Thesis Title

Institution Name

Author Name

Day Month Year

Abstract

Abstract goes here

Dedication

To mum and dad

Acknowledgements

I want to thank...

Declarations

All mouse samples used in chapter X were obtained from Eli Lilly & Co.Ltd., Windleesham (United Kingdom).

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

- RNA extractions from mouse samples was performed by Dr Isabel Castanho
- Short-read RNA Sequencing was prepared by Dr Isabel Castanho, Audrey Farbos and Dr Karen Moore at the University of Exeter Sequencing Service
- Sample loading and machine operation for Iso-Seq targeted sequencing of the final two batches (described in Chapter X) by Dr Stefania Policicchio and Dr Aaron Jeffries at the University of Exeter Sequencing Service
- Nanopore targeted sequencing (described in Chapter X) was performed with Dr Aaron Jeffries at the University of Exeter Sequencing Service

Contents

List of Figures

List of Tables

List of equations

Abbreviations

3'SS	3' Splice Site
5'SS	5' Splice Site
A3SS	Alternative 3' Splice Site
A5SS	Alternative 5' Splice Site
AD	Alzheimer's disease
APA	Alternative Poly-Adenylation
APOE	Apolipoprotein E
APP	Amyloid Precursor Protein
AS	Alternative Splicing
ATI	Alternative Transcription Initiation
BACE	Beta-secretase
BIN1	Bridging Integrator
BPP	Branchpoinnt Binding Protein
BPS	Branch Point Sequence
CLU	Clusterin
CPM	Counts per million

CR1	Complement Receptor 1
DIE	Differential Isoform Expression
DS	Differential Splicing
EOAD	Early Onset Alzheimer's Disease
EST	Expressed Sequence Tags
FAD	Familial's Alzheimer's Disease
FDR	False Discovery Rate
GFAP	Glial Fibrillary Acidic Protein
GWAS	Genome-wide association studies
IR	Intron Retention
IR-isoforms	Intron-retained isoforms
Iso-Seq	Isoform Sequencing
lncRNA	Long non-coding RNA
LOAD	Late Onset Alzheimer's Disease
miRNA	micro RNA
NATs	Natural Antisense Transcripts
NFT	Neurofibrillary tangles
NMD	Nonsense Mediated Decay
NMD-isoforms	Isoforms characterised with nonsense mediated decay
ONT	Oxford Nanopore Technologies
ORF	Open Reading Frame
PacBio	Pacific Biosciences

PICALM	Phosphatidylinositol Binding Clathrin Assembly Protein
PPT	Polypyrimidine Tract
PSEN1	Presenilin 1
PSEN2	Presenilin 2
PSI	Percent-Spliced In
RNA-Seq	RNA-Sequencing
RPKM	Reads of a transcript sequence per Millions
SAGE	Serial Analysis of Gene Expression
SE	Skipped Exon
SLR	Synthetic Long Read
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
snRNPs	Small Nuclear Ribonucleoproteins
spISO-seq	Sparse Isoform Sequencing
TMM	Trimmed Mean of M-values
ToFU	Transcript isOforms: Full-length and Unassembled
TSS	Transcription Start Sites
TTS	Transcription Termination Sites

Chapter 1

Introduction

1.1 Alzheimer's Disease

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 XXX worldwide with numbers expecting to increase to X by 2050, ensuing both a heavy economic and social burden amounting to £XXX each year. Despite international efforts to better understand the disorder for drug discovery and development, there are currently no cure and existing medication only act to reduce symptoms.

1.1.1 Pathology

The symptoms of AD are underpinned by both morphological and molecular changes in the brain, initially in the temporal lobes (hippocampus and entorhinal cortex) and later in the frontal lobes. Conversely, the occipital lobes, motor cortex, and the cerebellum are relatively resistant to neuronal degeneration even in advanced stages of AD.

Neuroimaging scans and post-mortem brain analysis from patients reveal significant brain atrophy caused by neuronal and synaptic loss. Further microscopic examination reveal accumulation of beta-amyloid (Abeta) in amyloid plaques and aggregation of tau in neurofibrillary

tangles, which are now believed to manifest years before presentation of clinical symptoms and diagnosis. In addition to these neuropathological changes, there is increasing evidence for the causative role of the innate immune system. Despite the well characterisation of these neuropathological hallmarks, the exact biological mechanisms driving AD onset and pathogenesis are still widely unknown.

1.1.2 The two hallmarks: plaques and tangles

Amyloid plaques are extracellular deposits of amyloid-beta, short fragments produced from sequential cleavage of APP (amyloid precursor protein), a transmembrane protein involved in synapse formation and stability, by beta- and gamma-secretase (BACE). It is thought that in AD, the processing of APP is altered resulting in imbalanced ratio of longer (and more aggregating) and shorter forms of Abeta, with an increased propensity to form plaques that disrupt synaptic transmission and cause neuronal apoptosis.

Neurofibrillary tangles (NFT) are dense intracellular aggregates of misfolded and hyperphosphorylated tau, which is a microtubule-associated protein involved in microtubule maintenance and stability. It is thought that in AD, the increased phosphorylation of tau induces detachment from microtubule with an increased propensity to form tangles of paired helical filaments that disrupt microtubule function and subsequent axonal growth and transport. The degree/amount of neurofibrillary tangle formation is further found to be closely associated to the severity of AD, allowing AD classification into 6 stages (BRAAK stages) that are defined by the spread of NFTs.

1.1.3 Genetic Component

AD is commonly known to affect people who are aged 65 and above (termed late-onset Alzheimer's disease, LOAD), with younger patients accounting for 5% of total AD cases (termed early-onset Alzheimer's disease, EOAD). While LOAD is complex with a heterogeneous genetic composition and a heritability of 50-80%, EOAD is almost completely genetically determined with EOAD patients presenting a clear familial autosomal dominant pattern of inheritance (Familial Alzheimer's disease, FAD) (Jarmolowicz et al. 2015); to date, more than 160 highly-penetrant, causative mutation have been identified in EOAD, all located within three genes involved in amyloid plaque formation: APP, PSEN1 and PSEN2 (presenilin 1 and 2, which are components

of BACE) (Chai, 2007).

Despite challenges to identity causative mutations in LOAD, being a complex disorder with a heterogeneous etiology, the emergence of genome-wide association studies (GWAS) and subsequent meta-analyses has facilitated the identification of multiple genetic loci that are associated with an increased risk of developing LOAD. These genetic loci are typically changes or variants of single DNA base-pair (single-nucleotide polymorphisms – SNPs) that are more commonly found in individuals with LOAD than without.

To date, the most recent GWAS meta-analysis of 74,000 AD individuals identified over XX significant LOAD risk loci, many of which were annotated to the non-coding cis regulatory regions of gene (Lambert et al., 2013). At least 42 genes/loci have been associated with LOAD at genome-wide significance in at least one GWAS.[1] These genes included BIN1 (bridging integrator 1), CLU (clusterin), CR1 (complement receptor 1), PICALM (phosphatidylinositol binding clathrin assembly protein), with the most significant genetic locus annotated to APOE (apolipoprotein E); inheritance of both APOE allele increases the risk of AD development by X%. Common biological pathways emerging from these GWAS studies are immune response, lipid metabolism, endocytosis, and cell adhesion molecule (CAM) pathways ([1]).

Collectively, these common but low penetrant variants, with the exception of APOE, contribute modestly to the risk of developing AD, highlighting the polygenic nature of AD. The mechanisms behind these variants currently remain poorly understood, however they typically fall into three main biological pathways that may play an important role: the immune system and inflammatory responses, cholesterol and lipid metabolism, and endosomal vesicle recycling. Comprehensive case-control examination of genes proximal to these LOAD-associated variants have further revealed significant differential changes in gene expression and splicing (Humphries et al. 2015), implicating the role of transcriptomic dysregulation in AD pathogenesis. The very fact that most variants lie within the introns rather than exons suggest that it is the fine tune balance of gene expression and regulation that is at play, emphasising the importance epigenomic and transcriptomic studies.

1.1.4 Mouse Models

Molecular changes in both genes and regulatory regions are highly conserved between human AD and mouse model neurodegeneration,

1.1.5 Currently available mouse models in AD

1.2 Gene expression and regulation

Common observation from gene expression analysis is that genes typically express multiple isoforms, and the greater the number of annotated isoforms, the greater the number of expressed isoforms (with a plateau of 12 isoforms).[2] These isoforms are defined as mRNA transcripts that are transcribed from the same gene locus, but generated and processed in a different manner either through alternative transcriptional start sites (Alternative transcription initiation (ATI)), alternative polyadenylation sites (APA) or alternative splicing (AS)). However, as most studies are performed on bulk-tissues, it is unclear whether this is a consequence of multiple isoforms in one single cell or from multiple isoforms from multiple single cells. Perhaps assumed but the expression of alternative isoform is also not consistent, with usually a dominant isoform.[2]

MicroRNA), 22 nucleotides long, involved in regulation of gene expression through various ways, including promotion of transcript degradation and inhibition of translation machinery. This is typically achieved by the contact of miRNA with the 3'UTR of mRNA. It is estimated that up to XX% of genes are regulated by miRNAs, and has been found to multiple roles in immune functions.

1.3 Transcriptional profiling

Transcriptome profiling by the identification of full landscape of transcribed elements is critical to elucidate the functional relationship between the genomic loci and molecular mechanisms that drive development and diseases. Transcriptome profiling of disease-relevant tissue has enabled discovery of pathogenic coding and non-coding splicing variants in rare diseases, that would have otherwise been missed by exome and whole-genome sequencing in Mendialian disease diagnosis (,[34])

With Mendialian diseases such as Duchenne muscular dystrophy, pathogenic variants that result in aberrant splicing (exon inclusion, exon skipping, exon extension, intronic splice gain, exonic splice gain) can have significant downstream impacts (i.e. loss of function). A genetic variant can result in aberrant splicing in the following ways ([3]):

- variant at the splicing donor or acceptor site resulting in a masked splicing site and downstream alternative site used for splicing, thus exonic extension
- variant at the splicing donor or acceptor site resulting in masked splicing site, exon skipping
- variant within an intron (cryptic splice site), resulting in a strong splicing site and thus intronic splice gain

1.3.1 Alternative Splicing

Alternative splicing and polyadenylation is a widespread phenomenon that facilitates generation of multiple distinct mRNA transcripts or isoforms from one gene, which are subsequently translated to different protein isoforms with unique, and potentially, antagonistic functions.[5] AS further regulates gene expression through various mechanisms: non-sense mediated decay, miRNA-mediated mRNA degradation, altered translational efficiency of isoforms. In contrast, alternative polyadenylation regulates RNA transportation, localization, stability, and translation by generating splice isoforms with different cleavage sites.

Alternative splicing is essential in shaping transcriptome and proteome diversity - over 95% of 22,000 protein-coding multi-exonic human genes are estimated to undergo alternative splicing,[6] with up to 70% containing multiple polyadenylation sites and 60% with two or more promoters from alternative transcription start sites.[?] Each gene is estimated to have on av-

erage six transcript isoforms,[7] and this figure is likely to increase with more transcriptomic studies. It occurs most prevalently in the brain implicating its role in neuronal development and maintenance (Pan et al., 2008) (Mazin et al., 2014) (Raj, Blencowe, 2015). It is predicted that a single cell, with a transcription of 600,000 molecules, will have generated 5 - 15 conservative isoforms per gene, and 2-4 exon cassette isoforms ([8]) (a single oligodendrocyte contained 2000 conservative transcripts associated with 700 genes, and 1000 unique isoforms).

Mechanism

Nuclear pre-mRNA splicing involves the removal of non-coding sequences (introns) from the mRNA precursors and ligation of coding sequences (exons). This relies on a concerted and regulated assembly of the spliceosome - a multimegaton, dynamic ribonucleoprotein complex - by its recognition and stepwise-bindingo sequence elements within the pre-mRNA (cis-elements), and a group of regulating splicing factor proteins (trans-elements). There are two types of spliceosome - major and minor - both of which involve the activity of five uridine-rich small nuclear ribonucleoproteins (snRNP) and numerous non-snRNP proteins.[9] Using a similar mechanism but composed of different snRNPs, the minor spliceosome removes less than 1% (0.4%) of introns[10] and is thus referred to as "U12-dependent non-canonical splicing", as opposed to "U2-dependent canonical splicing" with major spliceosome.

Correct splicing first requires the recognition of short sequence motifs upstream (5' splice site - 5'SS - or donor site) and downstream (3' splice site - 3'SS - or acceptor site) of the intron/exon boundary and the branch point sequence[11,12] (BPS (Figure 1.1a). The 5'SS is typically defined by a conserved nine-nucleotide sequence that predominant contains a GU(T) dinucleotide, whereas the 3'SS is defined by a variable length polypyrimidine tract (PPT) followed by a conserved AG dinucleotide.[9] Almost all introns in both human and mouse are flanked by the GT-AG splice site dinucleotides[13] (termed splice junctions), with other variations such as GC-AG and AT-AC known to exist in very minute proportions; GC-AG and AT-AC comprises ~0.9%[14] and ~0.09%[14] of human splice sites and are processed by the major and minor spliceosome, respectively. The branch point sequence is defined by a highly conserved sequence, distinguished by an adenosine, and is typically found within 18-40 nucleotides upstream of the 3'SS.[9]

In the classic splicing model (Figure 1.1), spliceosome assembly first involves the identification of 5' and 3' splice sites, followed by sequential assembly of the spliceosome components as the spliceosome matures and activates for catalysis[15] - intron excision is primarily executed by two transesterification reactions. Recent studies suggest that this process occurs co-transcriptionally, such that the intron can be identified and removed as soon as it is synthesised by the RNA polymerase (Pol II).

Events

Isoforms can differ at the 5' (alternative transcript start sites - TSSs), exons (alternative splicing) and 3' end (alternative transcription termination sites - TTSs)). Exon splicing can be further divided into alternative splice sites (alternative 5'-splice site, alternative 3'-splice site), exon skipping and intron retention. AS events can be classified into five different types:

- Intron retention , defined by the presence of an exon which overlaps with the intron of another transcript within the same gene. IR can introduce stop codons, subsequently prompting non-sense mediated decay but can also change open reading frame , generating functionally different variant
- Skipped exon , defined by the presence a missed exon which is completely overlapped with an intron of another transcript
- Alternative 5' splice site
- Alternative 3' splice site
- Mutually Exclusive Exon

In addition to above five common categories, many other complex types, such as alternative position, i.e., alternative 3' and 5' site (Wang and Brendel, 2006), AS and transcriptional initiation (ASTI) (Nagasaki et al., 2006) alternative first exons (Chen et al., 2007), and composite patterns (Wang and Rio, 2018), can occur."

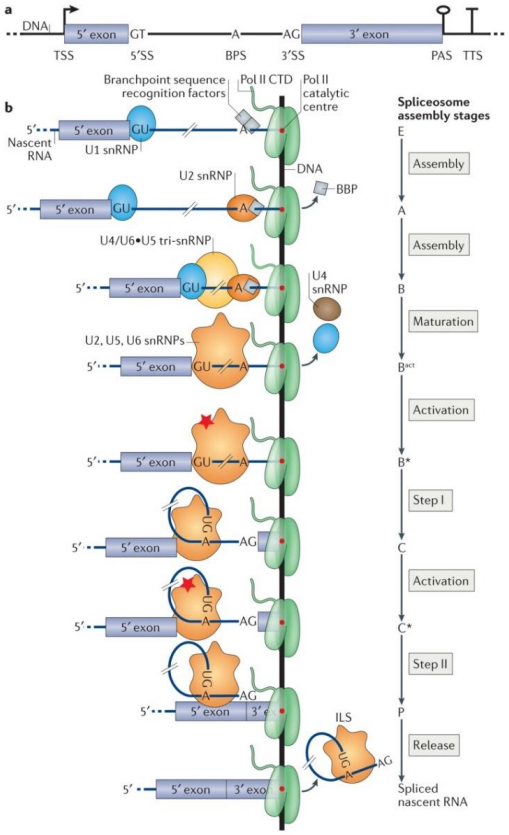


Figure 1.1: Intron removal is catalysed by the assembly and complex rearrangement of spliceosome. a): Consensus sequence of splice sites that demarcate the intron/exon boundary and are essential for recruitment of spliceosomal snRNPs. **b)** Co-transcriptional assembly of spliceosome with stepwise interaction of spliceosomal snRNPs, with 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. The "A" here denotes to the adenosine of BPS
3. B pre-catalytic spliceosome complex with recruitment of U4,U5 and U6 snRNPs
4. Bact 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 (first step of transesterification)
6. C catalytically-active complex with further conformational changes in the U2 snRNA to C* complex, with nucleophilic attack of the 5'SS to 3'SS (second step of transesterification)
7. P (Post-spliceosome complex). The mRNA product is then released from remaining spliceosome (ILS), now bound to the intron lariat. The snRNPs can then disassociate and be recycled for next cycle of splicing.

BPS - Branch point sequence, CTD - carboxyl-terminal domain, ILS - Intron lariat spliceosome, PAS - poly(A) site, SS - Splice site, TSS - Transcription start site,TTS - Transcription termination site. Figure is taken from Herzel et al. 2017.[15]

Nonsense mediated decay (NMD) products are alternatively spliced isoforms that are not translated into proteins, by containing an early stop codon. A premature termination-translation codon highly supportive of NMD is defined by a stop codon within at least 50-55 base pairs upstream of splice junctions.

Fusion Transcripts are a consequence of trans-splicing event of merging two separately encoded pre-mRNA into one transcript

Long non-coding RNA are polyadenylated RNA with more than 200 nucleotides.

Natural Antisense Transcripts

3'Polyadenylation Polyadenylation of 3'end of mRNA regulates mRNA stability and translation efficiency. Studies using long-read sequencing of human transcriptome have revealed differences in poly(A) length distribution between genes, and even between isoforms of the same gene with protein-coding isoforms having shorter poly-A tails than intron-retaining isoforms (.[16] This is line with studies showing that hyperadenylation targets intron-retaining transcripts for degradation ([17])

Allele-specific expression Preferential transcription of RNA from the paternal or maternal copy, which can be assessed using long-read sequencing from coverage of heterozygous SNP.

1.3.2 Short-read RNA-sequencing

Transcriptomic profiling of AD in human and mouse models (determination of changes in splicing patterns) have been traditionally performed using exon microarrays and more recently, RNA-Sequencing (RNA-Seq) (Table X). Multiple methods for transcriptome profiling in the past:

- One of the first methods of transcriptome profiling is to use multiple expressed sequence tags (EST)), short oligonucleotide tags, that can be sequenced - Serial Analysis of Gene Expression).
- Hybridisation of cDNA to oligonucleotides on an array (microarray) i.e Affymetrix's GeneChips, also allowing examination of individual exons

- Quantitative PCR for validation of expression data

Through massively-parallel sequencing of amplified DNA templates in a "sequence-by-synthesis" fashion to generate short-reads (Figure X) rather than relying on hybridization of target and probe, RNA-Seq allows deep surveying of the entire transcriptome, with transcript identification and quantification, and interrogation of alternative splicing events by discovery of splice variants and polymorphisms. With greater signal-to-noise ratio and higher nucleotide-level resolution, has been effective in identifying AS events such as exon skipping and intron retention, with the establishment of its role in diseases (E. T. Wang et al., 2008).

Typically, several millions of short reads (ranging in length between 200bp - 700bp depending on the sequencing platform) are generated and aligned to genome to identify transcribed sequences. Major advances, including generation of reads that retain information on transcript orientation, allowing input of low yield or quality, have revolutionised the field. Also, now possible to sequence transcriptome *de novo*, allowing characterisation of novel organisms.

However, despite its power to identify and quantify gene expression (transcriptional profiling at a gene level) even with paired-end sequencing of both ends of a library fragment, RNA-Seq is severely limited in assembling and reconstructing transcripts due to the reliance of short-reads that are only able to span a small part of the transcript rather than the full length (Figure X)[18];[19] short reads have an average length of 100-500bp, whereas transcripts are on average 2-3kb - 50% of human transcripts are > 2.5Kb[20] and ranges from 60bp to 103kb[20,21] - the longest known human processed transcript to date is Titin with 363 exons and spanning 106kb.[22] In particular, there are three transcriptional features that are difficult to characterise with short reads:[23]

1. Transcript start sites (TSS) and Transcript termination sites (TTS) ,for which any interior multiple TSS and TTS sites within a transcribed locus would be undetected due to overlapping exons and splicing junctions, and low coverage
2. Exon chaining given that short-reads typically only span one splice junction. Thus, while short-reads may able to accurately identify the exons present, the exact sequence and linking of the exons are predicted by short-read assemblers with challenges (Figure **??**).

3. Transcriptional Noise, particularly of reads in intronic regions that are falsely identified as intron retention, or of reads in intergenic regions that are erroneously classified as fusion gene.

It is therefore unclear which combination of exons are spliced in, and whether alternative (distant) exons pairs are included in mutually exclusive or independent fashion (i.e. whether events are coordinated though some distant alternative exons have shown to be correlated included ([24])Furthermore, short-read RNA-sequencing fails to capture the connectivity of exons and informs whether the alternative processive events are coordinated (coordination is defined by two or more alternative RNA processing events are dependent of each other and the probability of this occurrence is greater than the observation of the sole event). –> Molecular co-association of distant human alternative exons

Various bioinformatic packages have been developed to assemble these short reads into transcripts, by probabilistically assigning and mapping reads to isoforms and exon-exon boundary or XXX, to identify and estimate transcript abundance (Figure X) (Trapnell et al., 2010)(Kingsford, Schatz, & Pop, 2010)(Au et al., 2013). This, however, requires complex computational analysis and has resulted in conflicting outcomes and limited success, compounded by the fact that alternative transcripts often have significant overlaps and only a minor proportion of reads span splicing junctions. These tools further rely heavily on reference annotation libraries (RefSeq/Ensembl) or predefined splicing events, which may be inaccurate or incomplete; resulting in prediction of transcripts that do not exist (false positives) or fails to detect true transcripts (false negatives) particularly with genes that have large number of variants (Au et al., 2013). Pre-defined models are particularly limiting when comparing splicing profiles between different conditions, such as control versus transgenic mice, as any splicing changes observed are likely to be AD-specific. While there are tools that are de novo, these typically generate different and often conflicting results [Table X].

Attempts to overcome challenges with transcriptome assembly included generation of "synthetic long reads", by tagging full-length complementary DNAs with unique molecular identifiers (UMIs) before cluster amplification and sequencing on Illumina (Tilgner et al., 2015). With the presence of UMIs, transcript isoforms can be reconstructed for up to 4Kb for isoform discovery and expression analysis (Stark, Grzelak, & Hadfield, 2019). [However…] RNA-Seq

is thus impaired to profile the transcriptome at an isoform-level, investigate cis-acting mechanisms with transcripts, and characterise the functional aspects of isoform diversity (Tardaguila et al., 2018)(Hayer et al., 2015).

1.3.3 Long-read sequencing approaches

The limitations with RNA-Seq were addressed with the emergence of long-read, third-generation sequencing approaches, which generated longer reads that were able to span the full-length transcript. Rather than massively-parallel sequencing of templates in "wash-and-scan" fashion that resulted in de-phasing and subsequently shorter reads, both platforms allowed real-time sequencing of templates in an uninterrupted and processive manner. Two technologies currently dominate this space: Single Molecule Real Time (SMRT) from Pacific Biosciences (PacBio) and protein nanopore sequencing technology from Oxford Nanopore Technologies (ONT). Both platforms have been able to generate very long reads (~15kb for PacBio and >30kb for ONT). The performance and cost specifications of these two platforms are outlined in Table X. Other long read sequencing methods and protocols, synthetic long read (SLR) ([25]) or sparse isoform sequencing (spISO-seq) ([26]), however these require more complex workflows.

The consequent generation of longer reads, ranging from 300 – 20,000 bases provided unprecedented ability to sequence entire or new entire lengths of transcripts from 5' end to polyA tail, relinquishing the need for transcriptome assembly and resolving splicing junctions. Allowing greater accuracy at transcript identification, an increasing number of studies have used such technologies to characterise isoform diversity and splicing with unprecedented success (Table X). Generally in comparison with RNA-Seq, Iso-Seq encapsulates longer transcripts, identifies novel gene locus, and correction of gene model. "Long transcript reads provide better support and higher accuracy in splice junctions than short reads, when these reads are aligned back to the genome. Thus gene models predicted from long reads yield more accurate exon/intron structure and can merge two or more misannotated adjacent genes."

