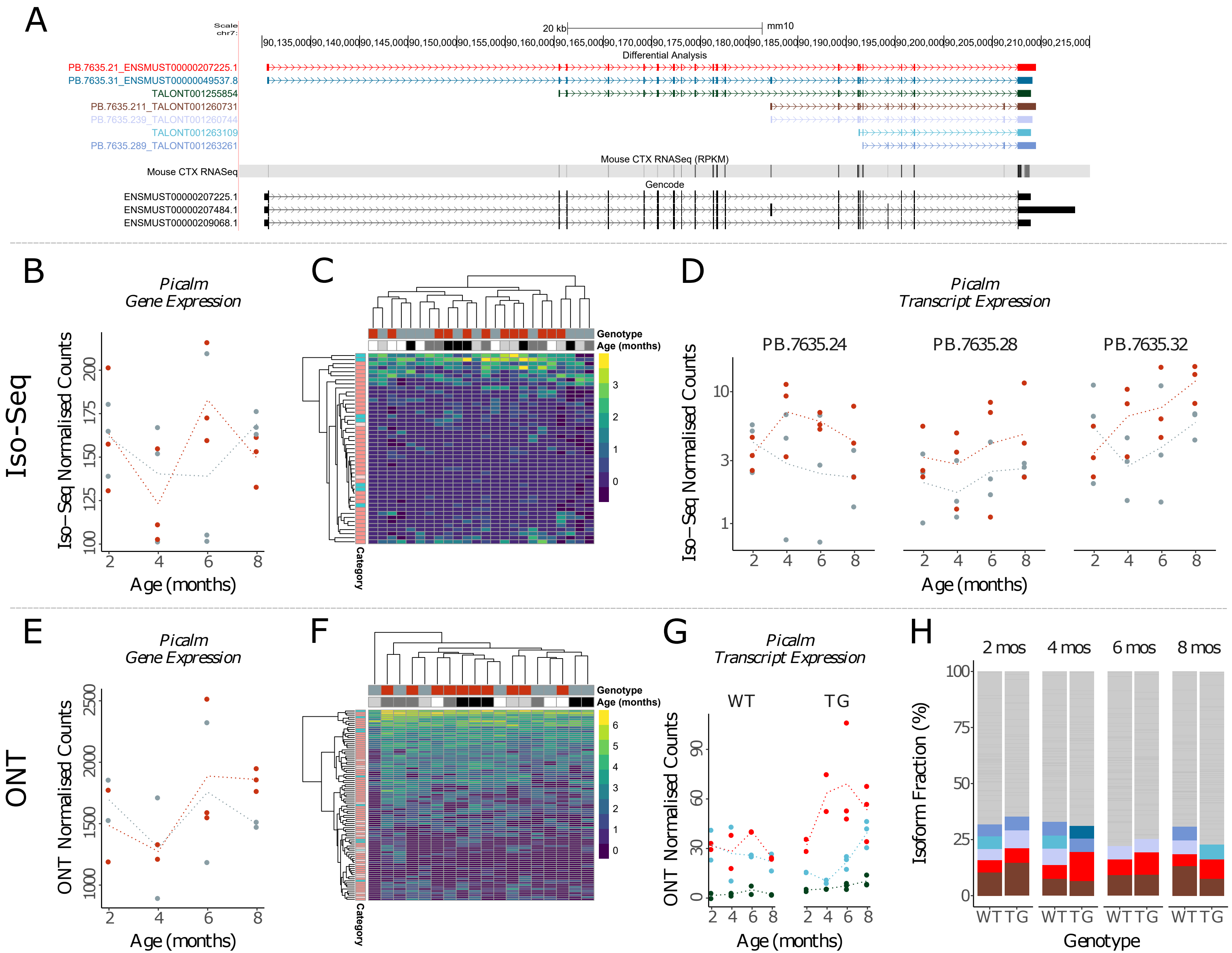


Figure E.10: Differential *Picalm* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Picalm* in the rTg4510 cortex. *Refer to **Figure 6.58** for the same legend.*

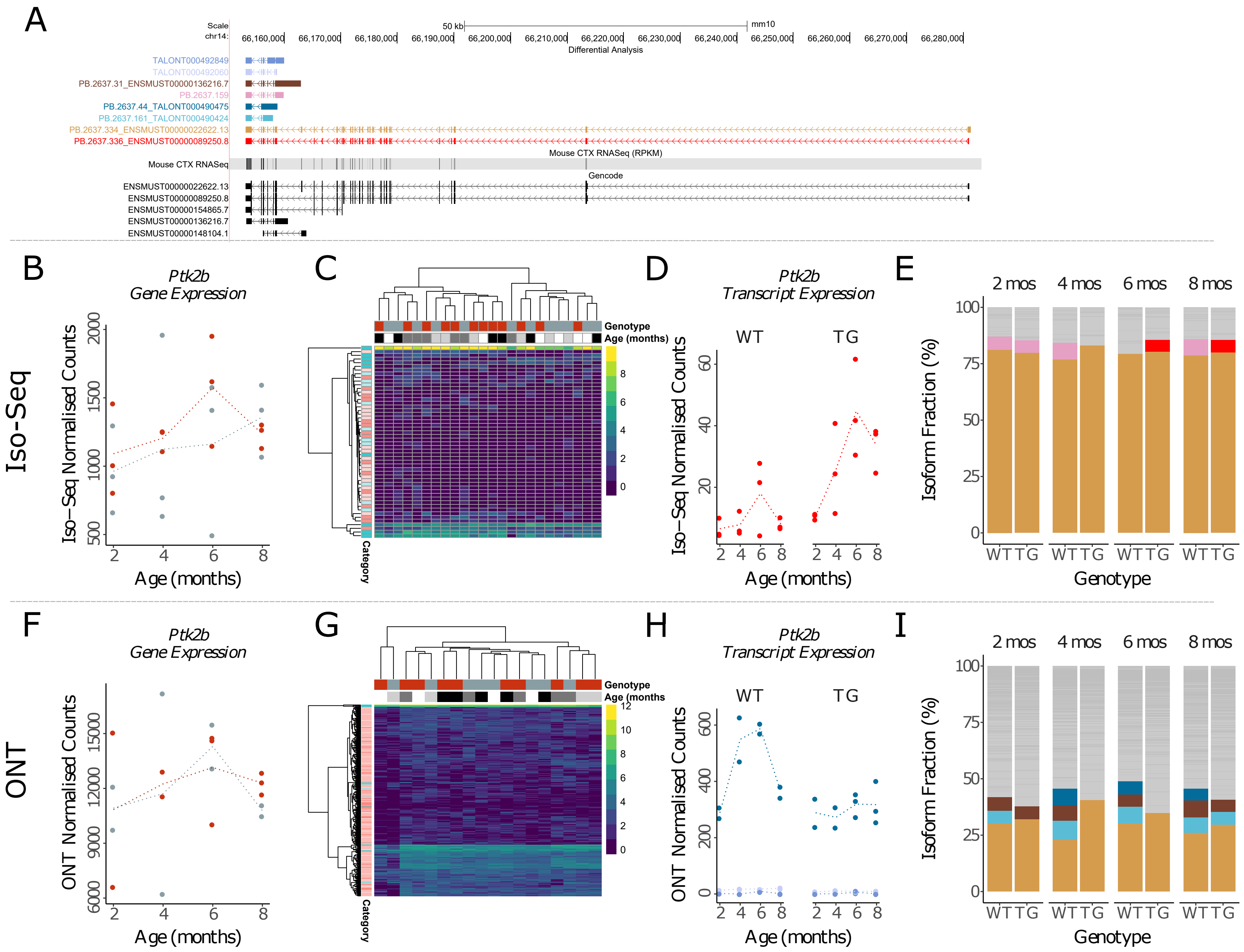


Figure E.11: Differential *Pt2kb* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Pt2kb* in the rTg4510 cortex. *Refer to* ***Figure 6.58*** *for the same legend.*

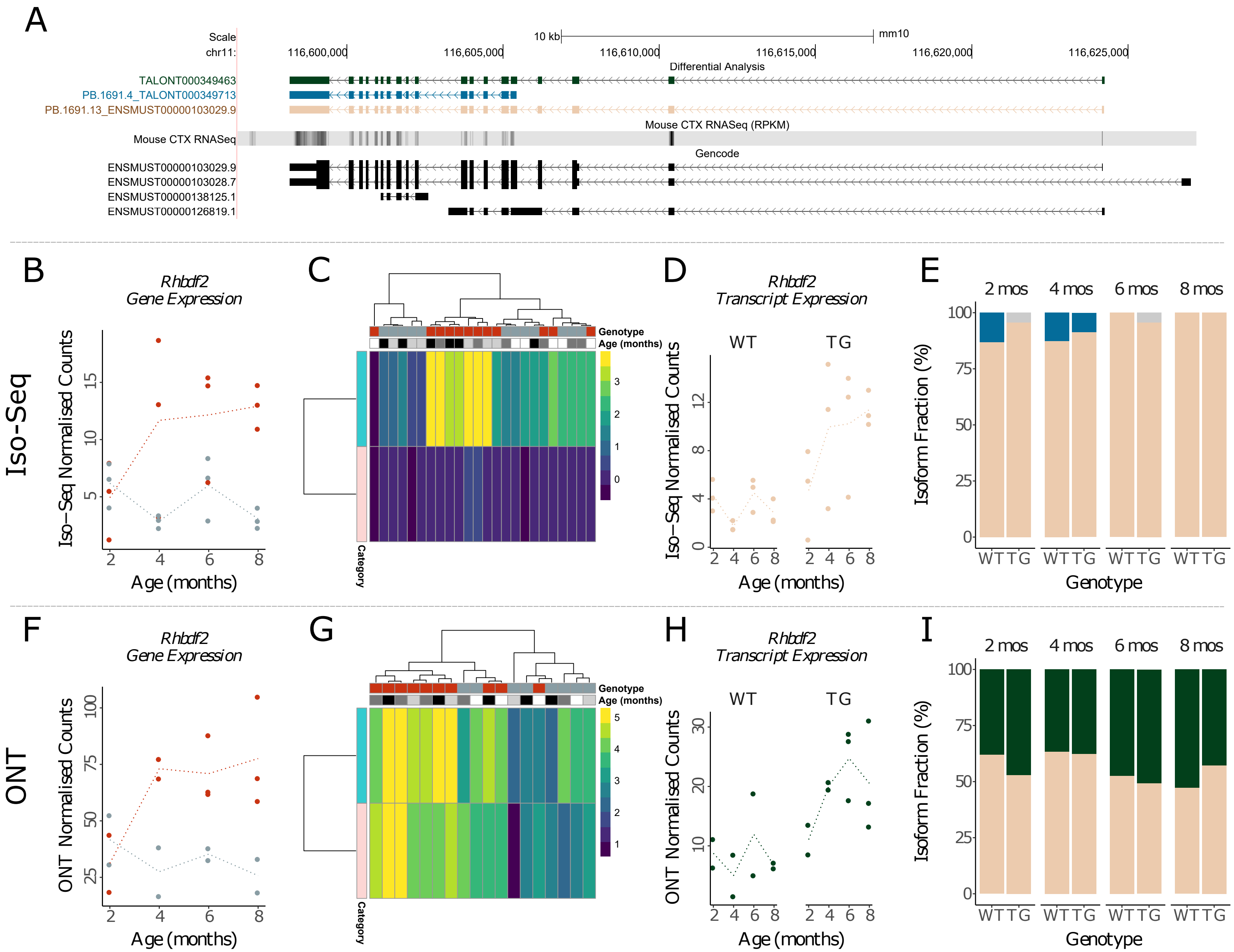


Figure E.12: Differential *Rhbdf2* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Rhbdf2* in the rTg4510 cortex. *Refer to* ***Figure 6.58*** *for the same legend.*

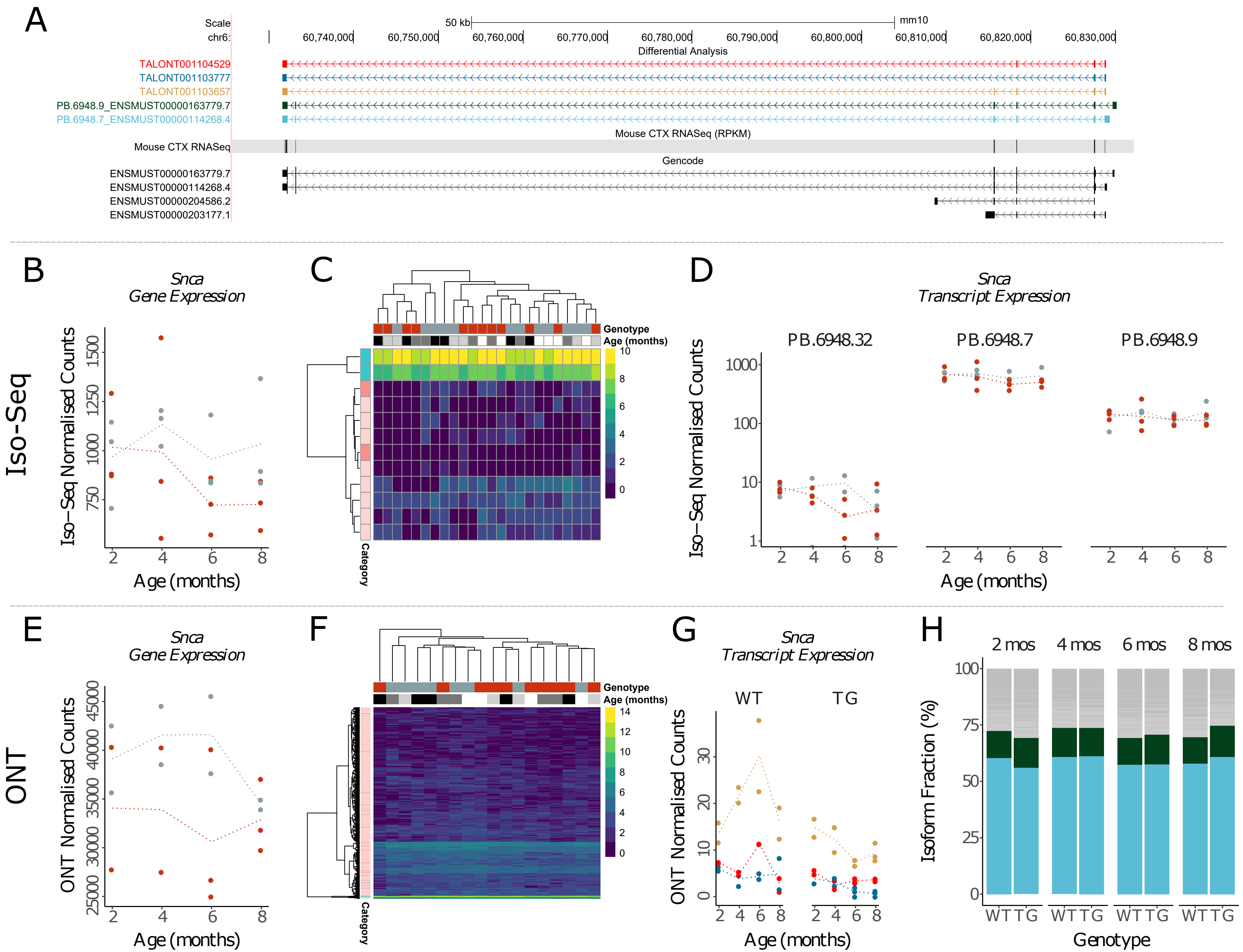


Figure E.13: Differential *Snca* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Snca* in the rTg4510 cortex. *Refer to **Figure 6.58** for the same legend.*

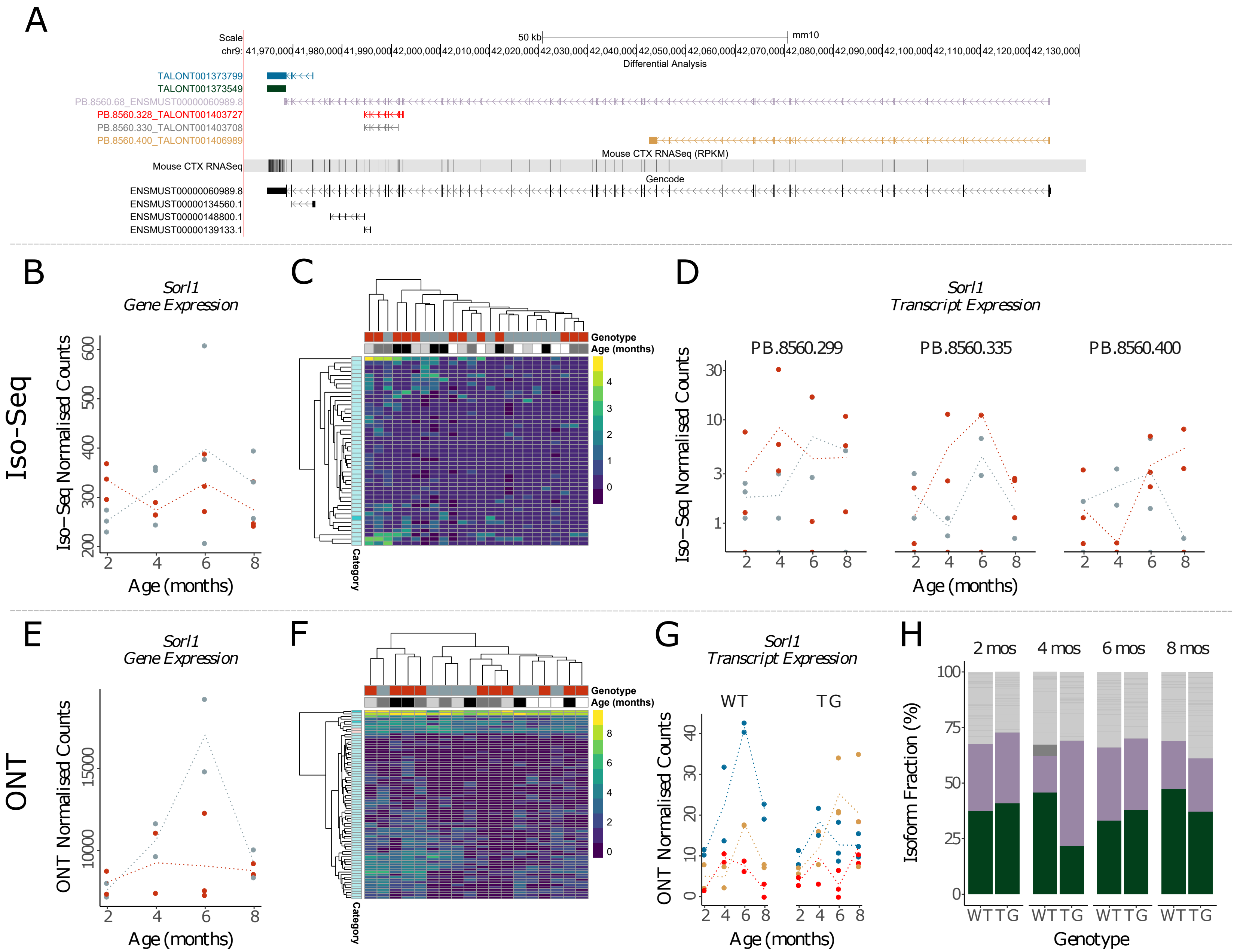


Figure E.14: Differential *Sorl1* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Sorl1* in the rTg4510 cortex. *Refer to **Figure 6.58** for the same legend.*

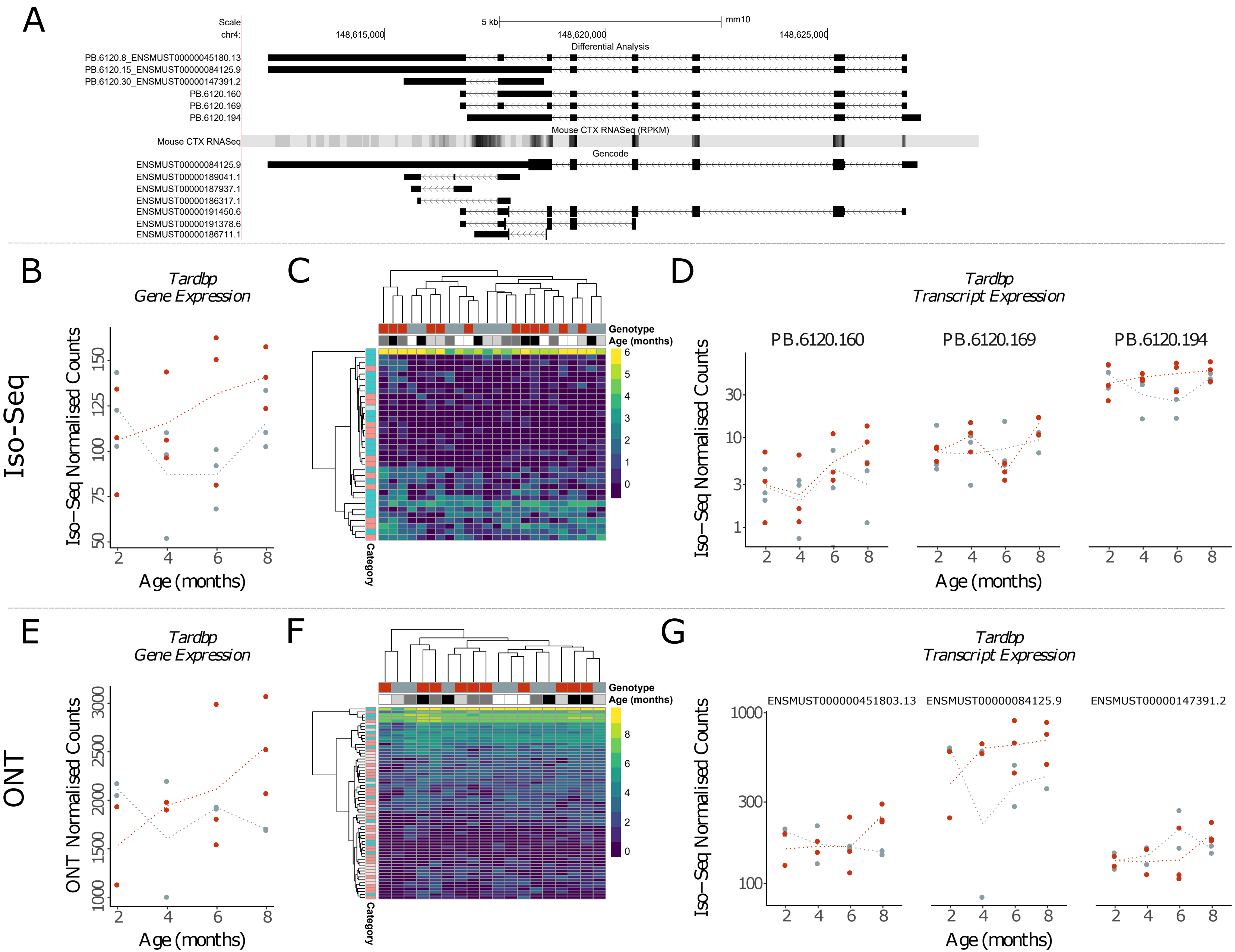


Figure E.15: Differential *Tardbp* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Tardbp* in the rTg4510 cortex. *Refer to **Figure 6.58** for the same legend.*

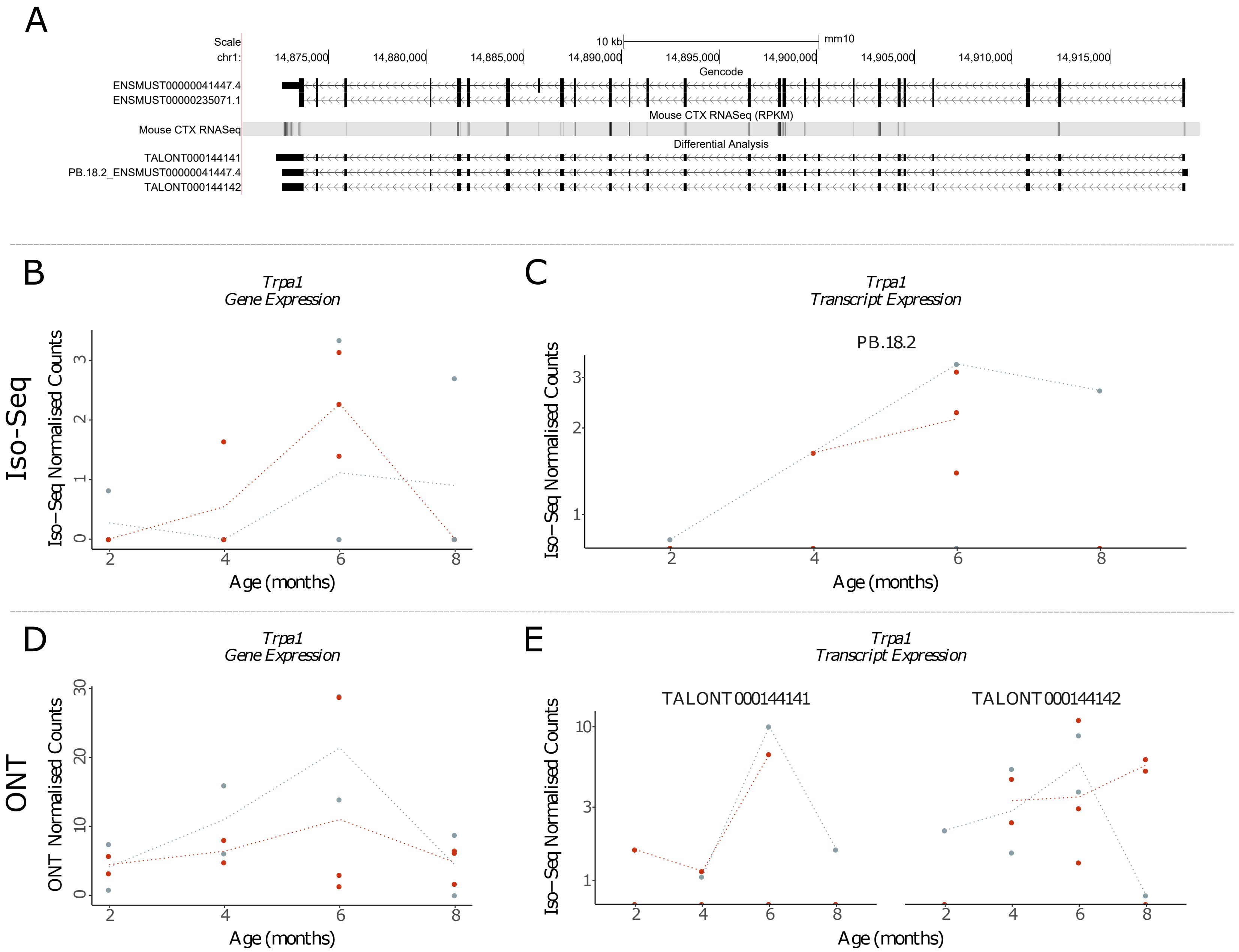


Figure E.16: Differential *Trpa1* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Trpa1* in the rTg4510 cortex. *Refer to* ***Figure 6.58*** *for the same legend.*

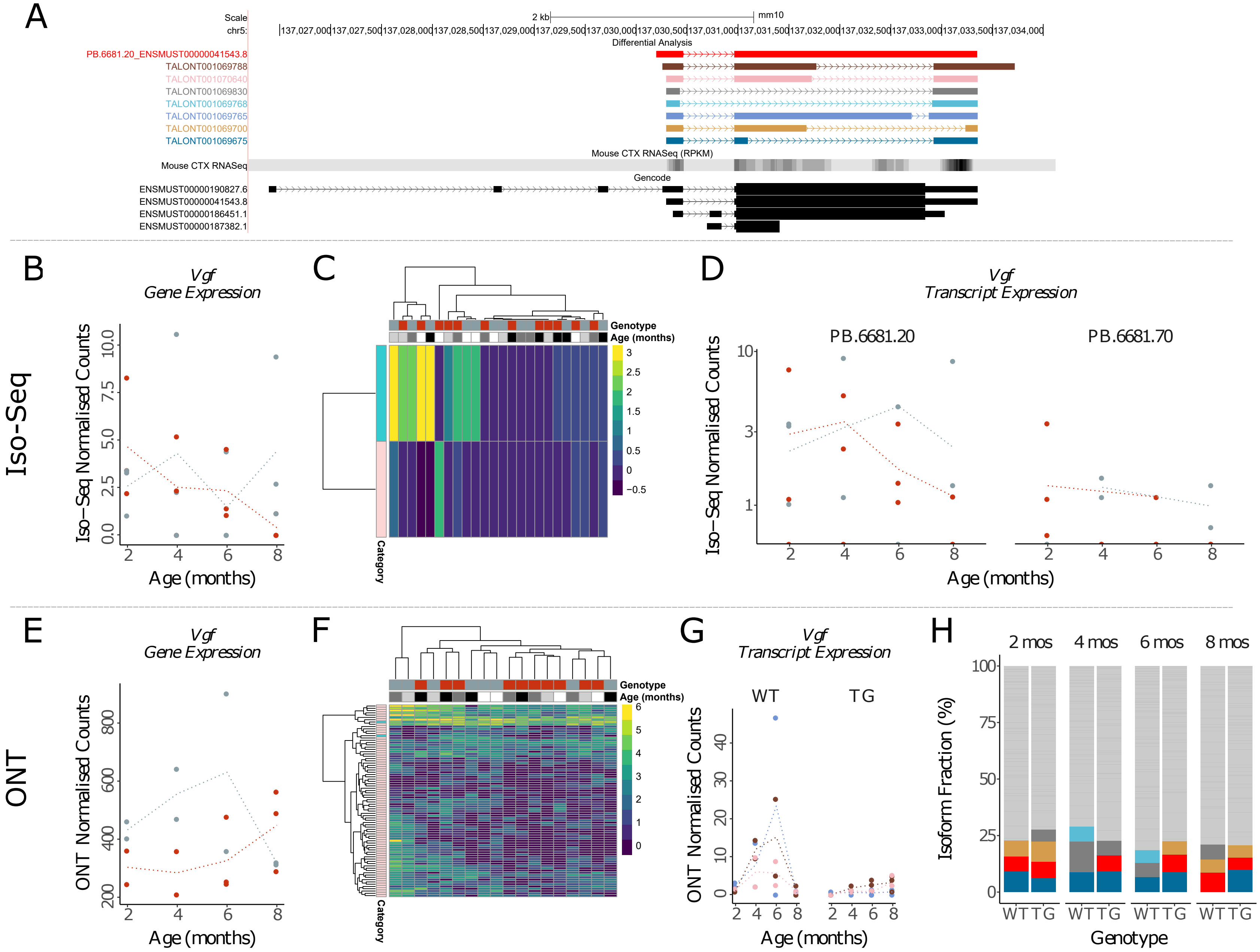


Figure E.17: Differential *Vgf* transcript expression and usage: Shown are plots generated from the differential expression and splicing analyses of *Vgf* in the rTg4510 cortex. *Refer to* ***Figure 6.58*** *for the same legend.*

Appendix F

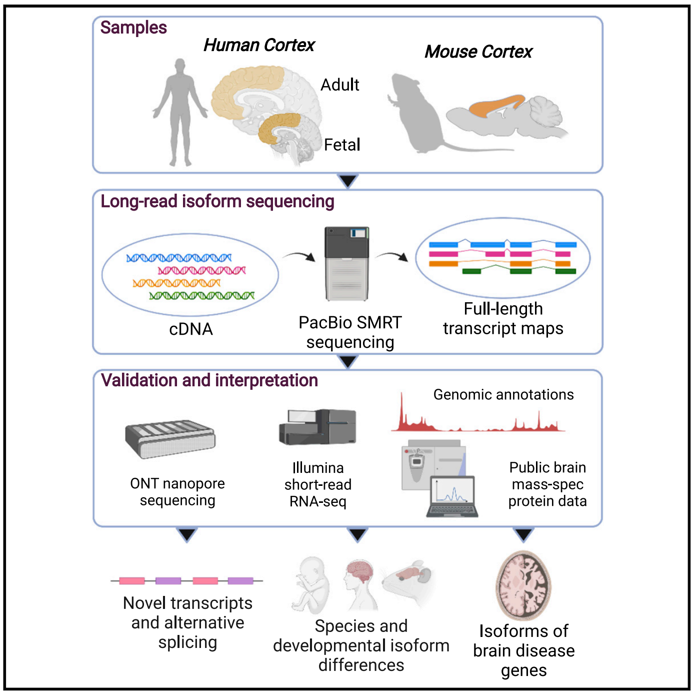
Leung et al. (2021) - Cell Reports paper

Cell Reports

Resource

Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing

Graphical abstract


Samples
Human Cortex
Adult
Fetal
Mouse Cortex
Long-read isoform sequencing
cDNA
PacBio SMRT sequencing
Full-length transcript maps
Validation and interpretation
ONT nanopore sequencing
Illumina short-read RNA-seq
Genomic annotations
Public brain mass-spec protein data
Novel transcripts and alternative splicing
Species and developmental isoform differences
Isoforms of brain disease genes

Highlights

- There is widespread transcript diversity in the cortex and many novel transcripts
- Some genes display big differences in isoform number between human and mouse cortex
- There is evidence of differential transcript usage between human fetal and adult cortex
- There are many novel isoforms of genes associated with human brain disease

Authors

Szi Kay Leung, Aaron R. Jeffries,
Isabel Castanho, ...,
Gloria M. Sheynkman, Eilis Hannon,
Jonathan Mill

Correspondence

J.mill@exeter.ac.uk

In brief

Leung et al. use long-read sequencing to annotate RNA isoforms in the human and mouse cortex. They identify novel transcripts and evidence for differential transcript usage between the fetal and adult cortex. Their data confirm the importance of alternative splicing as a mechanism underpinning gene regulation in the brain.

Leung et al., 2021, Cell Reports *37*, 110022
November 16, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.celrep.2021.110022

Cell Reports

CellPress
OPEN ACCESS

Resource

Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing

Szi Kay Leung,[1,13] Aaron R. Jeffries,[1,13] Isabel Castanho,[1,2,3] Ben T. Jordan,[4] Karen Moore,[1] Jonathan P. Davies,[1] Emma L. Dempster,[1] Nicholas J. Bray,[5] Paul O'Neill,[1] Elizabeth Tseng,[6] Zeshan Ahmed,[7] David A. Collier,[7] Erin D. Jeffery,[4] Shyam Prabhakar,[8] Leonard Schalkwyk,[9] Connor Jops,[10] Michael J. Gandal,[10] Gloria M. Sheynkman,[4,11,12] Eilis Hannon,[1] and Jonathan Mill[1,14,*]

[1]University of Exeter, Exeter, UK
[2]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA, USA
[3]Harvard Medical School, Boston, MA, USA
[4]Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA
[5]School of Medicine, Cardiff University, Cardiff, UK
[6]Pacific Biosciences, Menlo Park, CA, USA
[7]Eli Lilly & Co., Windlesham, UK
[8]Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore
[9]School of Life Sciences, University of Essex, Colchester, UK
[10]Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA, USA
[11]Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
[12]UVA Cancer Center, University of Virginia, Charlottesville, VA, USA
[13]These authors contributed equally
[14]Lead contact
*Correspondence: J.mill@exeter.ac.uk

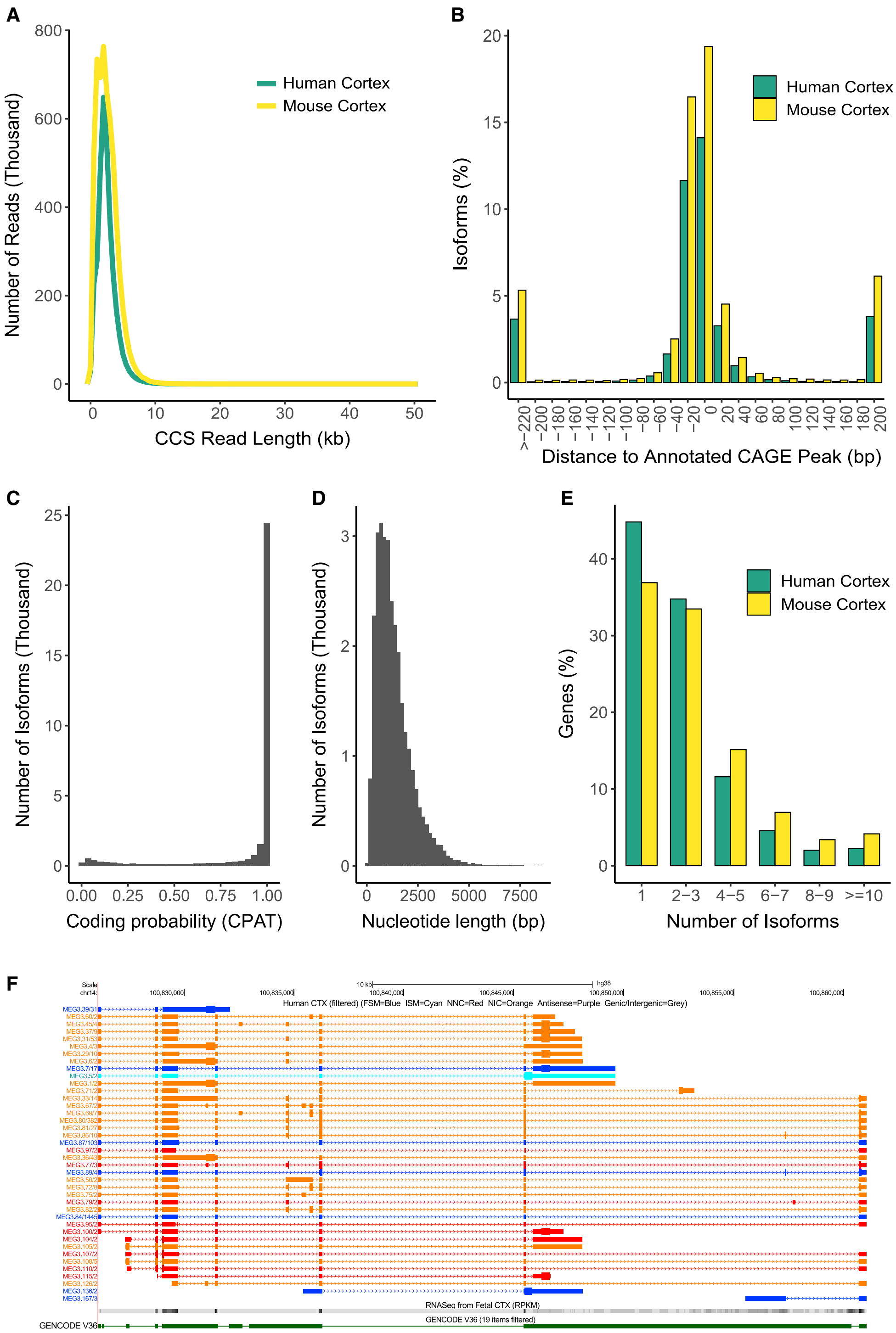
https://doi.org/10.1016/j.celrep.2021.110022

SUMMARY

Alternative splicing is a post-transcriptional regulatory mechanism producing distinct mRNA molecules from a single pre-mRNA with a prominent role in the development and function of the central nervous system. We used long-read isoform sequencing to generate full-length transcript sequences in the human and mouse cortex. We identify novel transcripts not present in existing genome annotations, including transcripts mapping to putative novel (unannotated) genes and fusion transcripts incorporating exons from multiple genes. Global patterns of transcript diversity are similar between human and mouse cortex, although certain genes are characterized by striking differences between species. We also identify developmental changes in alternative splicing, with differential transcript usage between human fetal and adult cortex. Our data confirm the importance of alternative splicing in the cortex, dramatically increasing transcriptional diversity and representing an important mechanism underpinning gene regulation in the brain. We provide transcript-level data for human and mouse cortex as a resource to the scientific community.

INTRODUCTION

Alternative splicing (AS) is a post-transcriptional regulatory mechanism producing multiple RNA isoforms from a single mRNA precursor. In eukaryotes, AS dramatically increases transcriptomic and proteomic diversity from the coding genome and is an important mechanism in the developmental control of gene expression. The mechanisms involved in AS include the use of alternative first (AF) and last (AL) exons, exon skipping (SE), alternative 5′ (A5′) and A3′ splice sites, mutually exclusive exons (MX), and intron retention (IR) (Wang et al., 2008). These phenomena are common, influencing the transcription of >95% of human genes (Pan et al., 2009). Because alternatively spliced transcripts from a single gene can produce proteins with different functions (Eksi et al., 2013; Yang et al., 2016), there is increasing interest in their role in human disease (Wang and Cooper, 2007). Of note, the correction of AS deficits has been shown to have therapeutic benefit in several disorders including spinal muscular atrophy (Wan and Dreyfuss, 2017). AS appears to be particularly important and prevalent in the central nervous system (CNS) (GTEx Consortium, 2015), where it impacts neurodevelopment (Mazin et al., 2013), aging (Tollervey et al., 2011), and key neural functions (Raj and Blencowe, 2015). AS is a common feature of many neuropsychiatric and neurodegenerative diseases (Mills and Janitz, 2012), with recent studies highlighting splicing differences associated with autism (Parikshak et al., 2016),



(legend on next page)

schizophrenia (SZ) (Takata, Matsumoto and Kato, 2017), and Alzheimer's disease (AD) (Raj et al., 2018).

Characterizing the full complement of isoforms across tissues and development is important for understanding transcriptional variation in health and disease. For example, transcript-level annotation can be used to improve the understanding of the functional consequences of rare genetic variants (Cummings et al., 2020). However, efforts to fully characterize RNA isoform diversity are constrained by standard RNA sequencing (RNA-seq) approaches, which generate short reads that cannot span full-length transcripts (Steijger et al., 2013). Recent advances in long-read sequencing have addressed these challenges; Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies (ONT) nanopore sequencing can generate reads >10 Kb, enabling direct assessment of alternatively spliced transcripts (Amarasinghe et al., 2020).

In this study, we systematically characterize RNA isoform diversity in the cerebral cortex, a key region of the brain involved in perception, cognition, and consciousness. We first use the PacBio isoform sequencing (Iso-Seq) approach (Gordon et al., 2015) to generate full-length cDNA sequences from the human and mouse cortex. We identify widespread transcript diversity with the detection of novel transcripts not previously described in existing genomic annotations, including in genes associated with neuropsychiatric and neurodegenerative disease. We subsequently use ONT nanopore sequencing and short-read RNA-seq to validate and complement our Iso-Seq data. We find widespread evidence of different AS events and examples of fusion genes representing read-through transcription between adjacent genes. A comparison of human and mouse cortex identified species-specific transcript diversity, and a comparison of fetal and adult human cortex highlighted developmental changes in AS and transcript expression. Our data confirm the importance of AS in the cortex, dramatically increasing annotated transcriptional diversity and representing an important mechanism underpinning gene regulation in the brain. Our transcript annotations and sequencing data are available as a resource to the research community via browsable tracks and a searchable transcript visualization database.

RESULTS

Methodological overview

An overview of the methods and datasets used in our study is given in Figure S1. PacBio Iso-Seq data were generated on RNA isolated from human cortex tissue (n = 7) dissected from fetal (n = 3, mean age = 16 weeks post-conception [WPC], range = 14–17 WPC) and adult (n = 4, mean age = 61.8 years, range = 24–89 years) donors (Table S1). Raw reads were processed using the *Iso-Seq* pipeline (Gordon et al., 2015), mapped to the genome, and clustered using *cDNA Cupcake* followed by *SQANTI2* (Tardaguila et al., 2018) annotation (Table S2). In parallel, we generated a mouse cortex Iso-Seq dataset (n = 12, mean age = 5 months, range = 2–8 months; Table S1) and also profiled tissue from two additional human brain regions (hippocampus and striatum). Rarefaction curves confirmed that our coverage of RNA isoform diversity is representative of the population of transcripts present (Figures S2A–S2F). All downstream analyses and statistics reported in our manuscript are based on the subset of *SQANTI2*-filtered transcripts unless otherwise indicated, although the extended (unfiltered) datasets are available as genome browser tracks as a resource. To validate the transcripts identified using Iso-Seq, we generated short-read RNA-seq data (human: n = 3; mouse: n = 12) and additional full-length transcriptome data using nanopore sequencing (ONT) in a subset of samples (human: n = 2; Table S1). Taken together, our analysis represents the most comprehensive characterization yet undertaken of full-length transcripts and transcript diversity in the human and mouse cortex.

Iso-Seq identifies widespread transcript diversity in the human cortex

We obtained a total of 3.30 M (million) circular consensus sequence (CCS) reads from the human cortex samples (Table S3), with the majority of reads 2 to 3 kb in length (mean length = 2.46 kb; Figure 1A; Figures S3A–S3C), corresponding to the mean length of mRNA in the human genome (Piovesan et al., 2019). Following stringent quality control (QC), these reads mapped to 12,910 "annotated" genes (Table 1) with expression patterns reflecting those expected for the cortex; using the Human Gene Atlas database (Kuleshov et al., 2016), the 500 most abundantly expressed genes were most enriched for "prefrontal cortex" (odds ratio = 5.99, adjusted p = 9.18×10^{-24}) (Table S4). In total, we identified 32,802 unique transcripts (mean length = 2.77 kb, SD = 1.29 kb, range = 0.104–11.8 kb) in the human cortex (Table 1); as expected, these were enriched near Cap Analysis Gene Expression (CAGE) peaks from the FANTOM5 dataset (Lizio et al., 2019) (median distance from a CAGE peak = −1 bp with 25,762 [78.5%] of transcripts located within 50 bp of a CAGE peak) (Figure 1B) and were also located proximally to annotated transcription start sites and transcription termination sites (Figures S4B and S4C). Using the Coding-Potential Assessment Tool (CPAT) (Wang et al., 2013) to characterize open reading frames (ORFs) among detected transcripts, we identified a high level of coding potential: 29,998 (91.5%) of the

Figure 1. Generation of high-quality long-read transcriptome datasets for human and mouse cerebral cortex

(A) The distribution of CCS read lengths in our human (n = 7 biologically independent samples) and mouse (n = 12 biologically independent samples) cortex Iso-Seq datasets. The distribution of CCS read lengths for individual samples can be found in Figure S3.

(B) Distance between transcription start site (TSS) and closest annotated CAGE peak. A negative value refers to a CAGE peak located upstream of a TSS.

(C) The distribution of coding potential scores for all transcripts detected in the human cortex.

(D) The ORF lengths for transcripts predicted to be protein-coding. Equivalent plots for mouse cortex can be found in Figures S7A and S7B.

(E) The number of isoforms identified per gene detected in the human and mouse cortex.

(F) UCSC genome browser track of transcripts annotated to *MEG3* in the human cortex. Transcripts are colored based on *SQANTI2* classification categories (blue = FSM; cyan = ISM; red = NIC; orange = NNC).

Table 1. An overview of the whole-transcriptome Iso-Seq datasets generated on human and mouse cerebral cortex

	Human cortex	Mouse cortex	Adult cortex	Fetal cortex
Unique genes	12964	14684	11021	9679
Annotated genes (%)	12910 (99.58)	14482 (98.62)	10987 (99.69)	9660 (99.8)
Novel genes (%)	54 (0.42)	202 (1.38)	34 (0.31)	19 (0.2)
Isoforms	32802	46626	22048	18612
Genes with >1 isoform (%)	7155 (55.19)	9266 (63.1)	5003 (45.4)	4200 (43.39)
Genes with >10 isoforms (%)	205 (1.58)	466 (3.17)	66 (0.6)	50 (0.52)
Protein-coding transcripts (%)	30411 (92.71)	43530 (93.36)	20537 (93.15)	17464 (93.83)
Non-protein-coding transcripts (%)	2391 (7.29)	3096 (6.64)	1511 (6.85)	1148 (6.17)
Known transcripts (FSM, ISM) (%)	20832 (63.51)	23530 (50.47)	15659 (71.02)	13177 (70.8)
Novel transcripts (%)	11970 (36.49)	23096 (49.53)	6389 (28.98)	5435 (29.2)
FSM (%)	17080 (52.07)	19803 (42.47)	13007 (58.99)	11346 (60.96)
ISM (%)	3752 (11.44)	3727 (7.99)	2652 (12.03)	1831 (9.84)
NIC (%)	8721 (26.59)	13763 (29.52)	4464 (20.25)	4315 (23.18)
NNC (%)	3021 (9.21)	8751 (18.77)	1796 (8.15)	1041 (5.59)
Genic genomic (%)	35 (0.11)	62 (0.13)	20 (0.09)	8 (0.04)
Antisense (%)	31 (0.09)	119 (0.26)	22 (0.1)	7 (0.04)
Fusion (%)	136 (0.41)	297 (0.64)	74 (0.34)	51 (0.27)
Intergenic (%)	26 (0.08)	104 (0.22)	13 (0.06)	13 (0.07)
Genic intron (%)	0 (0)	0 (0)	0 (0)	0 (0)

FSM = full splice match; ISM = incomplete splice match; NIC = novel in catalogue; NNC = novel not in catalogue.

detected transcripts were predicted to be protein-coding (Figure 1C) with an average ORF length of 1,327 nucleotides (Figure 1D). A wide range in the number of multi-exonic RNA isoforms was identified per gene (n = 1–40; Table S5), with over half of all detected genes (n = 7,155 [55.2%]) characterized by more than one isoform, and a notable proportion characterized by more than ten isoforms (n = 205 [1.58%]) (Figure 1E). *MEG3,* a maternally expressed imprinted long non-coding RNA (lncRNA) gene involved in synaptic plasticity (Tan et al., 2017), displayed the greatest isoform diversity in human cortex (40 isoforms; Figure 1F). Gene Ontology (GO) analysis showed that the most enriched molecular function among the 100 most isoformic genes in human cortex was ‘‘pre-mRNA binding’’ (human cortex: odds ratio = 31.8, adjusted $p = 2.39 \times 10^{-3}$) (Table S4), an interesting observation given the role that RNA-binding proteins (RBPs) themselves play in regulating tissue-specific patterns of AS (Fu and Ares, 2014). The number of detected isoforms was correlated with both gene length (corr = 0.19, $p = 1.52 \times 10^{-106}$; Figure S5A) and the number of exons (corr = 0.24, $p = 7.97 \times 10^{-155}$; Figure S5E), with these relationships being stronger among ‘‘highly expressed’’ (> 2.5 Log_{10} transcripts per million [TPM]) genes (correlation between isoform number and gene length = 0.49; $p = 1.39 \times 10^{-33}$; correlation between isoform number and number of exons = 0.45, $p = 7.42 \times 10^{-28}$; Figures S5C and S5G), reflecting the additional sensitivity for detecting transcripts of highly expressed genes.

Novel transcripts were detected for a large proportion of expressed genes in the human cortex

Among full-length transcripts annotated to known genes (n = 32,745 transcripts) in the human cortex, the majority were characterized either as a complete full splice match (FSM: n = 17,080 [52.2%]) or incomplete splice match (ISM: n = 3,752 [11.4%]) to existing annotations in GENCODE (hg38) (Figure 2B; Table S6). A significant proportion of transcripts, however, represented ‘‘novel’’ transcripts not present in existing annotation databases (Table S7): 11,913 transcripts (36.4%) associated with 5,327 (41.5%) genes were classified as ‘‘novel’’ (mean size = 2.84 kb, SD = 1.2 kb, range = 0.104–11.2 kb, mean number of exons = 11.1) (Figure 2B; Figure S6A). Most of these novel transcripts contained a combination of known donor and acceptor splice sites and were classified as ‘‘novel in catalog’’ (NIC: n = 8,721, 73.2% of all novel transcripts of known genes). The remaining novel transcripts were predominantly classified as ‘‘novel not in catalog’’ (NNC), with at least one novel donor or acceptor site (n = 3,021, 25.4% of all novel transcripts of known genes). Novel transcripts were generally less abundant than annotated transcripts (Mann-Whitney-Wilcoxon test: $W = 1.62 \times 10^{8}$, $p < 2.23 \times 10^{-308}$; Figures S6C and S6D) and presumably harder to detect using standard RNA-seq approaches (Conesa et al., 2016). Novel transcripts were also longer ($W = 1.10 \times 10^{8}$, $p = 4.04 \times 10^{-25}$) and had more exons ($W = 8.84 \times 10^{7}$, $p < 2.23 \times 10^{-308}$) (Figures S6E and S6F). Finally, the majority of novel transcripts (n = 9,538 transcripts, 80% of novel transcripts) were within 50 bp of an annotated CAGE peak from the FANTOM5 database (Figure S4B).

NIC, NNC, and ISM transcripts were characterized by a similar distribution of predicted ORF lengths and CPAT coding probability scores to FSM transcripts, although the protein coding potential of NIC and NNC transcripts was marginally lower (Figures 2C and 2D). We used public mass spectrometry (MS)-based human cortex proteomics data to look for evidence of translation of NIC

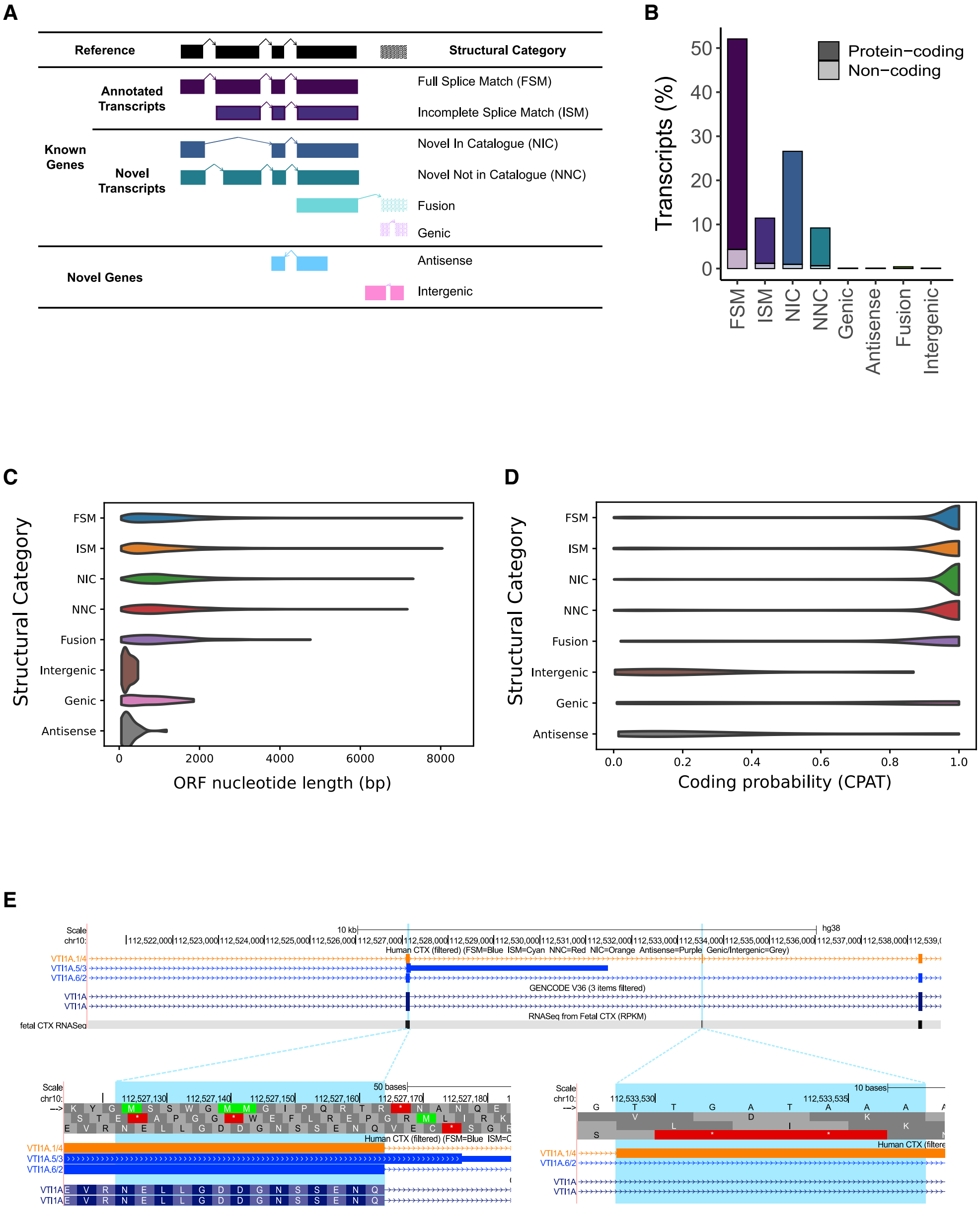


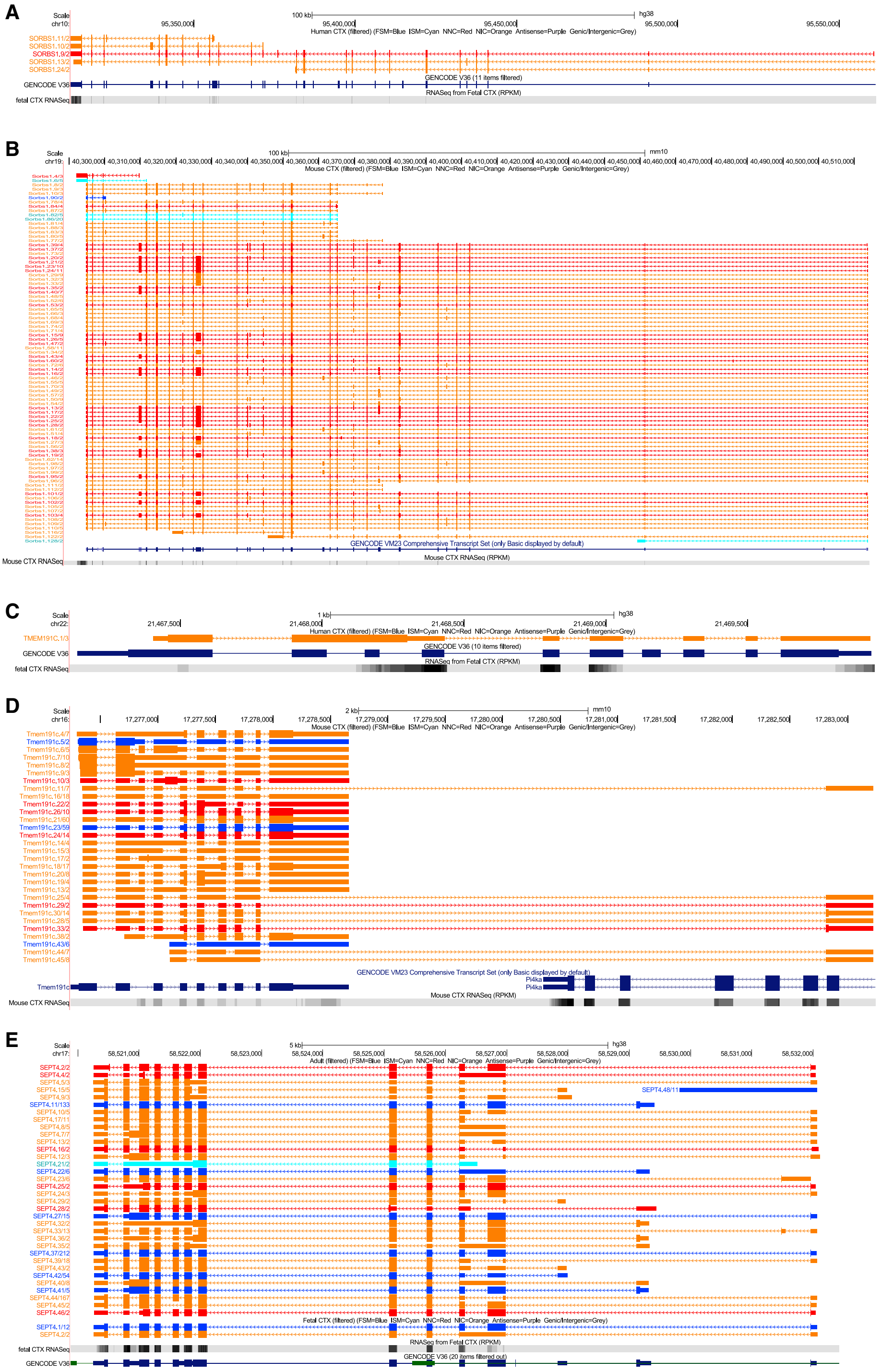
Figure 2. A large proportion of cortical transcripts are not described in existing annotations

(A) A transcript was classified as “FSM” if it aligned with the reference genome with the same splice junctions and contained the same number of exons; “ISM” if it contained fewer 5′ exons than the reference genome; “NIC” if it represented a novel transcript containing a combination of known donor or acceptor sites; and “NNC” if it represented a novel transcript with at least one novel donor or acceptor site.