1.3.4 Hybrid approach of short and long read sequencing

Despite the ability of long-read sequencing (particularly, Iso-Seq) to discover large number of novel and longer transcripts and identify complex splicing events such as alternative adenylation, there are inherent biases to sequencing the more highly-expressed and relatively shorter

transcripts. Consequently, while the new chemistry has improved the error rate and increased throughput, the coverage is still insufficient for accurate transcript quantification and sensitive differential transcript analysis based on long reads alone (Koren et al., 2012). Furthermore, there is currently no consensus to validate or functionally characterise these transcripts (B. Wang, Kumar, Olson, & Ware, 2019). The current standard for such application is thus a hybrid approach of aligning the short-reads to the long-reads to improve alignment and assemblage, and for downstream isoform quantification.

1.3.5 Isoform quantification

Isoform-specific expression can be deduced from short-reads alone using statistical models if the gene is well annotated (i.e. all isoforms are known) based on i) reads aligning to contiguous genomic segment (exonic reads) and ii) reads aligning to two contiguous segments with a single gap of 60-400bp (junction reads)(Jiang and Wong, 2009)).

Various bioinformatic tools and computational models have been developed to quantify isoform quantification from RNA-Seq data. There are currently two main methods:

1. Inclusion level, calculated for a regulated exon by aligning reads either to candidate alternative exons and its junctions (inclusion reads), or to flanking exons and subsequently skipping the candidate alternative exon (skipping/exclusion reads) (Chen et al. 2012)
2. Percent-Spliced-In (PSI), calculated by proportion of isoforms that include the exon (Venables et al. 2008)(Katz et al. 2010). If the PSI value is calculated for a particular splicing event, it can be considered equivalent to the inclusion level.

Isoform quantification can either be expressed as a global measure of expression, which provides a global gene expression ranking in one sample (measured by RPKM: Reads of a transcript sequence per Millions mapped read), or as a relative measure of expression, which is normalized per gene locus and comparable across conditions (measured by inclusion level or PSI value).

Isoform abundance calculated by aligning short-reads to transcriptome is preferential to alignment with reference annotation library (RefSeq/GENCODE) in narrowing down the isoforms expressed and thus subsequently enabling more reliable abundance quantification. Reference annotation library is constructed on all data from the same species, and inclusion of annotated

but not truly expressed isoforms can increase variability of abundance estimates. Finally, if the reference library is incomplete, then truly expressed isoforms would be completely missed and RNA-Seq reads would be incorrectly assigned to annotated isoform (?u2013)

Differential Isoform Usage

When analyzing splicing patterns between multiple conditions, changes in isoform abundance can be defined in two ways:

1. Differential Isoform Expression (DIE): changes in absolute expression of an isoform, evaluated using count matrixes
2. Differential Isoform Usage (DIU): changes in relative expression of an isoform from the same gene, resulting in a change in isoform proportion and is evaluated using changes in gene exon usage

Figure X shows an example of a change in DIE but no change in DS: A two-fold increase of both isoforms from the same gene results in a change in absolute but not relative expression to one another. A change in DIE but not in DS may indicate a transcription-related mechanism. If a change in DS is observed, a change in DIE of one of the isoforms would also be observed. A change in multiple isoforms would also be observed, as long as the change is not in the same direction (upregulated/downregulated) with the same magnitude. Any changes in DS/relative abundance of isoforms indicate a splicing-related mechanism.

In addition to exploring differential splicing in terms of isoform abundance, which typically involves an exon-based approach that focuses on differential exon usage (i.e. DEXSeq), a splicing based approach can also be taken. This involves analyzing individual splicing events (exon skipping, alternative donor and acceptor) for systematic changes between conditions. rMATS, SUPP2, LeafCutter and Majiq are such tools that identify and quantify splicing events using junction reads.

Applications of these novel bioinformatic approaches to study alternative splicing in AD human post-mortem brains revealed over 2000 genes with differential transcript usage, including *APP* and *BIN1*.[27]

1.4 Aims and Objectives

1. Whole transcriptome analysis of AD post-mortem brain tissues as reference dataset, shed light on differential isoform expression
2. Particular interest on 19 loci identified from meta-analysis of GWAS studies on AD (Lambert et al. 2013) Targeted transcriptome analysis
3. Classification of AS events, which most commonly observed/dominant? Isoforms derived from transcriptional regulation (alternative promoters) vs post-transcriptional regulation?
4. Impact of AS events on protein domains. Non-sense mediated decay?
5. Integration with other (epi)genetic analysis on same samples, i.e. DNA methylation, lysine acetylation, gene expression
6. Protein analysis? Integration with any publicly available mass-spec datasets

Figure 1.2: Mouse samples for Whole and Targeted Transcriptome Iso-Seq.

Gene expression and mRNA isoforms vary widely across tissues ([5]), thus sequencing the disease-relevant tissue (in this case entorhinal cortex) is important for understanding the pathology of AD. However, it is consequently important to note that other tissues may have to be considered to fully grasp the whole picture of AD development.

While human post-mortem brain tissues remain to be the gold standard for transcriptomic studies, important to highlight that post-mortem interval and storage conditions of brain material highly influence transcriptome stability, particularly affecting alternative splicing. Furthermore changes in gene/transcript expression can be due to differences in cellular composition (i.e. neuronal loss/reactive gliosis) rather than indicative of disease-associated transcriptional regulation.

1.5 Future Directions

At the time of writing, there have been other major advances in the field have unfortunately not be explored. This include, single cell transcriptomics and Direct RNA-Sequencing. Analysis of mRNA expression at the resolution of individual, "single", cells, allowing representation of cell-to-cell variation rather than taking the stochastic average from bulk measurements,

and thereby resolving heterogeneity. This is currently achieved by the capture and analysis of single cells using a microfluidic or droplet-based technology. Importance of single cell approaches highlighted in[8] with few isoforms shared between cells (7% of all detected isoforms shared between all cell-types, though this increased to 60% for exon-cassette isoforms).

To date, Direct RNA sequencing of the native RNA molecules rather than the cDNA is only possible on the ONT. This approach offers several benefits over standard cDNA sequencing, in i) eliminating the risk of generating library artefacts from reverse transcription and PCR, ii) removing length bias toward shorter abundant fragments that could skew the population of transcripts, and finally iii) elucidation of RNA epigenetic modifications.

Single-cell studies have highlighted the difference in transcriptome diversity at a single cell level, with small overlap of isoforms between cells ([8]). Previous methods on quantifying transcripts at a single cell level have relied on RNA-fluorescence in-situ Hybridisation (RNA-FISH), which is limited in terms of throughput and characterisation of complex splicing events ([28])

While the methods I have adopted for long-read sequencing in this thesis allows interrogation of full-length transcripts, this is reliant on the generation and amplification of cDNA from mRNA, which can produce artefacts (template switching), introduce bias (distortion of relative cDNA abundance) and lose RNA modifications. In 2018, ONT showed that it was able to sequence RNA directly using the minION by adding poly(T) adapters directly to the mRNA, with a translocase that was able to bind and process RNA efficiently,[29] achieving coverage and accuracy comparable to that with ONT-cDNA method.

Chapter 2

Long-read Sequencing

2.1 Pacific Biosciences: Isoform Sequencing

2.1.1 Introduction

For successful DNA polymerisation, the DNA polymerase requires high concentration of nucleotides to allow high accuracy and processivity. However for sequencing, this limits sensitivity to detect each labelled base incorporation and respective fluorophore emission, due to high background noise level. In the past, second-generation sequencing technologies have circumvented this issue by the step-wise addition, scan and wash of each set of labelled nucleotides, but at a compromise of read length.

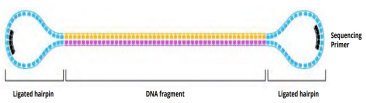
Unlike RNA-Sequencing, Pacific Bioscience's Single Molecule Real Time sequencing (SMRT) is able to generate long reads is due to its ability to mimic natural, uninterrupted, processive DNA synthesis, through three important innovations:[30]

1. Creation of a circular template, SMRTbell, enclosed with hairpin adapters at end of the inserted target double-stranded DNA, allowing uninterrupted DNA polymerisation[31] (Figure 2.1a).
2. Sequencing of each SMRTbell with a bound polymerase at the bottom of a nanometre-wide well (zero-mode-waveguide - ZMW), and all wells contained within a single SMRT chip.[32] Due to the very nanoscale size of the ZMW and reduced detection volume, a

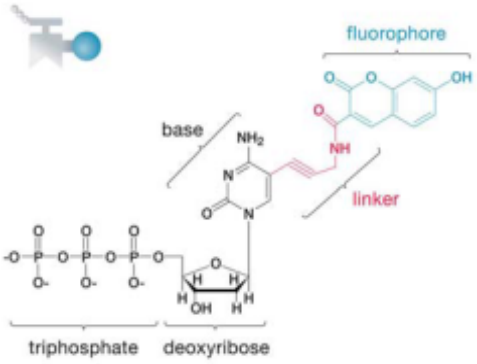
single nucleotide incorporation can be sensitively detected against the high background of labelled nucleotides, achieving a high-signal-to-noise ratio (Figure 2.1c)).

3. Addition of phospholinked nucleotides, each labelled with a different colour fluorophore corresponding to the four different bases (A, C, G and T), which allows for natural, accurate and processive DNA synthesis[33] (Figure 2.1b).

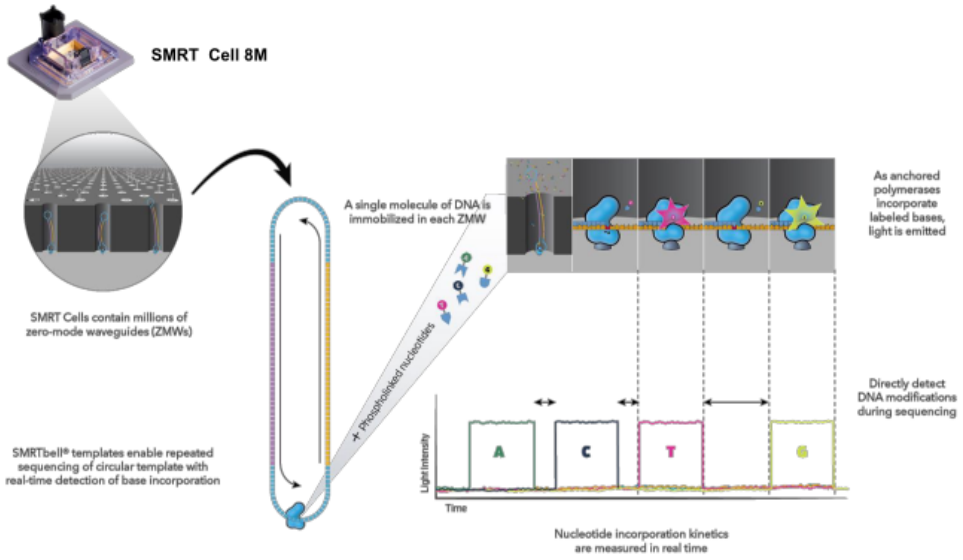
In summary, SMRT sequencing detect fluorescence events that correspond to addition of one specific nucleotide by a polymerase attached to the bottom of a tiny well.



(a) SMRT-bell template



(b) Phospholinked Nucleotides



(c) PacBio Single-molecule real time sequencing

Figure 2.1: PacBio SMRT: At time of writing, PacBio released Sequel II with the provision of an 8M chip, containing 8 million wells, each capable of sequencing one single molecule. Figures adapted from PacBio

Currently, PacBio offers two sequencers: Sequel I and Sequel II; RSII was the first commercially available sequencer, but is no longer supported. With Sequel v2 chemistry from 2017, fragments longer than 10kbp were typically only read once and had a single pass accuracy of 58-87%. Last 3 years have seen teh release of 1 instrument (Sequel II), 4 chemistries (Sequel v2,v3, Sequel II v1, v2) and 4 versions of the SMRT-Link analysis suite.

2.1.1.1 Mechanism

Due to the circular nature of the SMRT bell, the polymerase can continually read through the insert, sequence the same DNA template multiple times and generate a continuous sequence of bases (continuous long read, CLR or polymerase read), which contains the hairpin adapter sequences. Pending on the polymerase lifetime and insert length, both strands can be sequenced multiple times, or "passes" in a CLR, which can then be delineated by the adapter sequences and resolved to multiple reads (subreads). These subreads can be further collapsed to yield a highly-accurate Circular Consensus Sequence (CCS), increasing the accuracy of individual raw subreads from 85% to 99% of the consensus read, proportional to the number of passes.[31]

Reads from PacBio are therefore not a set length, as with short-reads generated by RNA-Seq, but a distribution of lengths dependent on the library size and the polymerase activity.[34,35] However with previous chemistries, there was a bias towards sequencing molecules of a certain length due to preferential loading of SMRTbell templates - loading by diffusion favoured shorter molecules[36] whereas loading using Magbeads allowed proportional loading to the concentration rather than by length, but prevented sequencing of molecules <1kb. Previous attempts to mitigate this loading problem have been to fractionate the library by length (size selection) and enrich for longer cDNA molecules before sequencing,[37] but this approach is more expensive and laborious. Thankfully, recent improvement in the technology and chemistry (v3.0) with usage of molecular crowding agents have alleviated the short-read bias and resulting the magbead loading method obsolete.[38]

2.1.1.2 Performance and Run Quality Metric

In an ideal situation, all the wells will contain an insert that will generate a positive signal. However, because XXXX, there will be some wells that are empty (quality metric denoted as P0:

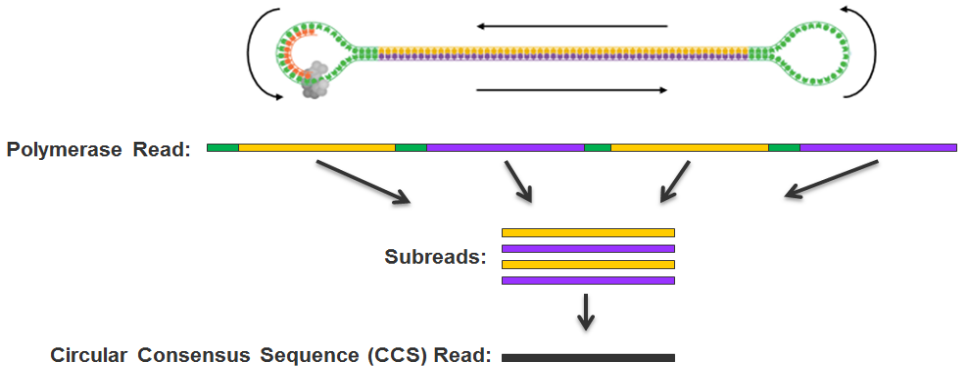


Figure 2.2: Generation of Circular Consensus Sequence: CCS is generated by the collapse of multiple subreads, which sequence correspond to the double-stranded cDNA of interest. The greater the number of "passes" sequenced by the polymerase, the longer the polymerase read, the more subreads generated, and subsequently the higher the quality of CCS. Picture adapted from PacBio

Productivity 0), and some wells that will be overloaded with multiple inserts with more than one polymerase (quality metric denoted as P2: Productivity: P2). Thus only wells that contain one polymerase (denoted as P1, Productivity 1) will generate a positive signal. Overloading may lead to increase in output of yield per SMRT cell, but increases the chance of P2 (multi-loaded ZMWs), resulting in shortened read lengths and lower accuracy compared to single-loaded ZMW. Loading can be optimised through titration.

A good run is defined by 50-70% P1, a >XX kB polymerase read-length. Over-loading (>70%) may result in reduced base quality (noisy base-calling), whereas under-loading (<50%) results in lower throughput. A short polymerase read-length indicates sequencing/library preparation issues. These metrics are dependent on chemistry, pre-extension, and movie-runtime.

2.1.2 Lab Pipeline

The Iso-Seq lab protocol, as outlined in Figure 2.3, involved three main steps by first converting total RNA transcripts to full-length complementary DNA (cDNA) using the Clontech SMARTer PCR cDNA synthesis kit, which was then subsequently amplified and purified to generate double-stranded cDNA, which was then constructed to a SMRT bell library for sequencing. Size selection was not performed with full-length transcript detection of up to 4 kB. For targeted sequencing using IDT probes, all the steps in the Iso-Seq protocol are the same with an additional step of target capture post ds-DNA amplification and pre SMRT bell library, and usage of barcodes to allow multiplexing (Figure 2.4).

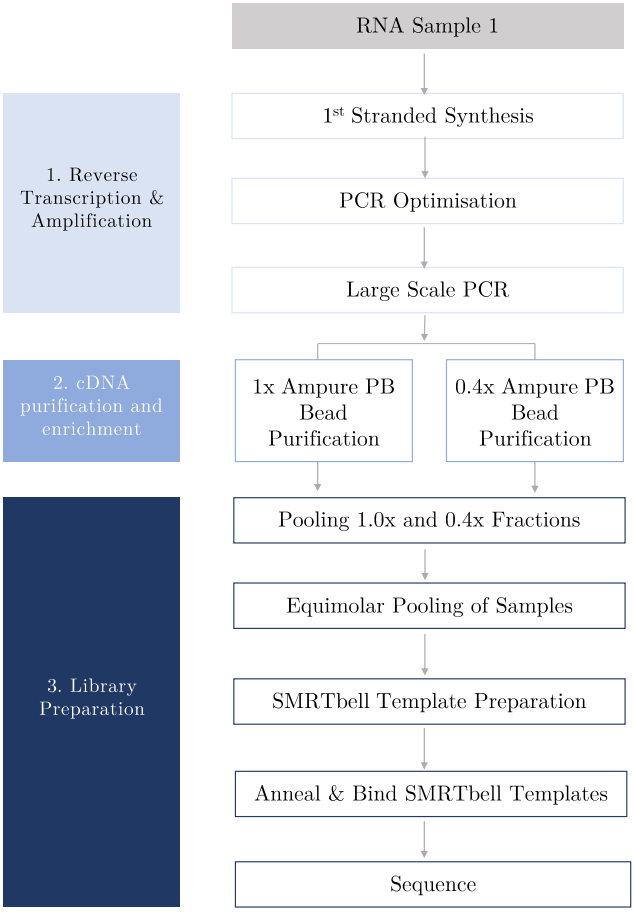


Figure 2.3: An overview of the lab Iso-Seq pipeline used for whole transcriptome profiling. The lab pipeline, as adapted from official Iso-Seq protocol, involves three main steps: 1) reverse transcription and amplification of cDNA (Section 2.1.2.1), 2) cDNA purification with ampure beads (Section 2.1.2.6) and 3) library preparation involving ligation of SMRT bell templates, and primer and polymerase binding (Section 2.1.2.10)

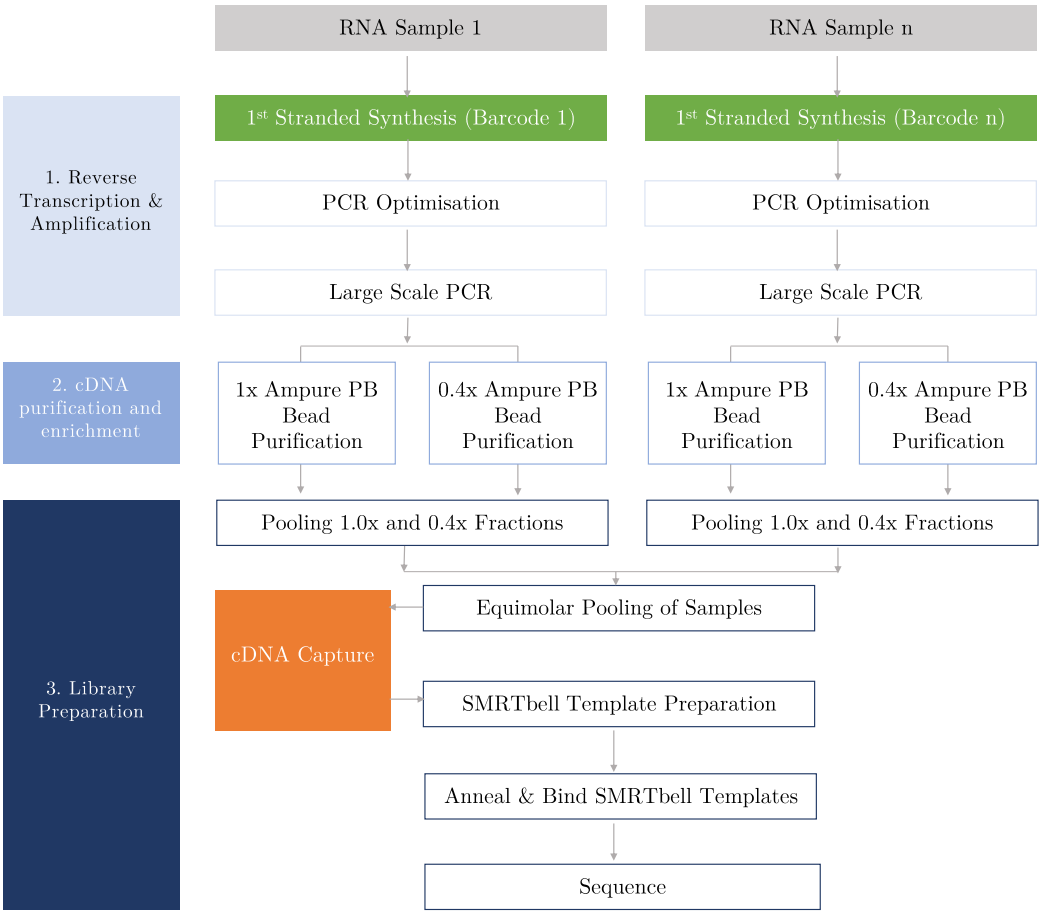


Figure 2.4: An overview of the lab Iso-Seq pipeline used for whole transcriptome profiling. The lab pipeline for targeted transcriptome profiling involves all the steps in the standard Iso-Seq lab pipeline (Figure 2.3), with the addition of the target capture step (Boxed orange, Section 2.1.2.9) and the use of barcoded primers in reverse transcription (Boxed green and denoted here as Barcode 1 and Barcode n) to allow sample multiplexing (denoted here as Sample n, PacBio recommends 6-8 multiplexed samples per run). The list of barcodes can be found in Table 2.1

2.1.2.1 Complementary DNA synthesis

As part of the official Iso-Seq protocol, SMARTer PCR cDNA Synthesis Kit (Clontech) was used to convert 200ng extracted total RNA to complementary DNA by first strand cDNA synthesis, as outlined in Figure X. In brief, the polyA+ tails of RNA transcripts is first primed by a modified oligo (dT) primer, transcribed by SMARTScribe Reverse Transcriptase to generate a first single-stranded DNA, which is then diluted and subsequently amplified.[39] All reagents were provided with the kit, except for the Pacific Bioscience's barcodes, with all reagents and consumables used being sterile and DNAse and RNAse free. In order to sequence samples simultaneously ("multiplex"), as exploited for targeted sequencing, unique barcoded oligo (dT) primer was used in place of the standard oligo (dT) primer (Table 2.1). With new Sequel system, cDNA can be sequenced without size selection.

While this kit is advantageous in preferentially enriching for full-length cDNA sequences, as a template switching oligo is required to ensure complete reverse transcription, it cannot differentiate between intact and truncated RNA; which, present in poor-quality samples will be amplified as a potential source of contamination in the final cDNA library. One alternative is to exploit the 5'-cap that is present only in intact RNA and not truncated RNA (5-cap refers to the addition of 7-methylguanosine to the 5'-end of mRNA during transcription, to protect nascent mRNA from degradation and assist in protein translation). Alternative reverse transcriptase have been explored that only converts 5'capped mRNAs to cDNA, however, these have been found to negatively affect read length on the ONT platform (Cartolano et al. 2016). An alternative method, Full-Length cDNA Amplification (Teloprime), relies on a double-stranded adapter that recognises and ligates to the 5'cap at the end of first strand synthesis (Section X, Chapter 2)([40]).