(B) Approximately half of all transcripts identified in the human cortex were FSM, with a large proportion of transcripts assigned as being novel (NIC, NNC).

(C and D) Distribution of (C) ORF length and (D) coding probability of transcripts by category. A similar ORF length and CPAT probability score profile was observed for FSM, NIC, and NNC transcripts. Equivalent plots for mouse cortex can be found in Figures S7C and S7D.

(E) Shown is a UCSC genome browser track of *VTI1A* in the human cortex. Interrogation of human protein data identified a peptide (NELLGDDGNSSENQLIK, highlighted blue) that confirmed inclusion of a novel exon.



(legend on next page)

and NNC transcripts. Briefly, using the ORFs predicted from CPAT, we assembled a cortex-specific full-length protein database and searched the results against a bottom-up proteomics dataset generated from adult and fetal human brain cortex samples. We found examples of novel peptides, each mapping uniquely to one or more novel transcript(s), providing evidence for the stable translation of these isoforms in the cortex (Table S8); Figure 2E shows a peptide assigned to a novel transcript of *VTI1A*—a gene encoding a soluble N-ethylmaleimide-sensitive factor attachment protein receptor with neuron-specific functions (Tang, 2020)—providing evidence for translation of a protein isoform with a novel exon.

Overall patterns of transcript diversity are similar between human and mouse cortex

We generated a parallel Iso-Seq dataset on mouse cortex, obtaining 5.66 M CCS reads with similar size profiles (mean length = 2.57 kb; Figure 1A; Figure S3B) to those seen in human cortex. These reads mapped to 14,482 annotated genes (Table 1), with the 500 most abundantly expressed genes being primarily enriched for "cerebral cortex" genes in the Mouse Gene Atlas database (Kuleshov et al., 2016) (odds ratio = 6.07, adjusted $p = 6.8 \times 10^{-17}$; Table S4). We identified 46,626 unique transcripts (mean length = 3.18 kb, SD = 1.68 kb, range = 0.083–15.9 kb) in the mouse cortex (Table 1), which were again enriched near CAGE peaks (median distance from CAGE peak = -1 bp, 35,262 [75.6%] transcripts located within 50 bp of a CAGE peak; Figure 1B). A wide range in the number of multi-exonic RNA isoforms was also identified per gene (1 to 86) (Table S5), with a similar distribution to that observed in the human cortex (n = 9,266 genes [63.1%] with more than one isoform, n = 466 [3.17%] with more than ten isoforms) (Figure 1E). The number of detected RNA isoforms was also correlated with both gene length (corr = 0.25, $p = 1.33 \times 10^{-197}$; Figure S5B) and exon number (corr = 0.25, $p = 4.02 \times 10^{-193}$; Figure S5F), with a stronger relationship observed among "highly expressed" genes (Figures S5D and S5H). As in the human cortex, we identified a large proportion of novel transcripts associated with known genes (n = 22,873 [49.3%], mean size = 3.28 kb, SD = 1.61 kb, range = 0.182–15.0 kb, mean number of exons = 12.4), with the vast majority identified as either NIC (n = 13,763 [60.2%]) or NNC (n = 8,751 [38.3%]) (Figure S6A). They were also less abundant ($W = 3.66 \times 10^8$, $p < 2.23 \times 10^{-308}$), longer ($W = 2.37x\ 10^8$, $p = 2.13 \times 10^{-42}$), and had more exons ($W = 1.94 \times 10^8$, $p < 2.23 \times 10^{-308}$) than already known transcripts (Figures S6C–S6H), with the majority (n = 17,252 [75.4%]) mapping to within 50 bp of an annotated CAGE peak (Figure S4A). Finally, predicted coding potential across different transcript categories reflected those observed in human cortex (Figures S7A–S7D).

A subset of genes is characterized by major differences in transcript diversity between human and mouse cortex

Although previous studies have highlighted evidence of major splicing diversity between human and mouse (Ule and Blencowe, 2019), we found that among multi-exonic genes, for which transcripts were detected in both human and mouse cortex (n = 10,202 genes; Figure S8A), the number of isoforms detected for each gene was significantly correlated between species (corr = 0.51, $p < 2.23 \times 10^{-308}$; Figure S8C). There was a stronger relationship among highly expressed genes (> 2.5 Log_{10} TPM in both species, corr = 0.64, $p = 1.21 \times 10^{-25}$; Figure S8E), a possible reflection of a deeper sequencing coverage of these genes. Despite the overall stability in cortical RNA isoform diversity between human and mouse, there were striking exceptions for specific genes (Table S5). *SORBS1*(Figures 3A and 3B) and *ARPP21* (Figures S9A and S9B) had the largest absolute difference in numbers of isoforms detected between human and mouse. *LPAR2* had the highest *relative* number of isoforms detected in human cortex (n = 12 isoforms; Figure S9C) compared to mouse cortex (1 isoform; Figure S9D) Figures S9, whereas *Tmem191c* had the highest *relative* number of isoforms in mouse cortex (n = 30 isoforms; Figure 3C) compared to human cortex (1 isoform; Figure 3D).

Comparisons with short-read RNA-seq data and nanopore sequencing confirms the accuracy and sensitivity of Iso-Seq

Although Iso-Seq is accurate at characterizing RNA diversity (Wang et al., 2019), its sensitivity for quantifying gene expression has not been systematically explored. We generated highly parallel RNA-seq data on a subset of samples (Table S9), finding a strong correlation between gene-level expression quantified using the two methods in both datasets (human fetal cortex: n = 9,221 genes, corr = 0.54, $p < 2.23 \times 10^{-308}$; mouse cortex: n = 13,923 genes, corr = 0.71, $p < 2.23 \times 10^{-308}$; Figures S10A and S10C). To further assess the quantitative accuracy of Iso-Seq, we included External RNA Controls Consortium (ERCC) spike-in control molecules into our mouse cDNA libraries. Among the detected ERCC transcripts, we found a near-perfect correlation between full-length Iso-Seq reads and the actual amount of control used (corr = 0.98, $p = 1.42 \times 10^{-41}$; Figure S10F), highlighting the power of Iso-Seq to accurately quantify the abundance of highly expressed transcripts. The vast majority of unique splice junctions identified in our Iso-Seq data were supported by RNA-seq in both human (n = 89,975 [99.4%] junctions) and mouse (n = 152,872 [98.1%] junctions). For transcripts that could be recapitulated in the matched RNA-seq data, there was a significant correlation between transcript expression levels quantified using both sequencing

Figure 3. A subset of genes are characterized by dramatic differences in cortical transcript diversity between species (human and mouse) and between developmental stages (fetal and adult)

(A–E) UCSC genome browser tracks showing transcripts detected for (A) *SORBS1* in human cortex (n = 5 transcripts); (B) *Sorbs1* in mouse cortex (n = 86 transcripts); (C) *TMEM191C* in human cortex (n = 1 transcript); (D) *Tmem191c* in mouse cortex (n = 30 transcripts); and (E) *SEPT4* in human adult cortex (n = 34 transcripts) and human fetal cortex (n = 2 transcripts).

Additional examples of genes with considerable differences in the number of transcripts between human and mouse cortex are shown in Figures S9A–S9D. Additional examples of genes with considerable differences in the number of transcripts between fetal and adult cortex are shown in Figures S16A and S16B. For each gene, RNA-seq data tracks from human cortex (n = 3 samples) and mouse cortex (n = 12 samples) are also displayed. Transcripts are colored based on *SQANTI2* classification categories (blue = FSM; cyan = ISM; red = NNC; orange = NIC).

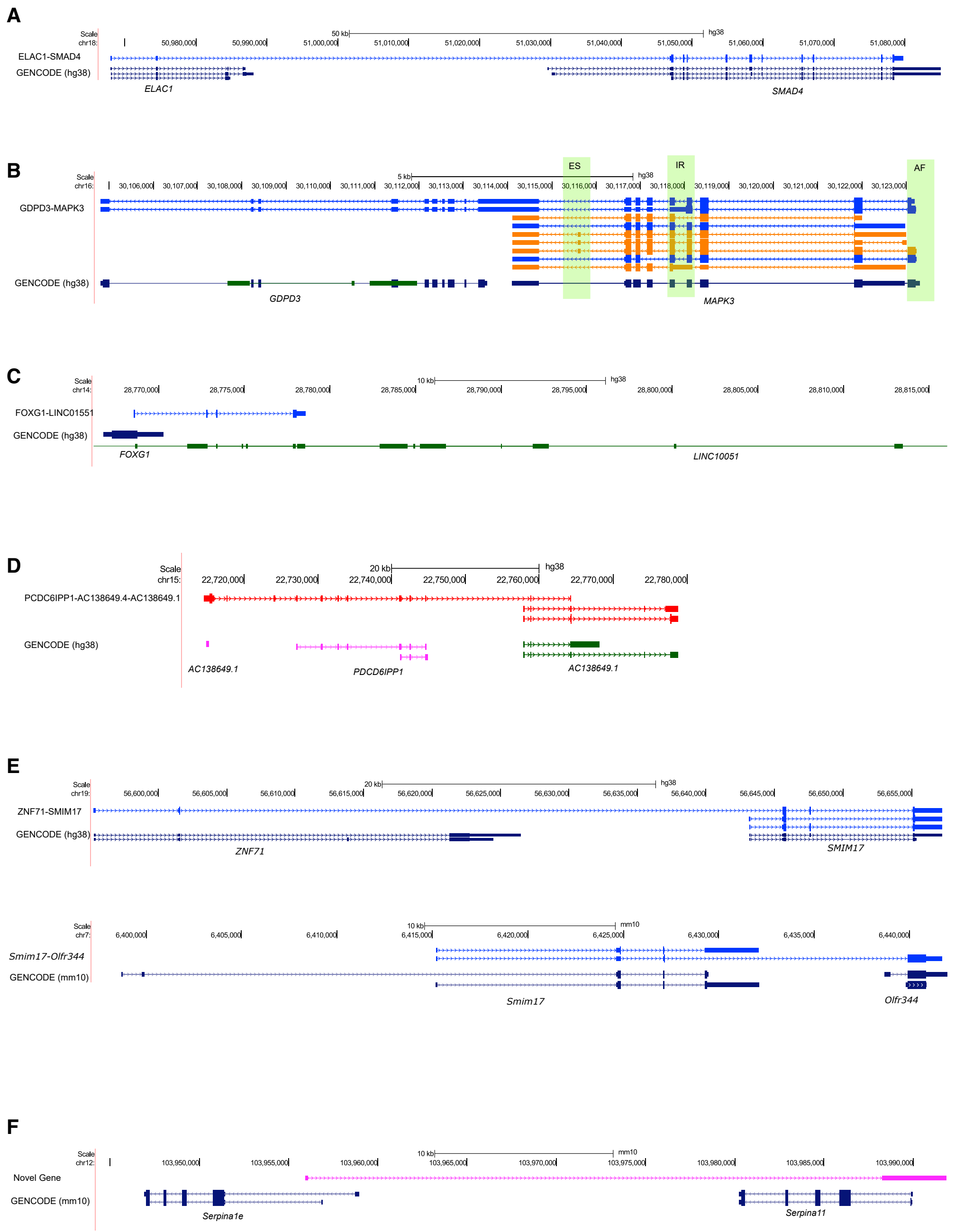


Figure 4. Examples of fusion transcripts in the cortex

(A) A fusion transcript incorporating exons from *ELAC1* and *SMAD4* in the human cortex.

(B) Two read-through transcripts incorporating exons from *MAPK3* and *GDPD3* in the human cortex. Of note, one of the fusion transcripts is characterized by intron retention, as observed in another novel isoform of *MAPK3*.

(C) A fusion transcript incorporating exons from *FOXG1* and *LINC01551* in the human cortex.

(D) A fusion transcript incorporating exons across three pseudogenes in the human cortex.

(legend continued on next page)

approaches (human cortex: n = 17,583 transcripts, corr = 0.40, $p < 2.23 \times 10^{-308}$; mouse cortex: n = 41,488 transcripts, corr = 0.48, $p < 2.23 \times 10^{-308}$; Figures S10B and S10D), further highlighting that transcript abundance can be reliably quantified using Iso-Seq.

Using our Iso-Seq data as a scaffold, we generated a reference-guided transcriptome assembly from our mouse cortex RNA-seq data using *Stringtie* (Pertea et al., 2015). Many of the isoforms reconstructed from RNA-seq reads appeared to represent incomplete fragments of full-length transcripts identified in Iso-Seq. Overall, isoforms assembled using RNA-seq reads had a significantly shorter mean length (RNA-seq: 2.31 kb versus Iso-Seq: 3.18 kb, t = 71.9, $p < 2.2 \times 10^{-16}$), lower average number of exons (RNA-seq: 7.30 versus Iso-Seq: 10.8, t = 76.7, $p < 2.2 \times 10^{-16}$), and were less likely to be located within a CAGE peak (RNA-seq: 34.0% versus Iso-Seq: 71.9%, Fisher's exact test = $p < 2.2 \times 10^{-16}$, odds ratio = 4.97) (Figures S11A and S11B). Importantly, more than 50% of isoforms robustly detected using Iso-Seq could not be readily recapitulated using standard RNA-seq, highlighting the advantage of long-read sequencing for characterizing isoform diversity (Figure S11C). Finally, a large proportion of novel transcripts identified using Iso-Seq (n = 6,417 [53.78%]) were also detected with ONT nanopore sequencing (40.7 M reads) from a subset of samples.

Several cortex-expressed transcripts represent fusion events between neighboring genes

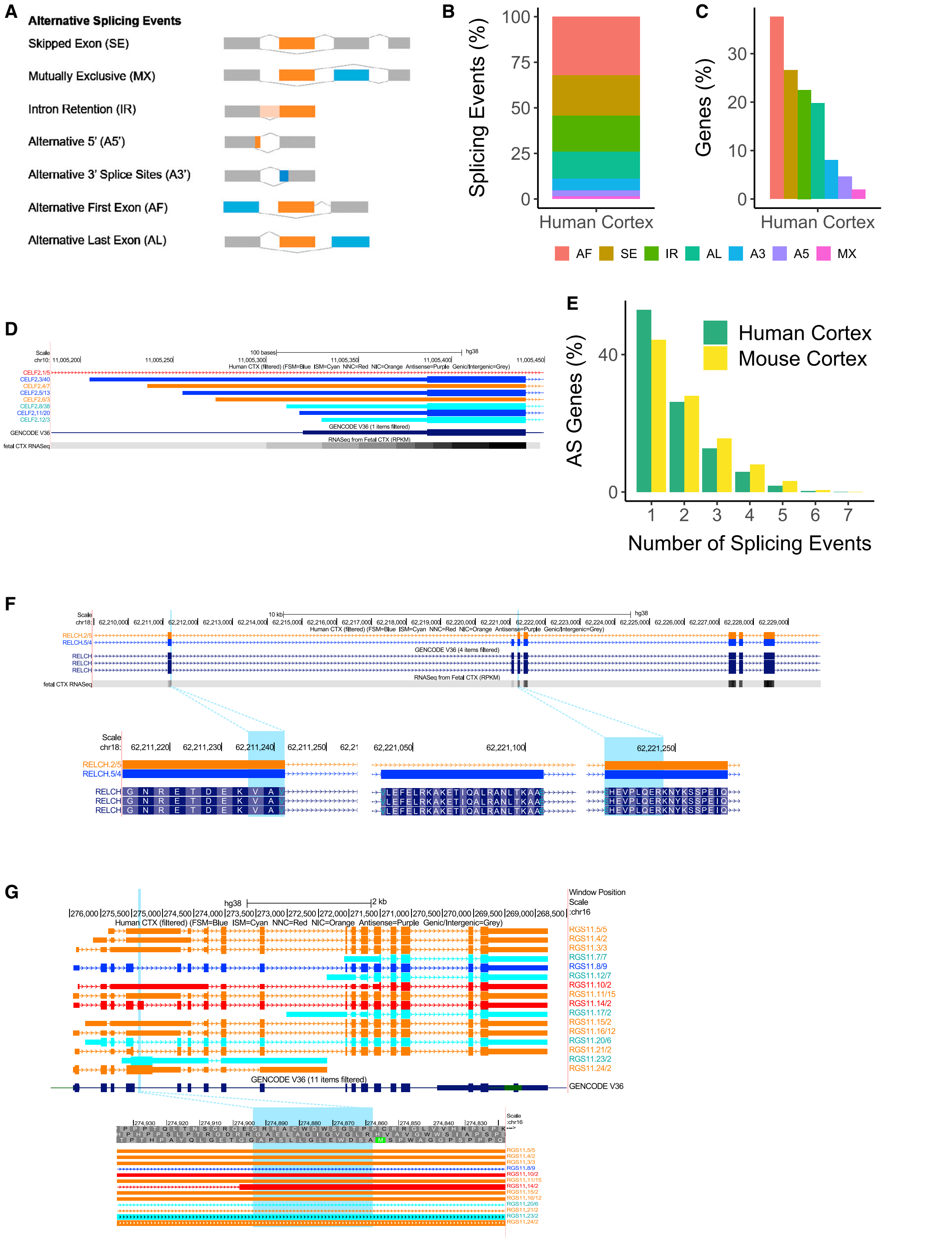
Transcriptional read-through between two or more adjacent genes can produce "fusion transcripts" that represent an important class of mutation in several types of cancer (McCartney et al., 2019). Although fusion events are thought to be rare (Akiva et al., 2006), we found evidence of fusion transcripts in both the human (n = 136 fusion transcripts [0.41% of all transcripts] associated with 108 genes [0.83% of total genes]); mouse cortex (n = 297 fusion transcripts [0.64% of all transcripts] associated with 218 genes [1.48% of total genes]) (Figure 4A–4E). A number of these genes were associated with more than one fusion transcript (human: n = 22 genes [20.3% of fusion genes]; mouse: n = 53 genes [24.3% of fusion genes]), and we identified examples of fusion transcripts encompassing more than two genes, e.g., a fusion transcript incorporating exons from three adjacent pseudogenes in the human cortex *AC138649.4-AC138649.1-PDCD6IPP1* (Figure 4D). The vast majority of the fusion transcripts identified were supported by RNA-seq data generated on both mouse (n = 282 [95%] transcripts) and human fetal (n = 51 [100%] transcripts) cortex. We also confirmed a significant proportion (n = 46 [33.8%] transcripts) of the human cortex fusion events using our ONT nanopore data. Several of the fusion transcripts identified in the human (n = 4 [2.94% of all fusion transcripts]) and mouse cortex (n = 11 [3.7% of all fusion transcripts]) were predicted as potential "conjoined genes" in the *ConjoinG* database (Prakash et al., 2010). Although the majority of fusion events were specific to the human or mouse datasets, we found evidence of potential protein-coding fusion transcripts incorporating exons from *SMIM17* (*Smim17*) in both species (Figure 4E; Figures S12A–S12D).

Identification of novel cortex-expressed genes using long-read sequencing

Although the vast majority of transcripts identified in both the human and mouse cortex were assigned to annotated genes (human: 99.8% of total transcripts; mouse: 99.5% of total transcripts), a small number represent expression from potentially novel genes (human: n = 57 novel transcripts mapping to 54 novel genes; mouse: n = 223 novel transcripts mapping to 202 novel genes) (Figure 4F; Table S10). These novel genes were either intergenic or antisense to existing annotated genes and were all multi-exonic (human: mean length = 2.09 kb, SD = 1.01 kb, range = 0.254–4.9kb, mean number of exons = 2.9; mouse: mean length = 1.75 kb, SD = 1.21 kb, range = 0.098–6.86 kb, mean number of exons = 2.5). Most transcripts from these novel genes were predicted to be non-coding (human: n = 34 [59.7%] transcripts; mouse: n = 143 [64.1%] transcripts), were generally shorter (human: $W = 1.18 \times 10^{6}$, $p = 7.71 \times 10^{-5}$; mouse: $W = 7.79 \times 10^{6}$, $p = 5.22 \times 10^{-45}$), and less abundant than transcripts of annotated genes (human: $W = 5.28 \times 10^{5}$, $p = 1.72 \times 10^{-19}$; mouse: $W = 2.29 \times 10^{6}$, $p = 1.5 \times 10^{-73}$). Although the majority of these novel genes did not show high sequence homology with other genomic regions, BLAST analysis revealed that 18 (31.6%) of the human cortex novel-gene transcripts and 31 (13.9%) of the mouse cortex novel-gene transcripts showed relatively high similarity (greater than 500 bp, more than 90% identity) to other genomic regions (Table S11). Of the 57 novel-gene transcripts identified in the human cortex, 27 (47.4%) demonstrated evidence of transcription in data from the GTEx consortium (CHESS v2.2 annotation) (Pertea et al., 2018). Further evidence of transcription from a large proportion of the human novel-gene transcripts (n = 28 [49.1%]) was provided by our ONT nanopore sequencing dataset. We used the FANTOM5 CAGE dataset to show that around a quarter of the novel-gene transcripts (human: n = 14 [24.6%]; mouse: n = 58 [26.0%]) were located within 50 bp of a CAGE peak (Table S10). There was an enrichment of antisense transcripts among those mapping to novel genes (human cortex: n = 31 transcripts [54.4%] mapping to 28 novel genes; mouse cortex: n = 119 transcripts [53.4%] mapping to 97 novel genes) (Table S10). The majority of these antisense novel genes were found within an annotated gene (human: n = 25 [89.2% of antisense novel genes], mouse: n = 95 [97.9% of antisense novel genes]), with a relatively large proportion of these sharing exonic regions (human: n = 12 [48%], mouse: n = 72 [74.2%]) reflecting sense-antisense (SAS) pairs (Galante et al., 2007). Finally, there were several striking examples of antisense novel genes overlapping two known genes in the mouse cortex (Figure 4F).

(E) Fusion transcripts with exons from *SMIM17*/*Smim17* were identified in both human and mouse cortex. Additional examples of overlapping fusion transcripts between human and mouse cortex are shown in Figures S12A–S12D.

(F) An example of a novel antisense transcript spanning *Serpina1e* and *Serpina11* in the mouse cortex. Transcripts are colored based on *SQANTI2* classification categories (blue = FSM; cyan = ISM; red = NNC; orange = NIC).



(legend on next page)

Many transcripts map to lncRNA genes with a subset containing predicted ORFs

Although the majority of transcripts were classified as protein-coding by the presence of an ORF, a relatively large number of transcripts were annotated as encoding lncRNA (human: n = 1,197 transcripts associated with 792 genes; mouse: n = 1,141 transcripts associated with 734 genes). These lncRNA transcripts were shorter than non-lncRNA transcripts (human: mean length of lncRNA transcripts = 2.32 kb [SD = 1.14 kb, range = 0.104–7.78 kb], mean length of non-lncRNA transcripts = 2.78 kb [SD = 1.29 kb, range = 0.107–11.8 kb], W = 2.28 × 10^{7}, p = 3.22 × 10^{-34}; mouse: mean length of lncRNA transcripts = 2.22 kb [SD = 1.36 kb, range = 0.148–.49 kb], mean length of non-lncRNA transcripts = 3.21 kb [SD = 1.68 kb, range = 0.083–15.9 kb], W = 3.52 × 10^{7}, p = 8.24 × 10^{-98}). As reported previously they alsocontained fewer exons (Statello et al., 2021) (human: W = 3.31 × 10^{7}, p < 2.23 × 10^{-308}; mouse: W = 4.56 × 10^{7}, p < 2.23 × 10^{-308}), with a dramatic enrichment of monoexonic molecules (Kuo et al., 2017) (human: n = 348 [29.1%]; mouse: n = 273 [23.9%]) compared to non-lncRNA transcripts (human: n = 583 [1.85%]; mouse: n = 914 [2.02%]) (Figures S13A–S13D). They were also characterized by lower transcript expression than non-lncRNA transcripts (Statello et al., 2021; Liu et al., 2016) (human: W = 2.27 × 10^{7}, p = 9.44 × 10^{-35}; mouse: W = 3.16 × 10^{7}, p = 5.67 × 10^{-40}), with fewer isoforms identified per lncRNA gene compared to non-lncRNA genes (human: mean n = 1.51 versus 2.6, W = 6.63 × 10^{6}, p = 1.21 × 10^{-80}; mouse: mean n = 1.55 versus 3.29, W = 7.40 × 10^{6}, p = 5.76 × 10^{-107}) (Figures S13E–S13H). A small proportion of these annotated lncRNA transcripts contained a putative ORF (human: n = 235 [19.6%]; mouse: n = 153 [13.4%]), supporting recent observations that some lncRNA have potential protein coding capacity (Kageyama, Kondo and Hashimoto, 2011), although the majority of such ORFs are unlikely to code for proteins (Guttman et al., 2013); of note, these ORFs were shorter than those identified in non-lncRNA transcripts (human: mean length = 133 bp versus 441 bp, W = 1.41 × 10^{7}, p = 3.12 × 10^{-221}; mouse: mean length = 139 bp versus 519 bp, W = 1.75 × 10^{7}, p = 8.33 × 10^{-195}).

AS events make a major contribution to RNA isoform diversity in the cortex

AS, the process by which different combinations of splice sites within a mRNA precursor are selected to produce variably spliced mRNAs, is the primary mechanism underlying transcript diversity in eukaryotes (Park et al., 2018) and a major source of transcriptional diversity in the CNS (Raj and Blencowe, 2015). Numerous types of AS have been described (Figure 5A), and we used both *SUPPA2* (Trincado et al., 2018) and custom analysis scripts to identify transcripts associated with (1) SE, (2) MX, (3) AF and AL exons, (4) A3′ and A5′ splice sites, and (5) IR in our cortical Iso-Seq data. The overall frequency of these specific AS events was similar in human and mouse cortex, with AF and SE being the most prevalent AS events in both species (human: AF = 8,546 [32.2%] events associated with 4,879 [37.6%] genes, SE: 5,776 [22.0%] events associated with 3,446 [26.6%] genes; mouse: AF = 12,853 [31.9%] events associated with 6,476 [44.1%] genes, SE = 8,686 [21.6%] events associated with 4,570 [31.1%] genes) (Figures 5B and 5C; Figure S14A; Table S12). Using publicly available human brain proteomic data, we found evidence of translated isoforms with novel SE events (Table S8); for example, we identified a novel peptide that was annotated to *RELCH* that spanned across exons 2 and 4 but skipped exon 3 (Figure 5F).