The general structure of barcoded oligo-dT primer is as follows:

Primer Sequence 16-bp barcode oligo-dT

5'AAGCAGTGGTATCAACGCAGAGTACtcagacgatgcgtcatTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN3'

Barcode Name	Sequence
Barcode 1	AAGCAGTGGTATCAACGCAGAGTACCACATATCAGAGTGCGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 2	AAGCAGTGGTATCAACGCAGAGTACACACACAGACTGTGAGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 3	AAGCAGTGGTATCAACGCAGAGTACACACATCTCGTGAGAGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 4	AAGCAGTGGTATCAACGCAGAGTACCACGCACACACGCGCGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 5	AAGCAGTGGTATCAACGCAGAGTACCACTCGACTCTCGCGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 6	AAGCAGTGGTATCAACGCAGAGTACCATATATATCAGCTGTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 7	AAGCAGTGGTATCAACGCAGAGTACTCTGTATCTCTATGTGTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 8	AAGCAGTGGTATCAACGCAGAGTACACAGTCGAGCGCTGCGTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 9	AAGCAGTGGTATCAACGCAGAGTACACACACGCGAGACAGATTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Barcode 10	AAGCAGTGGTATCAACGCAGAGTACACGCGCTATCTCAGAGTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN

Table 2.1: Barcoded oligo-dT primers were used for multiplexing samples in targeted transcriptome sequencing. Each of the barcoded primers contain the same 5' primer sequence and oligo-dT for reverse transcription of first strand cDNA synthesis using Clontech kit SMARTer PCR cDNA Synthesis Kit. The different internal 16bp sequence allows tagging and differentiation of samples in the same sequencing run. The barcodes are recommended from official PacBio's multiplex protocol.

2.1.2.2 ERCC-RNA Spike-In Controls

To evaluate the performance of library preapration and the sequencing runs, and to validate the Iso-Seq pipeline to accurately characterise the transcriptome using long reads, a set of external RNA Spike-In controls, External RNA Controls Consortium (ERCC), was used. ERCC consists of 92 polyadenylated synthetic transcripts (250 to 2000 nucleotides) of known sequences from the ERCC plasmid library, which are added in pre-determined amounts to the sample prior to first-strand cDNA synthesis. The addition of ERCC would allow assessment of the quantitative power of long-read sequencing approaches in addition to providing absolute quantification of mRNA isoform with the sample by generating a standard curve. It can further validate that the bioinformatics pipeline only identifies 1 isoform per ERCC gene.

The amount determined of spike-in control was calculated using the below equation:[41]

$$mass_{RNAspike} = fraction_{spikedreads} * fraction_{targetRNA} * mass_{RNAinput} \quad (2.1)$$

$$concentration_{RNAspike} = mass_{RNAspike} * volume_{RNAspike} \quad (2.2)$$

where:

$mass_{RNAspike}$ = mass of RNA spike-in to be added to sample
$concentration_{RNAspike}$ = final diluted concentration (ngs/ul) of the RNA spike-in
$fraction_{spikedreads}$ = desired proportion of sequenced spike-in RNA 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 *(0.1uL)*

Equation 2.1: Determining amount of ERCC-RNA Spike-In Control. 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 RNA Transcriptomics 2018 Course[41] 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 2.5a), reflecting significant enrichment of ERCC transcripts at these two respective lengths as is expected (Figure 2.5b). The stark contrast of these two bands, however, to the smear of cDNA, suggests that the ERCC transcripts are predominant - this could be due to an overestimation of assumed proportion of mRNA to total RNA,

which is likely in reality to be lower than 3%. To reduce unnecessary sequencing and coverage of ERCC transcripts, a lower ERCC RNA-spike in concentration was used (final concentration of 0.6ng/μL and a dilution factor of 1:50.5, Figure 2.5c).

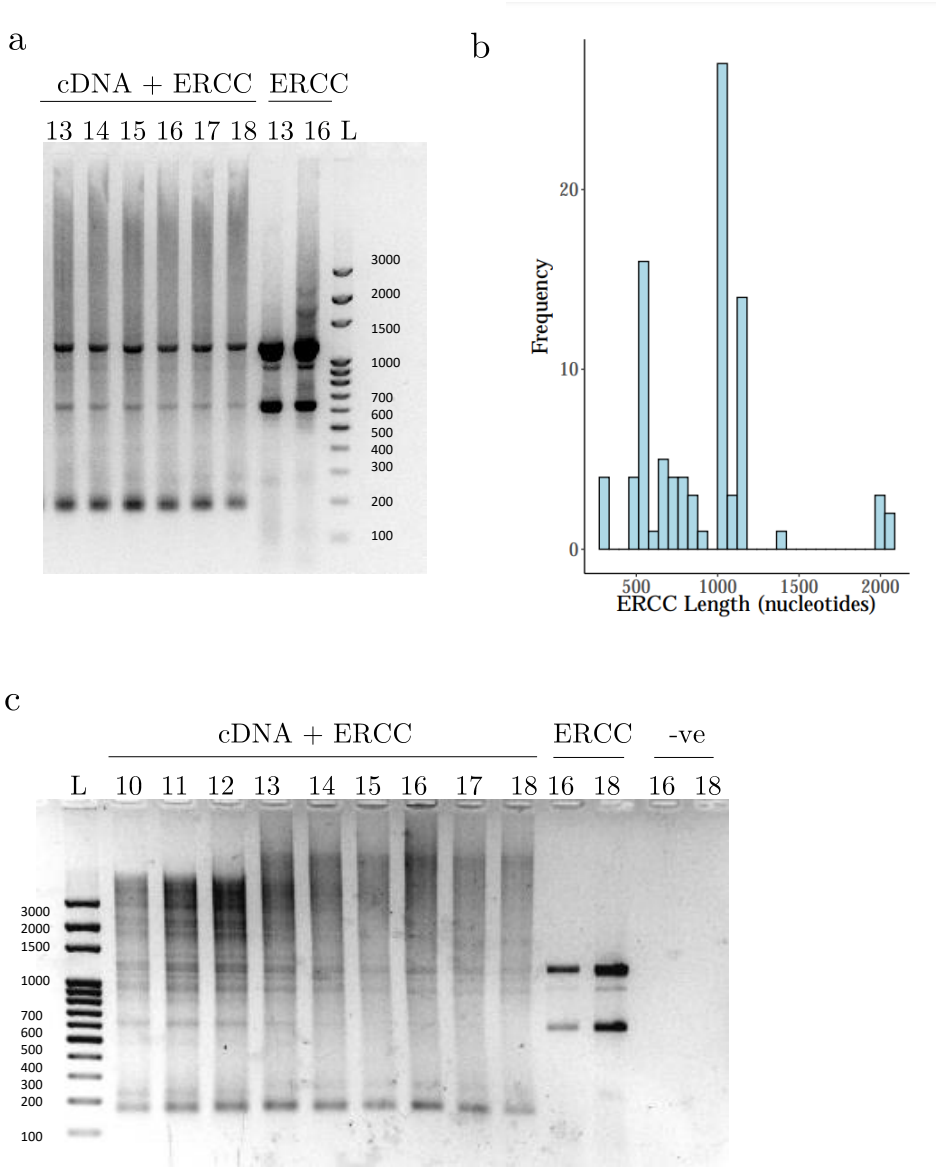


Figure 2.5: Successful addition of ERCC to first-strand cDNA synthesis a). Agarose gel image taken from PCR amplification of cDNA and ERCC (1.8ng/μL determined from equation 2.1), and ERCC alone as a positive control. 5μL of PCR aliquots were taken every cycle (13 - 18) and then run on gel electrophoresis. The two bands at 600bp and 1000bp refer to the enrichment of ERCC transcripts at these two lengths as would be expected. **b)** Distribution of known ERCC length, with a significant proportion of transcripts sized at 500-600bp and 1000-1200bp. **c)** Agarose gel image after a repeat of PCR amplification of cDNA and ERCC at a lower concentration (0.6ng/μL), ERCC as positive and water as negative control respectively. The numbers above the lane refer to the number of cycles, L denotes to 100bp Ladder.

2.1.2.3 PCR optimisation and DNA Amplification

To minimise PCR bias (under or over-amplification), which can result in under or over representation of the different cDNA library size, the optimal number of PCR cycles for amplification of first-strand synthesis products with PrimeSTAR GXL DNA Polymerase (Clontech) was determined through collection of 5uL PCR aliquots during every two cycles (cycle 10, 12, 14, 16, 18, 20) and assessed a 1.5% Agarose gel electrophoresis with ethidium bromide. Large scale PCR amplification was subsequently performed using the optimal number of cycles.

2.1.2.4 Polymerase Chain Reaction (PCR)

To generate sufficient DNA for sequencing, single-stranded DNA was amplified using Polymerase Chain Reaction (PCR), a well-established method of 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 various other components required for polymerisation (deoxynucleotides , buffers). This reaction is then subjected to a series of heating and cooling steps:

1. Denaturation at 96C, to separate any double-stranded DNA
2. Annealing, typically between 55 to 65C, 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 72C to allow the polymerase to extend the primers, consequently synthesising a new complementary DNA strand using dNTPs

These three steps are then repeated for a number of times, "cycles", for an exponential generation of the DNA template of interest.

2.1.2.5 Agarose Gel Electrophoresis

Agarose gel electrophoresis allows the separation of (double-stranded) DNA molecules based on its length. It is most commonly used to determine DNA quality and quantity, and assess the efficiency of molecular biology techniques such as PCR amplification. It works on the principle that by applying an electrical charge, negatively-charged DNA migrates through a gel matrix towards the positive anode at a rate 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 DNA structure and fluoresces under ultraviolet light.

2.1.2.6 AMPure Bead Purification

Post large scale amplification, the resulting PCR product was divided into two fractions and purified with 0.4X and 1X AMPure PB beads (PacBio). Double-stranded DNA was bound to the beads in either 1:1 or 1:0.4 ratio, which were then isolated on a magnetic rack, and washed with 70% ethanol. DNA purification with 0.4x AMPure beads allows for enrichment of longer DNA fragments to provide a more representative library given that shorter fragments diffuse quicker into ZMW and are more likely to be sequenced. The ability to enrich for longer fragments is due to the preferential binding of beads to more negatively-charged, and subsequently larger molecular weight DNA, and thus displacement of shorter fragments. Quantification and size distribution of each fraction was then determined using Qubit DNA High sensitivity assay (Invitrogen) and Bioanalyzer assays on the 2100 Bioanalyzer (Agilent). Two fractions per sample were then recombined at equimolar quantities and library preparation performed using SMRTbell Template Prep Kit v1.0 (PacBio).

The molarity was calculated by the following equation:

$$\frac{concentration(\frac{ng}{ul}) \times 10^6}{660(\frac{g}{mol}) \times average\ library\ size\ in\ bp^*} = concentration\ in\ nM \tag{2.3}$$

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

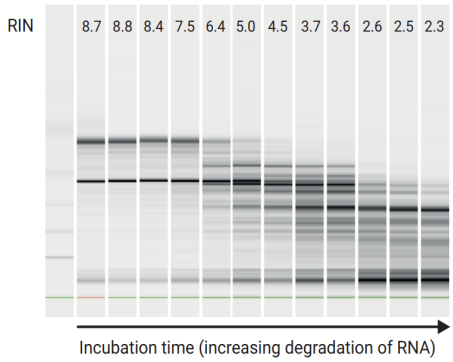
2.1.2.7 Bioanalyzer

ScreenTape and Bioanalyzer assays are commonly used to provide accurate assessment of nucleic acid quality and size, prior to proceeding with downstream experiments. As an automative alternative to agarose gel electrophoresis, both assays similarly take advantage of nucleic acid's inclination to migrate in response to an electrical field. While the Bioanalyzer assay is more sensitive than the ScreenTape assay, it is more expensive to run as it uses a chip consisting of 12 sample wells rather than independent lanes on the ScreenTape.

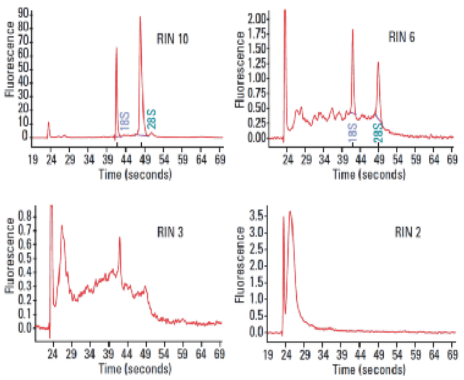
For this thesis, most of the assessments of DNA quality in the Iso-Seq and ONT protocol were performed on the DNA 12000 Kit (Agilent) on the 2100 Bioanalyzer (Agilent)for accurate

determination of library molarity (Section X). However, the D5000 ScreenTape (Agilent) was used on 4200 Tapestation (Agilent) in a few of the quality control steps where it is optional to assess for DNA quality (Section X).

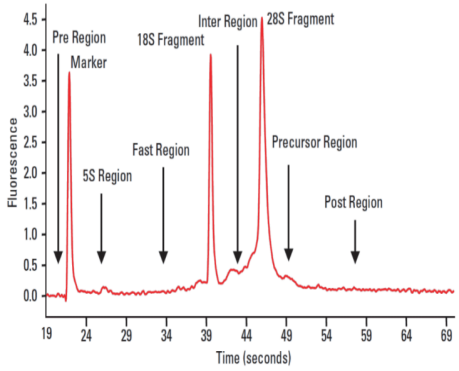
RNA extracted by Dr Isabel Castanho was also run on RNA ScreenTape assay and the Bioanalyzer RNA analysis to provide accurate evaluation of 5' RNA degradation asses using a RNA Integrity Number (RIN) - this uses features from an electrophoretic trace (ratio of 28S to 18S area) to give a number between 1 and 10, where 1 is indicative of high degradation, and 10 of low degradation and thus high integrity (Figure 2.6). As a pre-requisite for good sequencing yield on Sequel and MinION, only samples with RIN $>$ 8 were selected for long-read sequencing on Iso-Seq and ONT protocol.



(a) Automated gel electrophoresis of RNA degradation



(b) Electropherogram of RNA degradation



(c) Electropherogram of regions that are indicative of RNA quality

Figure 2.6: Evaluation of RNA integrity with Bioanalyzer and Tapestation: Total RNA degradation can be observed by a shift towards shorter fragment size as depicted in Figure a, after prolonged incubation. The degree of degradation is represented by a RNA integrity number (RIN), ranging from intact (RIN = 10) to degraded (RIN = 2) RNA, and is calculated by the relative ratio of the fast region and 18S, 28S fragment (Figure c). Figures and legends are adapted from Mueller et al. 2016.

2.1.2.8 Qubit

Qubit assays (Invitrogen) allow accurate nucleic acid quantification by the selective binding of fluorescent Qubit dyes to double-stranded DNA (dsDNA) or RNA, making it more sensitive and specific than UV absorbance used in NanoDrop 8000 spectrophotometer (Thermo Fisher Scientific). It is commonly performed to determine the average concentration of DNA or RNA prior to proceeding with downstream experiments. Many of the steps in the Iso-Seq protocol and ONT protocol thus require performing Qubit assays, particularly post bead purification, and are detailed in Section A.0.2.2.

2.1.2.9 Target Capture using IDT Probes

For targeted sequencing, we used the official PacBio protocol "cDNA Capture Using IDT xGen® Lockdown Probes" (an adaptation of the official IDT protocol "xGen hybridisation capture of DNA libraries"), which slotted as an additional step to the standard protocol between cDNA amplification and ligation. Enrichment of target genes involved hybridisation of dsDNA using pre-designed, complementary 5' biotinylated DNA 120nt-long oligonucleotides (hereby referred as probes). The hybridised library fragments were then washed, isolated with magnetic streptavidin beads, amplified using Takara Hot-Start polymerase and then further purified with AMPure beads. After assessing the quality and quantity of the target cDNA using the bionanalyzer and qubit, SMRT Bell template preparation, primer and polymerase annealing were proceeded as per standard Iso-Seq protocol. Given the samples were multiplexed for targeted sequencing, the samples were first pooled in equal molarity before probe hybridisation.

Selection of probes

Probes were designed to a panel of 20 AD-associated genes: Bin1, Trem2, Cd33, Vgf, Fyn-Mapt, Trpa1, Picalm, Sorl1, Abca7, Snca, Apoe, Abca1, App, Ank1, Clu, Fus, Ptk2b, Rhbdf2, Tardbp. Two separate pools of the equal molar probes were created using the mouse genome (GRCm28/mm10) and human genome (GRCh37/hg19). While IDT provided a pre-designed set of probes to the target genes, many of them were found to overlap with the intronic regions of the target gene with contiguous coverage.

Given that previous studies with targeted sequencing have found that the target gene can be

successfully enriched with a few unique probes to the exonic regions, I manually assessed the list of probes for each target gene using the following criteria:

- Ensured each exon in every gene is covered at least once (exons > 500bp has >1 probe)
- Removed any probes to intronic regions
- Within each exon, removed any contiguous probes (as seen in the 1x tiling density) and ensured probes spaced 300-500bp (equivalent to 0.2x – 0.3x tiling density)
- From the contiguous "cluster", selected probes with the highest GC content (40-65% GC content)/minimal number of blast hits

The coverage of each target gene can be found in Appendix.

2.1.2.10 SMRT Bell Template Preparation

The library preparation post pooling the two fractions at equimolar quantities with the SMRTbell Template Prep Kit v1.0 (PacBio) involved several steps. DNA Damage and End Repair was first performed on the pooled library to polish ends of fragments for ligation of blunt hairpin adapters, necessary to generate high quality library of closed, circular SMRTBell templates. Any abasic sites were filled-in, thymine dimers resolved, and deaminated cytosine are alkylated. 3' overhangs were removed, whereas 5' overhangs were filled-in by T4 DNA Polymerase and phosphorylated by T4 PNK. Following 1x AMPure purification of repaired dsDNA, hairpin adapters were then ligated to the blunt ends for up to 24hours. Any fragments failed to ligate were removed with exonuclease III and VII. The repaired, ligated SMRT bell library was then purified twice with 1x AMPure beads, and assessed with Qubit DNA High sensitivity assay (Invitrogen) and Bioanalyzer 2100 (Analyzer) before proceeding to primer annealing and polymerase binding (Figure X).

2.1.2.11 Primer Annealing and Polymerase Binding

Post ligation of hairpin adapters, sequencing primer and polymerase were bound to both ends of the SMRTbell templates. The primer and polymerase to template ratio was critical to minimise under or –over loading, thus the concentration was sample specific.

Prior to XXX chemistry, MagBead Loading was only recommended for IsoSeq SMRTbell libraries, whereas Diffusion Loading was recommended for all other applications with insert sizes from 250 – 100001bp. As in the name, Diffusion Loading involves immobilization of

polymerase-bound SMRTbells to ZMW by diffusion, whereas Magbead Loading involves immobilization by attachment to paramagnetic beads. Diffusion loading thus preferentially loads longer transcripts, whereas magbead loading preferentially loads shorter transcripts of 700bp as it rolls across nanowells.

2.1.2.12 Sequencing

Sequencing was performed on the PacBio Sequel 1M SMRT cell. Samples were processed using either the version 3 chemistry (parameters: diffusion loading at 5pM, pre-extension 4 hours, Capture time 20 hours) or version 2.1 chemistry (parameters: magbead loading at 50pM with a 2 hour pre-extension and 10 hour capture).

2.1.3 Bioinformatics Pipeline

2.1.3.1 Introduction

While the official PacBio bioinformatics tool (Iso-Seq) has been revised multiple times during the scope of this PhD, there were two main steps with the aim of generating high-quality (HQ) isoforms de novo (Figure X), namely:

- Classify to identify full-length non-chimeric (FLNC), and non-FLNC reads
- Cluster reads derived from the same isoform to generate consensus sequence

Bioinformatic analysis of Iso-Seq raw data can be performed using PacBio SMRT Link Suite (ref), a web-based end-to-end user interface. However, for optimisation of parameters and parallelisation of samples, an end-to-end command line was developed and used. Since the development of Iso-Seq, a myriad of bioinformatics tools have been released, as outlined in Table X.

Analysing long-read sequencing data requires a different approach to short-read, as the initial processing focuses on reducing the high error rate (due to low read coverage relative to short reads). Currently there were three methods of correcting long reads:[42]

- Hybrid error correction strategy using short-reads: LSC[43] which maps short reads, and

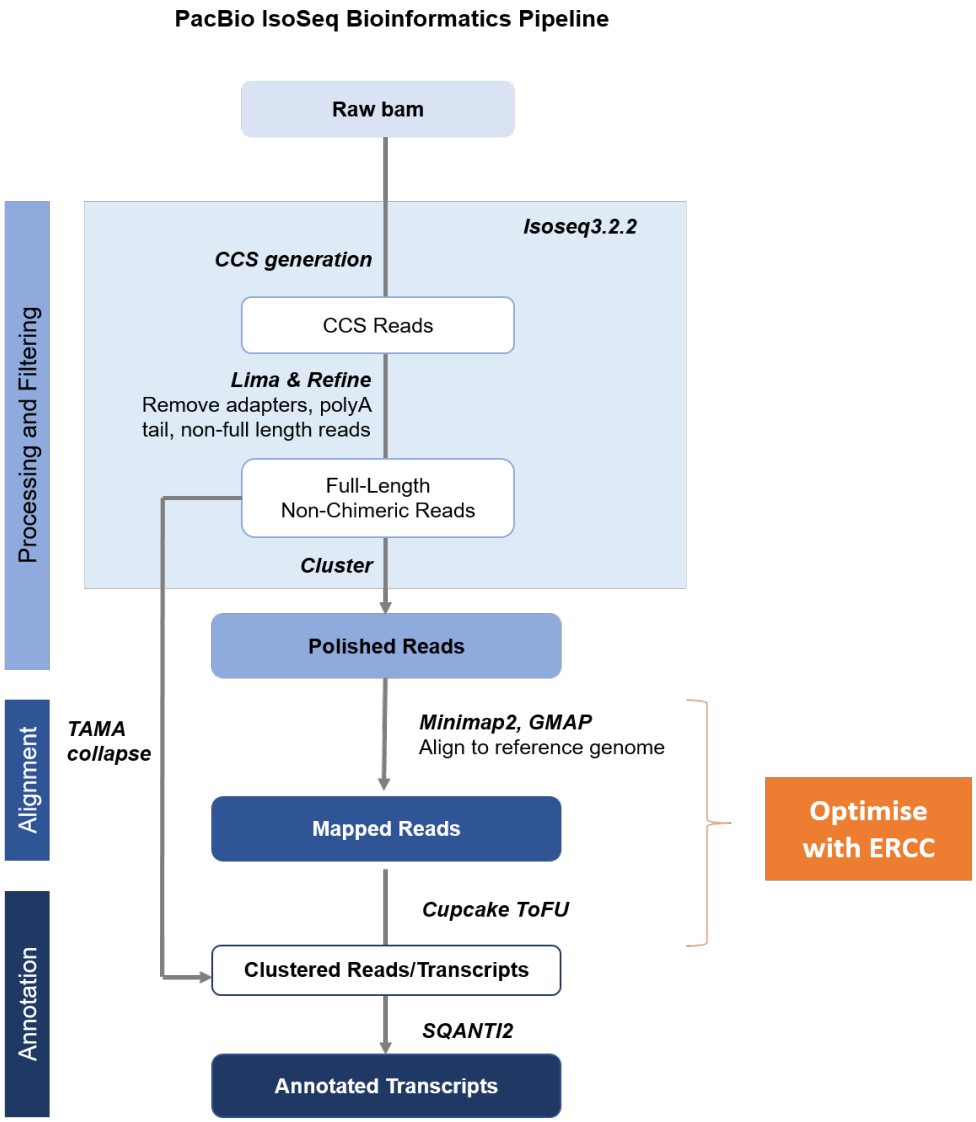


Figure 2.7: PacBio Isoseq Bioinformatics Pipeline: Pipeline is adapted from ToFU [18]

LoRDEC which build De Brujin graph of short reads[44]

- Self-correction using long reads only: Long-read multiple aligner (LoRMA)[45]
- Reference-based correction by alignment of reads to reference genome by spliced-awwere aligners: Minimap2. GMAP and STAR can also be used for alignment, however, they do not perform error correction during alignment and further capture non-canonical splice sites.

Although the raw error rate of PacBio sequencing is 10-14%, this is greatly reduced by the use of circular template and subsequent generation of circular consensus sequence.