IR is a relatively common form of AS in the cortex that is associated with reduced expression and nonsense-mediated mRNA decay (NMD)

IR, the process by which specific introns remain unspliced in polyadenylated transcripts, is the least understood AS mechanism but is hypothesized to be an important mechanism of transcriptional control in the brain (Jacob and Smith, 2017; Ameur et al., 2011). We found evidence for IR in a relatively large proportion of genes (IR-genes) in both the human (n = 5,231 IR-transcripts associated with 2,566 [19.8%] detected genes) and mouse cortex (n = 6,803 IR transcripts associated with 3,375 [23.0%] genes) (Table S13), with IR-genes themselves enriched for biological processes related to mRNA splicing in human cortex (odds ratio = 3.24, adjusted p = 3.28 × 10^{-12}) and mRNA processing in mouse cortex (odds ratio = 2.97, p = adjusted 7.74 × 10^{-13}, Table S4). The majority of IR-transcripts were supported by matched short-read RNA-seq data from both human (n = 2,713 [97.5%] IR-transcripts) and mouse cortex samples (n = 6,454 [94.9%] IR-transcripts). Most IR-genes were found to express more than one IR-transcript (human cortex: n = 1,463 [72%] IR-genes; mouse cortex: n = 1,872 [72.4%] IR-genes), with *MEG3* having the largest number of IR-transcripts in human cortex (30 isoforms [75% of *MEG3* isoforms]; Figure 1F) and *Entr1* having the largest number of IR-transcripts in mouse cortex (31 isoforms [91.2% of *Entr1* isoforms]). A small number of genes were found to *only* express transcripts characterized by IR (Table S14) (human: n = 197 [7.68% of genes with IR-transcripts, 1.52% of total detected genes]; mouse: n = 150 [4.44% of genes with IR-transcripts, 1.02% of total detected genes]). Overall, there was considerable overlap in the list of IR-genes detected between human and mouse cortex (Figure S15A), with 1,078 homologous genes showing evidence of IR in both the

Figure 5. Alternative splicing (AS) events make a major contribution to transcript diversity in the cortex

(A) An overview of the different types of AS considered in our analysis.

(B) Alternative first (AF) exon use is the most prevalent AS event in both the human cortex and mouse cortex (Figure S14A).

(C) The majority of human cortex-expressed genes are predominantly characterized by AF and SE.

(D) AF events are supported by RNA-seq data. The differing lengths of first exon of *CELF2* in human cortex correspond to differing RNA-seq coverage.

(E) A large proportion of AS genes in human and mouse cortex are characterized by more than one type of splicing event.

(F) Shown is a UCSC genome browser track of *RELCH* with a novel peptide (VAEHEVPLQER, highlighted blue) spanning across exons 2 and 4 of *RELCH* while skipping exon 3, confirming exon skipping in a novel transcript.

(G) A novel peptide (GAELAGIGVGLR, highlighted blue) confirms translation of a retained intronic region observed in a transcript of *RGS11*.

human (48.4% of IR-genes) and mouse (35.3% of IR-genes). Importantly, a larger proportion of lowly expressed genes showed evidence for IR than highly expressed genes in both human (< 2.5 Log_{10} TPM, n = 2,269 [88.4%] genes; > 2.5 Log_{10} TPM, n = 297 [11.6%] genes) and mouse (< 2.5 Log_{10} TPM, n = 3,039 [90.04%] genes; > 2.5 Log_{10} TPM, n = 336 [9.96%] genes; Figure S15G) cortex, corroborating previous analyses suggesting that IR is associated with reduced transcript abundance (Braunschweig et al., 2014). Although most IR-containing transcripts are associated with reduced protein expression, IR-transcripts can produce a stable protein, especially if the intron is relatively short and does not disrupt the translational frame (Grabski et al., 2021). For example, we found evidence for a novel translated IR event involving the 4th intron in *RGS11* in our analysis of MS-based human brain proteomic data (Figure 5G).

NMD acts to reduce transcriptional errors by degrading transcripts containing premature stop codons (Hug, Longman and Cáceres, 2015) and is one mechanism by which IR can influence gene expression (Pan et al., 2006). Overall, >10% of transcripts mapping to annotated genes were predicted to undergo NMD (NMD-transcripts), characterized by the presence of an ORF and a coding sequence (CDS) end motif before the last junction (human cortex: n = 4,370 [13.4%] transcripts associated with 2,323 [18%] of annotated genes; mouse cortex: n = 6,014 [13.0%] transcripts associated with 2,945 [20.3%] of annotated genes). These NMD-transcripts were found to be less abundant than non-NMD-transcripts (human: mean expression of NMD-transcripts = 15.2 TPM, SD = 63.0 TPM, mean expression of non-NMD-transcripts = 33.1 TPM, SD = 261 TPM, W = 4.40×10^{7}, p = 3.59×10^{-114}; mouse: mean expression of NMD-transcripts = 11.2 TPM, SD = 85.0 TPM, mean expression of non-NMD-transcripts = 23.1 TPM, SD = 143.1 TPM, W = 8.72×10^{7}, p = 6.15×10^{-156}).

NMD was found to be particularly enriched among IR-transcripts that were predicted to be protein-coding (human: n = 1,930 [38.7%] IR-transcripts associated with 1,104 [8.55%] genes; mouse: n = 2,341 [36.2%] IR-transcripts associated with 1,380 [9.53%] genes), and transcripts with both IR and predicted NMD were particularly lowly expressed (human: W = 4.77×10^{6}, p = 3.81×10^{-12}; mouse: W = 7.50×10^{6}, p = 1.67×10^{-42}). Only a small number of genes were associated with transcripts where IR and NMD were mutually exclusive (human: n = 163 [1.26%] genes; mouse: n = 277 [1.91%] genes; Figures S15C–S14F), providing additional support for the hypothesized relationship between these two transcriptional control mechanisms (Ge and Porse, 2014).

Developmental changes in cortical RNA isoform abundance

Our human cortical Iso-Seq dataset included samples derived from both fetal and adult donors, and as expected, there was considerable overlap in the set of genes detected in each (total overlap = 8,111 [84.0% of fetal annotated genes, 73.8% of adult annotated genes]; Figure S8B). Using the Human Gene Atlas database (Kuleshov et al., 2016), we found that the 500 most abundant genes in the fetal cortex dataset were most significantly enriched for ''fetal brain'' (odds ratio = 6.98, adjusted p = 6.75×10^{-20}), and those in the adult cortex were most significantly enriched for ''prefrontal cortex'' genes (odds ratio = 6.75, adjusted p = 1.27×10^{-28}; Table S4). In total, we detected 18,592 transcripts mapping to 9,660 annotated genes in the fetal cortex (mean length = 2.90 kb, SD = 1.30 kb, range = 0.132–11.8 kb) and 22,013 transcripts mapping to 10,987 annotated genes in the adult cortex (mean length = 2.53 kb, SD = 1.18 kb, range = 0.104–10.0 kb) (Table S6). Overall patterns of RNA isoform diversity were similar between fetal and adult cortex with a similar number of genes characterized by more than one isoform (fetal: n = 4,200 [43.5%]; adult: 5,003 [45.5%]). A strong correlation was observed between the number of isoforms detected in fetal and in adult human cortex datasets (corr = 0.53, p < 2.23×10^{-308}), which was stronger among highly expressed genes (> 2.5 Log_{10} TPM in both fetal and adult cortex, corr = 0.72, p = 2.54×10^{-42}; Figures S8D and S8F). Despite these similarities, there were some notable exceptions with certain genes characterized by large differences in isoform number between fetal and adult cortex; *SEPT4* had the highest relative number of isoforms detected in adult cortex compared to fetal cortex (34 versus 2 isoforms) (Figure 3E), whereas *CELF3* had the highest relative number of isoforms in fetal cortex compared to adult cortex (11 versus 1 isoforms) (Table S15). *SEPT4*, *RAP1GAP* (adult cortex: n = 25 isoforms; fetal cortex, n = 3 isoforms), and *RUNX1T1* (adult cortex: n = 5 isoforms; fetal cortex: n = 21 isoforms) had the largest absolute difference in isoform numbers detected between human fetal and adult cortex (Figure S16). A similar proportion of novel transcripts were detected in both fetal (n = 5,415 [29.1%] transcripts associated with 3,027 [31.3%] annotated genes) and adult cortex (n = 6,354 [28.9%] associated with 3,468 [31.6%] annotated genes), with 1,670 genes characterized by novel transcripts in both fetal cortex (55.2% of genes with novel transcripts) and adult (48.1% of genes with novel transcripts) cortex. Characterization of ORFs using CPAT revealed a similar distribution of predicted coding potential across different transcript categories between adult and fetal cortex (Figure S7).

We identified 206 transcripts (associated with 189 genes) that were classified as ''fetal-specific'' and not detected in the adult cortex, and 185 transcripts (associated with 174 genes) that were classified as ''adult-specific.'' We also identified examples of significant differential transcript usage—a switch of dominant isoform expression—between fetal and adult cortex (Table S16). *RTN4*, which encodes a neurite outgrowth inhibitor specific to the CNS (GrandPré et al., 2000), was characterized by the largest expression difference in dominant transcripts between adult- and fetal-specific isoforms (Figure S17A).

A similar frequency of AS events was observed in the human adult and fetal cortex (adult: 4,963 unique AS genes with 14,793 AS events; fetal: 4,231 unique AS genes associated with 11,955 AS events) (Figure S14B; Table S12), with considerable overlap between both datasets (2,812 annotated genes [56.6% of AS genes in adult cortex, 66.5% of AS genes in fetal cortex]). IR was significantly more prevalent in the fetal cortex (2,783 transcripts associated with 1,589 genes [16.4% of annotated genes]) than adult cortex (2,383 transcripts associated with 1,422 genes [12.9% of annotated genes]; odds ratio = 1.45, p = 1.06×10^{-35}, Fisher's exact test), corroborating previous studies suggesting that IR plays a role in the developmental regulation of gene transcription in the brain (Nellore et al., 2016). Furthermore, although genes with IR-transcripts were generally more lowly expressed, they were more highly expressed in the fetal than the adult cortex ($W = 1.01 \times 10^{6}$, $p = 7.71 \times 10^{-7}$).

Differential transcript usage across human fetal brain regions

We next generated Iso-Seq data on two additional fetal brain regions (hippocampus and striatum) from matched donors (Table S1). Although the sequencing depth for these additional brain regions was lower than that of the fetal cortex (Table S17), we were able to explore fetal transcriptional differences across fetal hippocampus, striatum, and cortex using a merged dataset (incorporating 24,989 transcripts annotated to 11,072 genes). As expected, there was considerable overlap in genes detected across the three fetal brain regions (2,312 transcripts associated with 2,096 genes with TPM > 20), although a notable subset of transcripts was uniquely expressed in each brain region (cortex: n = 122; hippocampus: n = 25; striatum: n = 58 with TPM > 20). We further identified robust evidence for differential transcript usage across brain regions for a subset of genes (cortex and hippocampus: n = 9 genes; cortex and striatum: n = 10 genes; striatum and hippocampus n = 18 genes) (Table S18). For example, *APLP1* was found to express different isoforms in the cortex and hippocampus; a ~2.0 kb transcript consisting of 16 exons (ENST00000586861.5) was detected in the hippocampus, whereas a ~2.3 kb novel transcript also consisting of 16 exons was detected in the cortex (Figure S17B).

Widespread isoform diversity in genes associated with brain disease

AS has been increasingly implicated in health and disease and is recognized to play a prominent role in brain disorders hypothesized to involve the cerebral cortex including autism, SZ, and AD. There has been considerable progress in identifying genes associated with these disorders using genome sequencing and genome-wide association study (GWAS) approaches (Tam et al., 2019). However, the full repertoire of RNA isoforms transcribed from these genes in the cortex has not been systematically characterized. First, we used the human GWAS catalog database (Kuleshov et al., 2016) to interrogate the most transcriptionally diverse genes in the human cerebral cortex, finding them to be enriched for genes implicated in relevant GWAS datasets (‘‘AD (late onset)’’: odds ratio = 10.06, p = 0.004; ‘‘autism spectrum disorder or SZ’’: odds ratio = 1.94, p = 0.083: ‘‘SZ’’: odds ratio = 2.70, p = 0.005; Table S4). Second, we assessed RNA isoform diversity in genes robustly associated with AD (three familial AD genes [Bekris et al., 2010] and 59 genes nominated from a recent GWAS meta-analysis [Andrews et al., 2020; Sims et al., 2020]), autism (393 genes nominated as being category 1 [high confidence] and category 2 [strong candidate] from the SFARI Gene database, https://gene.sfari.org/), and SZ (339 genes nominated from the a recent GWAS meta-analysis [Pardiñas et al., 2018]). Among disease-associated genes detected in the cortex, we found evidence for considerable isoform diversity (human cortex: 2,016 transcripts were mapped to 610 disease-associated genes; mouse cortex: 3,218 transcripts were mapped to 670 disease-associated genes; Table S19). The vast majority of disease-associated genes detected in the cortex were characterized by more than one RNA isoform in both the human (n = 420 [68.9%] genes) and mouse (n = 538 [80.3%] genes) cortex. *TCF4* (autism- and SZ-associated) was the most ‘‘isoformic’’ disease gene in both human (n = 33 isoforms) and mouse (n = 57 isoforms) cortex; of note, both genes have been shown to be key members of transcriptional networks associated with neuropsychiatric disease (Li et al., 2018). Importantly, a large number of the transcripts mapping to disease-associated genes had not been previously annotated in existing databases in human (n = 790 [39.2%] isoforms) and mouse (n = 1,825 [56.7%] isoforms) cortex, identifying novel transcripts that may have potential relevance to understanding neurodegenerative and neuropsychiatric disorders. Interestingly, transcripts from disease-associated genes were characterized by a relatively high level of IR in the human cortex (AD: n = 9 [27.3%]; autism: n = 62 [19.6%]; SZ: n = 75 [26.0%]), with a large proportion of these annotated IR-transcripts being predicted for NMD (AD: n = 4 [44.4% of IR-genes]; autism: n = 24 [38.7% of IR-genes]; SZ: n = 29 [38.6%]; Table S20). There are known links between fusion transcripts and disease (Oliver et al., 2019), and a number of disease-associated genes were involved in fusion events (autism: n = 8, e.g., *ELAC1-SMAD4*; Figure 4A; SZ: n = 5, e.g., *GDPD3-MAPK3*; Figure 4B; autism- and SZ-associated: n = 1, e.g., *FOXG1-LINC01551*; Figure 4C). Given the hypothesized role of neurodevelopment and aging in autism, SZ, and AD, it is notable that we found large differences in isoform diversity between human adult and human fetal cortex for many disease-associated genes (Table S20).

DISCUSSION

We used long-read Iso-Seq to characterize full-length cDNA sequences and generate detailed maps of AS in the human and mouse cortex. We identify considerable RNA isoform diversity among expressed genes in the cortex across both species, including many novel transcripts not present in existing genome annotations. The majority of these isoforms have high coding potential, with the analysis of cortical proteomic data confirming the translation of several novel transcripts. Of note, we detect full-length transcripts from several previously unannotated genes in both the human and mouse cortex and many examples of fusion transcripts incorporating exons from multiple genes. Although global patterns of isoform diversity appear to be similar between both species, we identified some notable exceptions, with certain genes showing species-specific transcriptional complexity. Furthermore, we identify some striking developmental changes in transcript diversity, with certain genes characterized by differential transcript usage between fetal and adult cortex. Importantly, we show that genes associated with autism, SZ, and AD are characterized by considerable RNA isoform diversity, identifying novel transcripts that might play a role in pathology. Our data confirm the importance of AS in the cortex and highlight its role as an important mechanism underpinning gene regulation in the brain.

Our findings highlight the power of long-read sequencing approaches for transcriptional profiling. By generating reads spanning entire transcripts, it is possible to systematically characterize the repertoire of expressed RNA isoforms and fully assess the prevalence of AS. To our knowledge, our analysis represents the most comprehensive characterization of full-length transcripts and isoform diversity in the cerebral cortex yet undertaken. Several findings are particularly notable. First, we highlight that existing gene annotations are incomplete and that novel transcripts are likely to exist for a large proportion of expressed genes. Our data show examples of novel exons and even entire genes not currently annotated in existing databases. Importantly, it has been shown that such incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders (Zhang et al., 2020). Our resource enhances our understanding of the repertoire of expressed transcripts in the cerebral cortex. Second, we show that read-through transcripts (or gene fusion transcripts)—formed when exons from two genes fuse together—occur at detectable levels in the cortex. Although many of these fusion transcripts appear to be associated with NMD, some have the potential to be translated into proteins or may have a regulatory effect at the RNA level. Despite gene-fusion transcripts having a well-documented role in several human cancers (Futreal et al., 2004), the systematic analysis of gene fusion and read-through transcripts has been limited to date given the limitations of existing short-read sequencing technologies (Haas et al., 2019). Our data support recent data suggesting that read-through transcripts occur naturally (Mehani et al., 2020) and suggest that some fusion transcripts may have protein-coding potential, with important implications for brain disease. Third, we are able to highlight the significant extent to which AS events contribute to isoform diversity in the cortex. In particular, we show that IR is a relatively common form of AS in the cortex that is associated with reduced expression and NMD. Importantly, IR was more prevalent in the human fetal cortex than adult cortex, supporting previous studies that implicate a role of IR in the developmental regulation of gene transcription in the brain (Ameur et al., 2011). Finally, we highlight major developmental changes in cortical isoform abundance in the human brain. In particular, we identify striking examples of transcript usage between fetal and adult cortex and also significant differences in isoform expression between different regions of the human brain.

Our results should be interpreted in the context of several limitations. First, we profiled tissue from a relatively small number of human and mouse donors. Although we found highly consistent patterns of AS across these biological replicates and rarefaction curves confirmed our sequencing dataset was close to saturation, we were unable to explore inter-individual variation in AS. Recent studies have highlighted considerable evidence for genetic influences on isoform diversity in the human cortex, with splicing quantitative trait loci (sQTL) widely implicated in health and disease (Takata, Matsumoto and Kato, 2017). Future work will aim to extend our analyses to larger numbers of samples to explore population-level variation in transcript abundance in the cerebral cortex and differences associated with pathology. Second, despite the advantages of long-read sequencing approaches for the characterization of novel full-length transcripts, these methods are often assumed to be less quantitative than traditional short-read RNA sequencing methods (Zhao et al., 2019). We implemented a stringent QC pipeline and undertook considerable filtering of our data, finding high consistency across biological replicates and validating our findings using complementary approaches (i.e., nanopore sequencing, RNA-seq, and by comparison to existing genomic databases). We show that transcriptional profiles generated using Iso-Seq reflect those expected from the tissues we assessed (i.e., the cerebral cortex), and we found a strong correlation with both gene- and transcript-level expressions measured using short-read RNA-seq on the same samples. We also observed a strong correlation between expected and detected levels of ERCC spike-in control molecules, highlighting the power of Iso-Seq to accurately quantify the abundance of highly expressed transcripts. Given that we have adopted stringent QC approaches, many true transcripts from our final dataset—particularly lowly-expressed transcripts—are likely to have been filtered out. Our analyses therefore probably underestimate the extent of RNA isoform diversity in the cerebral cortex so we also provide a less conservatively filtered dataset for download from our online track hub. Third, our analyses were performed on "bulk" cortex tissue containing a heterogeneous mix of neurons, oligodendrocytes, and other glial cell types. It is likely that these different cell types express a specific repertoire of RNA isoforms, and we are not able to explore these differences in our data. Of note, novel approaches for using long-read sequencing approaches in single cells will enable a more granular approach to exploring transcript diversity in the cortex. Although such approaches are currently limited by technological and analytical constraints, a recent study used long-read transcriptome sequencing to identify cell-type-specific transcript diversity in the mouse hippocampus and prefrontal cortex (Joglekar et al., 2021). Finally, although we explored the extent to which novel transcripts contained ORFs, the extent to which they are actually translated and contribute to cortical proteomic diversity is not known.

In summary, our data confirm the importance of AS and AF exon usage in the cerebral cortex, dramatically increasing transcriptional diversity and representing an important mechanism underpinning gene regulation in the brain. We highlight the power of long-read sequencing for completing our understanding of human and mouse gene annotation, and our transcript annotations, isoform data, and Iso-Seq analysis pipeline are available as a resource to the research community.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
    - ○ Lead contact
    - ○ Materials availability
    - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
    - ○ Brain samples

Bibliography

[1] A. D. International, "Dementia statistics | Alzheimer's Disease International (ADI)," 2020.

[2] D. J. Selkoe, "The molecular pathology of Alzheimer's disease," apr 1991.

[3] D. P. Perl, "Neuropathology of Alzheimer's disease," jan 2010.

[4] R. A. Sperling, C. R. Jack, and P. S. Aisen, "Testing the right target and right drug at the right stage," nov 2011.

[5] M. T. Heneka, M. J. Carson, J. E. Khoury, G. E. Landreth, F. Brosseron, ..., D. T. Golenbock, and M. P. Kummer, "Neuroinflammation in Alzheimer's disease," apr 2015.

[6] H. Braak and E. Braak, "Neuropathological stageing of Alzheimer-related changes," sep 1991.

[7] J. Xu, S. Patassini, N. Rustogi, I. Riba-Garcia, B. D. Hale, A. M. Phillips, ..., G. J. Cooper, and R. D. Unwin, "Regional protein expression in human Alzheimer's brain correlates with disease severity," Commun. Biol., vol. 2, pp. 1–15, dec 2019.

[8] C. Hulette, S. Mirra, W. Wilkinson, A. Heyman, G. Fillenbaum, and C. Clark, "The consortium to establish a registry for alzheimer's disease (CERAD):Part IX. A prospective cliniconeuropathologic study of parkinson's features in alzheimer's disease," Neurology, vol. 45, no. 11, pp. 1991–1995, 1995.

[9] L. Parkkinen, H. Soininen, and I. Alafuzoff, "Regional distribution of α-synuclein pathology in unimpaired aging and Alzheimer disease," J. Neuropathol. Exp. Neurol., vol. 62, pp. 363–367, apr 2003.

[10] K. Wakabayashi, K. Tanji, F. Mori, and H. Takahashi, "The Lewy body in Parkinson's disease: Molecules implicated in the formation and degradation of α-synuclein aggregates," Neuropathology, vol. 27, pp. 494–506, oct 2007.

[11] M. G. Spillantini, M. L. Schmidt, V. M.-Y. Lee, J. Q. Trojanowski, R. Jakes, and M. Goedert, "α-synuclein in Lewy bodies [8]," 1997.

[12] A. King, F. Sweeney, I. Bodi, C. Troakes, S. Maekawa, and S. Al-Sarraj, "Abnormal TDP-43 expression is identified in the neocortex in cases of dementia pugilistica, but is mainly confined to the limbic system when identified in high and moderate stages of Alzheimer's disease," Neuropathology, vol. 30, no. 4, pp. 408–419, 2010.

[13] K. E. McAleese, L. Walker, D. Erskine, A. J. Thomas, I. G. McKeith, and J. Attems, "TDP-43 pathology in Alzheimer's disease, dementia with Lewy bodies and ageing," Brain Pathol., vol. 27, no. 4, pp. 472–479, 2017.

[14] T. Arai, I. R. Mackenzie, M. Hasegawa, T. Nonoka, K. Niizato, K. Tsuchiya, S. Iritani, M. Onaya, and H. Akiyama, "Phosphorylated TDP-43 in Alzheimer's disease and dementia with Lewy bodies," Acta Neuropathol., vol. 117, no. 2, pp. 125–136, 2009.

[15] G. S. Pesiridis, V. M. Lee, and J. Q. Trojanowski, "Mutations in TDP-43 link glycine-rich domain functions to amyotrophic lateral sclerosis," Hum. Mol. Genet., vol. 18, no. R2, 2009.

[16] A. M. Palmer, "Alzheimer' s Disease : Neurobiology and Drug Targets," tech. rep., 2015.

[17] D. R. Thal, U. Rüb, M. Orantes, and H. Braak, "Phases of Aβ-deposition in the human brain and its relevance for the development of AD," Neurology, vol. 58, pp. 1791–1800, jun 2002.

[18] C. L. Masters, R. Bateman, K. Blennow, C. C. Rowe, R. A. Sperling, and J. L. Cummings, "Alzheimer's disease," Nat. Rev. Dis. Prim., vol. 1, pp. 1–18, 2015.

[19] A. I. Jarmolowicz, H. Y. Chen, and P. K. Panegyres, "The patterns of inheritance in early-onset dementia: Alzheimer's disease and frontotemporal dementia," Am. J. Alzheimers. Dis. Other Demen., vol. 30, pp. 299–306, may 2015.

[20] L. M. Bekris, C. E. Yu, T. D. Bird, and D. W. Tsuang, "Review article: Genetics of Alzheimer disease," dec 2010.

[21] C. K. Chai, "The genetics of Alzheimer ' s disease," Am. J. Alzheimer's Dis. Other Dementias, vol. 22, pp. 37–41, dec 2007.

[22] M. Gatz, C. A. Reynolds, L. Fratiglioni, B. Johansson, J. A. Mortimer, S. Berg, A. Fiske, and N. L. Pedersen, "Roles of Genes and Environments for Explaining Alzheimer Disease," Arch. Gen. Psychiatry, vol. 63, no. 2, p. 168, 2006.

[23] C. Bellenguez, F. Küçükali, I. Jansen, V. Andrade, and S. Moreno-grau, "New insights on the genetic etiology of Alzheimer's and related dementia," MedRxiv, pp. 1–35, oct 2020.

[24] A. C. Naj, J. Sha, Y. Zhao, G. Leonenko, X. Jian, B. Grenier-Boley, ..., J. Williams, and G. D. Schellenberg, "Genome-wide meta-analysis of late-onset Alzheimer's disease using rare variant imputation in 65,602 subjects identifies risk loci with roles in memory, neurodevelopment, and cardiometabolic traits: The international genomics of Alzheimer's project (IGAP)," Alzheimer's Dement., vol. 16, p. e044193, dec 2020.

[25] B. W. Kunkle, B. Grenier-Boley, R. Sims, J. C. Bis, V. Damotte, A. C. Naj, ..., J.-C. Lambert, and M. A. Pericak-Vance, "Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing," Nat. Genet. 2019 513, vol. 51, pp. 414–430, feb 2019.

[26] I. E. Jansen, J. E. Savage, K. Watanabe, J. Bryois, D. M. Williams, S. Steinberg, ..., O. A. Andreassen, and D. Posthuma, "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk," Nat. Genet., vol. 51, pp. 404–413, jan 2019.

[27] J. C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, ..., L. J. Launer, and S. Seshadri, “Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer’s disease,” Nat. Genet., vol. 45, no. 12, pp. 1452–1458, 2013.

[28] A. C. Naj, G. Jun, G. W. Beecham, L. S. Wang, B. N. Vardarajan, J. Buros, ..., L. A. Farrer, and G. D. Schellenberg, “Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer’s disease,” Nat. Genet., vol. 43, pp. 436–443, may 2011.

[29] P. Hollingworth, D. Harold, R. Sims, A. Gerrish, J. C. Lambert, M. M. Carrasquillo, ..., P. Amouyel, and J. Williams, “Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer’s disease,” Nat. Genet., vol. 43, no. 5, pp. 429–436, 2011.

[30] D. Harold, R. Abraham, P. Hollingworth, R. Sims, A. Gerrish, M. L. Hamshere, ..., M. J. Owen, and J. Williams, “Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer’s disease,” Nat. Genet., vol. 41, pp. 1088–1093, oct 2009.

[31] J. C. Lambert, S. Heath, G. Even, D. Campion, K. Sleegers, ..., M. Lathrop, and P. Amouyel, “Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer’s disease,” Nat. Genet., vol. 41, no. 10, pp. 1094–1099, 2009.

[32] L. Bertram, C. Lange, K. Mullin, M. Parkinson, M. Hsiao, M. F. Hogan, ..., K. D. Becker, and R. E. Tanzi, “Genome-wide Association Analysis Reveals Putative Alzheimer’s Disease Susceptibility Loci in Addition to APOE,” Am. J. Hum. Genet., vol. 83, no. 5, pp. 623–632, 2008.

[33] D. H. Mauch, K. Nägier, S. Schumacher, C. Göritz, E. C. Müller, A. Otto, and F. W. Pfrieger, “CNS synaptogenesis promoted by glia-derived cholesterol,” Science (80-.)., vol. 294, pp. 1354–1357, nov 2001.

[34] L. A. Farrer, “Effects of Age, Sex, and Ethnicity on the Association Between Apolipoprotein E Genotype and Alzheimer Disease,” JAMA, vol. 278, p. 1349, oct 1997.

[35] Z. S. Nagy, M. M. Esiri, K. A. Jobst, C. Johnston, S. Litchfield, E. Sim, and A. D. Smith, “Influence of the apolipoprotein E genotype on amyloid deposition and neurofibrillary tangle formation in Alzheimer’s disease,” Neuroscience, vol. 69, pp. 757–761, dec 1995.

[36] E. H. Corder, A. M. Saunders, N. J. Risch, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, J. B. Rimmler, P. A. Locke, P. M. Conneally, K. E. Schmader, G. W. Small, A. D. Roses, J. L. Haines, and M. A. Pericak-Vance, “Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease,” Nat. Genet., vol. 7, pp. 180–184, jun 1994.

[37] I. de Rojas, S. Moreno-Grau, N. Tesi, B. Grenier-Boley, V. Andrade, I. E. Jansen, ..., S. J. van der Lee, and A. Ruiz, “Common variants in Alzheimer’s disease and risk stratification by polygenic risk scores,” Nat. Commun., vol. 12, pp. 1–16, jun 2021.

[38] J. A. Hardy and G. A. Higgins, “Alzheimer’s disease: The amyloid cascade hypothesis,” apr 1992.

[39] J. Kang, H. G. Lemaire, A. Unterbeck, J. M. Salbaum, C. L. Masters, K. H. Grzeschik, G. Multhaup, K. Beyreuther, and B. Müller-Hill, “The precursor of Alzheimer’s disease amyloid A4 protein resembles a cell-surface receptor,” Nature, vol. 325, no. 6106, pp. 733–736, 1987.

[40] A. Asami-Odaka, Y. Ishibashi, T. Kikuchi, C. Kitada, and N. Suzuki, "Long Amyloid β-Protein Secreted from Wild-Type Human Neuroblastoma IMR-32 Cells," Biochemistry, vol. 34, no. 32, pp. 10272–10278, 1995.

[41] N. M. Li, K. F. Liu, Y. J. Qiu, H. H. Zhang, H. Nakanishi, and H. Qing, "Mutations of beta-amyloid precursor protein alter the consequence of Alzheimer's disease pathogenesis," Neural Regen. Res., vol. 14, pp. 658–665, apr 2019.

[42] D. Scheuner, C. Eckman, M. Jensen, X. Song, M. Citron, N. Suzuki, T. D. Bird, J. Hardy, M. Hutton, W. Kukull, E. Larson, E. Levy-Lahad, M. Viitanen, E. Peskind, P. Poorkaj, G. Schellenberg, R. Tanzi, W. Wasco, L. Lannfelt, D. Selkoe, and S. Younkin, "Secreted amyloid β-protein similar to that in the senile plaques of Alzheimer's disease is increased in vivo by the presenilin 1 and 2 and APP mutations linked to familial Alzheimer's disease," Nat. Med., vol. 2, no. 8, pp. 864–870, 1996.

[43] J. T. Jarrett, E. P. Berger, and P. T. Lansbury, "The Carboxy Terminus of the β Amyloid Protein Is Critical for the Seeding of Amyloid Formation: Implications for the Pathogenesis of Alzheimer's Disease," Biochemistry, vol. 32, pp. 4693–4697, may 1993.

[44] Z. P. Van Acker, M. Bretou, and W. Annaert, "Endo-lysosomal dysregulations and late-onset Alzheimer's disease: Impact of genetic risk factors," jun 2019.

[45] K. S. Kosik, C. L. Joachim, and D. J. Selkoe, "Microtubule-associated protein tau (τ) is a major antigenic component of paired helical filaments in Alzheimer disease," Proc. Natl. Acad. Sci. U. S. A., vol. 83, no. 11, pp. 4044–4048, 1986.

[46] M. E. Orr, A. C. Sullivan, and B. Frost, "A Brief Overview of Tauopathy: Causes, Consequences, and Therapeutic Strategies," jul 2017.

[47] I. Grundke-Iqbal, K. Iqbal, M. Quinlan, Y. C. Tung, M. S. Zaidi, and H. M. Wisniewski, "Microtubule-associated protein tau. A component of Alzheimer paired helical filaments," J. Biol. Chem., vol. 261, pp. 6084–6089, may 1986.

[48] I. Grundke-Iqbal, K. Iqbal, Y. C. Tung, M. Quinlan, H. M. Wisniewski, and L. I. Binder, "Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology.," Proc. Natl. Acad. Sci. U. S. A., vol. 83, pp. 4913–4917, jul 1986.

[49] E. M. Coomans, D. N. Schoonhoven, H. Tuncel, S. C. Verfaillie, E. E. Wolters, R. Boellaard, R. Ossenkoppele, A. den Braber, W. Scheper, P. Schober, S. P. Sweeney, J. M. Ryan, R. C. Schuit, A. D. Windhorst, F. Barkhof, P. Scheltens, S. S. Golla, A. Hillebrand, A. A. Gouw, and B. N. van Berckel, "In vivo tau pathology is associated with synaptic loss and altered synaptic function," Alzheimer's Res. Ther., vol. 13, pp. 1–13, feb 2021.

[50] A. d. C. Alonso, A. Mederlyova, M. Novak, I. Grundke-Iqbal, and K. Iqbal, "Promotion of hyperphosphorylation by frontotemporal dementia tau mutations," J. Biol. Chem., vol. 279, pp. 34873–34881, aug 2004.

[51] A. Serrano-Pozo, J. Qian, A. Muzikansky, S. E. Monsell, T. J. Montine, M. P. Frosch, R. A. Betensky, and B. T. Hyman, "Thal amyloid stages do not significantly impact the correlation between neuropathological change and cognition in the Alzheimer disease continuum," J. Neuropathol. Exp. Neurol., vol. 75, pp. 516–526, jun 2016.

[52] P. Giannakopoulos, F. R. Herrmann, T. Bussière, C. Bouras, E. Kövari, D. P. Perl, J. H.

Morrison, G. Gold, and P. R. Hof, “Tangle and neuron numbers, but not amyloid load, predict cognitive status in Alzheimer’s disease,” Neurology, vol. 60, pp. 1495–1500, may 2003.

[53] A. PV, G. JH, H.-W. ET, and H. BT, “Neurofibrillary tangles but not senile plaques parallel duration and severity of Alzheimer’s disease,” Neurology, vol. 42, no. 3 Pt 1, pp. 631–639, 1992.

[54] D. Trabzuni, S. Wray, J. Vandrovcova, A. Ramasamy, R. Walker, C. Smith, C. Luk, J. R. Gibbs, A. Dillman, D. G. Hernandez, S. Arepalli, A. B. Singleton, M. R. Cookson, A. M. Pittman, R. De silva, M. E. Weale, J. Hardy, and M. Ryten, “MAPT expression and splicing is differentially regulated by brain region: Relation to genotype and implication for tauopathies,” Hum. Mol. Genet., vol. 21, pp. 4094–4103, sep 2012.

[55] K. R. Brunden, J. Q. Trojanowski, and V. M.-Y. Lee, “Advances in tau-focused drug discovery for Alzheimer’s disease and related tauopathies,” 2009.

[56] S. S. Sisodia, “β-Amyloid precursor protein cleavage by a membrane-bound protease,” Proc. Natl. Acad. Sci. U. S. A., vol. 89, pp. 6075–6079, jul 1992.

[57] A. Peric and W. Annaert, “Early etiology of Alzheimer’s disease: tipping the balance toward autophagy or endosomal dysfunction?,” jan 2015.

[58] T. Tomiyama, S. Matsuyama, H. Iso, T. Umeda, H. Takuma, K. Ohnishi, K. Ishibashi, R. Teraoka, N. Sakama, T. Yamashita, K. Nishitsuji, K. Ito, H. Shimada, M. P. Lambert, W. L. Klein, and H. Mori, “A mouse model of amyloid β oligomers: Their contribution to synaptic alteration, abnormal tau phosphorylation, glial activation, and neuronal loss in vivo,” J. Neurosci., vol. 30, pp. 4845–4856, apr 2010.

[59] M. Knobloch, U. Konietzko, D. C. Krebs, and R. M. Nitsch, “Intracellular Aβ and cognitive deficits precede β-amyloid deposition in transgenic arcAβ mice,” Neurobiol. Aging, vol. 28, pp. 1297–1306, sep 2007.

[60] L. M. Billings, S. Oddo, K. N. Green, J. L. McGaugh, and F. M. LaFerla, “Intraneuronal Aβ causes the onset of early Alzheimer’s disease-related cognitive deficits in transgenic mice,” Neuron, vol. 45, pp. 675–688, mar 2005.

[61] D. Z. Christensen, S. L. Kraus, A. Flohr, M. C. Cotel, O. Wirths, and T. A. Bayer, “Transient intraneuronal Aβ rather than extracellular plaque pathology correlates with neuron loss in the frontal cortex of APP/PS1KI mice,” Acta Neuropathol., vol. 116, pp. 647–655, oct 2008.

[62] V. Schmidt, A. Subkhangulova, and T. E. Willnow, “Sorting receptor SORLA: cellular mechanisms and implications for disease,” nov 2017.

[63] S. B. Dumanis, T. Burgert, S. Caglayan, A. Füchtbauer, E. M. Füchtbauer, V. Schmidt, and T. E. Willnow, “Distinct functions for anterograde and retrograde sorting of SORLA in amyloidogenic processes in the brain,” J. Neurosci., vol. 35, pp. 12703–12713, sep 2015.

[64] G. Cisbani and S. Rivest, “Targeting innate immunity to protect and cure Alzheimer’s disease: opportunities and pitfalls,” apr 2021.

[65] A. Griciuc and R. E. Tanzi, “The role of innate immune genes in Alzheimer’s disease,” Curr. Opin. Neurol., vol. 34, pp. 228–236, apr 2021.

[66] G. R. Frost, L. A. Jonas, and Y. M. Li, “Friend, Foe or Both? Immune Activity in Alzheimer’s Disease,” dec 2019.

[67] L. Qin, Y. Liu, C. Cooper, B. Liu, B. Wilson, and J. S. Hong, “Microglia enhance β-amyloid peptide-induced toxicity in cortical and mesencephalic neurons by producing reactive oxygen species,” J. Neurochem., vol. 83, pp. 973–983, nov 2002.

[68] W. Y. Wang, M. S. Tan, J. T. Yu, and L. Tan, “Role of pro-inflammatory cytokines released from microglia in Alzheimer’s disease,” jun 2015.

[69] C. H. Chen, W. Zhou, S. Liu, Y. Deng, F. Cai, M. Tone, Y. Tone, Y. Tong, and W. Song, “Increased NF-κB signalling up-regulates BACE1 expression and its therapeutic potential in Alzheimer’s disease,” Int. J. Neuropsychopharmacol., vol. 15, pp. 77–90, feb 2012.

[70] P. L. McGeer, S. Itagaki, H. Tago, and E. G. McGeer, “Reactive microglia in patients with senile dementia of the Alzheimer type are positive for the histocompatibility glycoprotein HLA-DR,” Neurosci. Lett., vol. 79, pp. 195–200, aug 1987.

[71] K. G. Mawuenyega, W. Sigurdson, V. Ovod, L. Munsell, T. Kasten, J. C. Morris, K. E. Yarasheski, and R. J. Bateman, “Decreased clearance of CNS β-amyloid in Alzheimer’s disease,” Science (80-.)., vol. 330, p. 1774, dec 2010.

[72] F. Leng and P. Edison, “Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here?,” dec 2021.

[73] J. M. Castellano, J. Kim, F. R. Stewart, H. Jiang, R. B. DeMattos, B. W. Patterson, A. M. Fagan, J. C. Morris, K. G. Mawuenyega, C. Cruchaga, A. M. Goate, K. R. Bales, S. M. Paul, R. J. Bateman, and D. M. Holtzman, “Human apoE isoforms differentially regulate brain amyloid-β peptide clearance,” Sci. Transl. Med., vol. 3, pp. 89ra57–89ra57, jun 2011.

[74] R. M. Koffie, T. Hashimoto, H. C. Tai, K. R. Kay, A. Serrano-Pozo, D. Joyner, S. Hou, K. J. Kopeikina, M. P. Frosch, V. M. Lee, D. M. Holtzman, B. T. Hyman, and T. L. Spires-Jones, “Apolipoprotein E4 effects in Alzheimer’s disease are mediated by synaptotoxic oligomeric amyloid-β,” Brain, vol. 135, no. 7, pp. 2155–2168, 2012.

[75] R. Koldamova, N. F. Fitz, and I. Lefterov, “The role of ATP-binding cassette transporter A1 in Alzheimer’s disease and neurodegeneration,” aug 2010.

[76] D. M. Hatters, N. Zhong, E. Rutenber, and K. H. Weisgraber, “Amino-terminal Domain Stability Mediates Apolipoprotein E Aggregation into Neurotoxic Fibrils,” J. Mol. Biol., vol. 361, pp. 932–944, sep 2006.

[77] N. F. Fitz, K. N. Nam, C. M. Wolfe, F. Letronne, B. E. Playso, B. E. Iordanova, T. D. Kozai, R. J. Biedrzycki, V. E. Kagan, Y. Y. Tyurina, X. Han, I. Lefterov, and R. Koldamova, “Phospholipids of APOE lipoproteins activate microglia in an isoform-specific manner in preclinical models of Alzheimer’s disease,” Nat. Commun., vol. 12, pp. 1–18, jun 2021.

[78] D. E. Schmechel, A. M. Saunders, W. J. Strittmatter, B. J. Crain, C. M. Hulette, S. H. Joo, M. A. Pericak-Vance, D. Goldgaber, and A. D. Roses, “Increased amyloid β-peptide deposition in cerebral cortex as a consequence of apolipoprotein E genotype in late-onset Alzheimer disease,” Proc. Natl. Acad. Sci. U. S. A., vol. 90, pp. 9649–9653, oct 1993.

[79] E. Kok, S. Haikonen, T. Luoto, H. Huhtala, S. Goebeler, H. Haapasalo, and P. J. Karhunen,

"Apolipoprotein E-dependent accumulation of alzheimer disease-related lesions begins in middle age," Ann. Neurol., vol. 65, pp. 650–657, jun 2009.

[80] G. Di Paolo and T. W. Kim, "Linking lipids to Alzheimer's disease: Cholesterol and beyond," mar 2011.

[81] A. Sierksma, V. Escott-Price, and B. De Strooper, "Translating genetic risk of Alzheimer's disease into mechanistic insight and drug targets," oct 2020.

[82] N. Sakae, C. C. Liu, M. Shinohara, J. Frisch-Daiello, L. Ma, Y. Yamazaki, ..., G. Bu, and T. Kanekiyo, "ABCA7 deficiency accelerates amyloid-βgeneration and alzheimer's neuronal pathology," J. Neurosci., vol. 36, no. 13, pp. 3848–3859, 2016.

[83] A. M. Hall and E. D. Roberson, "Mouse models of Alzheimer's disease," may 2012.

[84] C. Cook, J. H. Dunmore, M. E. Murray, K. Scheffel, N. Shukoor, J. Tong, M. Castanedes-Casey, V. Phillips, L. Rousseau, M. S. Penuliar, A. Kurti, D. W. Dickson, L. Petrucelli, and J. D. Fryer, "Severe amygdala dysfunction in a MAPT transgenic mouse model of frontotemporal dementia," Neurobiol. Aging, vol. 35, no. 7, pp. 1769–1777, 2014.

[85] J. Gamache, K. Benzow, C. Forster, L. Kemper, C. Hlynialuk, E. Furrow, K. H. Ashe, and M. D. Koob, "Factors other than hTau overexpression that contribute to tauopathy-like phenotype in rTg4510 mice," Nat. Commun., vol. 10, pp. 1–12, jun 2019.

[86] T. Blackmore, S. Meftah, T. K. Murray, P. J. Craig, A. Blockeel, K. Phillips, B. Eastwood, M. J. O'Neill, H. Marston, Z. Ahmed, G. Gilmour, and F. Gastambide, "Tracking progressive pathological and functional decline in the rTg4510 mouse model of tauopathy," Alzheimer's Res. Ther., vol. 9, no. 1, 2017.

[87] "APOE4/Trem2*R47H Mouse Model."

[88] C. T. Lewandowski, J. Maldonado Weng, and M. J. LaDu, "Alzheimer's disease pathology in APOE transgenic mouse models: The Who, What, When, Where, Why, and How," jun 2020.

[89] Y. W. Wan, R. Al-Ouran, C. G. Mangleburg, T. M. Perumal, T. V. Lee, K. Allison, ..., L. M. Mangravite, and B. A. Logsdon, "Meta-Analysis of the Alzheimer's Disease Human Brain Transcriptome and Functional Dissection in Mouse Models," Cell Rep., vol. 32, jul 2020.

[90] I. Castanho, T. K. Murray, E. Hannon, A. Jeffries, E. Walker, E. Laing, H. Baulf, J. Harvey, L. Bradshaw, A. Randall, K. Moore, P. O'Neill, K. Lunnon, D. A. Collier, Z. Ahmed, M. J. O'Neill, and J. Mill, "Transcriptional Signatures of Tau and Amyloid Neuropathology," Cell Rep., vol. 30, no. 6, pp. 2040–2054.e5, 2020.

[91] M. Kikuchi, N. Hara, M. Hasegawa, A. Miyashita, R. Kuwano, T. Ikeuchi, and A. Nakaya, "Enhancer variants associated with Alzheimer's disease affect gene expression via chromatin looping," BMC Med. Genomics, vol. 12, pp. 1–16, sep 2019.

[92] T. Raj, Y. I. Li, G. Wong, J. Humphrey, M. Wang, S. Ramdhani, Y.-C. Wang, B. Ng, I. Gupta, V. Haroutunian, E. E. Schadt, T. Young-Pearse, S. Mostafavi, B. Zhang, P. Sklar, D. A. Bennett, and P. L. De Jager, "Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility," Nat. Genet., vol. 50, pp. 1584–1592, 2018.

[93] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, “Alternative isoform regulation in human tissue transcriptomes,” Nature, vol. 456, no. 7221, pp. 470–476, 2008.

[94] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, “Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing,” Nat. Genet., vol. 40, no. 12, pp. 1413–1415, 2008.

[95] G. Yeo, D. Holste, G. Kreiman, and C. B. Burge, “Variation in alternative splicing across human tissues.,” Genome Biol., vol. 5, pp. 1–15, sep 2004.

[96] P. Mazin, J. Xiong, X. Liu, Z. Yan, X. Zhang, M. Li, L. He, M. Somel, Y. Yuan, Y.-P. Phoebe Chen, N. Li, Y. Hu, N. Fu, Z. Ning, R. Zeng, H. Yang, W. Chen, M. Gelfand, and P. Khaitovich, “Widespread splicing changes in human brain development and aging,” Mol. Syst. Biol., vol. 9, no. 1, pp. 633–633, 2014.

[97] B. Raj and B. J. Blencowe, “Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles,” 2015.

[98] M. J. Gandal, P. Zhang, E. Hadjimichael, R. L. Walker, C. Chen, S. Liu, ..., D. Pinto, and D. H. Geschwind, “Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder,” Science (80-.)., vol. 362, no. 6420, 2018.

[99] R. L. Walker, G. Ramaswami, C. Hartl, N. Mancuso, M. J. Gandal, L. de la Torre-Ubieta, B. Pasaniuc, J. L. Stein, and D. H. Geschwind, “Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms,” Cell, vol. 179, pp. 750–771.e22, oct 2019.

[100] L. Herzel, D. S. Ottoz, T. Alpert, and K. M. Neugebauer, “Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function,” 2017.

[101] C. L. Will and R. Lührmann, “Spliceosome structure and function,” Cold Spring Harb. Perspect. Biol., vol. 3, pp. 1–2, jul 2011.

[102] N. Sheth, X. Roca, M. L. Hastings, T. Roeder, A. R. Krainer, and R. Sachidanandam, “Comprehensive splice-site analysis using comparative genomics,” Nucleic Acids Res., vol. 34, no. 14, pp. 3955–3967, 2006.

[103] G. E. Parada, R. Munita, C. A. Cerda, and K. Gysling, “A comprehensive survey of non-canonical splice sites in the human transcriptome,” sep 2014.

[104] H. Li, Z. Wang, T. Ma, G. Wei, and T. Ni, “Alternative Splicing in Aging and Age-related Diseases,” Transl. Med. Aging, vol. 1, pp. 32–40, oct 2017.

[105] B. P. Lewis, R. E. Green, and S. E. Brenner, “Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans,” Proc. Natl. Acad. Sci., vol. 100, no. 1, pp. 189–192, 2003.

[106] A. Nickless, J. M. Bailis, and Z. You, “Control of gene expression through the nonsense-mediated RNA decay pathway,” may 2017.

[107] J. Weischenfeldt, J. Waage, G. Tian, J. Zhao, I. Damgaard, J. S. Jakobsen, K. Kristiansen,

A. Krogh, J. Wang, and B. T. Porse, “Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns,” Genome Biol., vol. 13, may 2012.

[108] J. J. Wong, W. Ritchie, O. A. Ebner, M. Selbach, J. W. Wong, Y. Huang, D. Gao, N. Pinello, M. Gonzalez, K. Baidya, A. Thoeng, T. L. Khoo, C. G. Bailey, J. Holst, and J. E. Rasko, “XOrchestrated intron retention regulates normal granulocyte differentiation,” Cell, vol. 154, pp. 583–595, aug 2013.

[109] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, ..., D. A. Hume, and Y. Hayashizaki, “Genome-wide analysis of mammalian promoter architecture and evolution,” Nat. Genet., vol. 38, pp. 626–635, jun 2006.

[110] A. Reyes and W. Huber, “Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues,” Nucleic Acids Res., vol. 46, no. 2, pp. 582–592, 2018.

[111] S. A. Shabalina, A. N. Spiridonov, N. A. Spiridonov, and E. V. Koonin, “Connections between alternative transcription and alternative splicing in mammals,” Genome Biol. Evol., vol. 2, no. 1, pp. 791–799, 2010.

[112] C. De Jonghe, M. Cruts, E. A. Rogaeva, C. Tysoe, A. Singleton, H. Vanderstichele, W. Meschino, B. Dermaut, I. Vanderhoeven, H. Backhovens, E. Vanmechelen, C. M. Morris, J. Hardy, D. C. Rubinsztein, P. H. St George-Hyslop, and C. Van Broeckhoven, “Aberrant splicing in the presenilin-1 intron 4 mutation causes presenile Alzheimer’s disease by increased Aβ42 secretion,” Hum. Mol. Genet., vol. 8, pp. 1529–1540, aug 1999.

[113] I. D’Souza, P. Poorkaj, M. Hong, D. Nochlin, V. M.-Y. Lee, T. D. Bird, and G. D. Schellenberg, “Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements,” Proc. Natl. Acad. Sci. U. S. A., vol. 96, pp. 5598–5603, may 1999.

[114] B. Ghetti, A. L. Oblak, B. F. Boeve, K. A. Johnson, B. C. Dickerson, and M. Goedert, “Invited review: Frontotemporal dementia caused by microtubule-associated protein tau gene (MAPT) mutations: A chameleon for neuropathology and neuroimaging,” feb 2015.

[115] S. J. Adams, M. A. de Ture, M. McBride, D. W. Dickson, and L. Petrucelli, “Three repeat isoforms of tau inhibit assembly of four repeat tau filaments,” PLoS One, vol. 5, no. 5, p. e10810, 2010.

[116] K. M. M. Schoch, S. L. L. DeVos, R. L. L. Miller, S. J. J. Chun, M. Norrbom, D. F. F. Wozniak, H. N. N. Dawson, C. F. Bennett, F. Rigo, and T. M. M. Miller, “Increased 4R-Tau Induces Pathological Changes in a Human-Tau Mouse Model,” Neuron, vol. 90, pp. 941–947, jun 2016.

[117] B. Bai, C. M. Hales, P.-C. Chen, Y. Gozal, E. B. Dammer, J. J. Fritz, X. Wang, Q. Xia, D. M. Duong, C. Street, G. Cantero, D. Cheng, D. R. Jones, Z. Wu, Y. Li, I. Diner, C. J. Heilman, H. D. Rees, H. Wu, L. Lin, K. E. Szulwach, M. Gearing, E. J. Mufson, D. A. Bennett, T. J. Montine, N. T. Seyfried, T. S. Wingo, Y. E. Sun, P. Jin, J. Hanfelt, D. M. Willcock, A. Levey, J. J. Lah, and J. Peng, “U1 small nuclear ribonucleoprotein complex and RNA splicing alterations in Alzheimer’s disease.,” Proc. Natl. Acad. Sci. U. S. A., pp. 6–11, 2013.

[118] D. Marques-Coelho, L. d. C. C. Iohan, A. R. Melo de Farias, A. Flaig, F. Letournel, M. L. Martin-Négrier, F. Chapon, M. Faisant, C. Godfraind, C. A. Maurage, V. Deramecourt, M. Duchesne, D. Meyronnet, N. Streichenberger, A. M. de Paula, V. Rigau, F. Vandenbos-

Burel, C. Duyckaerts, D. Seilhean, S. Milin, D. C. Chiforeanu, A. Laquerrière, F. Marguet, B. Lannes, J. C. Lambert, and M. R. Costa, "Differential transcript usage unravels gene expression alterations in Alzheimer's disease human brains," npj Aging Mech. Dis., vol. 7, pp. 1–15, dec 2021.

[119] N. A. Twine, K. Janitz, M. R. Wilkins, and M. Janitz, "Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease," PLoS One, vol. 6, no. 1, 2011.

[120] J. D. Mills, T. Nalpathamkalam, H. I. L. Jacobs, C. Janitz, D. Merico, P. Hu, and M. Janitz, "RNA-Seq analysis of the parietal cortex in Alzheimer's disease reveals alternatively spliced isoforms related to lipid metabolism," Neurosci. Lett., vol. 536, pp. 90–95, mar 2013.

[121] M. K. P. Lai, M. M. Esiri, and M. G. K. Tan, "Genome-wide profiling of alternative splicing in Alzheimer's disease," Genomics Data, vol. 2, pp. 290–292, 2014.

[122] J. D. Mills, P. J. Sheahan, D. Lai, J. J. Kril, M. Janitz, and G. T. Sutherland, "The alternative splicing of the apolipoprotein E gene is unperturbed in the brains of Alzheimer's disease patients," Mol. Biol. Rep., vol. 41, pp. 6365–6376, oct 2014.

[123] C. Humphries, M. A. Kohli, P. Whitehead, D. C. Mash, M. A. Pericak-Vance, and J. Gilbert, "Alzheimer disease (AD) specific transcription, DNA methylation and splicing in twenty AD associated loci," Mol. Cell. Neurosci., vol. 67, pp. 37–45, 2015.

[124] M. Magistri, D. Velmeshev, M. Makhmutova, and M. A. Faghihi, "Transcriptomics Profiling of Alzheimer's Disease Reveal Neurovascular Defects, Altered Amyloid-β Homeostasis, and Deregulated Expression of Long Noncoding RNAs," J. Alzheimer's Dis., vol. 48, no. 3, pp. 647–665, 2015.

[125] R. Alkallas, L. Fish, H. Goodarzi, and H. S. Najafabadi, "Inference of RNA decay rate from transcriptional profiling highlights the regulatory programs of Alzheimer's disease," Nat. Commun., vol. 8, no. 1, 2017.

[126] A. Annese, C. Manzari, C. Lionetti, E. Picardi, D. S. Horner, M. Chiara, M. F. Caratozzolo, A. Tullo, B. Fosso, G. Pesole, and A. M. D'Erchia, "Whole transcriptome profiling of Late-Onset Alzheimer's Disease patients provides insights into the molecular changes involved in the disease," Sci. Rep., vol. 8, no. 1, 2018.

[127] E. C. Johnson, E. B. Dammer, D. M. Duong, L. Yin, M. Thambisetty, J. C. Troncoso, J. J. Lah, A. I. Levey, and N. T. Seyfried, "Deep proteomic network analysis of Alzheimer's disease brain reveals alterations in RNA binding proteins and RNA splicing associated with disease," Mol. Neurodegener., vol. 13, pp. 1–22, oct 2018.

[128] S. Han, J. E. Miller, S. Byun, D. Kim, S. L. Risacher, A. J. Saykin, Y. Lee, and K. Nho, "Identification of exon skipping events associated with Alzheimer's disease in the human hippocampus," BMC Med. Genomics, vol. 12, pp. 51–61, jan 2019.

[129] S. Adusumalli, Z. K. Ngian, W. Q. Lin, T. Benoukraf, and C. T. Ong, "Increased intron retention is a post-transcriptional signature associated with progressive aging and Alzheimer's disease," Aging Cell, vol. 18, p. e12928, jun 2019.

[130] C. Fan, K. Chen, J. Zhou, P. pui Wong, D. He, Y. Huang, X. Wang, T. Ling, Y. Yang, and H. Zhao, "Systematic analysis to identify transcriptome-wide dysregulation of Alzheimer's

disease in genes and isoforms," Hum. Genet., vol. 140, pp. 609–623, apr 2021.

[131] M. Yang, Y. Ke, P. Kim, and X. Zhou, "ExonSkipAD provides the functional genomic landscape of exon skipping events in Alzheimer's disease," Brief. Bioinform., vol. 2021, pp. 1–13, jan 2021.

[132] V. García-Escudero, D. Ruiz-Gabarre, R. Gargini, M. Pérez, E. García, R. Cuadros, I. H. Hernández, J. R. Cabrera, R. García-Escudero, J. J. Lucas, F. Hernández, and J. Ávila, "A new non-aggregative splicing isoform of human Tau is decreased in Alzheimer's disease," Acta Neuropathol., vol. 142, pp. 159–177, may 2021.

[133] H. D. Li, C. C. Funk, K. McFarland, E. B. Dammer, M. Allen, M. M. Carrasquillo, Y. Levites, P. Chakrabarty, J. D. Burgess, X. Wang, D. Dickson, N. T. Seyfried, D. M. Duong, J. J. Lah, S. G. Younkin, A. I. Levey, G. S. Omenn, N. Ertekin-Taner, T. E. Golde, and N. D. Price, "Integrative functional genomic analysis of intron retention in human and mouse brain with Alzheimer's disease," Alzheimer's Dement., vol. 17, pp. 984–1004, jun 2021.