2.1.3.2 Classify

CCS Generation: In the first stage, the raw subreads (stored as a BAM file, unaligned.bam) from each "productive" ZMW were processed individually and collapsed to generate a CCS (Figure 2.2), according to:

- The number of full "passes" from the polymerase, and subsequently number of subreads generated; a full pass is defined by the presence of both SMRT adapters at both ends (Default: 3 passes)
- The minimum base accuracy across all subreads (Default: 99%)
- Length of the subreads (Default: minimum 10 bases, maximum 21000 bases)
- Quality of Subread predicted by the CCS model (Default: Z-score of -3.5), and proportion of total subreads meeting the quality score (Default: >30%)

Across literature and PacBio scientific community, different parameter settings were recommended, particularly with *number of full passes* and *minimum base accuracy*, which had the greatest effect on the number of CCS reads generated for downstream analyses. Taking a subset of raw data from 10 randomised samples, a range of values across these two parameters were tested. CCS were then classified to full-length (FL, determined by the presence of 3'/5' primers and poly-A tail) and non-full-length (NFL) reads.

Lima: With successfully-generated CCS, cDNA primers and PacBio barcodes were identified and then removed using lima. CCS with unwanted orientations were removed and were oriented 5' to 3'. A barcode score is calculated for each barcode pair (leading and trailing barcode), and is based on accuracy alignment to input cDNA primer sequences. The proportion of FL reads (number of FL reads over the number of CCS reads) varies on the insert transcript size; for Iso-Seq, a non-size selected library with a library distribution of 1-3kB typically has a 60-70% FL.

Refine: Finally, full-length reads were refined by trimming of polyA tails, of a least a length of 20 bases, and removal of artificial concatemers to generate full-length non-chimeric (FLNC) reads. Artificial concatemers were defined as cDNA sequences with internal runs of polyA and polyT sequences, due to insufficient amount of blunt adapters during library preparation - this is typically rwere (<0.5%). Conversely, it is challenging to differentiate and remove PCR-induced artificial chimera from true biological chimera. PCR-induced artefacts were defined as

cDNA sequences that appear to be fusion transcripts, but were actually a result of non-optimal PCR reaction conditions. The number of FLNC reads should be very close to the number of FL reads, and any significant loss implicates issues at the SMRT bell library preparation. Note Tama works on FLNC reads from Classify

2.1.3.3 Cluster

In the second stage, Iso-Seq uses an iterative isoform-clustering algorithm (ICE – iterative clustering for error, called Quiver for PacBio RSII data and Arrow for PacBio Sequel data) to group all FLNC reads that were thought to be derived from the same isoform if:

- They differ less than 100bp on the 5' end
- Differ less 30bp on the 3'end
- Do not contain internal gaps that exceed 10bp

By collapsing transcripts with differing 5' start [due to cDNA synthesis not preserving 5' end], some transcripts with alternative transcription start sites were lost while preserving those with alternative splicing and alternative polyadenylation. The representative transcript from those clustered is the longest one.

A minimum of two FLNC reads were further required for a cluster. Consequently each FLNC read is classified to only one cluster, which is comprised of two or more FLNC reads. Two possible issues: reads belong to incorrect clusters, and reads that belong together were in separate clusters. [Briefly it first does clique-finding based on a similarity graph, then calls consensus using the Directed Acyclic Graph Consensus method and finally reassign sequences to different clusters based on their likelihood (Gordon et al. 2015)]. In previous Iso-Seq bioinformatic versions, NLF reads were used to increase the coverage of each consensus isoform. However, with increasing throughput with Sequel I and Sequel II, this has been foregone. Cluster outputs the high-quality isoforms (HQ-isoforms), which have a consensus accuracy >=99%.

So in summary, each productive ZMW generates one polymerase read, which is collapsed to give a circular consensus sequence (CCS) assuming the requirements were met. CCS were then trimmed and processed for primer and poly-A sequence removal to generate full-length non-chimeric (FLNC) reads, which were clustered if they were thought to be derived from the same isoform. The number of associated full-length (FL) reads of each isoform therefore rep-

resents the number of ZMWs that sequenced the isoform of interest, and can infer abundance of mRNA isoform. However, Iso-Seq is only semi-quantitative due to preferential loading and sequencing bias of shorter fragments. It is worthy to note that all the steps up to now have een processed without a reference genome or transcriptome.

Iso-Seq Versions In response to a much higher experimental throughput of Sequel compwered to RSII, each subsequent version of the official PacBio Iso-Seq tool saw a reduction in runtime, but an improvement of sensitivity to recover transcripts and specificity to reduce artefacts.

Iso-Seq 1 Iso-Seq 2

In previous versions of official PacBio IsoSeq tool, non-FLNC reads were re-incorporated at this stage to polish the consensus isoforms. Short reads from RNA-Seq can also be incorporated for error correction using various tools such as LoRDEC, LSC and Proovread.

Since the introduction of Iso-Seq protocol, 3 versions of the informatics pipeline has been developed. Iso-Seq2 has an extra pre-clustering step to bin full length non-chimeric reads based on gene families. The latest version Iso-Seq3 is used in response to the much higher throughput of Sequel compwered to RSII by using faster clustering algorithms. Using a more conservative primer removal and barcode demultiplexing step (with tool named LIMA), the Iso-Seq3 pipeline generates fewer but higher quality polished transcripts.

High confidence transcripts can be determined by 1) presence of open reading frame (ORF), CDS length, interpro domain coverage, annotation edit distance

2.1.3.4 Genome/Transcriptome Alignment

High quality isoforms were then aligned to the reference genome (as opposed to transcriptome as otherwise miss novel isoforms using BLASR) using splice-awwere aligner Minimap2. Various long-read studies have used Minimap2 and GMAP (Križanovic et al. 2018 demonstrated marked success of GMAP vs other RNA-Seq Aligners). Tang et al. 2020, using subset of Oxford Nanopore reads evaluated number of splice sites mapped relative to known junctions, found Minimap2 to be more precise than GMAP.

Using the –secondary=no parameter restricts the output to the best alignment, -x splice assumes read orientation relative to transcript strand unknown, and thus tries two rounds of alignment to infer orientation. As a splice-awwere alignment, -x splice prefers GT[A/G]…[C/T]AG over GT[C/T]…[A/G]AG over other splicing signals (main donor/acceptor motifs). –u f forces minimap2 to consider forward transcript strand only for alignment, slightly improving accuracy. –c 5 to accept non-canonical GT/AG splice junctions.

–splice-flank=yes for human/mouse data in reads with relatively high sequencing error rate (necessary for ONT), but not for high quality IsoSeq reads (99% - 100%).

2.1.3.5 Genome Mapping

HQ-isoforms from the pooled dataset were aligned to mouse genome using Minimap2, and a total of XXX reads (XX%) were mapped. Errors for substitution, insertion and deletion were X%, X% and X% respectively. XX% of transcripts (polished) could not be mapped to reference genome, thus representing genes that fall into gaps in the assembly (mouse genome should be quite updated though)

2.1.3.6 ERCC

One source of error from long-read sequencing can occur at reverse transcription (RT), whereby a premature termination in reverse transcription enzyme can result in a full-length cDNA, that is mistaken for a true isoform. To measure the degree of this technical error, ERCC, with known start and end positions can be used as benchmark. As detailed in,[8] most ERCC reads fell within +/- 5bp at both 5' and 3' ends, with 3' end slightly more accurate than 5' end. From,[20] drop in read length was observed for ERCC for molecules longer than 1.5kb (PacBio RSII). Interestingly, non-coding exon junctions were more variable than coding-exon junctions, suggesting that codon exon splicing has a stricter control with refined splice donor/acceptor sites ([8]) Of note, however, that while ERCC has been used as a standard for RNA-Seq method validation, the longest molecule is only 2kB, thus limiting is usage to validate longer molecules. Given that XX of RNA transcripts in human and mouse transcriptome were >2kB, there is a need for longer control sequences.

To assess the sensitivity across Iso-Seq runs to detect ERCC, a merged analysis of whole tran-

scriptome samples (n = 10, WT = 5, TG = 5) was performed with ERCC alignment and further collapse using Cupcake. The counts of full-length transcripts pertaining to each sample were then obtained using a custom demultiplexed script, which classifies and counts the merged data based on the unique sequencing run id. Post SQANTI annotation and filtering, only a third of ERCCs (unique number of ERCC = 37, 40.22%) were identified from both WT (mean number of ERCC: 32.4 (35%)) and TG (mean number of ERCC: 32.2 (35.22%)), with no difference in number of ERCC detected between WT and TG, although there were some ERCC that were detected in WT but not in TG, and vice versa. A minority of ERCCs (n = 8, 8.7%) at higher concentration were further annotated with more than one "isoform", indicating the presence of technical artefacts and more stringent filtering or clustering required, with ERCC at a higher concentration more likely to be sequenced and annotated with multiple redundant "isoforms". Exploration of these "isoforms" revealed them to be shorter transcripts likely to be generated as a result of fragmentation of the original molecule, incomplete PCR synthesis and template-switching. Application of TAMA-GO's script, tama-remove-fragment-models.py, successfully removed these partial, redundant isoforms, while retaining the intact isoforms.

Deeper investigation into the low coverage of ERCCs further identified an additional 20 lowly-expressed ERCCs that were discarded from cupcake's collapse scripts under the default coverage (alignment identity) parameters at 99%. Exploration of these imperfect-aligned sequences revealed 5'prime degradation of XX-XX nucleotides - one of the limitations of not using a 5'cap protocol. Inclusion of these ERCCs using a lower minimum coverage threshold at 95% increased the number of ERCCs detected by 20% (unique number of ERCC = 57, 61.96%), and strengthening the relationship between full-length read count and known amount of ERCC (95% coverage: corr = 0.98, $p = 1.41 \times 10^{-41}$; 99% coverage: corr = 0.82, $p = 4.89 \times 10^{-10}$).

2.1.3.7 Cupcake

To avoid redundancy of transcripts, aligned and filtered HQ transcripts were further collapsed to obtain a final set of unique, full-length, high-quality isoforms using Cupcake (a set of publicly-available, supporting scripts). HQ transcripts were filtered out for lack of mapping and low coverage/identity before collapsing into unique isoforms.

The abundance of each unique isoform can be estimated from the number of associated FL and NFL reads during IsoSeq cluster (not accounting for HQ transcripts that have been filtered out).

Finally, isoforms were filtered by 5'degradation due to the lack of a cap protection employed in the cDNA synthesis step (Clontech SMARTer cDNA kit).

2.1.3.8 Validation of isoforms with RNASeq

Samples sequenced with paired-end reads, Illumina Hi-Seq, 125bases. Paired end reads as more accurate for identifying and sequencing junctions. RNASeq data through stringent filtering (plot of fastqc) and aligned to mouse genome (Gencode, version X) using STAR (see section X for parameters). Abundance in TPM was then calculated with Kallisto (v0.46.0)[46] as an input into SQANTI to identify coverage of splicing junctions with RNASeq.

Provides support of transcripts from RNA-Seq data, highest expression of RNA-Seq reads of the splice junctions The junction with lowest coverage from RNA-Seq, and its associated read count Standard deviation of read counts across all the junctions for each transcript

2.1.3.9 SQANTI2 classification and filtering of isoforms

High-quality, clustered, filtered isoforms from Cupcake were characterised using SQANTI2 (v7.4), a pipeline initially developed by Conesa et al. [ref] and refined by Elizabeth Tseng (Pacific Bioscience's specialist) [ref]. In combination with genome annotation, SQANTI2 performs a reference-based correction of sequences and classifies isoforms based on splice junctions. The curated transcriptome can be further filtered and annotated with public datasets and RNA-Seq data (Section 2.1.3.8). Public datasets include

- FANTOM5 Cap Analysis of Gene Expression (CAGE) peaks: map transcripts, transcription factors, transcriptional promoters and enhancers
- Intropolis[47] : a comprehensive human RNA-Seq dataset
- PolyA motifs

Transcriptome Annotation and Isoform Classification

Using SQANTI classifications based on splice junctions, the transcriptome was segregated into the following categories (Figure 2.8):

- Well-known annotated genes with known isoforms, further isoforms classified as
  - Full Splice Match (FSM) if reference and query isoform have the same number of exons with matching internal junction. The 5' and 3' end, however, can differ

    - Incomplete Splice Match (ISM) if query isoform has fewer 5' exons than the reference, but the 3' exons and internal junctions match. The 5' and 3' end can also differ
- Well-known annotated genes with novel isoforms, with isoforms classified as
    - Novel in Catalog (NIC) if query isoform has different number and combination of exons to reference isoform, but is using a combination of known donor/acceptor splice sites
    - Novel Not In Catlog (NNC) if query isoform has different number and combination of exons to reference isoform like NNC, but also has at least one unannotated/novel donor or acceptor site
    - Genic Intron: the query isoform is completely contained within an annotated intron.
    - Genic Genomic: the query isoform overlaps with introns and exons.
- Unannotated, novel genes with novel isoforms with isoforms classified as
    - Antisense: the query isoform does not have overlap a same-strand reference gene but is anti-sense to an annotated gene.
    - Intergenic: the query isoform is in the intergenic region

Based on the pair of dinucleotides framing the intron boundary, splice junctions were either categorised as canonical for GT-AG, GC-AG and AT-AC and all the other possible combinations as non-canonical.

Lastly it can provide further classification of transcripts: As protein-coding or non-protein-coding by the presence of coding sequence that may potentially undergo non-sense medicated decay by the presence of ORF but CDS ends before the last junction that contain one or multiple exons (mono-exonic or multi-exonic respectively) that contain intronic sequences (intron retention) as fusions. The criteria XXXX

Further filtering of isoforms from technical artifacts

SQANTI was further used to filter the curated transcriptome from any technical artifacts that were generated during library preparation, including artifacts from RT template switching (TS) and off-priming[48] (Figure 2.9. RT template switching is an intrinsic feature of RT whereby the enzyme can transit within (intramolecular TS) or across (intermolecular TS) DNA tem-

Figure 2.8: Isoforms were classified by SQANTI as novel or known, and annotated to novel or known genes based on splice junctions. An isoform was classified as 'FSM' if it aligned with 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 isoform containing a combination of known donor or acceptor sites, or 'NNC' if it represented a novel isoform with at least one novel donor or acceptor site. FSM – Full splice match, ISM – Incomplete splice match, NIC – Novel in catalogue, NNC – novel not in catalogue

plates without terminating cDNA synthesis, if the original DNA template harbours two or more direct repeats.[49] This can result in either chimeric cDNAs or short, incomplete cDNAs that can be misinterpreted as isoforms generated from non-canonical splicing[50] (Figure 2.9a). Capitalising on the fact that RT switching is homology-dependent, SQANTI can identify these RT-switching artifacts by finding these direct repeats.[48]

Finally, off-priming artifacts can be generated from the binding of oligo(dT) primer to other internal homopolymeric adenines (A) regions that can be located within cDNA template, and thereby generate truncated cDNAs[51] (Figure 2.9b). SQANTI can explore the likelihood of thee events by determining the percentage of adenines (A) within the 20 nucleotide window downstream from the genomic coordinates of the isoform 3'ends and remove any that have a percentage lower than the user-defined threshold[48] - the lower the percentage of As, the higher the likelihood of the presence of internal polyA and off-priming.

In summary, the isoforms were filtered by the following criteria:

1. FSM with a reliable 3' end by:

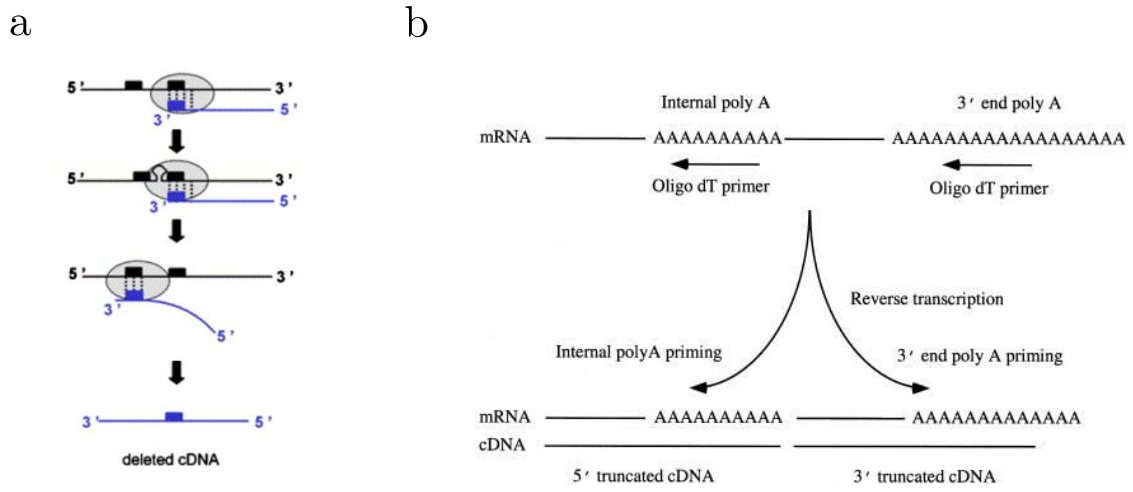


Figure 2.9: SQANTI identifies technical artifacts that were generated during first-strand cDNA synthesis a): Schematic diagram of reverse transcription template switching, taken from Cocquet et al.(2006).[49] The black and blue line represent the original cDNA and synthesising cDNA from RT respectively, the black box represent the direct repeats and the light grey sphere represent the RT enzyme. As exemplified, RT template switching is further facilitated by RNA secondary structures that could bring the repeats into proximity.[49] **b)** Schematic diagram of off-priming of oligo(dT) primer to internal A repeats, taken from Nam et al. (2002).[51] Oligo(dT) primer from first-strand cDNA synthesis can anneal to internal poly(A) sequence rather than the 3'end polyA, resulting in two truncated cDNAs.

- >60% of As in transcription termination site and no detected polyA motif, indicative of genomic contamination
- <Xbp 5' start and 3' end to reference transcript start end

2. Any other transcripts that have a reliable 3' end do not have any splice junctions were annotated as Reverse Transcription Switching.

2.1.3.10 Isoform expression from Iso-Seq

To control for sequencing bias in library depth, full-length (FL) read count for each isoform is normalized to transcripts per million (TPM)), which is calculated as:

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

With a cut-off lower than 0.5 TPM, a 0.5 - 10 TPM refers to low expression, a 11- 1000 refers to medium expression, and > 1000 TPM high expression [literature ref].

TPM is the most effective within-sample normalisation method to relatively quantify gene expression in a sample.[52] Other methods include RPKM (reads per kilobase of transcripts per million mapped reads), FPKM (fragments per kilobase of exon model per million mapped reads), which uses gene length to control for fragmentation in RNA-Seq protocol ("effective length normalisation") - however, this is not necessary in Iso-Seq.

Between-sample normalisation methods to relatively quantify expression of the same gene in different samples, remove technical variations due to presence of few highly expressed genes that make up a significant proportion of total reads, and due to different number of reads in each sample.

2.1.3.11 Limitations

While PacBio's Iso-Seq have major potential for transcriptome annotation, there were currently several major limitations that need to be addressed with further development of library preparation and bioinformatic data analyses:[23]

1. Lack of normalisation of RNA libraries, resulting in biased sequencing of high abundance transcripts and subsequent over-representation of such transcripts
2. Degradation of transcripts from 5' end, and thus lack of confidence in transcription start site and full-length structure

2.2 Iso-Seq: Optimisation

ERCC was used to assess the sensitivity and quality of whole transcriptome Iso-Seq runs and to optimise the bioinformatics pipeline (Figure X) by determining the number of ERCCs detected after:

1. varying the CCS parameters, which would affect the number of FL reads post Iso-Seq cluster
2. varying any additional parameters in cupcake collapse and usage of additional tools for filtering

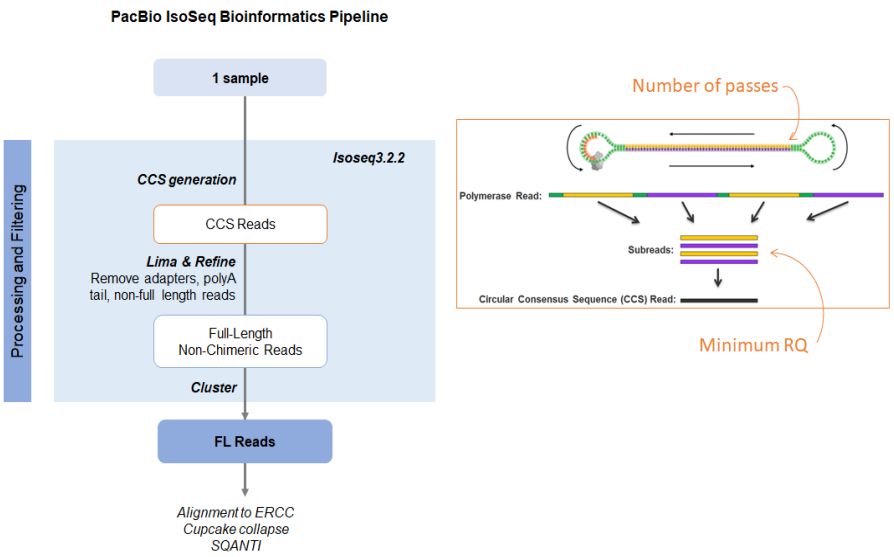
2.2.1 Varying CCS parameters

As described in Section 2.1.3.2, the proportion of raw subreads that can be successfully collapsed to generate a CCS is widely influenced by the number of passes (default: 3 passes) and minimum base accuracy (default: 99%), which settings are widely varied in scientific community.

To determine the most optimum parameters for CCS generation, CCS was generated on a subset of 1 sample using a combination of parameters, and then further validated with 2 whole samples (Figure 2.10):

Conclusion: 0.9 and 1 pass

- Results of number of reads and ERCCs detected (first and second round)



(a) Factors influencing successful CCS generation of raw subreads

Minimum RQ	Number of Passes: 1	2	3
0.8	12 different tests on 1 sample (10%)		
0.9			
0.95			
0.99			

(b) 1st CCS parameter optimisation

Minimum RQ	Number of Passes: 1	3
0.9	2 different tests on 2 samples (100%)	

(c) 2nd CCS parameter optimisation

Figure 2.10: Optimisation of CCS generation: Successful CCS generation of raw subreads is dependent on the number of polymerase passes and minimum RQ of subreads. A two-step approach was taken to determine the optimum parameters, using (b) whole range of parameters on 10% of one sample, and (c) extending analyses to two samples but with a more refined combination (as determined from the results of first step))

2.2.2 Additional parameters

To assess the sensitivity across Iso-Seq runs to detect ERCC, a merged analysis of whole transcriptome samples (n = 10, WT = 5, TG = 5) was performed with ERCC alignment and further collapse using Cupcake. The counts of full-length transcripts pertaining to each sample were then obtained using a custom demultiplexed script, which classifies and counts the merged data based on the unique sequencing run id. Post SQANTI annotation and filtering, only a third of ERCCs (unique number of ERCC = 37, 40.22%) were identified from both WT (mean number of ERCC: 32.4 (35%)) and TG (mean number of ERCC: 32.2 (35.22%)), with no difference in number of ERCC detected between WT and TG, although there were some ERCC that were detected in WT but not in TG, and vice versa. A minority of ERCCs (n = 8, 8.7%) at higher concentration were further annotated with more than one "isoform", indicating the presence of technical artefacts and more stringent filtering or clustering required, with ERCC at a higher

concentration more likely to be sequenced and annotated with multiple redundant "isoforms". Exploration of these "isoforms" revealed them to be shorter transcripts likely to be generated as a result of fragmentation of the original molecule, incomplete PCR synthesis and template-switching. Application of TAMA-GO's script, tama-remove-fragment-models.py, successfully removed these partial, redundant isoforms, while retaining the intact isoforms.

Deeper investigation into the low coverage of ERCCs further identified an additional 20 lowly-expressed ERCCs that were discarded from cupcake's collapse scripts under the default coverage (alignment identity) parameters at 99%. Exploration of these imperfect-aligned sequences revealed 5'prime degradation of XX-XX nucleotides - one of the limitations of not using a 5'cap protocol. Inclusion of these ERCCs using a lower minimum coverage threshold at 95% increased the number of ERCCs detected by 20% (unique number of ERCC = 57, 61.96%), and strengthening the relationship between full-length read count and known amount of ERCC (95% coverage: corr = 0.98, $p = 1.41 \times 10^{-41}$; 99% coverage: corr = 0.82, $p = 4.89 \times 10^{-10}$).

Several learnings were taken from analysis with ERCC: i) default parameters used in cupcake collapse, particularly alignment identity, are too stringent with removal of true transcripts, and ii) need for additional filtering using TAMA-GO's scripts to remove partial transcripts.

- Pipeline figure - a) unique ERCC b) isoform vs con, correlation - a) tama removal, b) tama removal further - a) mapping - a) readjustment, unique ERCC, lowly expressed transcipts, correlation

2.2.2.1 Application of additional parameters to Whole Transcriptome

A merged analysis of whole transcriptome samples (n = 12, WT = 6, TG = 6) was performed with alignment to the mouse genome (mm10), with 276,035 reads (99.3%) mapped to mouse genome, 365 reads (0.13%) mapped to ERCC and 1,568 reads (0.56%) unmapped. using cupcake's collapse default (99%) and reduced threshold (85%) for alignment identity.

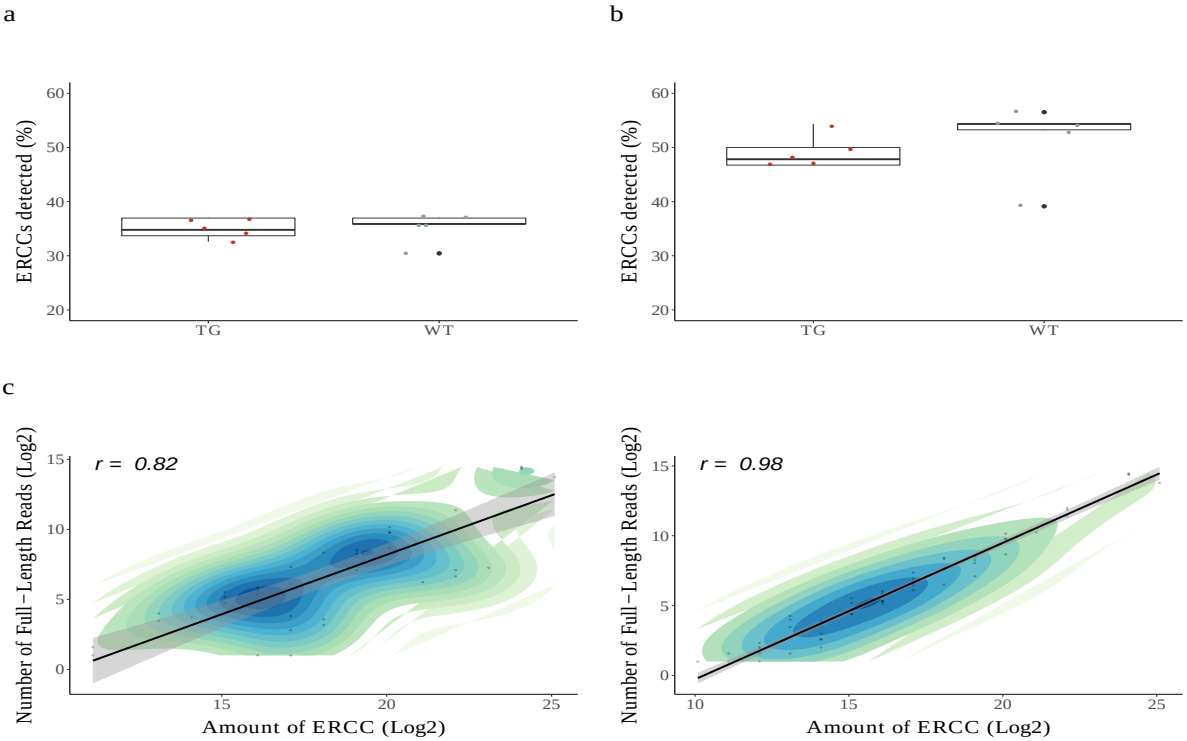


Figure 2.11: No significant correlation between RIN and Whole Transcriptome Iso-Seq run output: Samples with RIN >8 were selected for Whole Transcriptome Iso-Seq, with TG samples having distinctly lower RIN values than WT samples. However, no significant difference was observed for run output between WT and TG (Figure **??**)

2.3 Oxford Nanopore: cDNA Sequencing

2.3.1 Introduction

In 2014, Oxford Nanopore Technology (ONT) introduced another long-read sequencing technology akin to PacBio's SMRT with the ability to also generate long reads capable of resolving the exon structure of mRNA transcripts. However, rather than mimicking the natural DNA synthesis and measuring the incorporation events on the template strand (as is the focus in all major sequencing applications including the PacBio's SMRT) ONT's nanopore-based sequencing adopted an entirely novel approach - the DNA is directly inferred real-time from fluctuations in an electric current applied across a membrane as it passes through a protein pore.

2.3.1.1 Mechanism

In contrast to Pacbio's XXX by XXX Sequel, ONT's nanopore sequencing can be performed in a handheld MinION device (10 × 3 × 2 cm, 90 g), housing a flow cell at its centre where

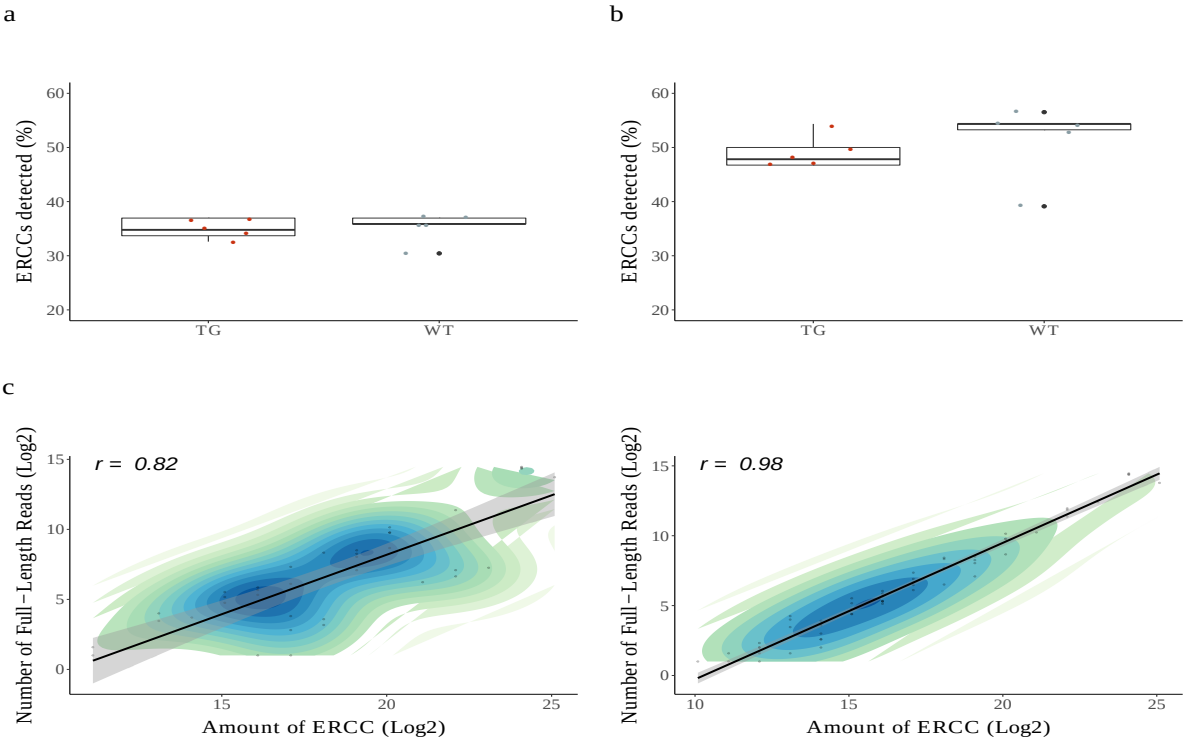


Figure 2.12: No significant correlation between RIN and Whole Transcriptome Iso-Seq run output: Samples with RIN >8 were selected for Whole Transcriptome Iso-Seq, with TG samples having distinctly lower RIN values than WT samples. However, no significant difference was observed for run output between WT and TG (Figure **??**)

the DNA sample is loaded. (Also on mobile device:[53]) Each flow cell contains a sensor array, consisting of 512 channels, each with 4 cells that can in turn house one nanopore (currently CsgG pore from *E.coli*[54]), which is inserted in an electrical resistant membrane surrounded by electrolytes. As an electric field is applied across the membrane, the negatively-charged DNA is driven across the nanpopore and subsequently interrupts the current. The pattern to the current disruption is unique and sensitive to the different nucleotides, which allows the sequence to be determined. However, while the current MinION contains a total of 2048 nanopores, only one of the four wells in each channel can be active at any time, as controlled by Application Specific Integrated Circuit (ASIC), allowing up to 512 independent DNA molecules to be sequenced simultaneously.

6-mer *flow cell and the different nanopore?*

2.3.1.2 Performance and Run Quality Metric

The ability of nanopore sequencing to directly read the DNA has both positive and negative implications. Inhibited mostly by the ability to deliver very high-molecular weight DNA to the pore, nanopore sequencing is able to generate much longer reads than SMRT sequencing (from 500bp to currently XXX), with theoretically no upper limit,[55] and with no bias towards length or GC content.[56,57] This offers great potential in transcriptomics profiling and genome assembly. However without the circular feature of PacBio SMRT sequencing, DNA strand cannot be sequenced multiple times and one of the major limitations of nanopore sequencing is its lower read accuracy. Low complexity stretches, including homopolymers, are furthermore difficult to resolve as translocation of homopolymers do not change the sequence of nucleotides within pore, thereby resulting in a constant signal.

However over recent years, major advances in both the basecalling algorithms, the chemistry and nanopore itself has drastically increased the initial accuracy from 60%[58] to 98.3% (vR.9.4.1 and Bonito). This includes the chemistry ($1D^2$) to sequence both the template and the complementary strand immediately after, thereby attaining a more accurate consensus read that increases the accuracy of template reads (1D) alone by 5%.[59] Further to mimic the circular consensus approach, two methodologies (INC-seq[60] and R2C2[61]) involving Rolling Circle Amplification and subsequent nanopore sequencing of circularised templates have been proposed with accuracy approaching 97.5%.

2.3.2 Lab Pipeline

Similar to the Iso-Seq library preparation (Chapter 2.1.2), nanopore sequencing of transcriptome also involved three main steps by first generating and amplifying cDNA, finally followed by ligation of adaptors. However unlike PacBio whereby the polymerase is bound after ligation of adaptors, the molecular motor (that drives the DNA through the pore) is already bound to the adaptors. At the time I trialled the nanopore sequencing, the Iso-Seq Express Library Preparation protocol had not yet been released, and the ONT rapid library preparation was much shorter and simpler than the Iso-Seq protocol I used.

2.3.2.1 cDNA synthesis

For a fair, direct comparison between ONT's MinION sequencing and PacBio's IsoSeq, 200ng total RNA extracted using AllPrep DNA/RNA Mini Kit (Qiagen) was likewise converted to single-stranded DNA using SMARTer PCR cDNA Synthesis(ClonTech).

2.3.2.2 ONT MinION Library Preparation

Despite a range of protocols available on the Oxford Nanopore community protocol that can be used pending on the source of sample, the SQK-LSK109 kit was used with cDNA as starting material. This kit is PCR-free and as such is dependent upon generation of high-quality and full-length cDNA, which would be provided using the SMARTer PCR cDNA synthesis kit rather than that detailed in 1D Strand switching cDNA by ligation protocol (SQK-LSK108).

2.3.2.3 Repair DNA and Ends

DNA calibration strand (DCS) is 3.6kb amplicon of Lambda genome, and is included in the sample library as a quality control of base-calling and sample preparation. End Repair prepare the ends of cDNA molecules for adapter attachment by addition of dA nucleotides

2.3.2.4 Adapter Ligation

Post DNA repair and end repair, ONT adapters with dT overhang are ligated to the 5' end of the dA-tailed cDNA molecules by hybridisation. The ONT adapters contain:

- motor protein (loaded processive enzyme?) that can bind to the nanopore and control/increase the speed of DNA translocation through the pore. While it is active in solution, it is inhibited from contacting the rest of the DNA through specialised bases in the adapter.
- cholesterol tether to facilitate DNA capture as (1) tethers the DNA molecule to the lipid bilayer (membrane) of the flow cell (2) reduces amount of diffusion of the DNA molecule from three dimensions (i.e the volume of whole buffer) to two dimensions (i.e. across the lipid bilayer)

2.3.2.5 Priming the Flow Cell

Sequencing buffer provides the optimal chemical conditions for powering DNA translocation through the Nanopore. This is the substrate cofactor of the motor enzyme that is used for DNA translocation process in the pore.

2.3.3 Bioinformatics Pipeline

Unlike the Iso-Seq bioinformatics pipeline which was largely established by PacBio, the pipeline for processing ONT reads was less defined and streamlined with constant emergence of new tools from research community. Much of the analysis, such as the usage of *Porechop* to remove primer sequences and of *TAMA* for transcript collapse, was taken from the Wellcome Trust Advanced Course: RNA Transcriptomics (2018), provided by J.Ragoussis (referred as WTAC), and refined using ERCC as benchmark.

2.3.3.1 Base-calling

The first analysis is to convert or "base-call" the electrical signals to the corresponding bases using Albacore, or a more recently developed package, Guppy, that requires information on the:

1. Chemistry of the run such as whether 1D or $1D^2$
2. Flow cell version used, to define the protein nanopore and subsequent 6-mer, which has different residual current
3. Sequencing kit used as this specifies the translocation speed, which informs the event segmentation algorithm how to recognise the corresponding bases from the electrical signal
4. use of barcoding to run multiple samples in one flow cells for downstream demultiplexing
5. type of output file, such as FASTQ or fast5

In contrast to PacBio's SMRT with the ability to generate consensus long reads, the raw accu-

racy of nanopore 1D cDNA sequencing is relatively low between 85–87%; however, significant improvements are made on reducing error rate by rapid development of both the technology and library preparation methods (Volden et al. 2018). Such high error rates, from frequent base deletions and insertions particularly near splice sites, can result in spurious alignments and in correct clustering of reads.

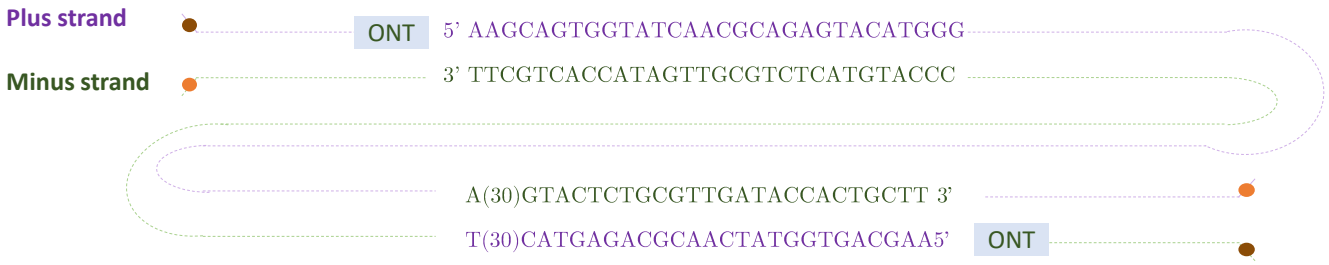
2.3.3.2 Quality Control of Run and Filtering of Base-called Reads

The quality characterisations of a single Nanopore sequencing run was assessed using *PycoQC*[62] and the official Nanopore QC tutorial,[63] by evaluating the performance of the run, the number of reads generated over time, and the length and quality score distribution of base-called reads. Reads with read quality score < 7 were removed using *Nanofilt*[64] with default parameters.

2.3.3.3 Removing of Nanopore and cDNA sequencing adapters

Similarly to the Iso-Seq bioinformatics pipeline, cDNA sequencing primer sequences and nanopore ligation adaptors were removed to prevent spurious alignment, using *Porechop*[65] (v0.2.4, parameters: –end_size 100, –adapter_threshold 90, –end_threshold 75, –min_trim_size 15, –discard_middle, –extra_end_trim 1). Under those parameters, a window of 100 nucleotides from each end of the reads were searched for a set of adaptors, which must have a minimum 90% identity to be considered present for trimming and a minimum 75% at the end of the reads; alignments smaller than 15bp or found within the middle of the reads (considered chimeric) were removed. By manually defining a unique set of adaptors that includes the cDNA primers (from Clontech SMARTer PCR cDNA synthesis kit), the ONT adaptors and corresponding polyA/T tail, it was possible to differentiate the strand orientation (Figure 2.13). Trimmed reads with adaptors present at both ends were retained, and reads corresponding to the minus strand were reverse complemented. Using *Cutadapt*[66] (v2.9, -a "A40"), the polyA sequence was then trimmed 40 nucleotides from the 3'end.

Of note, *Porechop* has been officially unsupported since 2018 and has been largely replaced by ONT's officially recommended tool, *Pychopper*;[67] however, *Pychopper* was not able to differentiate the strands for classification given that the 5'ends of both plus and negative strands have the same sequence (due to the nature of the cDNA synthesis kit).



Strand	Sequence
Plus strand start	AATGTACTTCGTTCAGTTACGTATTGCTAAGCAGTGGTATCAACGCAGAGTACATGGG
Plus strand end	AAAAAAAAGTACTCTGCGTTGATACCACTGCTT
Minus strand start	AATGTACTTCGTTCAGTTACGTATTGCTAAGCAGTGGTATCAACGCAGAGTACTTTTTTTT
Minus strand end	CCCATGTACTCTGCGTTGATACCACTGCTT

Figure 2.13: Structure of ONT library cDNA template: Shown is the final structure of cDNA molecules for ONT sequencing, after cDNA synthesis and adaptor ligation. The original cDNA molecules are outlined in purple and green, and the ONT boxes indicate the position of the ONT adaptors. The brown and orange circle refer to the motor protein and cholesterol moiety respectively.

2.3.3.4 Genome Alignment and Transcript Collapse

Trimmed reads were then aligned to the reference genome using Minimap2[68] (v2.17-r941, parameters: -ax splice). In the Iso-Seq bioinformatics pipeline, mapped transcripts were then processed using *Cupcake* for the removal of transcripts with low alignment identity and length, further collapse of high-quality transcripts to unique isoforms, and to obtain count information using output from *IsoSeq3 Cluster*.

As a similar comparison, mapped transcripts from ONT were processed using *TAMA* for removal of lowly-aligned transcripts and for further collapse to unique isoforms (script: tama_collapse.py, parameters: -e common_ends, -c 95, -i 80, -x capped -a 50, -z 50, -m 20, -d merge_dup) and to subsequently obtain count information (script: tama_read_support_levels.py). Under those parameters recommended by WTAC, transcripts were filtered for minimum alignment identity > 80% and alignment length >95% (same threshold as that applied in Iso-Seq pipeline using *Cupcake*) and collapsed by common exon start and end sites. 50 nucleotides at both 5'end and 3' end of the transcript, and 20 nucleotides at the exon/splice junction and tolerated for grouping transcripts to be collapsed. Of note, this is much more relaxed than the default *TAMA* parameters (-a 10 -m 10 -z 10), given that the 5'cap method was not used and the error rate of ONT reads were high.

Despite the growing emergence of various new tools developed for processing ONT reads - such as ONT's officially recommended pipeline with Pinfish and Stringtie, FLAIR, UNAGI - I chose to use *TAMA* due to greater flexibility and transparency with parameter usage, generation of multiple output files for quality control, and ability to subsequently obtain count information (script: tama_read_support_levels.py)

FLAIR: Full-Length Alternative Isoform analysis of RNA (FLAIR) Three steps are involved: Correct splice sites with short reads if incorrect splice site is within 10base pairs away from correct splice site, collapse reads to generate consensus sequences. This involves first grouping reads with identical splice junctions - "first pass nanopore isoform transcriptome"; the representative isoform within each group is determined by the most supported transcription and end site. All the reads, including reads that were aligned but not able to be fully corrected, are re-aligned to the "first-pass isoform" with the best alignment. First-pass isoforms that have

fewer than three supporting reads are filtered out; three supporting reads selected as threshold as this gave the highest base sensitivity without compromising on precision.

2.3.3.5 Isoform Quantification

In contrast to Iso-Seq, isoform quantification from ONT is relatively simpler in that each nanopore read corresponds to a single transcript (Tang et al. 2020). However, ambiguity still remains with assignment of truncated reads

2.3.3.6 Limitations of Oxford Nanopore

Refer to[16] for information on strand break etc

Chapter 3

Whole Transcriptome

This chapter is a modified version of the preprint, Jeffries et al. 2020,[69] currently under review for publication.

3.1 Introduction

3.1.1 Mouse model of AD amyloidopathy: J20

A mouse model of amyloidopathy, J20 overexpresses a mutant form human APP with two mutations identified by FAD, Indiana (V717F) and Swedish (K670N/M671L) mutations, directed by human platelet-growth-factor-beta promoter (PGRF-beta) with expression highest in the neocortex and hippocampus [Figure to show effects of mutations]. These mice exhibit defects in spatial memory and learning, with amyloid deposition by 5 – 7 moths, robust plaque formation by 8 – 10 months, and age-associated neuronal loss throughout the hippocampus. While J20 mouse closely resembles amyloidopathy development in human AD, insertion site of APP transgene has been shown to disrupt ZBTB20, a transcriptional repressor involved in hippocampal development.

3.1.2 Mouse model of AD tauopathy: rTg4510

Unlike with APP, there are currently no known mutations in MAPT linked to AD. Mouse models, such as rTG4510, that recapitulate AD tauopathy are therefore developed through

harbouring missense mutations in MAPT that are associated with tauopathy in familial frontotemporal dementia (FTD). In the case with rTg4510, the human tau transgene carrying the P301L mutation is over-expressed under the calcium calmodulin kinase II promotor (CaMK2a) and is largely restricted to the forebrain (such as hippocampus and cortex). These mice also exhibit cognitive and behavioural impairments, with neurofibrillary tangles developing as early as 2 months, and associated neuronal and synaptic loss evident by 9 months. Starting from the neocortex and progressing rapidly to the hippocampus, the age-dependent spread of neuropathology in rTG4510 mouse closely reflects the spread of NFTs in human AD, as classified into Braak stages. However, it is important to note that the genomic integration of CAMK2a and MAPT transgene has been to have off-target effects with disruption in the endogenous mouse genes, including XXX. [Figure X: rTg4510 with image of why it is called regulatable due to the mouse line]

3.2 Methods

Pacific Biosciences Iso-Seq dataset was generated with whole transcriptome approach using high-quality RNA from mouse entorhinal cortex of rTg4510 model (n = 12, WT = 6, TG = 6, mean age = 5 months, range = 2 - 8 months) (**??**). As a technological comparison and validation of the IsoSeq approach, a subset of samples were also sequenced on ONT (**??**). While both long-read sequencing approaches are superior to short-read RNA-Sequencing in the generation of full-length transcripts, there are major inherent batch biases due to the time-consuming and laborious protocol involved. The library preparation was standardised as much as possible, with the initial input of RNA for cDNA synthesis and the final library input for sequencing. However, due to the need for optimising each sample for library preparation and the rapid updates of sequencing chemistry throughout my PhD, each sample was effectively sequenced sequentially rather than as a batch.

3.2.1 RNA Extraction

Total RNA from mouse entorhinal cortex was extracted by Dr.Isabel Castanho (University of Exeter) using the AllPrep DNA/RNA Mini Kit (Qiagen), which is fully detailed in.[70] 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 D1000 ScreenTape System (Agilent).

3.2.2 RNA-Seq Library Preparation, Illumina Sequencing & raw data processing

In addition to long-read Iso-Seq, RNA from the same samples were also prepared for short-read RNA-sequencing by Dr. Isabel Castanho, which is also fully detailed in Castanho et al.(2020).[70] Raw sequencing reads, with Phred (Q) $\geq$ 35, were trimmed (ribosomal sequence removal, quality threshold 20, minimum sequence length 35) using fastqmcf (v1.0), yielding a mean untrimmed read depth of ~20 million reads/sample.