[134] B. F. Maziuk, D. J. Apicco, A. L. Cruz, L. Jiang, P. E. Ash, E. L. da Rocha, C. Zhang, W. H. Yu, J. Leszyk, J. F. Abisambra, H. Li, and B. Wolozin, "RNA binding proteins co-localize with small tau inclusions in tauopathy," Acta Neuropathol. Commun., vol. 6, p. 71, aug 2018.

[135] S. M. Rothman, K. Q. Tanis, P. Gandhi, V. Malkov, J. Marcus, M. Pearson, R. Stevens, J. Gilliland, C. Ware, V. Mahadomrongkul, E. O'Loughlin, G. Zeballos, R. Smith, B. J. Howell, J. Klappenbach, M. Kennedy, and C. Mirescu, "Human Alzheimer's disease gene expression signatures and immune profile in APP mouse models: A discrete transcriptomic view of Aβ plaque pathology," J. Neuroinflammation, vol. 15, no. 1, 2018.

[136] D. J. Apicco, C. Zhang, B. Maziuk, L. Jiang, H. I. Ballance, S. Boudeau, C. Ung, H. Li, and B. Wolozin, "Dysregulation of RNA Splicing in Tauopathies," Cell Rep., vol. 29, no. 13, pp. 4377–4388.e4, 2019.

[137] D. A. Salih, S. Bayram, S. Guelfi, R. H. Reynolds, M. Shoai, M. Ryten, J. W. Brenton, D. Zhang, M. Matarin, J. A. Botia, R. Shah, K. J. Brookes, T. Guetta-Baranes, K. Morgan, E. Bellou, D. M. Cummings, V. Escott-Price, and J. Hardy, "Genetic variability in response to amyloid beta deposition influences Alzheimer's disease risk," Brain Commun., vol. 1, jan 2019.

[138] D. Sharon, H. Tilgner, F. Grubert, and M. Snyder, "A single-molecule long-read survey of the human transcriptome," Nat. Biotechnol., vol. 31, no. 11, pp. 1009–1014, 2013.

[139] A. Piovesan, M. Caracausi, F. Antonaros, M. C. Pelleri, and L. Vitale, "GeneBase 1.1: A tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics," Database, vol. 2016, p. baw153, dec 2016.

[140] M. L. Bang, T. Centner, F. Fornoff, A. J. Geach, M. Gotthardt, M. McNabb, C. C. Witt, D. Labeit, C. C. Gregorio, H. Granzier, and S. Labeit, "The complete gene sequence of titin, expression of an unusual 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system," nov 2001.

[141] S. P. Gordon, E. Tseng, A. Salamov, J. Zhang, X. Meng, Z. Zhao, D. Kang, J. Underwood, I. V. Grigoriev, M. Figueroa, J. S. Schilling, F. Chen, and Z. Wang, "Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing," PLoS One, vol. 10, no. 7, 2015.

[142] B. Wang, E. Tseng, M. Regulski, T. A. Clark, T. Hon, Y. Jiao, Z. Lu, A. Olson, J. C. Stein, and D. Ware, "Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing," Nat. Commun., vol. 7, p. 11708, 2016.

[143] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," Nat. Biotechnol., vol. 28, no. 5, pp. 511–515, 2010.

[144] D. R. Koessler, D. J. Knisley, J. Knisley, and T. Haynes, "A predictive model for secondary RNA structure using graph theory and a neural network," BMC Bioinformatics, vol. 11, p. 21, dec 2010.

[145] K. F. Au, V. Sebastiano, P. T. Afshar, J. D. Durruthy, L. Lee, B. A. Williams, H. van Bakel, E. E. Schadt, R. A. Reijo-Pera, J. G. Underwood, and W. H. Wong, "Characterization of the human ESC transcriptome by hybrid sequencing," Proc. Natl. Acad. Sci., vol. 110, no. 50, pp. E4821–E4830, 2013.

[146] T. Steijger, J. F. Abril, P. G. Engström, F. Kokocinski, J. F. Abril, M. Akerman, T. Alioto, G. Ambrosini, S. E. Antonarakis, J. Behr, P. Bertone, R. Bohnert, P. Bucher, N. Cloonan, T. Derrien, S. Djebali, J. Du, S. Dudoit, P. G. Engström, M. Gerstein, T. R. Gingeras, D. Gonzalez, S. M. Grimmond, R. Guigó, L. Habegger, J. Harrow, T. J. Hubbard, C. Iseli, G. Jean, A. Kahles, F. Kokocinski, J. Lagarde, J. Leng, G. Lefebvre, S. Lewis, A. Mortazavi, P. Niermann, G. Rätsch, A. Reymond, P. Ribeca, H. Richard, J. Rougemont, J. Rozowsky, M. Sammeth, A. Sboner, M. H. Schulz, S. M. J. Searle, N. D. Solorzano, V. Solovyev, M. Stanke, T. Steijger, B. J. Stevenson, H. Stockinger, A. Valsesia, D. Weese, S. White, B. J. Wold, J. Wu, T. D. Wu, G. Zeller, D. Zerbino, M. Q. Zhang, T. J. Hubbard, R. Guigó, J. Harrow, and P. Bertone, "Assessment of transcript reconstruction methods for RNA-seq," Nat. Methods, vol. 10, no. 12, pp. 1177–1184, 2013.

[147] R. Stark, M. Grzelak, and J. Hadfield, "RNA sequencing: the teenage years," Nat. Rev. Genet., pp. 1–26, jul 2019.

[148] E. Tseng, W. J. Rowell, O. C. Glenn, T. Hon, J. Barrera, S. Kujawa, and O. Chiba-Falek, "The landscape of SNCA transcripts across synucleinopathies: New insights from long reads sequencing analysis," Front. Genet., vol. 10, no. JUL, 2019.

[149] A. D. Mays, M. Schmidt, G. Graham, E. Tseng, P. Baybayan, R. Sebra, M. Sanda, J. B. Mazarati, A. Riegel, and A. Wellstein, "Single-molecule real-time (SMRT) full-length RNA-sequencing reveals novel and distinct mRNA isoforms in human bone marrow cell subpopulations," Genes (Basel)., vol. 10, p. 253, mar 2019.

[150] K. K. Huang, J. Huang, J. K. L. Wu, M. Lee, S. T. Tay, V. Kumar, K. Ramnarayanan, N. Padmanabhan, C. Xu, A. L. K. Tan, C. Chan, D. Kappei, J. Göke, and P. Tan, "Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer," Genome Biol., vol. 22, pp. 1–24, jan 2021.

[151] H. Tilgner, F. Grubert, D. Sharon, and M. P. Snyder, "Defining a personal, allele-specific, and single-molecule long-read transcriptome," Proc. Natl. Acad. Sci., vol. 111, no. 27, pp. 9869–9874, 2014.

[152] B. Treutlein, O. Gokce, S. R. Quake, and T. C. Südhof, "Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing," Proc. Natl. Acad. Sci., vol. 111, no. 13, pp. E1291–E1299, 2014.

[153] D. Schreiner, T. M. Nguyen, G. Russo, S. Heber, A. Patrignani, E. Ahrné, and P. Scheiffele, "Targeted Combinatorial Alternative Splicing Generates Brain Region-Specific Repertoires of Neurexins," Neuron, vol. 84, no. 2, pp. 386–398, 2014.

[154] E. Tseng, H.-T. Tang, R. R. AlOlaby, L. Hickey, and F. Tassone, "Altered expression of the FMR1 splicing variants landscape in premutation carriers," Biochim. Biophys. Acta - Gene Regul. Mech., vol. 1860, no. 11, pp. 1117–1126, 2017.

[155] T. Aneichyk, W. T. Hendriks, R. Yadav, D. Shin, D. Gao, C. A. Vaine, R. L. Collins, A. Domingo, B. Currall, A. Stortchevoi, T. Multhaupt-Buell, E. B. Penney, L. Cruz, J. Dhakal, H. Brand, C. Hanscom, C. Antolik, M. Dy, A. Ragavendran, J. Underwood, S. Cantsilieris, K. M. Munson, E. E. Eichler, P. Acuña, C. Go, R. D. G. Jamora, R. L. Rosales, D. M. Church, S. R. Williams, S. Garcia, C. Klein, U. Müller, K. C. Wilhelmsen, H. T. Timmers, Y. Sapir, B. J. Wainger, D. Henderson, N. Ito, N. Weisenfeld, D. Jaffe, N. Sharma, X. O. Breakefield, L. J. Ozelius, D. C. Bragg, and M. E. Talkowski, "Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly," Cell, vol. 172, pp. 897–909.e21, feb 2018.

[156] M. Nattestad, S. Goodwin, K. Ng, T. Baslan, F. J. Sedlazeck, P. Rescheneder, T. Garvin, H. Fang, J. Gurtowski, E. Hutton, E. Tseng, C. S. Chin, T. Beck, Y. Sundaravadanam, M. Kramer, E. Antoniou, J. D. McPherson, J. Hicks, W. Richard McCombie, and M. C. Schatz, "Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line," Genome Res., vol. 28, pp. 1126–1135, aug 2018.

[157] A. Dainis, E. Tseng, T. A. Clark, T. Hon, M. Wheeler, and E. Ashley, "Targeted Long-Read RNA Sequencing Demonstrates Transcriptional Diversity Driven by Splice-Site Variation in MYBPC3," Circ. Genomic Precis. Med., vol. 12, p. e002464, may 2019.

[158] E. Flaherty, S. Zhu, N. Barretto, E. Cheng, P. J. Deans, M. B. Fernando, N. Schrode, N. Francoeur, A. Antoine, K. Alganem, M. Halpern, G. Deikus, H. Shah, M. Fitzgerald, I. Ladran, P. Gochman, J. Rapoport, N. M. Tsankova, R. McCullumsmith, G. E. Hoffman, R. Sebra, G. Fang, and K. J. Brennand, "Neuronal impact of patient-specific aberrant NRXN1α splicing," Nat. Genet., vol. 51, pp. 1679–1690, nov 2019.

[159] H. Chen, F. Gao, M. He, X. F. Ding, A. M. Wong, S. C. Sze, A. C. Yu, T. Sun, A. W.-H. Chan, X. Wang, and N. Wong, "Long-Read RNA Sequencing Identifies Alternative Splice Variants in Hepatocellular Carcinoma and Tumor-Specific Isoforms," Hepatology, vol. 70, pp. 1011–1025, sep 2019.

[160] B. Lian, X. Hu, and Z. ming Shao, "Unveiling novel targets of paclitaxel resistance by single molecule long-read RNA sequencing in breast cancer," Sci. Rep., vol. 9, pp. 1–10, dec 2019.

[161] M. T. Bolisetty, G. Rajadinakaran, and B. R. Graveley, "Determining exon connectivity in complex mRNAs by nanopore sequencing," Genome Biol., vol. 16, p. 204, sep 2015.

[162] A. De Roeck, T. Van den Bossche, J. van der Zee, J. Verheijen, W. De Coster, J. Van Dongen, L. Dillen, Y. Baradaran-Heravi, B. Heeman, R. Sanchez-Valle, A. Lladó, B. Nacmias, S. Sorbi, E. Gelpi, O. Grau-Rivera, E. Gómez-Tortosa, P. Pastor, S. Ortega-Cubero, M. A. Pastor, C. Graff, H. Thonberg, L. Benussi, R. Ghidoni, G. Binetti, A. de Mendonça, M. Martins, B. Borroni, A. Padovani, M. R. Almeida, I. Santana, J. Diehl-Schmid, P. Alexopoulos, J. Clarimon, A. Lleó, J. Fortea, M. Tsolaki, M. Koutroumani, R. Matěj, Z. Rohan, P. De Deyn, S. Engelborghs, P. Cras, C. Van Broeckhoven, K. Sleegers, V. Bessi, S. Bagnoli, F. S. do Couto, A. Verdelho, L. Fratiglioni, A. Padovani, Z. Rohan, C. Razquin, E. Lorenzo, E. Iglesias, M. Seijo-Martínez, R. Rene, J. Gascon, J. Campdelacreu, and R. Blesa, "Deleterious

ABCA7 mutations and transcript rescue mechanisms in early onset Alzheimer's disease," Acta Neuropathol., vol. 134, pp. 475–487, sep 2017.

[163] L. C. de Jong, S. Cree, V. Lattimore, G. A. Wiggins, A. B. Spurdle, A. Miller, M. A. Kennedy, and L. C. Walker, "Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events," Breast Cancer Res., vol. 19, no. 1, 2017.

[164] S. A. Hardwick, S. D. Bassett, D. Kaczorowski, J. Blackburn, K. Barton, N. Bartonicek, S. L. Carswell, H. U. Tilgner, C. Loy, G. Halliday, T. R. Mercer, M. A. Smith, and J. S. Mattick, "Targeted, high-resolution RNA sequencing of non-coding genomic regions associated with neuropsychiatric functions," Front. Genet., vol. 10, no. APR, p. 309, 2019.

[165] M. B. Clark, T. Wrzesinski, A. B. Garcia, N. A. Hall, J. E. Kleinman, T. Hyde, D. R. Weinberger, P. J. Harrison, W. Haerty, and E. M. Tunbridge, "Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain," Mol. Psychiatry, vol. 25, pp. 37–47, nov 2020.

[166] A. D. Tang, C. M. Soulette, M. J. van Baren, K. Hart, E. Hrabeta-Robinson, C. J. Wu, and A. N. Brooks, "Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns," Nat. Commun., vol. 11, no. 1, 2020.

[167] L. Tian, J. S. Jabbari, R. Thijssen, Q. Gouil, S. L. Amarasinghe, H. Kariyawasam, S. Su, X. Dong, C. W. Law, A. Lucattini, J. D. Chung, T. Naim, A. Chan, C. H. Ly, G. S. Lynch, J. G. Ryall, C. J. Anttila, H. Peng, M. A. Anderson, A. W. Roberts, D. C. Huang, M. B. Clark, and M. E. Ritchie, "Comprehensive characterization of single cell full-length isoforms in human and mouse with long-read sequencing," bioRxiv, p. 2020.08.10.243543, aug 2020.

[168] E. K. Robinson, P. Jagannatha, S. Covarrubias, M. Cattle, V. Smaliy, R. Safavi, B. Shapleigh, R. Abu-Shumays, M. Jain, S. M. Cloonan, M. Akeson, A. N. Brooks, and S. Carpenter, "Inflammation drives alternative first exon usage to regulate immune genes including a novel iron regulated isoform of aim2," Elife, vol. 10, may 2021.

[169] M. Oka, L. Xu, T. Suzuki, T. Yoshikawa, H. Sakamoto, H. Uemura, A. C. Yoshizawa, Y. Suzuki, T. Nakatsura, Y. Ishihama, A. Suzuki, and M. Seki, "Aberrant splicing isoforms detected by full-length transcriptome sequencing as transcripts of potential neoantigens in non-small cell lung cancer," Genome Biol. 2021 221, vol. 22, pp. 1–30, jan 2021.

[170] A. Bayega, S. Fahiminiya, S. Oikonomopoulos, and J. Ragoussis, "Current and future methods for mRNA analysis: A drive toward single molecule sequencing," in Methods Mol. Biol., vol. 1783, pp. 209–241, Humana Press, New York, NY, 2018.

[171] H. Mathys, J. Davila-Velderrain, Z. Peng, F. Gao, S. Mohammadi, J. Z. Young, M. Menon, L. He, F. Abdurrob, X. Jiang, A. J. Martorell, R. M. Ransohoff, B. P. Hafler, D. A. Bennett, M. Kellis, and L. H. Tsai, "Single-cell transcriptomic analysis of Alzheimer's disease," Nature, vol. 570, pp. 332–337, may 2019.

[172] A. Nott, I. R. Holtman, N. G. Coufal, J. C. Schlachetzki, M. Yu, R. Hu, C. Z. Han, M. Pena, J. Xiao, Y. Wu, Z. Keulen, M. P. Pasillas, C. O'Connor, C. K. Nickl, S. T. Schafer, Z. Shen, R. A. Rissman, J. B. Brewer, D. Gosselin, D. D. Gonda, M. L. Levy, M. G. Rosenfeld, G. McVicker, F. H. Gage, B. Ren, and C. K. Glass, "Brain cell type–specific enhancer–promoter interactome maps and disease-risk association," Science (80-.)., vol. 366, no. 6469, pp. 1134–1139, 2019.

[173] N. Thrupp, C. Sala Frigerio, L. Wolfs, N. G. Skene, N. Fattorelli, S. Poovathingal, Y. Fourne, P. M. Matthews, T. Theys, R. Mancuso, B. de Strooper, and M. Fiers, "Single-Nucleus RNA-Seq Is Not Suitable for Detection of Microglial Activation Genes in Humans," Cell Rep., vol. 32, p. 108189, sep 2020.

[174] M. Olah, V. Menon, N. Habib, M. F. Taga, Y. Ma, C. J. Yung, M. Cimpean, A. Khairallah, G. Coronas-Samano, R. Sankowski, D. Grün, A. A. Kroshilina, D. Dionne, R. A. Sarkis, G. R. Cosgrove, J. Helgager, J. A. Golden, P. B. Pennell, M. Prinz, J. P. G. Vonsattel, A. F. Teich, J. A. Schneider, D. A. Bennett, A. Regev, W. Elyaman, E. M. Bradshaw, and P. L. De Jager, "Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer's disease," Nat. Commun., vol. 11, pp. 1–18, nov 2020.

[175] K. Leng, E. Li, R. Eser, A. Piergies, R. Sit, M. Tan, N. Neff, S. H. Li, R. D. Rodriguez, C. K. Suemoto, R. E. P. Leite, A. J. Ehrenberg, C. A. Pasqualucci, W. W. Seeley, S. Spina, H. Heinsen, L. T. Grinberg, and M. Kampmann, "Molecular characterization of selectively vulnerable neurons in Alzheimer's disease," Nat. Neurosci., vol. 24, pp. 276–287, jan 2021.

[176] A. M. Young, N. Kumasaka, F. Calvert, T. R. Hammond, A. Knights, N. Panousis, J. S. Park, J. Schwartzentruber, J. Liu, K. Kundu, M. Segel, N. A. Murphy, C. E. McMurran, H. Bulstrode, J. Correia, K. P. Budohoski, A. Joannides, M. R. Guilfoyle, R. Trivedi, R. Kirollos, R. Morris, M. R. Garnett, I. Timofeev, I. Jalloh, K. Holland, R. Mannion, R. Mair, C. Watts, S. J. Price, P. J. Kirkpatrick, T. Santarius, E. Mountjoy, M. Ghoussaini, N. Soranzo, O. A. Bayraktar, B. Stevens, P. J. Hutchinson, R. J. Franklin, and D. J. Gaffney, "A map of transcriptional heterogeneity and regulatory variation in human microglia," Nat. Genet., vol. 53, no. 6, pp. 861–868, 2021.

[177] H. Keren-Shaul, A. Spinrad, A. Weiner, O. Matcovitch-Natan, R. Dvir-Szternfeld, T. K. Ulland, E. David, K. Baruch, D. Lara-Astaiso, B. Toth, S. Itzkovitz, M. Colonna, M. Schwartz, and I. Amit, "A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease," Cell, vol. 169, pp. 1276–1290.e17, jun 2017.

[178] H. Mathys, C. Adaikkan, F. Gao, J. Z. Young, E. Manet, M. Hemberg, P. L. De Jager, R. M. Ransohoff, A. Regev, and L. H. Tsai, "Temporal Tracking of Microglia Activation in Neurodegeneration at Single-Cell Resolution," Cell Rep., vol. 21, no. 2, pp. 366–380, 2017.

[179] C. Sala Frigerio, L. Wolfs, N. Fattorelli, N. Thrupp, I. Voytyuk, I. Schmidt, R. Mancuso, W. T. Chen, M. E. Woodbury, G. Srivastava, T. Möller, E. Hudry, S. Das, T. Saido, E. Karran, B. Hyman, V. H. Perry, M. Fiers, and B. De Strooper, "The Major Risk Factors for Alzheimer's Disease: Age, Sex, and Genes Modulate the Microglia Response to Aβ Plaques," Cell Rep., vol. 27, no. 4, pp. 1293–1306.e6, 2019.

[180] K. E. Tansey, D. Cameron, and M. J. Hill, "Genetic risk for Alzheimer's disease is concentrated in specific macrophage and microglial transcriptional networks," Genome Med., vol. 10, no. 1, 2018.

[181] G. Novikova, M. Kapoor, J. Tcw, E. M. Abud, A. G. Efthymiou, S. X. Chen, H. Cheng, J. F. Fullard, J. Bendl, Y. Liu, P. Roussos, J. L. Björkegren, Y. Liu, W. W. Poon, K. Hao, E. Marcora, and A. M. Goate, "Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes," Nat. Commun., vol. 12, pp. 1–14, mar 2021.

[182] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, L. Pinello, P. Skums, A. Stamatakis, C. S. O. Attolini, S. Aparicio, J. Baaijens, M. Balvert, B. de Barbanson, A. Cappuccio, G. Corleone, B. E. Dutilh, M. Florescu, V. Guryev, R. Holmer, K. Jahn, T. J. Lobo, E. M. Keizer,

I. Khatri, S. M. Kielbasa, J. O. Korbel, A. M. Kozlov, T. H. Kuo, B. P. Lelieveldt, I. I. Mandoiu, J. C. Marioni, T. Marschall, F. Mölder, A. Niknejad, L. Raczkowski, M. Reinders, J. de Ridder, A. E. Saliba, A. Somarakis, O. Stegle, F. J. Theis, H. Yang, A. Zelikovsky, A. C. McHardy, B. J. Raphael, S. P. Shah, and A. Schönhuth, "Eleven grand challenges in single-cell data science," 2020.

[183] A. Adil, V. Kumar, A. T. Jan, and M. Asger, "Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis," 2021.

[184] R. E. Workman, A. D. Tang, P. S. Tang, M. Jain, J. R. Tyson, R. Razaghi, P. C. Zuzarte, T. Gilpatrick, A. Payne, J. Quick, N. Sadowski, N. Holmes, J. G. de Jesus, K. L. Jones, C. M. Soulette, T. P. Snutch, N. Loman, B. Paten, M. Loose, J. T. Simpson, H. E. Olsen, A. N. Brooks, M. Akeson, and W. Timp, "Nanopore native RNA sequencing of a human poly(A) transcriptome," Nat. Methods, vol. 16, pp. 1297–1305, dec 2019.

[185] K. Karlsson and S. Linnarsson, "Single-cell mRNA isoform diversity in the mouse brain," BMC Genomics, vol. 18, no. 1, 2017.

[186] I. Gupta, P. G. Collier, B. Haase, A. Mahfouz, A. Joglekar, T. Floyd, F. Koopmans, B. Barres, A. B. Smit, S. A. Sloan, W. Luo, O. Fedrigo, M. E. Ross, and H. U. Tilgner, "Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells," Nat. Biotechnol., vol. 36, pp. 1197–1202, dec 2018.

[187] A. Byrne, A. E. Beaudin, H. E. Olsen, M. Jain, C. Cole, T. Palmer, R. M. DuBois, E. C. Forsberg, M. Akeson, and C. Vollmers, "Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells," Nat. Commun., vol. 8, 2017.

[188] R. Volden, T. Palmer, A. Byrne, C. Cole, R. J. Schmitz, R. E. Green, and C. Vollmers, "Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA," Proc. Natl. Acad. Sci. U. S. A., vol. 115, no. 39, pp. 9726–9731, 2018.

[189] D. R. Garalde, E. A. Snell, D. Jachimowicz, B. Sipos, J. H. Lloyd, M. Bruce, N. Pantic, T. Admassu, P. James, A. Warland, M. Jordan, J. Ciccone, S. Serra, J. Keenan, S. Martin, L. McNeill, E. J. Wallace, L. Jayasinghe, C. Wright, J. Blasco, S. Young, D. Brocklebank, S. Juul, J. Clarke, A. J. Heron, and D. J. Turner, "Highly parallel direct RN A sequencing on an array of nanopores," Nat. Methods, vol. 15, pp. 201–206, mar 2018.

[190] C. Sessegolo, C. Cruaud, C. Da Silva, A. Cologne, M. Dubarry, T. Derrien, V. Lacroix, and J.-M. Aury, "Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules," Sci. Reports 2019 91, vol. 9, pp. 1–12, oct 2019.

[191] M. Singh, G. Al-Eryani, S. Carswell, J. M. Ferguson, J. Blackburn, K. Barton, D. Roden, F. Luciani, T. Giang Phan, S. Junankar, K. Jackson, C. C. Goodnow, M. A. Smith, and A. Swarbrick, "High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes," Nat. Commun., vol. 10, pp. 1–13, jul 2019.

[192] A. Joglekar, A. Prjibelski, A. Mahfouz, P. Collier, S. Lin, A. K. Schlusche, J. Marrocco, S. R. Williams, B. Haase, A. Hayes, J. G. Chew, N. I. Weisenfeld, M. Y. Wong, A. N. Stein, S. A. Hardwick, T. Hunt, Q. Wang, C. Dieterich, Z. Bent, O. Fedrigo, S. A. Sloan, D. Risso, E. D. Jarvis, P. Flicek, W. Luo, G. S. Pitt, A. Frankish, A. B. Smit, M. E. Ross, and H. U. Tilgner, "A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain," Nat. Commun., vol. 12, pp. 1–16, jan 2021.

[193] A. Grubman, G. Chew, J. F. Ouyang, G. Sun, X. Y. Choo, C. McLean, R. K. Simmons, S. Buckberry, D. B. Vargas-Landin, D. Poppe, J. Pflueger, R. Lister, O. J. Rackham, E. Petretto, and J. M. Polo, “A single-cell atlas of entorhinal cortex from individuals with Alzheimer’s disease reveals cell-type-specific gene expression regulation,” Nat. Neurosci., vol. 22, pp. 2087–2097, nov 2019.

[194] M. Yue, A. Hanna, J. Wilson, H. Roder, and C. Janus, “Sex difference in pathology and memory decline in rTg4510 mouse model of tauopathy,” Neurobiol. Aging, vol. 32, pp. 590–603, apr 2011.

[195] M. Ramsden, L. Kotilinek, C. Forster, J. Paulson, E. McGowan, K. SantaCruz, A. Guimaraes, M. Yue, J. Lewis, G. Carlson, M. Hutton, and K. H. Ashe, “Age-dependent neurofibrillary tangle formation, neuron loss, and memory impairment in a mouse model of human tauopathy (P301L),” J. Neurosci., vol. 25, pp. 10637–10647, nov 2005.

[196] T. G. Heffner, J. A. Hartman, and L. S. Seiden, “A rapid method for the regional dissection of the rat brain,” Pharmacol. Biochem. Behav., vol. 13, no. 3, pp. 453–456, 1980.

[197] I. Castanho, Functional genomic characterisation of animal models of AD : relevance to human dementia
Doctoral thesis, University of Exeter, 2019.

[198] O. Mueller and A. Schroeder, “RNA Integrity Number (RIN) – Standardization of RNA Quality Control Application,” Nano, pp. 1–8, 2004.

[199] D. Ramsköld, S. Luo, Y. C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, and R. Sandberg, “Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells,” Nat. Biotechnol., vol. 30, no. 8, pp. 777–782, 2012.

[200] M. Cartolano, B. Huettel, B. Hartwig, R. Reinhardt, and K. Schneeberger, “cDNA library enrichment of full length transcripts for SMRT long read sequencing,” PLoS One, vol. 11, p. e0157779, jun 2016.

[201] R. Transcriptase, “SMARTer ™ Pico PCR cDNA Synthesis Kit User Manual,” Control, vol. 1, no. 634928, pp. 1–31, 2009.

[202] S. G. Acinas, R. Sarma-Rupavtarm, V. Klepac-Ceraj, and M. F. Polz, “PCR-induced sequence artifacts and bias: Insights from comparison of two 16s rRNA clone libraries constructed from the same sample,” Appl. Environ. Microbiol., vol. 71, no. 12, pp. 8966–8969, 2005.

[203] D. I. Ragoussis and D. S. Oikonomopoulos, “RNA Transcriptomics,” Wellcome Genome Campus Adv. Course 2018, pp. 1–111, 2018.

[204] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. DeWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, “Real-time DNA sequencing from single polymerase molecules,” Science (80-.)., vol. 323, pp. 133–138, jan 2009.

[205] K. J. Travers, C. S. Chin, D. R. Rank, J. S. Eid, and S. W. Turner, “A flexible and efficient

template format for circular consensus sequencing and SNP detection," Nucleic Acids Res., vol. 38, no. 15, p. e159, 2010.

[206] H. J. Levene, J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead, and W. W. Webb, "Zero-mode waveguides for single-molecule analysis at high concentrations," Science (80-.)., vol. 299, pp. 682–686, jan 2003.

[207] A. Mccarthy, "Third generation DNA sequencing: Pacific biosciences' single molecule real time technology," jul 2010.

[208] S. Ardui, A. Ameur, J. R. Vermeesch, and M. S. Hestand, "Single molecule real-time (SMRT) sequencing comes of age: Applications and utilities for medical diagnostics," mar 2018.

[209] A. Rhoads and K. F. Au, "PacBio Sequencing and Its Applications," Genomics. Proteomics Bioinformatics, vol. 13, pp. 278–289, oct 2015.