3.2.3 Iso-Seq Library Preparation

Following the Iso-Seq lab pipeline (**Section 2.1.2**), 200ng RNA from each sample was used for first strand cDNA synthesis (**Section 2.1.2.1**) and amplified using PCR with 14 cycles (**Fig-**

ure 3.1, **Section 2.1.2.4**). Purification with 0.4X and 1X AMPure PB beads selectively and successfully enriched cDNA with different molecular weights (**Figure 3.2**). The two fractions were then recombined at equimolar quantities and library preparation was successfully performed (**Figure 3.2**). Sequencing was performed for each sample on the PacBio Sequel using a 1M SMRT cell.

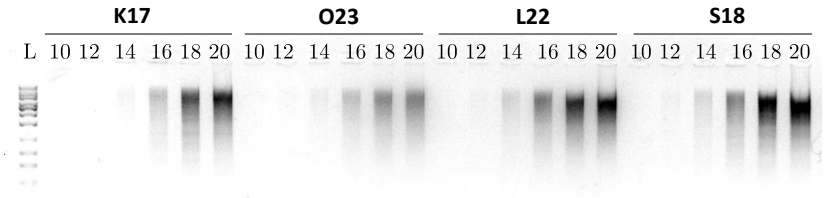


Figure 3.1: Samples were typically amplified using 14 cycles after performing PCR cycle optimisation: An example of gel image from PCR cycle optimisation of Samples K17, O23, L22 and S18. PCR aliquots were collected every two cycles (10, 12, 14, 16, 18, 20) and then run on gel electrophoresis. 14 cycles was determined to be optimal for large-scale amplification,as cycles below showed insufficient amplification whereas cycles above showed signs of over-amplification, which would result in biased sequencing representation. Ladder (L) shown is 1kb DNA ladder.

3.2.4 ONT Library Preparation

As a technological comparison (see **Section 2.3.2**), cDNA prepared from 2 samples for Iso-Seq were also sequenced on 2 separate ONT MinION flow cells (vXX). Briefly, following cDNA synthesis and amplification, library preparation was proceeded with ONT's Ligation Sequencing kit (SQK-LSK109) that follows the 1D cDNA sequencing protocol.

3.2.5 Iso-Seq Data Processing

Raw reads from each sample were processed using the Iso-Seq pipeline with optimised parameters (see Section X), and then merged to generate one complete transcriptome (**Figure 3.3**). In brief, the aim to identify poly-A full-length transcripts by the presence of both primers and polyA tail, and the clustering of similar transcripts to generate a unique,consensus isoform, which is then annotated by mapping to a reference genome. Briefly, circular consensus reads (CCS) were generated from a minimum of 1 pass and RQ X. cDNA primers and SMRT adapters were then removed using Lima (v1.9) to generate full-length (FL) reads, followed by removal of

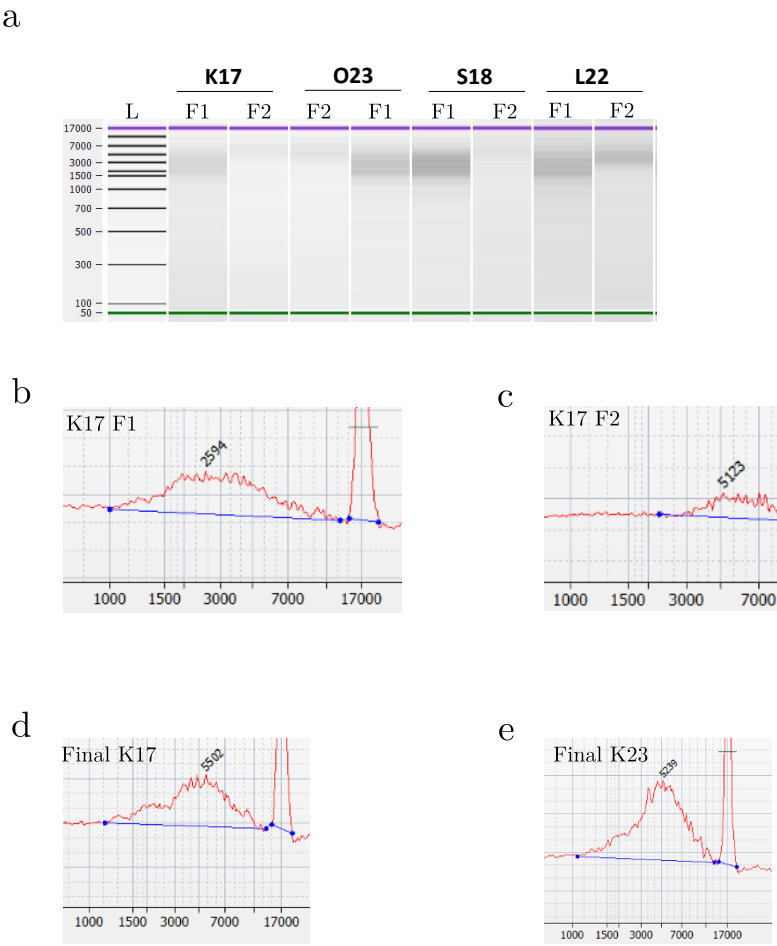


Figure 3.2: Library preparation was performed for each sample with successful cDNA purification and ligation with SMRT bell templates: Following large scale amplification using the optimum cycle number (as determined from **Figure 3.1**), the resulting cDNA was divided into two fractions (denoted here as F1 and F2) and purified with 1X (F1) and 0.4X (F2) Ampure beads. **a)** A bioanalyzer gel of amplified cDNA from the two fractions after ampure bead purification. **b)** A zoomed-in bioanalyzer electropherogram of Sample K17 Fraction 1 and **b)** a zoomed-in bioanalyzer electropherogram of Sample K17 Fraction 2, from the gel depicted in Figure a). **d)** A zoomed-in bioanalyzer electropherogram of Sample K17 and **e)** of Sample K23 recombining both fractions and performing SMRTbell template preparation. The samples at this point have been DNA-damage repaired, exonuclease treated, and ligated with SMRT bell adapters. The y-axis of the bioanalyzer electropherogram represents the size. The size distribution for each fraction was determined from the start to the end point of the smear, as in Figure a), or the equivalent peak, as depicted in Figure b) and Figure c).

As is evident from Figures a) - c), cDNA in Fraction 2 has a significantly higher molecular weight across all the samples as would be expected. As seen in Figures d) and e), pooling of both fractions have enriched high molecular weight cDNA fragments, which were still in intact after multiple processing in SMRTbell template preparation. Of note, despite the samples were prepared sequentially, the bioanalyzer profiles were consistent.
F1 - Fraction 1 containing cDNA purified with 1X Ampure beads; F2 - Fraction 2 containing cDNA purified with 0.4X Ampure beads

artificial concatemers reads and trimming of polyA tails in Iso-Seq3 Refine. Full-length, non-concatemers (FLNC) reads were then collapsed, according to default parameters in Iso-Seq3 Cluster, to high-quality transcripts with accuracy >99%, which were mapped to the reference mouse genome using minimap2 (v2.17). Transcripts were then further filtered based on mapping quality and clustered using Cupcake's collapse script, followed by SQANTI2 annotation to identify fusion transcripts, proximity to CAGE peaks derived from the FANTOM dataset, TSS and TTS sites and classification of lncRNA in combination with lcRNA gene annotation (vM22). Subsequent filtering by TAMA was then applied to remove potential artifacts. CAGE peaks facilitates the mapping of transcripts, transcription factors, transcriptional promoters and enhancers.

3.2.6 ONT Data Processing

In brief, raw reads were basecalled using Guppy and were classified as "pass" if the mean quality score was >7.

3.2.7 Characterisation of Alternative Splicing Events

Alternative splicing events were assessed using a range of packages and custom scripts: mutually exclusive exons (MX) and skipped exons (SE) were assessed using SUPPA with the parameter –f ioe, intron retention (IR) with 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.

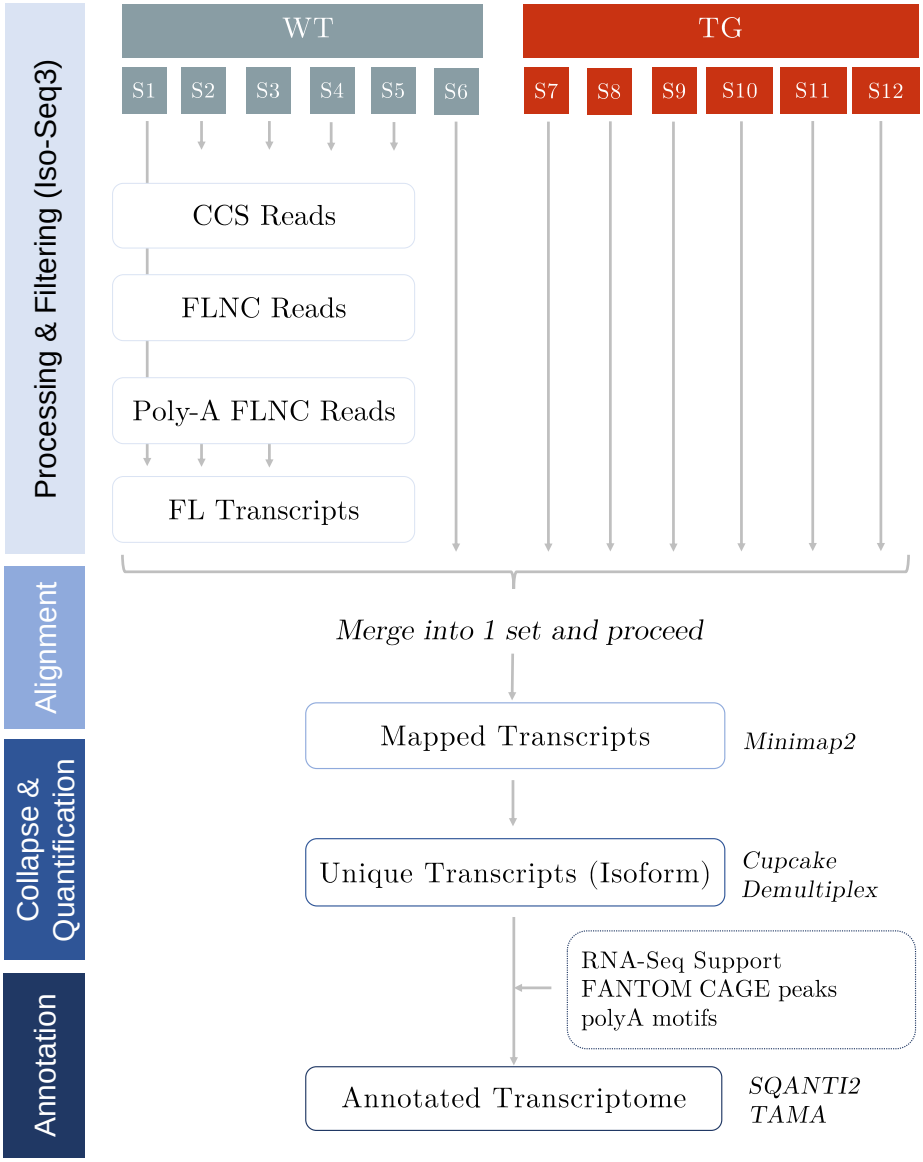


Figure 3.3: PacBio Isoseq Bioinformatics Pipeline: Pipeline is adapted from ToFU [18]

3.3 Results

12 mouse samples (6 WT and 6 TG) was sequenced using Iso-Seq approach on the PacBio Sequel 1 platform and analysed together for an accurate,deep characterisation of the full-length splice variants and identification of novel isoforms in the mouse transcriptome.

3.3.1 PacBio's Iso-Seq run performance and sequencing metrics

Following library preparation and single-molecule real time sequencing (SMRT), a total of 371Gb (s.d = 4.35Gb, range = 22.5Gb - 38.74Gb) and 8,082,647 polymerase reads (s.d = 63,013 reads, range = 530,974 - 733,495 reads) were obtained (**Table 3.1**). No significant difference was reported between WT and TG (n = 12 animals, two-tailed unpaired t-test, t(10) = -0.636, P = 0.539, **Figure 3.4a**), and no significant correlation was observed between run yield and RIN across samples (n = 12 animals, Pearson's correlation, t = -0.98, df = 10, P = 0.350, **Figure 3.4b**). Yield across all the samples were within the range as would expected from SMRT Iso-Seq library.

Sample	Age	Phenotype	RIN	Total Bases (GB)	Unique Yield (GB)
K17	2 months	WT	9.2	29.56	-
K18	2 months	TG	8.8	31.1	1.21
K23	8 months	WT	9.1	34.60	2.06
K24	8 months	TG	9.2	34.61	2.09
L22	8 months	TG	8.7	38.74	2.1
M21	2 months	WT	9.2	30.45	-
O18	2 months	TG	8.9	22.53	1.56
O23	8 months	WT	9	31.25	-
Q20	8 months	TG	8.6	33.16	2.27
Q21	2 months	WT	9.2	24.52	2.27
S18	2 months	TG	8.9	30.41	1.69
S23	8 months	WT	9.1	30.28	-

Table 3.1: Phenotypic information and Iso-seq run yield for each sample of Tg4510 sequenced using Whole Transcriptome approach

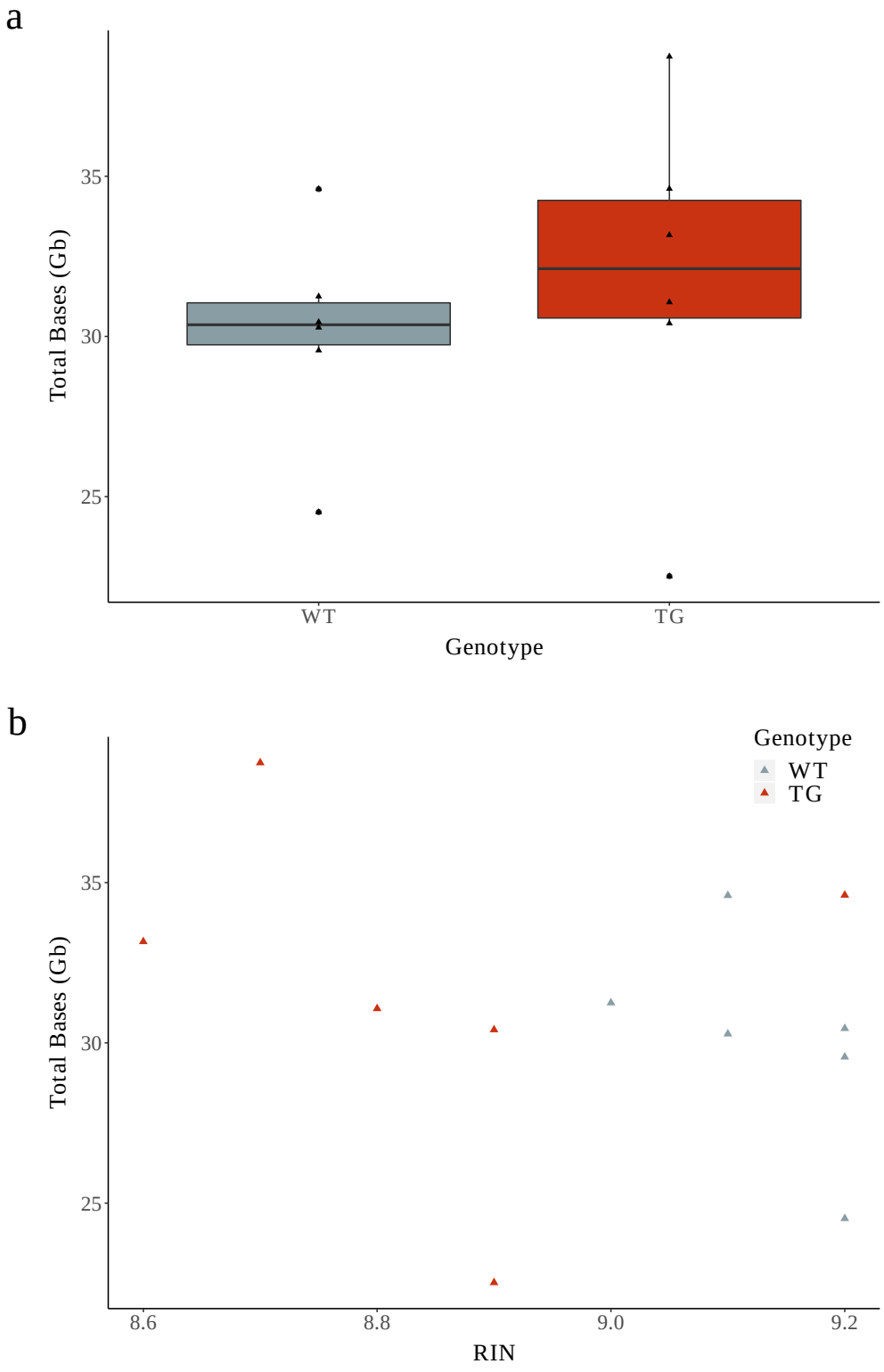


Figure 3.4: Whole Transcriptome Iso-Seq runs generated ~30Gb per sample, independent of RIN score: Sequential Iso-Seq run generated **a)** a range of 30-35Gb per sample of the whole transcriptome, with no significant difference observed between WT and TG Tg4510 mice. Of note, two samples with <25Gb in WT and TG refer to earlier samples sequenced with a lower chemistry. **b)** Despite TG samples having distinctly lower RIN values than WT samples, no significant difference in yield output was observed between WT and TG.

Sample	Polymerase Reads	Read Length						Productivity			Control				Local Base Rate	Template	
		Polymerase		Subread		Insert		P0	P1	P2	Total Reads	Pol RL Mean	Concordance			Adapter Dimer (0-10bp)	Short Insert (11- 100bp)
		Mean	N50	Mean	N50	Mean	N50						Mean	Mode			
B21	735598	39971	82100	1531	2125	3162	3896	8.71% (87817)	73.94% (745646)	18.33% (184883)	9940	34144	0.85	0.89	2.61	0	0
C20	749931	45670	91153	1426	2066	3204	4075	10.68% (107699)	75.36% (759912)	14.95% (150735)	9910	37019	0.85	0.89	2.75	0	0
C21	530395	44208	87750	2258	2794	3358	4250	38.0% (387661)	52.5% (535299)	9.4% (96275)	4880	50690	0.85	0.85	2.07	0.00	0.01
E18	545,272	41,036	83,295	2,467	3,049	3,588	4,335	38.88% (396026)	53.61% (546027)	7.58% (77181)	722	48,253	0.85	0.85	2	0	0
K17	673972	43856	90561	1253	2021	3336	4753	10.55% (106,736)	67.42% (681,794)	22.73% (229,816)	7036	34651	0.85	0.89	2.72	0.08	0.06
K18	566086	54892	101220	1256	1775	2863	3661	29.77% (299933)	57.25% (576863)	14.05% (141550)	10707	44640	0.87	0.89	3.05	0	0
K23	698178	49563	98801	1697	2670	3779	4779	16.1% (164308)	69.2% (704197)	14.7% (149841)	5951	40498	0.85	0.89	2.78	0	0
K24	711015	48675	97024	1714	2487	3834	5018	14.22% (144813)	70.49% (717880)	15.28% (155653)	6762	38363	0.85	0.87	2.671	0.01	0.01
L22	675283	57370	112630	1869	2867	3903	4793	17.41% (175439)	68.08% (686007)	15.58% (156900)	10647	44215	0.86	0.89	2.96	0.01	0
M21	660841	46082	91628	2234	2754	3952	4733	16.6% (168567)	65.9% (671224)	17.5% (178555)	10301	38690	0.85	0.87	2.79	0.01	0.01
O18	530974	42423	85331	2609	3146	3443	4082	41.8% (426378	52.6% (536435)	5.5% (56422)	5415	49778	0.86	0.85	2.05	0	0
O23	730733	42771	89372	1490	2347	3608	4878	9.37% (94536)	73.33% (740184)	18.19% (183626)	8908	34993	0.85	0.89	2.56	0.06	0.04
Q20	715206	46360	92519	1,999	2,926	3,978	4,954	11.51% (117223)	70.91% (722135)	17.58% (178988)	6855	37990	0.85	0.87	2.6	0.01	0.01
Q21	733495	33429	70750	2563	3286	3710	4750	15.9% (161679)	72.1% (735250)	12.0% (122305)	1668	44201	0.85	0.85	1.99	0.00	0.01
S18	682529	44549	90041	1435	2041	3282	4400	11.98% (121,055)	68.45% (691651)	20.35% (205,640)	7881	36541	0.86	0.89	2.85	0.11	0.07
S23	704335	42991	89160	1346	2020	3272	4383	7.02% (71074)	70.18% (710471)	23.39% (236801)	6019	35167	0.85	0.89	2.57	0.01	0.01

With the application of long-reads bioinformatics pipeline (as detailed in Section X), the raw reads were processed and clustered to unique consensus transcripts, which were then mapped and annotated as isoforms - low-quality, lowly-supported, unmapped and degraded reads were sequentially filtered at each stage. Across all 12 samples, a total of 5.66M CCS reads (mean = 471K,s.d = 46.8K, range = 353K - 512K) and 4.55 FLNC reads were successfully generated (mean = 379K, s.d = 47.0K, range = 270K - 412K) after multiple processing (**Figure 3.6a**). 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) (**Figure 3.6b**), and were mapped to 278K and 352 loci of the mouse reference (5K had multi-mapping) and ERCC annotations respectively. After filtering for 85% alignment identity and 95% length (**Figure 3.6c**), 266K transcripts were retained.

Showcasing the sensitivity of the sequencing platform and approach, only 62% (n = 57) of ERCCs were detected, those of which were more highly expressed and with a threshold concentration of XX (**Figure 3.7a**). However of those ERCCs detected, the number of FL reads detected was highly correlated to the known amount used (corr = 0.98, P = 1.42 x 10^{-41} **Figure 3.7b**), highlighting the power of Iso-Seq to quantify highly-expressed transcripts.

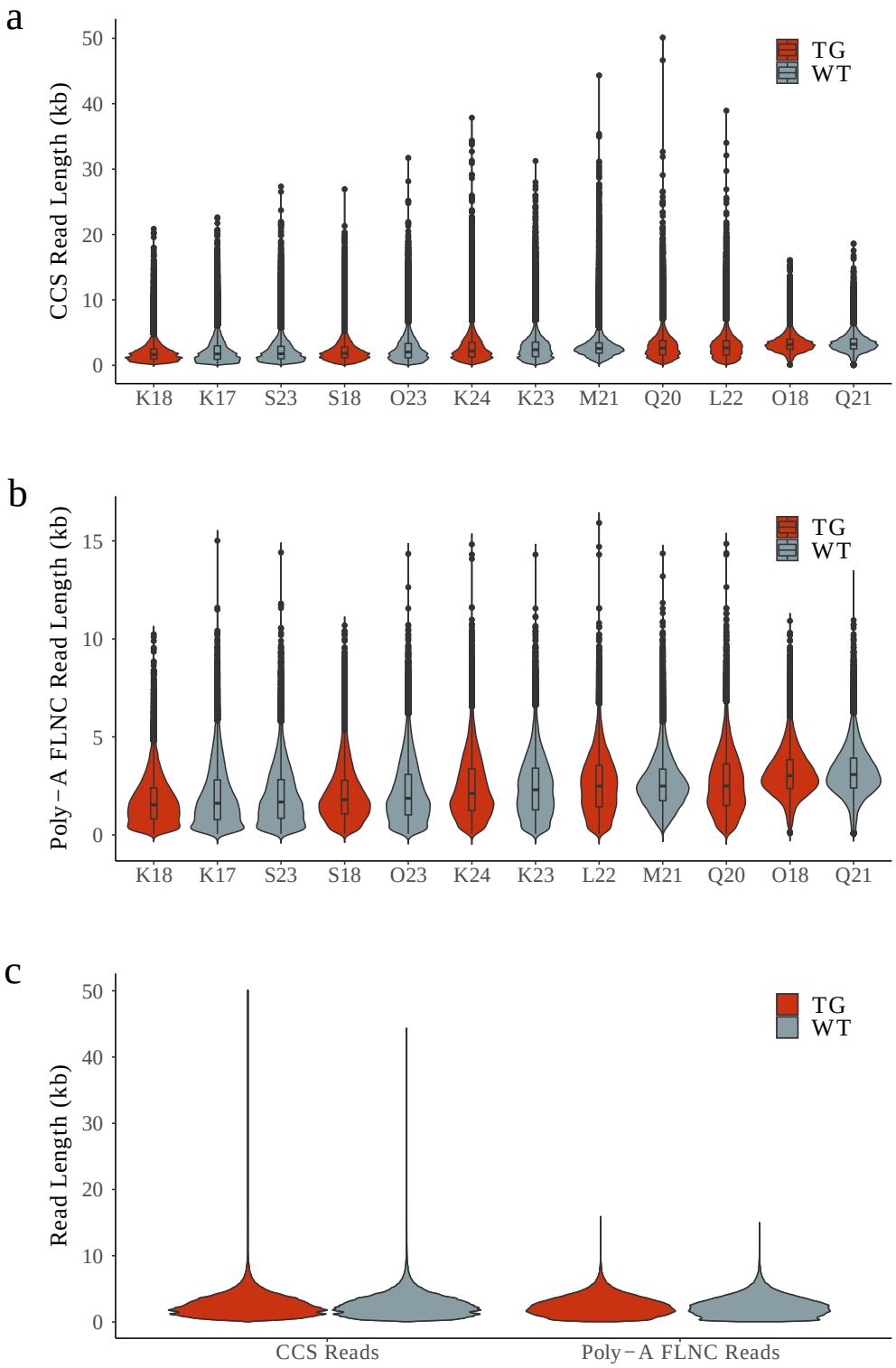


Figure 3.5: Whole Transcriptome Iso-Seq runs generated ~30Gb per sample, independent of RIN score: Sequential Iso-Seq run generated **a)** a range of 30-35Gb per sample of the whole transcriptome, with no significant difference observed between WT and TG Tg4510 mice. Of note, two samples with <25Gb in WT and TG refer to earlier samples sequenced with a lower chemistry. **b)** Despite TG samples having distinctly lower RIN values than WT samples, no significant difference in yield output was observed between WT and TG.

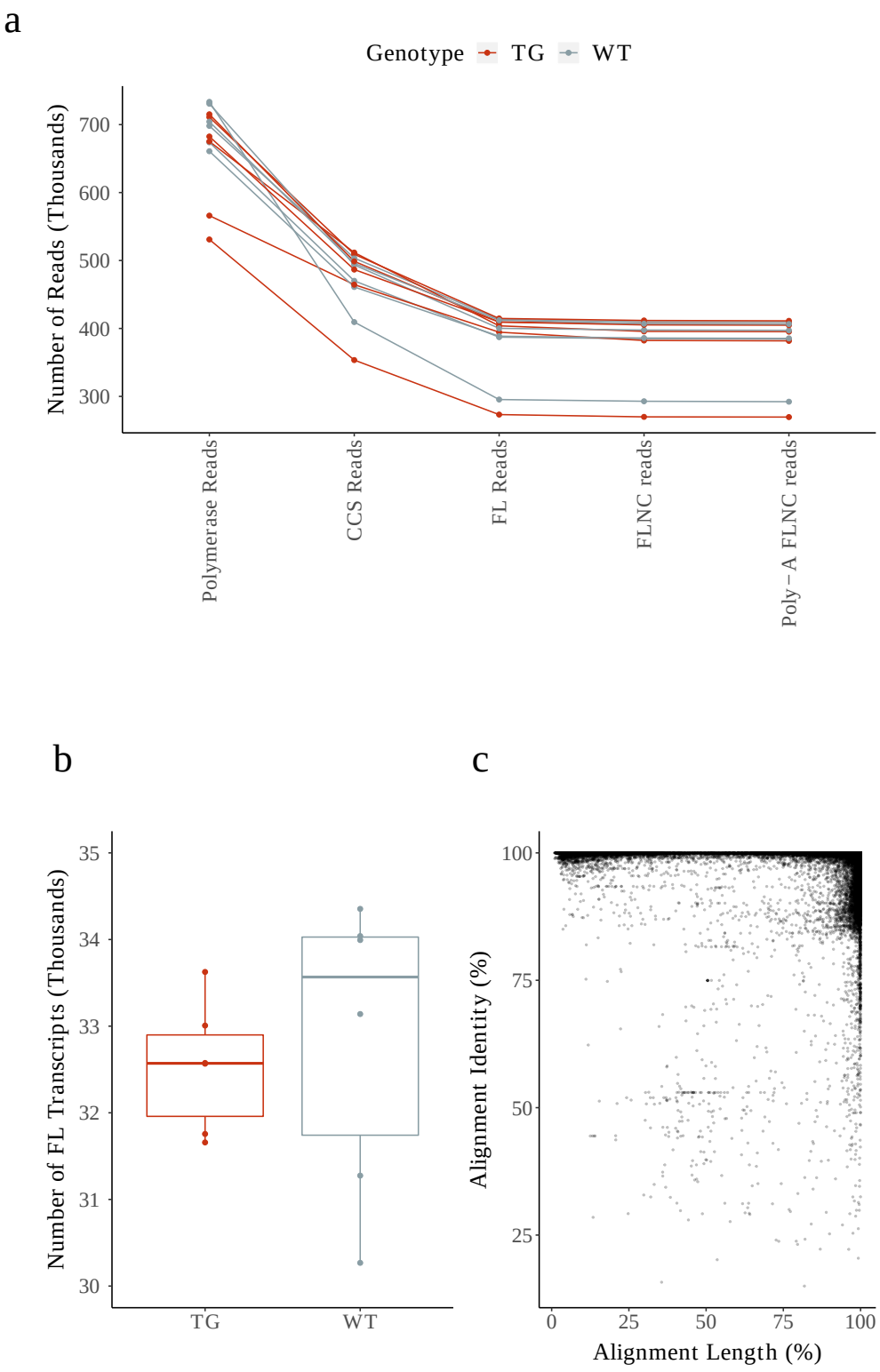


Figure 3.6: Sequential processing of Iso-Seq Reads generated around 32K transcripts per sample with good alignment to reference genome: **a)** Processing of Iso-Seq reads generated a similar number of reads across all sample throughout Iso-Seq3 bioinformatis pipeline, with the exception of 2 earlier samples. **b)** Despite this, all the samples had similar number of FL transcripts with no signficant difference observed between WT and TG. **c)**The majority of transcipts aligned to mouse reference genome (mm10) with >85% alignment identity and >95% length

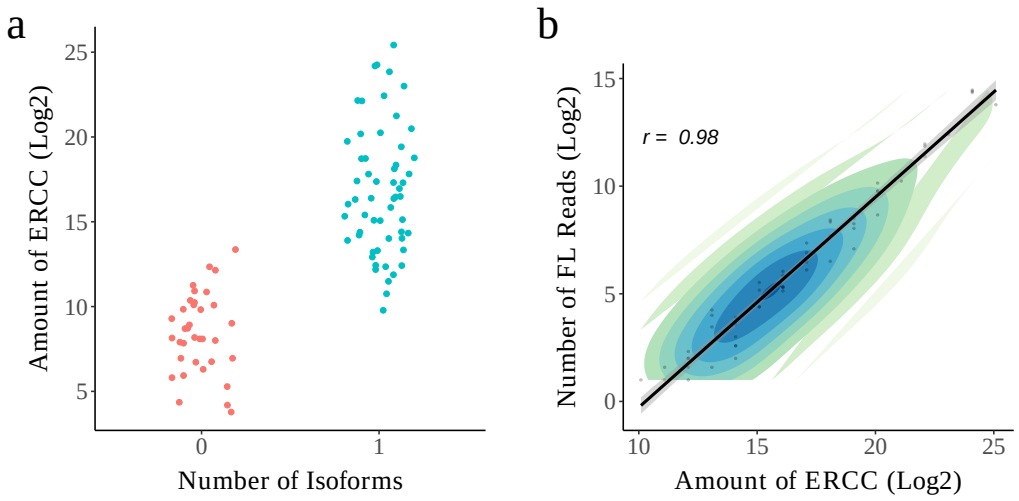


Figure 3.7: Over 60% of ERCCs were detected with highly accurate quantification **a** Highly-concentrated ERCCs were detected as single molecules, as expected, and **b** the number of full-length reads associated for each detected ERCC was highly correlated to known amount. FL - Full Length

3.3.2 Nanopore Sequencing run performance and sequencing metrics

As a technological comparison, two of the samples sequenced on the PacBio Sequel were also sequenced on the ONT MinION. In contrast to the Iso-Seq, the run performance and yield was significantly poorer and lower with greater variability observed between the two samples (K18: 10.1Gb from nanopore vs 31.1Gb from Iso-Seq, M21: 3.68Gb from nanopore vs 30.45Gb from Iso-Seq, **Table 3.2**). The significantly low performance of M21 was due to saturation and permanent blocking of pores (**Figure 3.8b**) with more rapid decline of sequencing activity than is expected (**Figure 3.9b**). No difference in the sequencing speed was observed across the course of the run (**Figure 3.10**), however, suggesting that the sequencing chemistry and flow cell was of good quality. The low performance output of the second sample is therefore likely due to introduction of air bubbles and contamination during library preparation.

After removing low-quality basecalled reads (QV < 7), a total of 6M reads were acquired (K18: 4.49M reads, 66% of total reads, M21: 1.68M reads, 79.2%, **Table 3.3**). Interestingly, despite the marked lower run performance of the second sample, the read quality was slightly higher (K18: mean QV = 9.5, M21: mean QV = 10.2) whereby the first sample had a large portion of very low-quality reads (QV < 2, **Figure 3.11a**). Nevertheless, both samples had a very similar read length distribution profile (**Figure 3.11c,d**), with a mean length around 1.8Kb. However in contrast to Iso-Seq, the distribution is skewed to the left with enrichment of smaller molecules

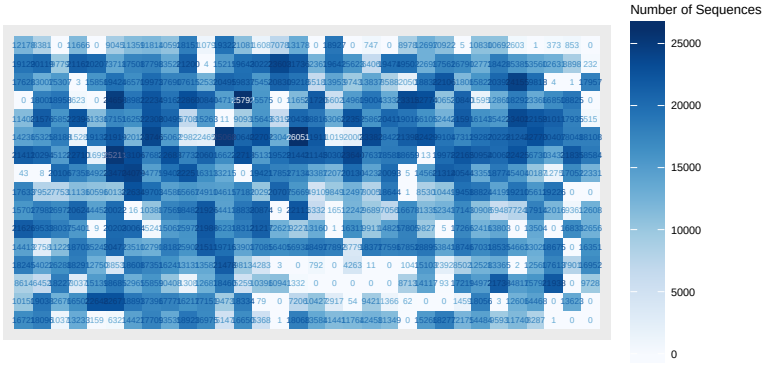
(1kb) - a likely reflection of the library size with no size enrichment during library preparation rather than a length bias of the technology.

Despite the relatively low run performance, more reads were acquired per flow cell per sample than per SMRT cell in Iso-Seq (mean number of basecalled reads from nanopore across 2 samples: 4.43M, mean number of polymerase reads from Iso-Seq across 12 samples: 0.67M). This is because while Iso-Seq was able to generate very long polymerase reads (mean length across 12 samples = 46kb), contributing to the run yield in gigabases, the number of reads generated per run was limited to 1M (number of SMRT cells). Conversely, there was no upper limit in the number of reads that can be generated with nanopore sequencing provided the pores remained active. A greater proportion of ERCCs (68 ERCCs, 73.9%) was therefore observed in one nanopore run compared to 12 Iso-Seq runs (57 ERCCs, 62%), and for those ERCCs detected, the number of FL reads detected was also highly correlated to the known amount used (corr = 0.98, P = 9.73×10^{-51}).

Sample	All Reads		Active channels	Run Duration
	Total Bases (Gb)	Number of Reads		
K18	10.1	6,752,951	479	48hours
M21	3.68	2,122,012	425	48hours

Table 3.2: Poorer run performance and lower yield output observed from Nanopore Sequencing of Whole Transcriptome. Two samples, sequenced on PacBio Sequel using Iso-Seq approach, were also sequenced on ONT MinION on two separate flow cells over 48hours. The number of total reads basecalled was less than a third of the reads generated on the same samples from the Iso-Seq approach (**Table 3.1**)

a



b

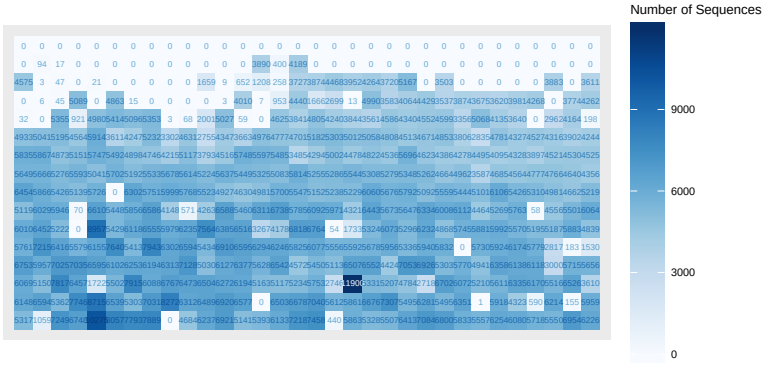


Figure 3.8: Sequencing channel activity plot from nanopore sequencing. Heatmap representation of channel productivity spatially for **a)** Sample K18 (10.1Gb) and **b)** Sample M21 (3.68Gb) as DNA is translocated through the pore and signal is collected. A stark contrast of activity can be seen between the two samples, with a significant number of inactive channels (white box) in Sample M21 - of the channels that are active, fewer DNA molecules are translocated and read. Of note, the activity shows the number of sequences generated per channel not per pore, given that each channel corresponds to four different pores.

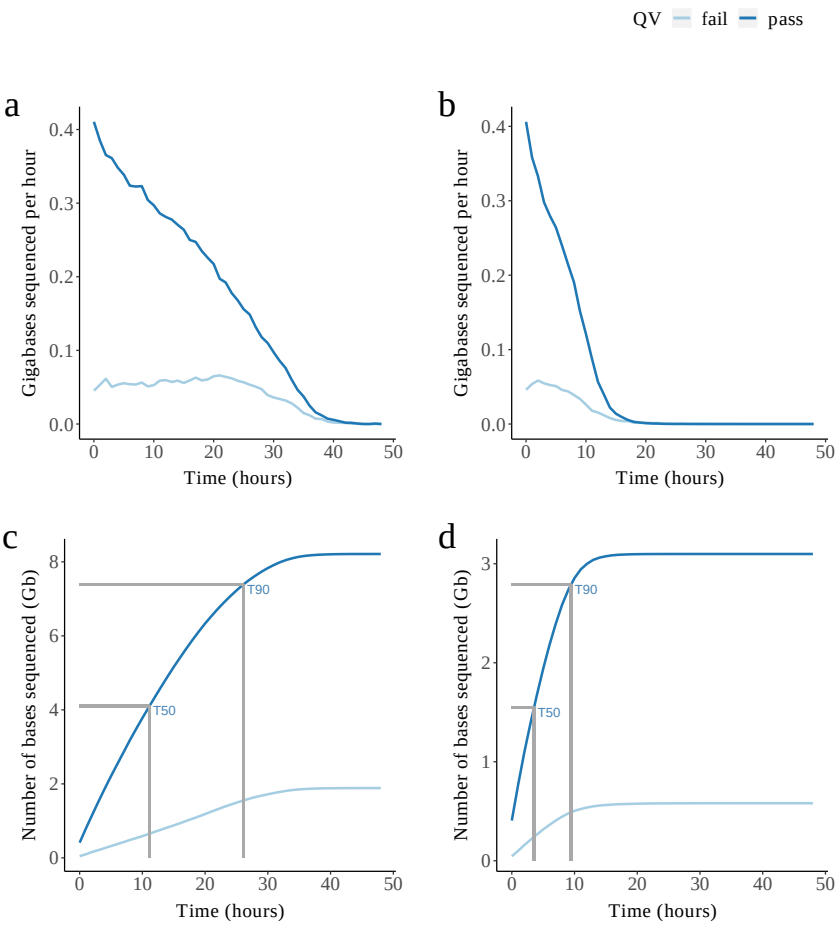


Figure 3.9: Temporal run performance from nanopore sequencing. Shown is the **a)** number of basses generated per hour over the course of the sequencing run from Sample K18 and from **b)** Sample M21, and **c)** cumulative reads generated from Sample K18 and from **d)** Sample M21. The reads are classified as "pass" (dark blue) if QV > 7 and "fail" (light blue) if QV < 7. The rapid decline of pore activity of Sample M21 is evident from Figure b) in contrast to Figure a), with 90% of the sequencing data acquired within the first 10hours of the run (Figure d). T50 and T90 refers to the time point at which 50% and 90% of total basecalled reads where acquired respectively. Gb - Gigabases

Sample	Total Bases (Gb)	Number of Reads	Read Length (bp)				Read Quality	
			Median	Mean	N50	Longest Read	Median	Mean
K18	8.21	4,468,629 (66.2%)	1400	1838	2521	48877	9.6	9.5
M21	3.1	1,679,931 (79.2%)	1410	1845	2644	82984	10.3	10.2

Table 3.3: Sequencing metrics of filtered high-quality reads. Basecalled reads were filtered on quality score with a QV threshold of 7. N50 refers to the sequence length at which 50% of reads are sized at or over. Gb - Gigabases

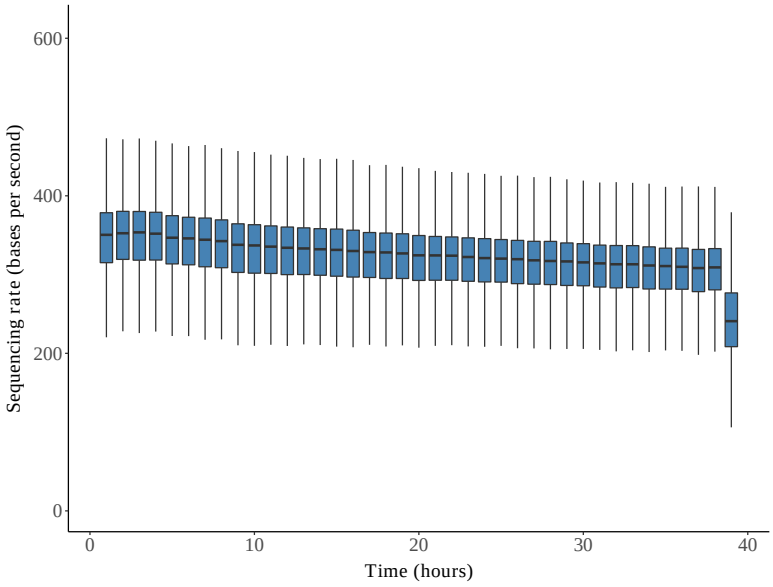


Figure 3.10: DNA translocation speed against time. A boxplot of the translocation speed (sequencing rate) for Sample M21 over the course of the 48-hour run.

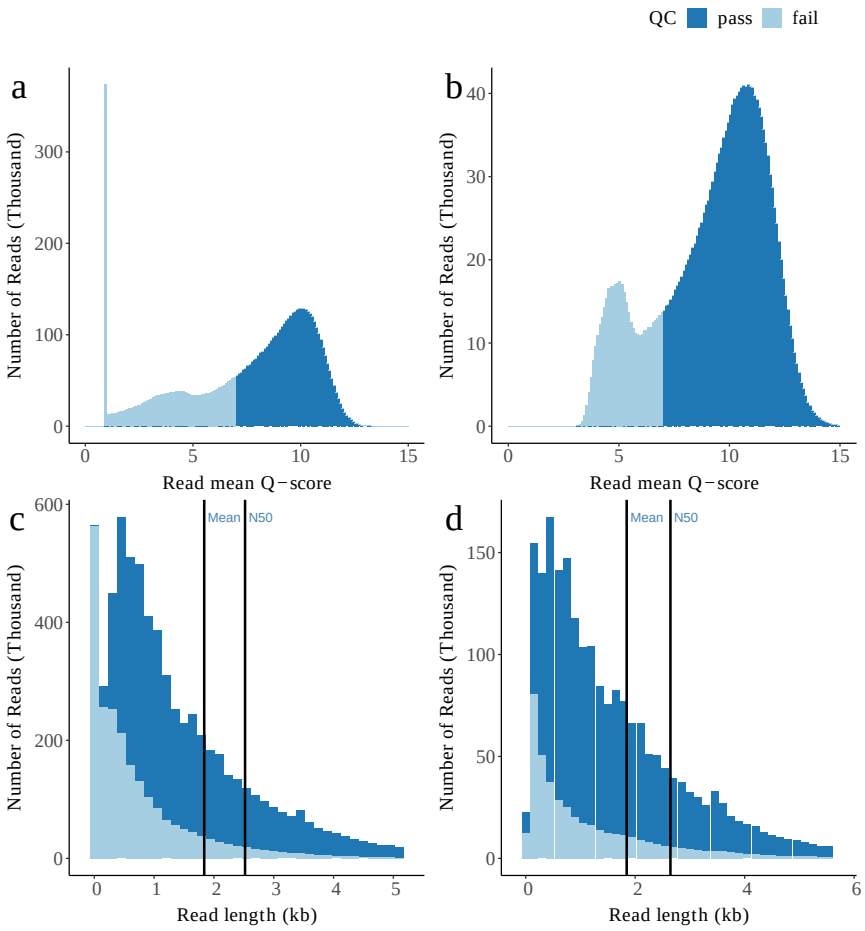


Figure 3.11: Length and quality distribution of ONT basecalled reads. Shown are histograms of the number of sequenced reads against **a)** mean read quality score of Sample K18, **b)** mean read quality score of Sample M21, and against **c)** length of Sample K18 and of **d)** Sample M21. The distribution has been shaded for reads that have passed or failed the quality filter (Q-score threshold of 7). N50 refers to the sequence length at which 50% of reads are sized at or over.

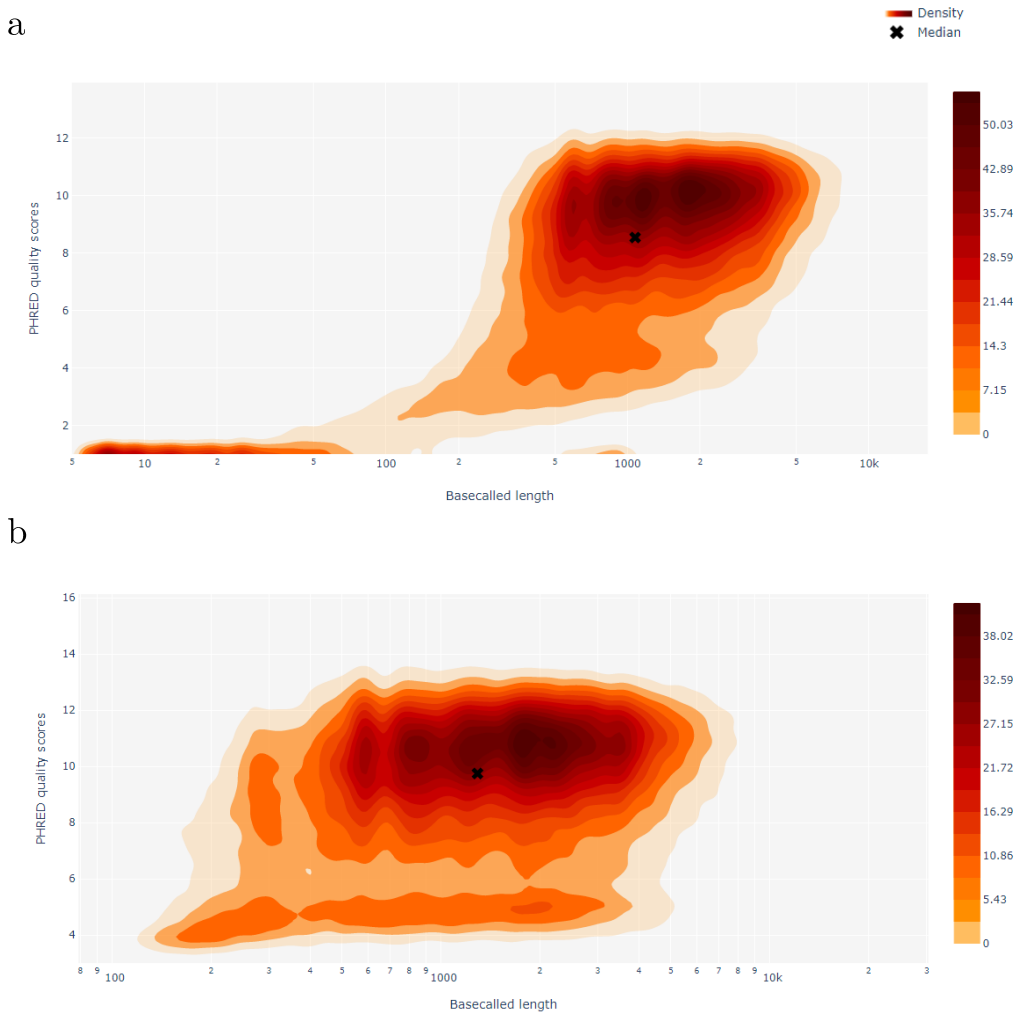


Figure 3.12: Distribution of quality scores against read lengths for all ONT basecalled reads. Shown are 2D density plots for **a)** Sample K18 and **b)** Sample M21 of mean sequence quality (Phred quality score) against read length (log10). Figures are generated from PycoQC.[62]

3.3.3 Transcriptome annotation

After further collapsing and filtering of transcripts using the Iso-Seq data, a total of 46,626 unique and intact isoforms were identified (mean = 27.5K, s.d = 2.32K, range = 24.2K - 31.2K) and annotated to 14,482 (98.6%) known and 202 (1.38%) novel genes. Gene expression patterns from Iso-Seq reflected expected transcriptional profiles for the brain regions profiled. Using the Mouse Gene Atlas database, the 500 most abundantly-expressed genes were most significantly enriched for 'cerebral cortex' (odds ratio = 6.07, adjusted P = 6.8×10^{-17}). Rarefaction curves confirmed that the dataset approached saturation, indicating that our coverage of isoform diversity was representative of the true population of transcripts (**Figure 3.13a**). Supporting the validity of these isoforms, the majority (n = 35,262, 75% of isoforms) were enriched near an annotated CAGE peaks (located within 50bp), and the vast majority of unique splice junctions (n = 138,032, 97.8% of junctions) were supported by RNA-Seq.