[210] E. W. Loomis, J. S. Eid, P. Peluso, J. Yin, L. Hickey, D. Rank, S. McCalmon, R. J. Hagerman, F. Tassone, and P. J. Hagerman, "Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene," Genome Res., vol. 23, pp. 121–128, jan 2013.

[211] S. Oikonomopoulos, A. Bayega, S. Fahiminiya, H. Djambazian, P. Berube, and J. Ragoussis, "Methodologies for Transcript Profiling Using Long-Read Technologies," jul 2020.

[212] G. M. Sheynkman, K. S. Tuttle, F. Laval, E. Tseng, J. G. Underwood, L. Yu, D. Dong, M. L. Smith, R. Sebra, L. Willems, T. Hao, M. A. Calderwood, D. E. Hill, and M. Vidal, "ORF Capture-Seq as a versatile method for targeted identification of full-length isoforms," Nat. Commun., vol. 11, no. 1, 2020.

[213] H. Li, "Minimap2: Pairwise alignment for nucleotide sequences," Bioinformatics, vol. 34, pp. 3094–3100, sep 2018.

[214] E. Tseng, "cDNA Cupcake."

[215] M. Tardaguila, L. De La Fuente, C. Marti, C. Pereira, F. J. Pardo-Palacios, H. Del Risco, M. Ferrell, M. Mellado, M. Macchietto, K. Verheggen, M. Edelmann, I. Ezkurdia, J. Vazquez, M. Tress, A. Mortazavi, L. Martens, S. Rodriguez-Navarro, V. Moreno-Manzano, and A. Conesa, "SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification," Genome Res., vol. 28, no. 3, pp. 396–411, 2018.

[216] "Pacific Biosciences IsoSeq v3."

[217] K. Križanović, A. Echchiki, J. Roux, and M. Šikić, "Evaluation of tools for long read RNA-seq splice-aware alignment," Bioinformatics, vol. 34, no. 5, pp. 748–754, 2018.

[218] S. Tang, A. Lomsadze, and M. Borodovsky, "Identification of protein coding regions in RNA transcripts," Nucleic Acids Res., vol. 43, pp. e78–e78, jul 2015.

[219] M. Lizio, I. Abugessaisa, S. Noguchi, A. Kondo, A. Hasegawa, C. C. Hon, M. De Hoon, J. Severin, S. Oki, Y. Hayashizaki, P. Carninci, T. Kasukawa, and H. Kawaji, "Update of the FANTOM web resource: Expansion to provide additional transcriptome atlases," Nucleic Acids Res., vol. 47, pp. D752–D758, jan 2019.

[220] A. Nellore, A. E. Jaffe, J. P. Fortin, J. Alquicira-Hernández, L. Collado-Torres, S. Wang, R. A. Phillips, N. Karbhari, K. D. Hansen, B. Langmead, and J. T. Leek, "Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive," Genome Biol., vol. 17, p. 266, dec 2016.

[221] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: Ultrafast universal RNA-seq aligner," Bioinformatics, vol. 29, pp. 15–21, jan 2013.

[222] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," Nat. Biotechnol., vol. 34, pp. 525–527, may 2016.

[223] J. Cocquet, A. Chong, G. Zhang, and R. A. Veitia, "Reverse transcriptase template switching and false alternative transcripts," Genomics, vol. 88, pp. 127–131, jul 2006.

[224] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, "A survey of best practices for RNA-seq data analysis," 2016.

[225] J. Houseley and D. Tollervey, "Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro," PLoS One, vol. 5, no. 8, 2010.

[226] D. K. Nam, S. Lee, G. Zhou, X. Cao, C. Wang, T. Clark, J. Chen, J. D. Rowley, and S. M. Wang, "Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription," Proc. Natl. Acad. Sci. U. S. A., vol. 99, pp. 6152–6156, apr 2002.

[227] R. I. Kuo, E. Tseng, L. Eory, I. R. Paton, A. L. Archibald, and D. W. Burt, "Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human," BMC Genomics, vol. 18, p. 323, dec 2017.

[228] M. Jain, I. T. Fiddes, K. H. Miga, H. E. Olsen, B. Paten, and M. Akeson, "Improved data analysis for the MinION nanopore sequencer," Nat. Methods, vol. 12, pp. 351–356, mar 2015.

[229] N. J. Loman and M. Watson, "Successful test launch for nanopore sequencing," mar 2015.

[230] S. Oikonomopoulos, Y. C. Wang, H. Djambazian, D. Badescu, and J. Ragoussis, "Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations OPEN," Nat. Publ. Gr., vol. 6, 2016.

[231] J. L. Weirather, M. de Cesare, Y. Wang, P. Piazza, V. Sebastiano, X.-J. Wang, D. Buck, and K. F. Au, "Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis," F1000Research, vol. 6, p. 100, 2017.

[232] H. Bayley, "Nanopore sequencing: From imagination to reality," Clin. Chem., vol. 61, no. 1, pp. 25–31, 2015.

[233] N. Ashkenasy, J. Sánchez-Quesada, H. Bayley, and M. R. Ghadiri, "Recognizing a single base in an individual DNA strand: A step toward DNA sequencing in nanopores," Angew. Chemie - Int. Ed., vol. 44, pp. 1401–1404, feb 2005.

[234] E. A. Manrao, I. M. Derrington, M. Pavlenok, M. Niederweis, and J. H. Gundlach, "Nucleotide discrimination with DNA immobilized in the MSPA nanopore," PLoS One, vol. 6, p. 25723, oct 2011.

[235] F. J. Rang, W. P. Kloosterman, and J. de Ridder, "From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy," jul 2018.

[236] Oxford Nanopore Technologies plc., "New nanopore sequencing chemistry in developers' hands; set to deliver Q20+ (99%+) "raw read" accuracy," 2021.

[237] "Nanopore Community Posts for Low Throughput with 1D2."

[238] C. Li, K. R. Chng, E. J. H. Boey, A. H. Q. Ng, A. Wilm, and N. Nagarajan, "INC-Seq: Accurate single molecule reads using nanopore sequencing," Gigascience, vol. 5, p. 34, dec 2016.

[239] A. Leger and T. Leonardi, "pycoQC, interactive quality control for Oxford Nanopore Sequencing," J. Open Source Softw., vol. 4, p. 1236, feb 2019.

[240] O. N. Technology, "Nanopore summary statistics and basic QC tutorial," 2019.

[241] R. R. Wick, L. M. Judd, and K. E. Holt, "Performance of neural network basecalling tools for Oxford Nanopore sequencing," Genome Biol., vol. 20, pp. 1–10, jun 2019.

[242] W. De Coster, S. D'Hert, D. T. Schultz, M. Cruts, and C. Van Broeckhoven, "NanoPack: Visualizing and processing long-read sequencing data," Bioinformatics, vol. 34, pp. 2666–2669, aug 2018.

[243] R. Wick, "rrwick/Porechop: Adapter Trimmer for Oxford Nanopore reads," 2017.

[244] "Pychopper: A tool to identify, orient, trim and rescue full length cDNA reads."

[245] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," EMBnet.journal, vol. 17, p. 10, may 2011.

[246] D. Wyman, G. Balderrama-Gutierrez, F. Reese, S. Jiang, S. Rahmanian, W. Zeng, B. Williams, D. Trout, W. England, S. Chu, R. C. Spitale, A. Tenner, B. Wold, and A. Mortazavi, "A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification," bioRxiv, 2019.

[247] B. Li and C. N. Dewey, "RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome," BMC Bioinformatics, vol. 12, 2011.

[248] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," Genome Biol., vol. 11, pp. 1–9, mar 2010.

[249] S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie, and Q. Gouil, "Opportunities and challenges in long-read sequencing data analysis," feb 2020.

[250] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel, "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data," Genome Biol., vol. 14, pp. 1–13, sep 2013.

[251] Y. Hu, L. Fang, X. Chen, J. F. Zhong, M. Li, and K. Wang, "LIQA: long-read isoform quantification and analysis," Genome Biol., vol. 22, pp. 1–21, dec 2021.

[252] C. Soneson, K. L. Matthes, M. Nowicka, C. W. Law, and M. D. Robinson, "Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage," Genome Biol., vol. 17, pp. 1–15, jan 2016.

[253] A. Mehmood, A. Laiho, M. S. Venälaïnen, A. J. McGlinchey, N. Wang, and L. L. Elo, "Systematic evaluation of differential splicing tools for RNA-seq studies," Brief. Bioinform., vol. 21, pp. 2052–2065, dec 2020.

[254] G. A. Merino, A. Conesa, and E. A. Fernández, "A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies," Brief. Bioinform., vol. 20, no. 2, pp. 471–481, 2019.

[255] L. De La Fuente, Á. Arzalluz-Luque, M. Tardáguila, H. Del Risco, C. Martí, S. Tarazona, P. Salguero, R. Scott, A. Lerma, A. Alastrue-Agudo, P. Bonilla, J. R. Newman, S. Kosugi, L. M. McIntyre, V. Moreno-Manzano, and A. Conesa, "TappAS: A comprehensive computational framework for the analysis of the functional impact of differential splicing," Genome Biol., vol. 21, pp. 1–32, may 2020.

[256] M. J. Nueda, S. Tarazona, and A. Conesa, "Next maSigPro: Updating maSigPro bioconductor package for RNA-seq time series," Bioinformatics, vol. 30, pp. 2598–2602, sep 2014.

[257] A. Conesa, M. J. Nueda, A. Ferrer, and M. Talón, "maSigPro: A method to identify significantly differential expression profiles in time-course microarray experiments," Bioinformatics, vol. 22, pp. 1096–1102, may 2006.

[258] A. Conesa and M. J. Nueda, "maSigPro User's Guide," tech. rep., University of Florida Genetics Institute, 2017.

[259] M. Gonzàlez-Porta, A. Frankish, J. Rung, J. Harrow, and A. Brazma, "Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene," Genome Biol., vol. 14, no. 7, p. R70, 2013.

[260] I. Ezkurdia, J. M. Rodriguez, E. Carrillo-De Santa Pau, J. Vázquez, A. Valencia, and M. L. Tress, "Most highly expressed protein-coding genes have a single dominant isoform," J. Proteome Res., vol. 14, pp. 1880–1887, apr 2015.

[261] S. K. Leung, A. R. Jeffries, I. Castanho, B. T. Jordan, K. Moore, J. P. Davies, E. L. Dempster, N. J. Bray, P. O'Neill, E. Tseng, Z. Ahmed, D. A. Collier, E. D. Jeffery, S. Prabhakar, L. Schalkwyk, C. Jops, M. J. Gandal, G. M. Sheynkman, E. Hannon, and J. Mill, "Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing," Cell Rep., vol. 37, p. 110022, nov 2021.

[262] J. L. Trincado, J. C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott, and E. Eyras, "SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions," Genome Biol., vol. 19, mar 2018.

[263] M. Pertea, G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell, and S. L. Salzberg, "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads," Nat. Biotechnol., vol. 33, no. 3, pp. 290–295, 2015.

[264] B. Wang, V. Kumar, A. Olson, and D. Ware, “Reviving the Transcriptome Studies: An Insight Into the Emergence of Single-Molecule Transcriptome Sequencing,” Front. Genet., vol. 10, p. 384, apr 2019.

[265] A. M. McCartney, E. M. Hyland, P. Cormican, R. J. Moran, A. E. Webb, K. D. Lee, J. Hernandez-Rodriguez, J. Prado-Martinez, C. J. Creevey, J. L. Aspden, J. O. McInerney, T. Marques-Bonet, M. J. O’Connell, and D. Pisani, “Gene Fusions Derived by Transcriptional Readthrough are Driven by Segmental Duplication in Human,” Genome Biol. Evol., vol. 11, pp. 2676–2690, sep 2019.

[266] P. Akiva, A. Toporik, S. Edelheit, Y. Peretz, A. Diber, R. Shemesh, A. Novik, and R. Sorek, “Transcription-mediated gene fusion in the human genome,” Genome Res., vol. 16, pp. 30–36, jan 2006.

[267] L. Statello, C. J. Guo, L. L. Chen, and M. Huarte, “Gene regulation by long non-coding RNAs and its biological functions,” dec 2021.

[268] S. J. Liu, T. J. Nowakowski, A. A. Pollen, J. H. Lui, M. A. Horlbeck, F. J. Attenello, D. He, J. S. Weissman, A. R. Kriegstein, A. A. Diaz, and D. A. Lim, “Single-cell analysis of long non-coding RNAs in the developing human neocortex,” Genome Biol., vol. 17, apr 2016.

[269] Y. Kageyama, T. Kondo, and Y. Hashimoto, “Coding vs non-coding: Translatability of short ORFs found in putative non-coding transcripts,” nov 2011.

[270] M. Guttman, P. Russell, N. T. Ingolia, J. S. Weissman, and E. S. Lander, “Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins,” Cell, vol. 154, pp. 240–251, jul 2013.

[271] N. Hug, D. Longman, and J. F. Cáceres, “Mechanism and regulation of the nonsense-mediated decay pathway,” dec 2015.

[272] Q. Pan, A. L. Saltzman, K. K. Yoon, C. Misquitta, O. Shai, L. E. Maquat, B. J. Frey, and B. J. Blencowe, “Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression,” Genes Dev., vol. 20, pp. 153–158, jan 2006.

[273] Y. Ge and B. T. Porse, “The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression,” BioEssays, vol. 36, pp. 236–243, mar 2014.

[274] B. Mehani, K. Narta, D. Paul, A. Raj, D. Kumar, A. Sharma, L. Kaurani, S. Nayak, D. Dash, A. Suri, C. Sarkar, and A. Mukhopadhyay, “Fusion transcripts in normal human cortex increase with age and show distinct genomic features for single cells and tissues,” Sci. Rep., vol. 10, dec 2020.

[275] L. Zhao, H. Zhang, M. V. Kohnen, K. V. Prasad, L. Gu, and A. S. Reddy, “Analysis of transcriptome and epitranscriptome in plants using pacbio iso-seq and nanopore-based direct RNA sequencing,” mar 2019.

[276] E. Tseng, J. G. Underwood, B. D. Evans Hutzenbiler, S. Trojahn, B. Kingham, O. Shevchenko, E. Bernberg, M. Vierra, C. T. Robbins, H. T. Jansen, and J. L. Kelley, “Long-read isoform sequencing reveals tissue-specific isoform expression between active and hibernating brown bears (Ursus arctos),” G3 Genes|Genomes|Genetics, dec 2021.

[277] F. Muramori, K. Kobayashi, and I. Nakamura, "A quantitative study of neurofibrillary tangles, senile plaques and astrocytes in the hippocampal subdivisions and entorhinal cortex in Alzheimer's disease, normal controls and non-Alzheimer neuropsychiatric diseases," Psychiatry Clin. Neurosci., vol. 52, no. 6, pp. 593–599, 1998.

[278] A. Ishiki, M. Kamada, Y. Kawamura, C. Terao, F. Shimoda, N. Tomita, H. Arai, and K. Furukawa, "Glial fibrillar acidic protein in the cerebrospinal fluid of Alzheimer's disease, dementia with Lewy bodies, and frontotemporal lobar degeneration," J. Neurochem., vol. 136, pp. 258–261, jan 2016.

[279] P. Chatterjee, S. Pedrini, E. Stoops, K. Goozee, V. L. Villemagne, P. R. Asih, I. M. Verberk, P. Dave, K. Taddei, H. R. Sohrabi, H. Zetterberg, K. Blennow, C. E. Teunissen, H. M. Vanderstichele, and R. N. Martins, "Plasma glial fibrillary acidic protein is elevated in cognitively normal older adults at risk of Alzheimer's disease," Transl. Psychiatry, vol. 11, pp. 1–10, jun 2021.

[280] E. Castillo, J. Leon, G. Mazzei, N. Abolhassani, N. Haruyama, T. Saito, T. Saido, M. Hokama, T. Iwaki, T. Ohara, T. Ninomiya, Y. Kiyohara, K. Sakumi, F. M. Laferla, and Y. Nakabeppu, "Comparative profiling of cortical gene expression in Alzheimer's disease patients and mouse models demonstrates a link between amyloidosis and neuroinflammation," Sci. Rep., vol. 7, pp. 1–16, dec 2017.

[281] K. T. Wirz, K. Bossers, A. Stargardt, W. Kamphuis, D. F. Swaab, E. M. Hol, and J. Verhaagen, "Cortical beta amyloid protein triggers an immune response, but no synaptic changes in the APPswe/PS1dE9 Alzheimer's disease mouse model," Neurobiol. Aging, vol. 34, pp. 1328–1342, may 2013.

[282] P. Langfelder and S. Horvath, "WGCNA: An R package for weighted correlation network analysis," BMC Bioinformatics, vol. 9, pp. 1–13, dec 2008.

[283] M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack, "Gene ontology analysis for RNA-seq: accounting for selection bias," Genome Biol., vol. 11, pp. 1–12, feb 2010.

[284] R. F. Roelofs, D. F. Fischer, S. H. Houtman, J. A. Sluijs, W. Van Haren, F. W. Van Leeuwen, and E. M. Hol, "Adult human subventricular, subgranular, and subpial zones contain astrocytes with a specialized intermediate filament cytoskeleton," Glia, vol. 52, pp. 289–300, dec 2005.

[285] W. Kamphuis, C. Mamber, M. Moeton, L. Kooijman, J. A. Sluijs, A. H. Jansen, M. Verveer, L. R. de Groot, V. D. Smith, S. Rangarajan, J. J. Rodríguez, M. Orre, and E. M. Hol, "GFAP isoforms in adult mouse brain with a focus on neurogenic astrocytes and reactive astrogliosis in mouse models of Alzheimer disease," PLoS One, vol. 7, no. 8, 2012.

[286] A. Ishigami, T. Ohsawa, M. Hiratsuka, H. Taguchi, S. Kobayashi, Y. Saito, S. Murayama, H. Asaga, T. Toda, N. Kimura, and N. Maruyama, "Abnormal accumulation of citrullinated proteins catalyzed by peptidylarginine deiminase in hippocampal extracts from patients with Alzheimer's disease," J. Neurosci. Res., vol. 80, pp. 120–128, apr 2005.

[287] H. Wang, K. K. Dey, P. C. Chen, Y. Li, M. Niu, J. H. Cho, X. Wang, B. Bai, Y. Jiao, S. R. Chepyala, V. Haroutunian, B. Zhang, T. G. Beach, and J. Peng, "Integrated analysis of ultra-deep proteomes in cortex, cerebrospinal fluid and serum reveals a mitochondrial signature in Alzheimer's disease," Mol. Neurodegener., vol. 15, pp. 1–20, jul 2020.

[288] J. R. McDermott and A. M. Gibson, "Degradation of Alzheimer's β-amyloid protein by

human cathepsin D," Neuroreport, vol. 7, no. 13, pp. 2163–2166, 1996.

[289] A. Kenessey, P. Nacharaju, L. W. Ko, and S. H. Yen, "Degradation of tau by lysosomal enzyme cathepsin D: Implication for Alzheimer neurofibrillary degeneration," J. Neurochem., vol. 69, no. 5, pp. 2026–2038, 1997.

[290] C. N. Suire, S. O. Abdul-Hay, T. Sahara, D. Kang, M. K. Brizuela, P. Saftig, D. W. Dickson, T. L. Rosenberry, and M. A. Leissring, "Cathepsin D regulates cerebral Aβ42/40 ratios via differential degradation of Aβ42 and Aβ40," Alzheimer's Res. Ther., vol. 12, pp. 1–13, jul 2020.

[291] C. Bieniossek, G. Papai, C. Schaffitzel, F. Garzoni, M. Chaillet, E. Scheer, P. Papadopoulos, L. Tora, P. Schultz, and I. Berger, "The architecture of human general transcription factor TFIID core complex," Nature, vol. 493, no. 7434, pp. 699–702, 2013.

[292] S. E. Wiley, A. N. Murphy, S. A. Ross, P. Van Der Geer, and J. E. Dixon, "MitoNEET is an iron-containing outer mitochondrial membrane protein that regulates oxidative capacity," Proc. Natl. Acad. Sci. U. S. A., vol. 104, pp. 5318–5323, mar 2007.

[293] A. Yamamoto, T. Nagano, S. Takehara, M. Hibi, and S. Aizawa, "Shisa promotes head formation through the inhibition of receptor protein maturation for the caudalizing factors, Wnt and FGF," Cell, vol. 120, pp. 223–235, jan 2005.

[294] T. Takafuta, M. Saeki, T. T. Fujimoto, K. Fujimura, and S. S. Shapiro, "A new member of the LIM protein family binds to filamin B and localizes at stress fibers," J. Biol. Chem., vol. 278, pp. 12175–12181, apr 2003.

[295] L. Jia, J. Piña-Crespo, and Y. Li, "Restoring Wnt/β-catenin signaling is a promising therapeutic strategy for Alzheimer's disease," dec 2019.

[296] Y. F. Chen, T. Y. Chou, I. H. Lin, C. G. Chen, C. H. Kao, G. J. Huang, L. K. Chen, P. N. Wang, C. P. Lin, and T. F. Tsai, "Upregulation of Cisd2 attenuates Alzheimer's-related neuronal loss in mice," J. Pathol., vol. 250, pp. 299–311, mar 2020.

[297] E. Drummond, G. Pires, C. MacMurray, M. Askenazi, S. Nayak, M. Bourdon, J. Safar, B. Ueberheide, and T. Wisniewski, "Phosphorylated tau interactome in the human Alzheimer's disease brain," Brain, vol. 143, no. 9, pp. 2803–2817, 2020.

[298] X. Li, L. Wang, M. Cykowski, T. He, T. Liu, J. Chakranarayan, A. Rivera, H. Zhao, S. Powell, W. Xia, and S. T. Wong, "OCIAD1 contributes to neurodegeneration in Alzheimer's disease by inducing mitochondria dysfunction, neuronal vulnerability and synaptic damages," EBioMedicine, vol. 51, 2020.

[299] E. L. D. Stefania S Policicchio, Jonathan P Davies, Barry Chioza, Joe Burrage, Jonathan Mill, "Fluorescence-activated nuclei sorting (FANS) on human post-mortem cortex tissue enabling the isolation of distinct neural cell populations for multiple omic profiling," protocols.io, 2020.

[300] G. Monti, M. Kjolby, A. M. G. Jensen, M. Allen, J. Reiche, P. L. Møller, R. Comaposada-Baró, B. E. Zolkowski, C. Vieira, M. M. Jørgensen, I. E. Holm, P. N. Valdmanis, N. Wellner, C. B. Vægter, S. J. Lincoln, A. Nykjær, N. Ertekin-Taner, J. E. Young, M. Nyegaard, and O. M. Andersen, "Expression of an alternatively spliced variant of SORL1 in neuronal dendrites is decreased in patients with Alzheimer's disease," Acta Neuropathol. Commun., vol. 9,

no. 1, 2021.

[301] R. De Paoli-Iseppi, J. Gleeson, and M. B. Clark, "Isoform Age - Splice Isoform Profiling Using Long-Read Technologies," 2021.

[302] L. T. Nordestgaard, A. Tybjærg-Hansen, B. G. Nordestgaard, and R. Frikke-Schmidt, "Loss-of-function mutation in ABCA1 and risk of Alzheimer's disease and cerebrovascular disease," Alzheimer's Dement., vol. 11, pp. 1430–1438, dec 2015.

[303] R. Koldamova, N. F. Fitz, and I. Lefterov, "ATP-binding cassette transporter A1: From metabolism to neurodegeneration," dec 2014.

[304] N. F. Fitz, A. A. Cronican, M. Saleem, A. H. Fauq, R. Chapman, I. Lefterov, and R. Koldamova, "Abca1 deficiency affects Alzheimer's disease-like phenotype in human ApoE4 but not in ApoE3-targeted replacement mice," J. Neurosci., vol. 32, pp. 13125–13136, sep 2012.

[305] S. Steinberg, H. Stefansson, T. Jonsson, H. Johannsdottir, A. Ingason, H. Helgason, P. Sulem, O. T. Magnusson, S. A. Gudjonsson, U. Unnsteinsdottir, A. Kong, S. Helisalmi, H. Soininen, J. J. Lah, D. Aarsland, T. Fladby, I. D. Ulstein, S. Djurovic, S. B. Sando, L. R. White, G. P. Knudsen, L. T. Westlye, G. Selbæk, I. Giegling, H. Hampel, M. Hiltunen, A. I. Levey, O. A. Andreassen, D. Rujescu, P. V. Jonsson, S. Bjornsson, J. Snaedal, and K. Stefansson, "Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease," Nat. Genet., vol. 47, pp. 445–447, may 2015.

[306] E. Cuyvers, A. De Roeck, T. Van den Bossche, C. Van Cauwenberghe, K. Bettens, S. Vermeulen, M. Mattheijssens, K. Peeters, S. Engelborghs, M. Vandenbulcke, R. Vandenberghe, P. P. De Deyn, C. Van Broeckhoven, and K. Sleegers, "Mutations in ABCA7 in a Belgian cohort of Alzheimer's disease patients: A targeted resequencing study," Lancet Neurol., vol. 14, pp. 814–822, aug 2015.

[307] K. L. Guennec, G. Nicolas, O. Quenez, C. Charbonnier, D. Wallon, C. Bellenguez, B. Grenier-Boley, S. Rousseau, A. C. Richard, A. Rovelet-Lecrux, D. Bacq, J. G. Garnier, R. Olaso, A. Boland, V. Meyer, J. F. Deleuze, P. Amouyel, H. M. Munter, G. Bourque, M. Lathrop, T. Frebourg, R. Redon, L. Letenneur, J. F. Dartigues, F. Pasquier, A. Rollin-Sillaire, E. Génin, J. C. Lambert, D. Hannequin, and D. Campion, "ABCA7 rare variants and Alzheimer disease risk," Neurology, vol. 86, pp. 2134–2137, jun 2016.

[308] K. E. Grear, I. F. Ling, J. F. Simpson, J. L. Furman, C. R. Simmons, S. L. Peterson, F. A. Schmitt, W. R. Markesbery, Q. Liu, J. E. Crook, S. G. Younkin, G. Bu, and S. Estus, "Expression of SORL1 and a novel SORL1 splice variant in normal and Alzheimers disease brain," Mol. Neurodegener., vol. 4, pp. 1–13, nov 2009.

[309] H. N. Cukier, B. W. Kunkle, B. N. Vardarajan, S. Rolati, K. L. Hamilton-Nelson, M. A. Kohli, P. L. Whitehead, B. A. Dombroski, D. Van Booven, R. Lang, D. M. Dykxhoorn, L. A. Farrer, M. L. Cuccaro, J. M. Vance, J. R. Gilbert, G. W. Beecham, E. R. Martin, R. M. Carney, R. Mayeux, G. D. Schellenberg, G. S. Byrd, J. L. Haines, and M. A. Pericak-Vance, "ABCA7 frameshift deletion associated with Alzheimer disease in African Americans," Neurol. Genet., vol. 2, jun 2016.

[310] A. R. Smith, R. G. Smith, J. Burrage, C. Troakes, S. Al-Sarraj, R. N. Kalaria, C. Sloan, A. C. Robinson, J. Mill, and K. Lunnon, "A cross-brain regions study of ANK1 DNA methylation in different neurodegenerative diseases," Neurobiol. Aging, vol. 74, pp. 70–76, 2019.

[311] K. Lunnon, R. Smith, E. Hannon, P. L. De Jager, G. Srivastava, M. Volta, C. Troakes, S. Al-Sarraj, J. Burrage, R. Macdonald, D. Condliffe, L. W. Harries, P. Katsel, V. Haroutunian, Z. Kaminsky, C. Joachim, J. Powell, S. Lovestone, D. A. Bennett, L. C. Schalkwyk, and J. Mill, "Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease," Nat. Neurosci., vol. 17, no. 9, pp. 1164–1170, 2014.

[312] D. Mastroeni, S. Sekar, J. Nolz, E. Delvaux, K. Lunnon, J. Mill, W. S. Liang, and P. D. Coleman, "ANK1 is up-regulated in laser captured microglia in Alzheimer's brain; the importance of addressing cellular heterogeneity," PLoS One, vol. 12, no. 7, 2017.

[313] A. M. Jablonski, L. Warren, M. Usenovic, H. Zhou, J. Sugam, S. Parmentier-Batteur, and B. Voleti, "Astrocytic expression of the Alzheimer's disease risk allele, ApoEϵ4, potentiates neuronal tau pathology in multiple preclinical models," Sci. Rep., vol. 11, no. 1, 2021.

[314] Y. W. Zhang, R. Thompson, H. Zhang, and H. Xu, "APP processing in Alzheimer's disease," jan 2011.

[315] P. K. Panegyres, K. Zafiris-Toufexis, and B. A. Kakulas, "Amyloid precursor protein gene isoforms in Alzheimer's disease and other neurodegenerative disorders," J. Neurol. Sci., vol. 173, pp. 81–92, feb 2000.

[316] J. Chapuis, F. Hansmannel, M. Gistelinck, A. Mounier, C. Van Cauwenberghe, K. V. Kolen, F. Geller, Y. Sottejeau, D. Harold, P. Dourlen, B. Grenier-Boley, Y. Kamatani, B. Delepine, F. Demiautte, D. Zelenika, N. Zommer, M. Hamdane, C. Bellenguez, J. F. Dartigues, J. J. Hauw, F. Letronne, A. M. Ayral, K. Sleegers, A. Schellens, L. V. Broeck, S. Engelborghs, P. P. De Deyn, R. Vandenberghe, M. O'Donovan, M. Owen, J. Epelbaum, M. Mercken, E. Karran, M. Bantscheff, G. Drewes, G. Joberty, D. Campion, J. N. Octave, C. Berr, M. Lathrop, P. Callaerts, D. Mann, J. Williams, L. Buée, I. Dewachter, C. Van Broeckhoven, P. Amouyel, D. Moechars, B. Dermaut, and J. C. Lambert, "Increased expression of BIN1 mediates Alzheimer genetic risk by modulating tau pathology," Mol. Psychiatry, vol. 18, pp. 1225–1234, feb 2013.

[317] R. J. Andrew, P. De Rossi, P. Nguyen, H. R. Kowalski, A. J. Recupero, T. Guerbette, S. V. Krause, R. C. Rice, L. Laury-Kleintop, S. L. Wagner, and G. Thinakaran, "Reduction of the expression of the late-onset Alzheimer's disease (AD) risk-factor BIN1 does not affect amyloid pathology in an AD mouse model," J. Biol. Chem., vol. 294, pp. 4477–4487, mar 2019.

[318] A. Crotti, H. R. Sait, K. M. McAvoy, K. Estrada, A. Ergun, S. Szak, G. Marsh, L. Jandreski, M. Peterson, T. L. Reynolds, I. Dalkilic-Liddle, A. Cameron, E. Cahir-McFarland, and R. M. Ransohoff, "BIN1 favors the spreading of Tau via extracellular vesicles," Sci. Rep., vol. 9, pp. 1–20, jul 2019.

[319] M. Taga, V. A. Petyuk, C. White, G. Marsh, Y. Ma, H. U. Klein, S. M. Connor, A. Kroshilina, C. J. Yung, A. Khairallah, M. Olah, J. Schneider, K. Karhohs, A. E. Carpenter, R. Ransohoff, D. A. Bennett, A. Crotti, E. M. Bradshaw, and P. L. De Jager, "BIN1 protein isoforms are differentially expressed in astrocytes, neurons, and microglia: neuronal and astrocyte BIN1 are implicated in tau pathology," Mol. Neurodegener., vol. 15, pp. 1–19, jul 2020.

[320] M. Sartori, T. Mendes, S. Desai, A. Lasorsa, A. Herledan, N. Malmanche, P. Mäkinen, M. Marttinen, I. Malki, J. Chapuis, A. Flaig, A. C. Vreulx, M. Ciancia, P. Amouyel, F. Leroux, B. Déprez, F. X. Cantrelle, D. Maréchal, L. Pradier, M. Hiltunen, I. Landrieu, D. Kilinc, Y. Herault, J. Laporte, and J. C. Lambert, "BIN1 recovers tauopathy-induced long-term memory deficits in mice and interacts with Tau through Thr348 phosphorylation," Acta Neuropathol., vol. 138, pp. 631–652, oct 2019.

[321] M. Malik, J. F. Simpson, I. Parikh, B. R. Wilfred, D. W. Fardo, P. T. Nelson, and S. Estus, “CD33 Alzheimer’s risk-altering polymorphism, CD33 expression, and exon 2 splicing,” J. Neurosci., vol. 33, no. 33, pp. 13320–13325, 2013.

[322] A. Griciuc, A. Serrano-Pozo, A. R. Parrado, A. N. Lesinski, C. N. Asselin, K. Mullin, B. Hooli, S. H. Choi, B. T. Hyman, and R. E. Tanzi, “Alzheimer’s disease risk gene cd33 inhibits microglial uptake of amyloid beta,” Neuron, vol. 78, no. 4, pp. 631–643, 2013.

[323] A. Griciuc, S. Patel, A. N. Federico, S. H. Choi, B. J. Innes, M. K. Oram, G. Cereghetti, D. McGinty, A. Anselmo, R. I. Sadreyev, S. E. Hickman, J. El Khoury, M. Colonna, and R. E. Tanzi, “TREM2 Acts Downstream of CD33 in Modulating Microglial Pathology in Alzheimer’s Disease,” Neuron, vol. 103, no. 5, pp. 820–835.e7, 2019.

[324] A. Bhattacherjee, J. Jung, S. Zia, M. Ho, G. Eskandari-Sedighi, C. D. St. Laurent, K. A. McCord, A. Bains, G. Sidhu, S. Sarkar, J. R. Plemel, and M. S. Macauley, “The CD33 short isoform is a gain-of-function variant that enhances Aβ1−42 phagocytosis in microglia,” Mol. Neurodegener., vol. 16, dec 2021.

[325] B. Padhy, B. Hayat, G. G. Nanda, P. P. Mohanty, and D. P. Alone, “Pseudoexfoliation and Alzheimer’s associated CLU risk variant, rs2279590, lies within an enhancer element and regulates CLU, EPHX2 and PTK2B gene expression,” Hum. Mol. Genet., vol. 26, pp. 4519–4529, nov 2017.

[326] K. Bettens, S. Vermeulen, C. Van Cauwenberghe, B. Heeman, B. Asselbergh, C. Robberecht, S. Engelborghs, M. Vandenbulcke, R. Vandenberghe, P. P. De Deyn, M. Cruts, C. Van Broeckhoven, and K. Sleegers, “Reduced secreted clusterin as a mechanism for Alzheimer-associated CLU mutations,” Mol. Neurodegener., vol. 10, pp. 1–12, jul 2015.

[327] R. J. Jackson, J. Rose, J. Tulloch, C. Henstridge, C. Smith, and T. L. Spires-Jones, “Clusterin accumulates in synapses in Alzheimer’s disease and is increased in apolipoprotein E4 carriers,” Brain Commun., vol. 1, jan 2019.

[328] I. F. Ling, J. Bhongsatiern, J. F. Simpson, D. W. Fardo, and S. Estus, “Genetics of clusterin isoform expression and Alzheimer’s disease risk,” PLoS One, vol. 7, no. 4, p. 33923, 2012.

[329] S. K. Herring, H. J. Moon, P. Rawal, A. Chhibber, and L. Zhao, “Brain clusterin protein isoforms and mitochondrial localization,” Elife, vol. 8, 2019.

[330] Y. Zhou, I. Hayashi, J. Wong, K. Tugusheva, J. J. Renger, and C. Zerbinatti, “Intracellular clusterin interacts with brain isoforms of the bridging integrator 1 and with the microtubule-associated protein Tau in Alzheimer’s Disease,” PLoS One, vol. 9, p. e103187, jul 2014.

[331] C. M. Karch, A. T. Jeng, P. Nowotny, J. Cady, C. Cruchaga, and A. M. Goate, “Expression of Novel Alzheimer’s Disease Risk Genes in Control and Alzheimer’s Disease Brains,” PLoS One, vol. 7, no. 11, p. 50976, 2012.

[332] S. Ishigaki, Y. Riku, Y. Fujioka, K. Endo, N. Iwade, K. Kawai, M. Ishibashi, S. Yokoi, M. Katsuno, H. Watanabe, K. Mori, A. Akagi, O. Yokota, S. Terada, I. Kawakami, N. Suzuki, H. Warita, M. Aoki, M. Yoshida, and G. Sobue, “Aberrant interaction between FUS and SFPQ in neurons in a wide range of FTLD spectrum diseases,” Brain, vol. 143, pp. 2398–2405, aug 2020.

[333] K. Bhaskar, G. A. Hobbs, S. H. Yen, and G. Lee, "Tyrosine phosphorylation of tau accompanies disease progression in transgenic mouse models of tauopathy," Neuropathol. Appl. Neurobiol., vol. 36, no. 6, pp. 462–477, 2010.

[334] J. Chin, J. J. Palop, J. Puoliväli, C. Massaro, N. Bien-Ly, H. Gerstein, K. Scearce-Levie, E. Masliah, and L. Mucke, "Fyn kinase induces synaptic and cognitive impairments in a transgenic mouse model of Alzheimer's disease," J. Neurosci., vol. 25, no. 42, pp. 9694–9703, 2005.

[335] C. Lee, C. Y. Low, P. T. Francis, J. Attems, P. T. Wong, M. K. Lai, and M. G. Tan, "An isoform-specific role of FynT tyrosine kinase in Alzheimer's disease," J. Neurochem., vol. 136, pp. 637–650, feb 2016.

[336] C. Y. Low, J. H. Lee, F. T. Lim, C. Lee, C. Ballard, P. T. Francis, M. K. Lai, and M. G. Tan, "Isoform-specific upregulation of FynT kinase expression is associated with tauopathy and glial activation in Alzheimer's disease and Lewy body dementias," Brain Pathol., vol. 31, pp. 253–266, mar 2021.

[337] K. R. Bowles, D. A. Pugh, L. M. Oja, B. M. Jadow, K. Farrell, K. Whitney, A. Sharma, J. D. Cherry, T. Raj, A. C. Pereira, J. F. Crary, and A. M. Goate, "Dysregulated coordination of MAPT exon 2 and exon 10 splicing underlies different tau pathologies in PSP and AD," Acta Neuropathol., vol. 143, pp. 225–243, feb 2022.

[338] I. Parikh, D. W. Fardo, and S. Estus, "Genetics of PICALM expression and Alzheimer's disease," PLoS One, vol. 9, p. e91242, mar 2014.

[339] K. Ando, R. De Decker, C. Vergara, Z. Yilmaz, S. Mansour, V. Suain, K. Sleegers, M. A. de Fisenne, S. Houben, M. C. Potier, C. Duyckaerts, T. Watanabe, L. Buée, K. Leroy, and J. P. Brion, "Picalm reduction exacerbates tau pathology in a murine tauopathy model," Acta Neuropathol., vol. 139, pp. 773–789, apr 2020.

[340] P. Narayan, G. Sienski, J. M. Bonner, Y. T. Lin, J. Seo, V. Baru, A. Haque, B. Milo, L. A. Akay, A. Graziosi, Y. Freyzon, D. Landgraf, W. R. Hesse, J. Valastyan, M. I. Barrasa, L. H. Tsai, and S. Lindquist, "PICALM Rescues Endocytic Defects Caused by the Alzheimer's Disease Risk Factor APOE4," Cell Rep., vol. 33, no. 1, 2020.

[341] K. Ando, K. Tomimura, V. Sazdovitch, V. Suain, Z. Yilmaz, M. Authelet, M. Ndjim, C. Vergara, M. Belkouch, M. C. Potier, C. Duyckaerts, and J. P. Brion, "Level of PICALM, a key component of clathrin-mediated endocytosis, is correlated with levels of phosphotau and autophagy-related proteins and is associated with tau inclusions in AD, PSP and Pick disease," Neurobiol. Dis., vol. 94, pp. 32–43, 2016.

[342] A. Giralt, B. de Pins, C. Cifuentes-Díaz, L. López-Molina, A. T. Farah, M. Tible, V. Deramecourt, S. T. Arold, S. Ginés, J. Hugon, and J. A. Girault, "PTK2B/Pyk2 overexpression improves a mouse model of Alzheimer's disease," Exp. Neurol., vol. 307, pp. 62–73, sep 2018.

[343] R. G. Smith, E. Pishva, G. Shireby, A. R. Smith, J. A. Roubroeks, E. Hannon, G. Wheildon, D. Mastroeni, G. Gasparoni, M. Riemenschneider, A. Giese, A. J. Sharp, L. Schalkwyk, V. Haroutunian, W. Viechtbauer, D. L. van den Hove, M. Weedon, D. Brokaw, P. T. Francis, A. J. Thomas, S. Love, K. Morgan, J. Walter, P. D. Coleman, D. A. Bennett, P. L. De Jager, J. Mill, and K. Lunnon, "A meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex," Nat. Commun., vol. 12, no. 1, 2021.

[344] P. L. De Jager, G. Srivastava, K. Lunnon, J. Burgess, L. C. Schalkwyk, L. Yu, ..., J. Mill, and D. A. Bennett, "Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci," Nat. Neurosci., vol. 17, no. 9, pp. 1156–1163, 2014.

[345] R. Lardenoije, J. A. Roubroeks, E. Pishva, M. Leber, H. Wagner, A. Iatrou, ..., A. Ramirez, and D. L. Van Den Hove, "Alzheimer's disease-associated (hydroxy)methylomic changes in the brain and blood," Clin. Epigenetics, vol. 11, pp. 1–15, nov 2019.

[346] R. Levy, C. Levet, K. Cohen, M. Freeman, R. Mott, F. Iraqi, and Y. Gabet, "A genome-wide association study in mice reveals a role for Rhbdf2 in skeletal homeostasis," Sci. Rep., vol. 10, no. 1, 2020.

[347] K. Beyer and A. Ariza, "α-Synuclein posttranslational modification and alternative splicing as a trigger for neurodegeneration.," aug 2013.

[348] K. Beyer, "α-Synuclein structure, posttranslational modification and alternative splicing as aggregation enhancers," sep 2006.

[349] M. V. Fernández, K. Black, D. Carrell, B. Saef, J. Budde, Y. Deming, B. Howells, J. L. Del-Aguila, S. Ma, C. Bi, J. Norton, R. Chasse, J. Morris, A. Goate, and C. Cruchaga, "SORL1 variants across Alzheimer's disease European American cohorts," Eur. J. Hum. Genet., vol. 24, pp. 1828–1830, sep 2016.

[350] A. Knupp, S. Mishra, R. Martinez, J. E. Braggin, M. Szabo, C. Kinoshita, D. W. Hailey, S. A. Small, S. Jayadev, and J. E. Young, "Depletion of the AD Risk Gene SORL1 Selectively Impairs Neuronal Endosomal Traffic Independent of Amyloidogenic APP Processing," Cell Rep., vol. 31, no. 9, 2020.

[351] A. Sobue, O. Komine, Y. Hara, F. Endo, H. Mizoguchi, S. Watanabe, S. Murayama, T. Saito, T. C. Saido, N. Sahara, M. Higuchi, T. Ogi, and K. Yamanaka, "Microglial gene signature reveals loss of homeostatic microglia associated with neurodegeneration of Alzheimer's disease," Acta Neuropathol. Commun., vol. 9, no. 1, 2021.

[352] N. Brouwers, K. Bettens, I. Gijselinck, S. Engelborghs, B. A. Pickut, H. Van Miegroet, A. G. Montoya, M. Mattheijssens, K. Peeters, P. P. De Deyn, M. Cruts, K. Sleegers, and C. Van Broeckhoven, "Contribution of TARDBP to Alzheimer's disease genetic etiology," J. Alzheimer's Dis., vol. 21, no. 2, pp. 423–430, 2010.

[353] S. A. Davis, K. A. Gan, J. A. Dowell, N. J. Cairns, and M. A. Gitcho, "TDP-43 expression influences amyloidβ plaque deposition and tau aggregation," Neurobiol. Dis., vol. 103, pp. 154–162, jul 2017.

[354] A. M. Herman, P. J. Khandelwal, B. B. Stanczyk, G. W. Rebeck, and C. E. Moussa, "β-Amyloid triggers ALS-associated TDP-43 pathology in AD models," Brain Res., vol. 1386, pp. 191–199, 2011.

[355] Y. Wang, M. Cella, K. Mallinson, J. D. Ulrich, K. L. Young, M. L. Robinette, S. Gilfillan, G. M. Krishnan, S. Sudhakar, B. H. Zinselmeyer, D. M. Holtzman, J. R. Cirrito, and M. Colonna, "TREM2 lipid sensing sustains the microglial response in an Alzheimer's disease model," Cell, vol. 160, pp. 1061–1071, mar 2015.

[356] D. L. Kober, J. M. Alexander-Brett, C. M. Karch, C. Cruchaga, M. Colonna, M. J. Holtzman, and T. J. Brett, "Neurodegenerative disease mutations in TREM2 reveal a functional surface

and distinct loss-of-function mechanisms," Elife, vol. 5, dec 2016.

[357] R. Guerreiro, A. Wojtas, J. Bras, M. Carrasquillo, E. Rogaeva, E. Majounie, C. Cruchaga, C. Sassi, J. S. Kauwe, S. Younkin, L. Hazrati, J. Collinge, J. Pocock, T. Lashley, J. Williams, J.-C. Lambert, P. Amouyel, A. Goate, R. Rademakers, K. Morgan, J. Powell, P. St. George-Hyslop, A. Singleton, and J. Hardy, "TREM2 Variants in Alzheimer's Disease," N. Engl. J. Med., vol. 368, pp. 117–127, jan 2013.

[358] K. Kiianitsa, I. Kurtz, N. Beeman, M. Matsushita, W. M. Chien, W. H. Raskind, and O. Korvatska, "Novel TREM2 splicing isoform that lacks the V-set immunoglobulin domain is abundant in the human brain," J. Leukoc. Biol., vol. 110, pp. 829–837, nov 2021.

[359] J. L. Del-Aguila, B. A. Benitez, Z. Li, U. Dube, K. A. Mihindukulasuriya, J. P. Budde, F. H. Farias, M. V. Fernández, L. Ibanez, S. Jiang, R. J. Perrin, N. J. Cairns, J. C. Morris, O. Harari, and C. Cruchaga, "TREM2 brain transcript-specific studies in AD and TREM2 mutation carriers," Mol. Neurodegener., vol. 14, pp. 1–13, may 2019.

[360] K. I. Lee, H. T. Lee, H. C. Lin, H. J. Tsay, F. C. Tsai, S. K. Shyue, and T. S. Lee, "Role of transient receptor potential ankyrin 1 channels in Alzheimer's disease," J. Neuroinflammation, vol. 13, pp. 1–16, apr 2016.

[361] M. Payrits, E. Borbely, S. Godo, D. Ernszt, A. Kemeny, J. Kardos, E. Szoke, and E. Pinter, "Genetic deletion of TRPA1 receptor attenuates amyloid beta- 1-42 (Aβ1-42)-induced neurotoxicity in the mouse basal forebrain in vivo," Mech. Ageing Dev., vol. 189, 2020.

[362] I. van Steenoven, B. Noli, C. Cocco, G. L. Ferri, P. Oeckl, M. Otto, M. J. Koel-Simmelink, C. Bridel, W. M. van der Flier, A. W. Lemstra, and C. E. Teunissen, "VGF peptides in cerebrospinal fluid of patients with dementia with lewy bodies," Int. J. Mol. Sci., vol. 20, no. 19, 2019.

[363] B. Bai, X. Wang, Y. Li, P. C. Chen, K. Yu, K. K. Dey, ..., G. Yu, and J. Peng, "Deep Multilayer Brain Proteomics Identifies Molecular Networks in Alzheimer's Disease Progression," Neuron, vol. 105, no. 6, pp. 975–991.e7, 2020.

[364] N. D. Beckmann, W. J. Lin, M. Wang, A. T. Cohain, A. W. Charney, P. Wang, ..., S. R. Salton, and E. E. Schadt, "Multiscale causal networks identify VGF as a key regulator of Alzheimer's disease," Nat. Commun., vol. 11, no. 1, 2020.

[365] D. Wyman and A. Mortazavi, "TranscriptClean: Variant-aware correction of indels, mismatches and splice junctions in long-read transcripts," Bioinformatics, vol. 35, no. 2, pp. 340–342, 2019.

[366] G. Pertea and M. Pertea, "GFF Utilities: GffRead and GffCompare," F1000Research, vol. 9, 2020.

[367] L. Wang, H. J. Park, S. Dasari, S. Wang, J. P. Kocher, and W. Li, "CPAT: Coding-potential assessment tool using an alignment-free logistic regression model," Nucleic Acids Res., vol. 41, apr 2013.

[368] U. Braunschweig, N. L. Barbosa-Morais, Q. Pan, E. N. Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia, and B. J. Blencowe, "Widespread intron retention in mammals functionally tunes transcriptomes," Genome Res., vol. 24, pp. 1774–1786, nov 2014.

[369] A. De Roeck, C. Van Broeckhoven, and K. Sleegers, “The role of ABCA7 in Alzheimer’s disease: evidence from genomics, transcriptomics and methylomics,” aug 2019.

[370] P. G. Gallagher and B. G. Forget, “An alternate promoter directs expression of a truncated, muscle-specific isoform of the human ankyrin 1 gene,” J. Biol. Chem., vol. 273, no. 3, pp. 1339–1348, 1998.

[371] R. Yan, S. Lai, Y. Yang, H. Shi, Z. Cai, V. Sorrentino, H. Du, and H. Chen, “A novel type 2 diabetes risk allele increases the promoter activity of the muscle-specific small ankyrin 1 gene,” Sci. Rep., vol. 6, 2016.

[372] E. M. Foster, A. Dangla-Valls, S. Lovestone, E. M. Ribe, and N. J. Buckley, “Clusterin in Alzheimer’s disease: Mechanisms, genetics, and lessons from other pathologies,” 2019.

[373] T. A. Shelkovnikova, H. K. Robinson, J. A. Southcombe, N. Ninkina, and V. L. Buchman, “Multistep process of FUS aggregation in the cell cytoplasm involves RNA-dependent and RNA-independent mechanisms,” Hum. Mol. Genet., vol. 23, no. 19, pp. 5211–5226, 2014.

[374] H. Seelaar, K. Y. Klijnsma, I. De Koning, A. Van Der Lugt, W. Z. Chiu, A. Azmani, A. J. Rozemuller, and J. C. Van Swieten, “Frequency of ubiquitin and FUS-positive, TDP-43-negative frontotemporal lobar degeneration,” J. Neurol., vol. 257, no. 5, pp. 747–753, 2010.

[375] X. Wang, J. C. Schwartz, and T. R. Cech, “Nucleic acid-binding specificity of human FUS protein,” Nucleic Acids Res., vol. 43, no. 15, pp. 7535–7543, 2015.

[376] B. de Pins, T. Mendes, A. Giralt, and J. A. Girault, “The Non-receptor Tyrosine Kinase Pyk2 in Brain Function and Neurological and Psychiatric Diseases,” oct 2021.

[377] Y. Yoshino, T. Mori, T. Yoshida, K. Yamazaki, Y. Ozaki, T. Sao, Y. Funahashi, J. I. Iga, and S. I. Ueno, “Elevated mRNA Expression and Low Methylation of SNCA in Japanese Alzheimer’s Disease Subjects,” J. Alzheimer’s Dis., vol. 54, no. 4, pp. 1349–1357, 2016.

[378] J. Verheijen, T. Van den Bossche, J. van der Zee, S. Engelborghs, R. Sanchez-Valle, A. Lladó, ..., C. Van Broeckhoven, and K. Sleegers, “A comprehensive study of the genetic impact of rare variants in SORL1 in European early-onset Alzheimer’s disease,” Acta Neuropathol., vol. 132, no. 2, pp. 213–224, 2016.

[379] A. Meneses, S. Koga, J. O’Leary, D. W. Dickson, G. Bu, and N. Zhao, “TDP-43 Pathology in Alzheimer’s Disease,” dec 2021.

[380] J. P. Quinn, S. E. Kandigian, B. A. Trombetta, S. E. Arnold, and B. C. Carlyle, “VGF as a biomarker and therapeutic target in neurodegenerative and psychiatric diseases,” Brain Commun., vol. 3, no. 4, 2021.

[381] J. Tapial, K. C. Ha, T. Sterne-Weiler, A. Gohr, U. Braunschweig, A. Hermoso-Pulido, M. Quesnel-Vallières, J. Permanyer, R. Sodaei, Y. Marquez, L. Cozzuto, X. Wang, M. Gómez-Velázquez, T. Rayon, M. Manzanares, J. Ponomarenko, B. J. Blencowe, and M. Irimia, “An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms,” Genome Res., vol. 27, pp. 1759–1768, oct 2017.

[382] R. I. Kuo, Y. Cheng, R. Zhang, J. W. Brown, J. Smith, A. L. Archibald, and D. W. Burt, “Illuminating the dark side of the human transcriptome with long read transcript sequencing,”

BMC Genomics, vol. 21, no. 1, 2020.

[383] F. Jiang, J. Zhang, Q. Liu, X. Liu, H. Wang, J. He, and L. Kang, "Long-read direct RNA sequencing by 5'-Cap capturing reveals the impact of Piwi on the widespread exonization of transposable elements in locusts," RNA Biol., vol. 16, no. 7, pp. 950–959, 2019.

[384] M. Loose, S. Malla, and M. Stout, "Real-time selective sequencing using nanopore technology," Nat. Methods, vol. 13, pp. 751–754, jul 2016.

[385] A. Payne, N. Holmes, T. Clarke, R. Munro, B. J. Debebe, and M. Loose, "Readfish enables targeted nanopore sequencing of gigabase-sized genomes," Nat. Biotechnol., vol. 39, pp. 442–450, nov 2021.

[386] S. Kovaka, Y. Fan, B. Ni, W. Timp, and M. C. Schatz, "Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED," Nat. Biotechnol., vol. 39, pp. 431–441, nov 2021.

[387] D. Zhang, S. Guelfi, S. Garcia-Ruiz, B. Costa, R. H. Reynolds, K. D'Sa, W. Liu, T. Courtin, A. Peterson, A. E. Jaffe, J. Hardy, J. A. Botía, L. Collado-Torres, and M. Ryten, "Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders," Sci. Adv., vol. 6, jun 2020.

[388] Y. Fu, P. H. Wu, T. Beane, P. D. Zamore, and Z. Weng, "Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers," BMC Genomics, vol. 19, pp. 1–14, jul 2018.

[389] M. L. Tress, F. Abascal, and A. Valencia, "Alternative Splicing May Not Be the Key to Proteome Complexity," feb 2017.

[390] B. J. Blencowe, "The Relationship between Alternative Splicing and Proteomic Complexity," 2017.

[391] M. L. Tress, F. Abascal, and A. Valencia, "Most Alternative Isoforms Are Not Functionally Important," 2017.

[392] M. Reixachs-Solé and E. Eyras, "Uncovering the impacts of alternative splicing on the proteome with current omics techniques," 2022.

[393] R. M. Miller, B. T. Jordan, M. M. Mehlferber, E. D. Jeffery, C. Chatzipantsiou, S. Kaur, R. J. Millikin, Y. Dai, S. Tiberi, P. J. Castaldi, M. R. Shortreed, C. J. Luckey, A. Conesa, L. M. Smith, A. Deslattes Mays, and G. M. Sheynkman, "Enhanced protein isoform characterization through long-read proteogenomics," Genome Biol., vol. 23, pp. 1–28, mar 2022.

[394] X. Wang, X. You, J. D. Langer, J. Hou, F. Rupprecht, I. Vlatkovic, C. Quedenau, G. Tushev, I. Epstein, B. Schaefke, W. Sun, L. Fang, G. Li, Y. Hu, E. M. Schuman, and W. Chen, "Full-length transcriptome reconstruction reveals a large diversity of RNA and protein isoforms in rat hippocampus," Nat. Commun., vol. 10, pp. 1–15, nov 2019.

[395] A. A. Davis, C. E. Leyns, and D. M. Holtzman, "Intercellular Sp read of Protein Aggregates in Neurodegenerative Disease," oct 2018.

[396] X. Yang, H. Han, D. D. DeCarvalho, F. D. Lay, P. A. Jones, and G. Liang, "Gene body methy-

lation can alter gene expression and is a therapeutic target in cancer," Cancer Cell, vol. 26, no. 4, pp. 577–590, 2014.

[397] S. Shukla, E. Kavak, M. Gregory, M. Imashimizu, B. Shutinoski, M. Kashlev, P. Oberdoerffer, R. Sandberg, and S. Oberdoerffer, "CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing," Nature, vol. 479, no. 7371, pp. 74–79, 2011.

[398] J. Zhang, Y. Z. Zhang, J. Jiang, and C. G. Duan, "The Crosstalk Between Epigenetic Mechanisms and Alternative RNA Processing Regulation," 2020.

[399] R. F. Luco, M. Allo, I. E. Schor, A. R. Kornblihtt, and T. Misteli, "Epigenetics in alternative pre-mRNA splicing," 2011.

[400] K. Nagata, M. Takahashi, Y. Matsuba, F. Okuyama-Uchimura, K. Sato, S. Hashimoto, T. Saito, and T. C. Saido, "Generation of App knock-in mice reveals deletion mutations protective against Alzheimer's disease-like pathology," Nat. Commun., vol. 9, pp. 1–7, may 2018.

[401] L. Serneels, D. T'Syen, L. Perez-Benito, T. Theys, M. G. Holt, and B. De Strooper, "Modeling the β-secretase cleavage site and humanizing amyloid-beta precursor protein in rat and mouse to study Alzheimer's disease," Mol. Neurodegener., vol. 15, pp. 1–11, dec 2020.

[402] D. C. Tan, S. Yao, A. Ittner, J. Bertz, Y. D. Ke, L. M. Ittner, and F. Delerue, "Generation of a New Tau Knockout (tau Δex1) Line Using CRISPR/Cas9 Genome Editing in Mice," J. Alzheimer's Dis., vol. 62, pp. 571–578, jan 2018.

[403] S. Picelli, O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg, "Full-length RNA-seq from single cells using Smart-seq2," Nat. Protoc., vol. 9, no. 1, pp. 171–181, 2014.