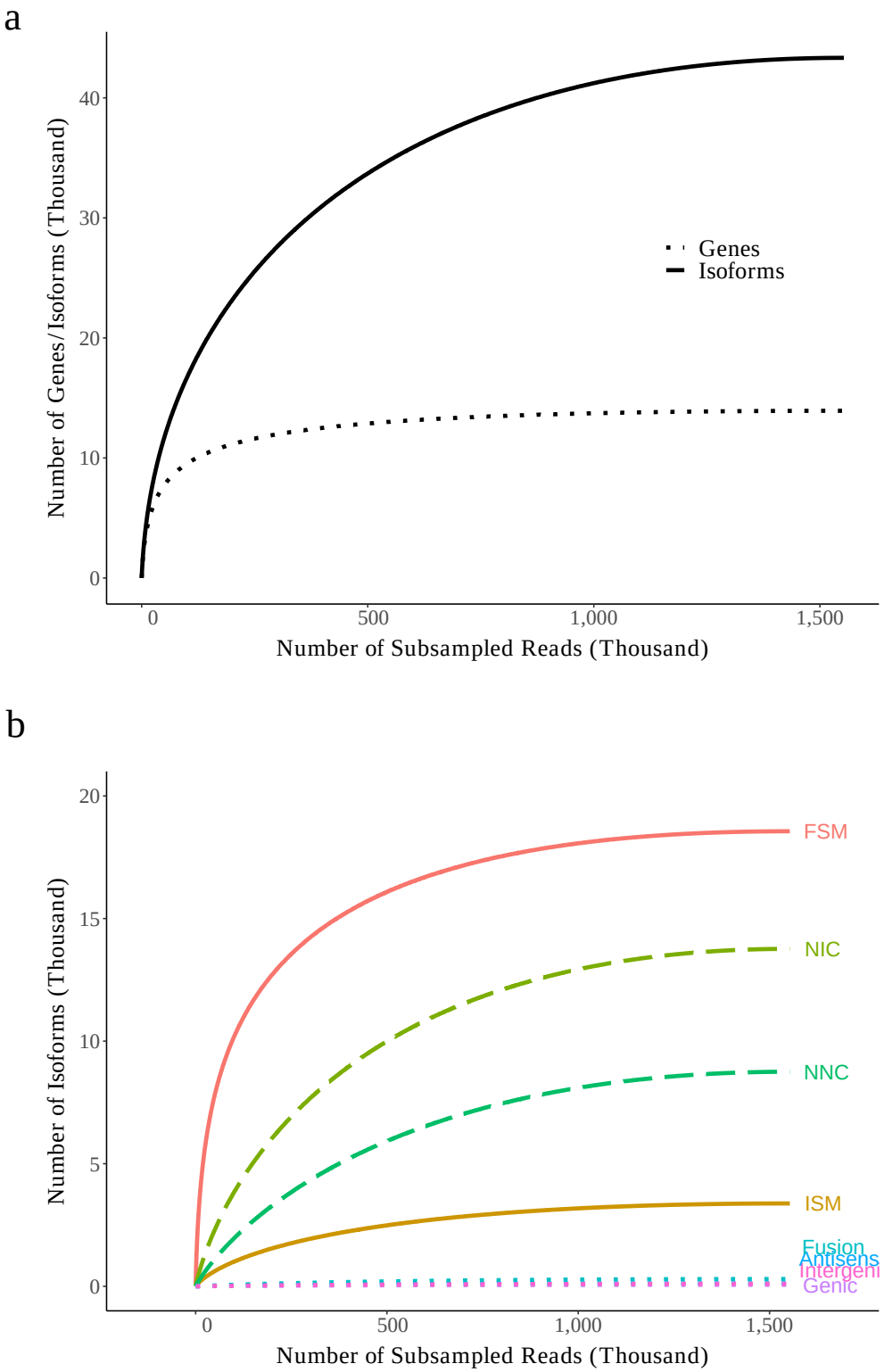


Figure 3.13: Rarefaction curve of Iso-Seq merged dataset indicated saturation and good coverage of genes and isoforms:

3.3.4 Isoform diversity

Compared with the mouse reference genome, there was a wider range in the number of isoforms identified per gene (1 – 86), with each gene associated with a median of 2 isoforms. Only 10% (n = 4,641) of isoforms were detected across all the samples (**Figure 3.14a**), with about half (47.8%) detected in 2 - 3 samples with very low transcript expression (**Figure 3.14b**).

Gene ontology (GO) analysis showed that the most enriched molecular function amongst the 100 most transcriptionally diverse genes in mouse cortex was 'tubulin binding' (odds ratio = 7.90, adjusted P = 6.70 x 10^{-4}), driven by the overexpression of MAPT in TG mice.

A significant proportion of isoforms (20,621, 45%) were sized 2 - 4kb in length (median length = 2.96kb, mean length = 3.18kb, s.d = 1.68kb, range = 0.083 - 15.9kb) (**Figure 3.15a**), corresponding to the mean length of mRNA mouse reference genome, with a wide range in the number of exons (1 - 89) observed per isoform (mean number of exons = 10.8). The number of isoforms per gene was correlated with gene length (corr = 0.25, P = 1.33 x 10^{-197}, **Figure 3.15**c), and exon number (corr = 0.24, P = 7.97 x 10^{-155}, **Figure 3.15**d).

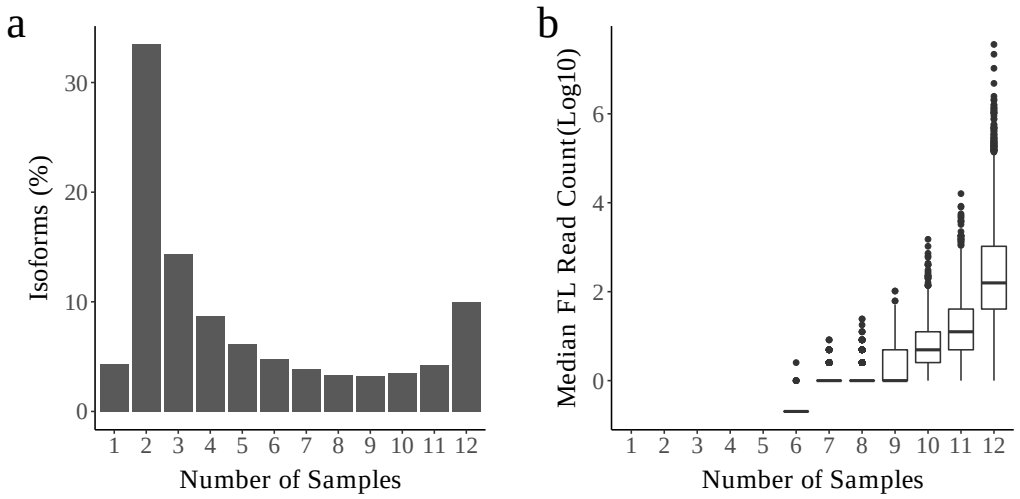


Figure 3.14: Highly-expressed isoforms are more likely to be sequenced across sampples and accurately quantified: Shown is **a)** the distribution of isoforms detected in the number of mouse samples, with a third detected in any two of the total 12 samples. However, **b)** quantification of these isoforms had very low expression (1-2 FL read), whereas those that were commonly detected across all 12 samples were very highly expressed. FL - Full Length

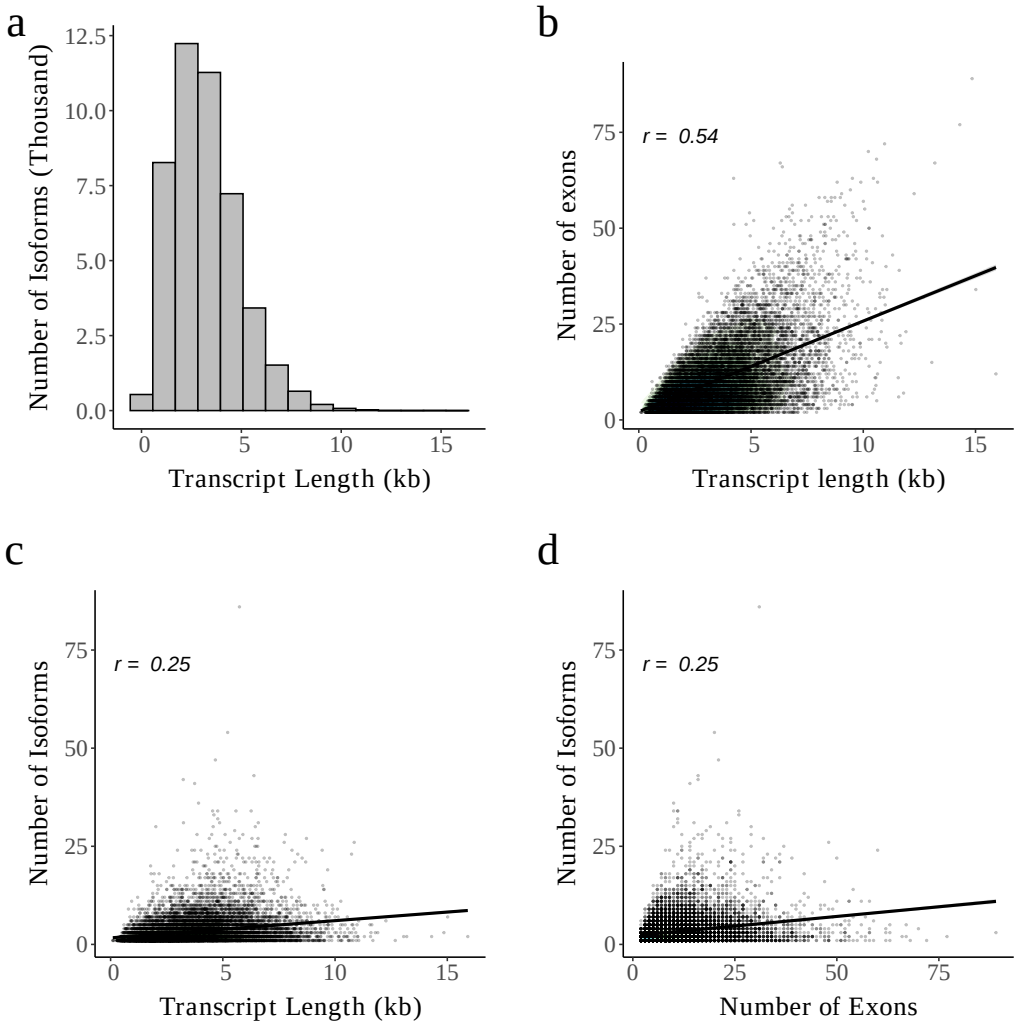


Figure 3.15: Longer genes with more exons were associated with more isoforms: **a** The majority of isoforms have a length between 1 - 5kb. **b)** The number of exons was correlated with the transcript length, and the **c)** the number of isoforms was correlated with the length and **d)** and the number of exons per gene. Gene length and exon number is represented by the longest transcript. kb - kilobases

3.3.5 Iso-Seq vs RNA-Seq

To compare the power of Iso-Seq versus RNA-Seq to detect full-length transcripts, a reference-guided transcriptome assembly using only Illumina's RNA-Seq reads of the same samples was generated with Stringtie. Using SQANTI to characterise isoforms similarly to the Iso-Seq analysis, RNA-Seq defined transcriptome revealed significantly more isoforms (156,253 isoforms vs 46,626 isoforms from Iso-Seq defined transcriptome, **Figure 3.16a**). However, upon further examination and comparison using gffcompare, majority of these isoforms were found to be incomplete fragments of isoforms identified in Iso-Seq, with significantly shorter isoform length (2.31kb vs 3.18kb of mean length of RNA-Seq and Iso-Seq defined isoforms respectively, two-tailed unpaired t-test, $t(203070) = 71.9$, $P < 2.2 \times 10^{-16}$, **Figure 3.16c**), fewer exons (7.30 vs 10.8 of mean number of exons of RNA-Seq and Iso-Seq defined isoforms respectively, two-tailed unpaired t-test, $t(203070) = 76.7$, $P < 2.2 \times 10^{-16}$, **Figure 3.16d**) and less supported by CAGE peaks (34.0% vs 71.9% of RNA-Seq and Iso-Seq defined isoforms within 50bp CAGE peak respectively, Fisher's Test: $P < 2.2 \times 10^{-16}$, odds ratio = 4.97, **Figure 3.16e**). Considering only isoforms that had a complete exact match as defined by gffcompare, more than 50% of isoforms detected from Iso-Seq dataset could not be readily recapitulated (**Figure 3.16a**), the majority of which were novel isoforms and genes (**Figure 3.16**f).

The isoform expression (TPM) was then compared using the following methods:

1. Iso-Seq data alone using FL read count
2. RNA-Seq data aligned to Iso-Seq defined transcriptome using Kallisto[46]
3. RNA-Seq data aligned to RNA-Seq defined transcriptome using Kallisto.[46] RNA-Seq transcriptome was generated using Stringtie.

Focusing only on the subset of isoforms that were commonly identified in both RNA-Seq and Iso-Seq defined transcriptomes (n = 23,761), the isoform expression using RNA-Seq data mapped to Iso-Seq (method 2) and RNA-Seq transcriptome (method 3) was highly correlated (Pearson's correlation = 0.77, $P < 2.2 \times 10^{-16}$). Conversely, the isoform expression derived from Iso-Seq data alone (method 1) was weekly correlated to the RNA-Seq derived isoform expression (method 3, Pearson's correlation = 0.45, $P < 2.2 \times 10^{-16}$). This highlights the power of Iso-Seq defined transcriptome to accurately identify and annotate isoforms as a scaffold, but is limited for quantitative analysis due to sequencing depth.

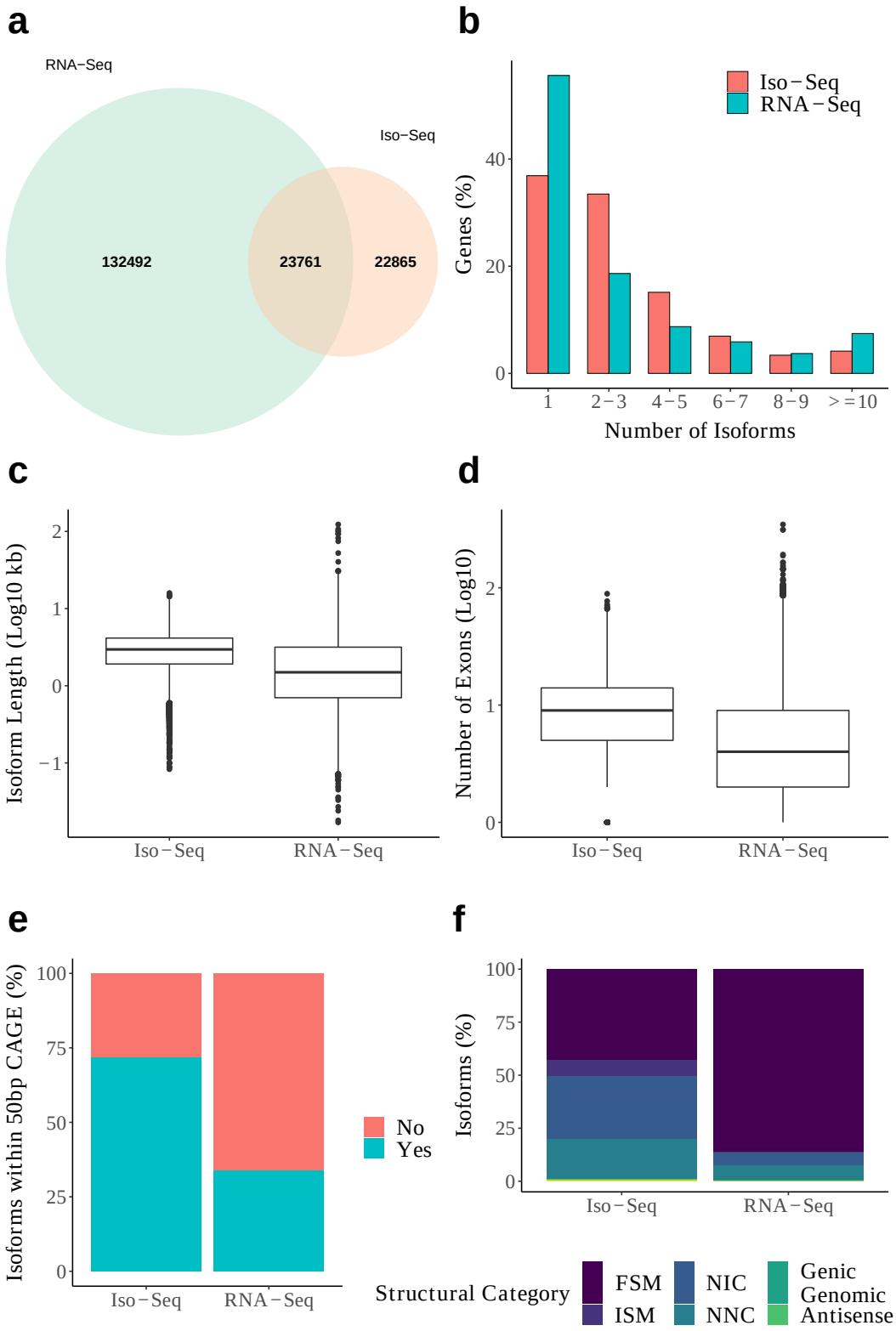


Figure 3.16: Iso-Seq identified more isoforms per gene, that were longer with more exons, and with a greater proportion of isoforms with CAGE peak: A reference-guided transcriptome using only RNA-Seq data (RNA-Seq defined transcriptome) was generated. **a)** RNA-Seq defined transcriptome identified more isoforms, as expected given the significantly higher sequencing depth. However, **b)** the isoform diversity was smaller than that from Iso-Seq defined transcriptome with the majority of genes associated with only one isoform. **c)** Isoforms identified from the RNA-Seq defined transcriptome were also more likely to be shorter and **d** contain fewer exons. **e)** Highlighting the power of Iso-Seq to identify true full-length isoforms in comparison to RNA-Seq, a significantly larger proportion of isoforms from Iso-Seq data were found within 50bp of a CAGE peak. **f** Approximately half of the isoforms identified using Iso-Seq were novel (NIC, NNC), which were not recapitulated using RNA-Seq. 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	
Annotated Genes	14482 (98.62%)	
Annotated 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

Table 3.4: Classification of annotated and novel genes and isoforms were based from SQANTI2, and from the merging of 12 samples. FSM - Full Splice Match, ISM - Incomplete Splice Match, NIC - Novel In Catalogue, NNC - Novel Not in Catalogue

3.3.6 Novel isoforms

Interestingly, the transcriptome was made up of 50% of isoforms that were known (23,350) and 50% that were novel (23,096) and were not present in existing annotation databases (**Table 3.4**). Benchmarking the accuracy and reliability of novel isoforms against known isoforms, no difference in the number supported within 50bp CAGE was observed (novel isoforms within CAGE: 17,252, 75.4%; known isoforms with CAGE: 17,842, 75.8%, Fisher's Test: P = 0.31, odds ratio = 0.978). Less RNA-Seq support was observed for novel isoforms compared to known isoforms (mean RNA-Seq expression for known isoforms = 8.95TPM, mean RNA-Seq expression for novel isoforms = 1.99TPM; two-tailed unpaired t-test: $t(46401) = 14.8$, $P = 1.37 \times 10^{-49}$); however, this is likely to reflect RNA-Seq's lack of power to detect novel isoforms rather than the validity of these isoforms.

Compared to known isoforms, these novel isoforms were less abundant (Mann-Whitney-Wilcoxon test, $W = 3.66 \times 108$, $P < 2.23 \times 10^{-308}$ **Figure 3.17a,b**) and longer (Mann-Whitney-Wilcoxon test, $W = 2.37 \times 10^{8}$, $P = 2.13 \times 10^{-42}$, **Figure 3.17c,d**) with more exons (Mann-Whitney-Wilcoxon test, $W = 1.94 \times 10^{8}$, $P < 2.23 \times 10^{-308}$, **Figure 3.17e,f**), suggesting that they would have been harder to detect using traditional short-read 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 (1,454 novel isoforms vs 1,154 annotated isoforms at least 1kb away from known TSS, Fisher's Test: P = 6.16 x 10^{-12}, odds ratio = 1.32) and termination sites (21,506 novel isoforms vs 21,434 annotated isoforms less than 1kb away from known TTS) than known isoforms.

The different types of splicing events were also compared between known and novel isoforms (see Section X). In total, 40,249 alternative splicing events were identified in annotated genes with AF (alternative TSS variation) and SE being the most prevalent events (AF: 12,853, 31.9%; SE: 8,686, 21.6%, **Figure 3.18**). It is important to note, however, that only around 30% of 5'end isoforms were located near (<5bp) any annotated 5' end whereas 70% of 3' ends were located near (<5bp) annotated 3'ends - this discrepancy is likely due to a combination of mRNA degradation, template switching artifacts during reverse transcription and true novel alternative TSS.

Except for AF and AL, all the other different splicing events, and in particularly intron retention, were more likely to be observed in novel isoforms than in known isoforms, implicating the power of Iso-Seq to detect full-length transcripts and the ability to recapitulate the usage of complex splicing events that would have otherwise been underestimated with only RNA-Seq data alone (Fisher's one-tailed 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; SE: P < 2.23 x 10^{-16}, odds ratio = 1.57, **Figure 3.18**).

3.3.7 Intron Retention and Nonsense mediated decay

For the majority of genes characterised by splicing, only one or two splicing events were observed (n = 10,708, 81.8% of AS genes, **Table 3.5**), suggesting that such events were often mutually independent. However, interestingly, Nonsense-mediated mRNA decay (NMD) - a mechanism that acts to reduce transcriptional errors by degrading transcripts containing premature stop codon - was found to be particularly enriched amongst isoforms characterised with intron retention (IR-isoforms). Of the 6,803 isoforms characterised with intron retention, 38.7% (n = 1,930) were also predicted to undergo NMD (NMD-isoforms), as characterised by the presence of an ORF and a coding sequence (CDS) end motif before the last junction. Novel isoforms, more likely to be characterised with intron retention, were also more likely

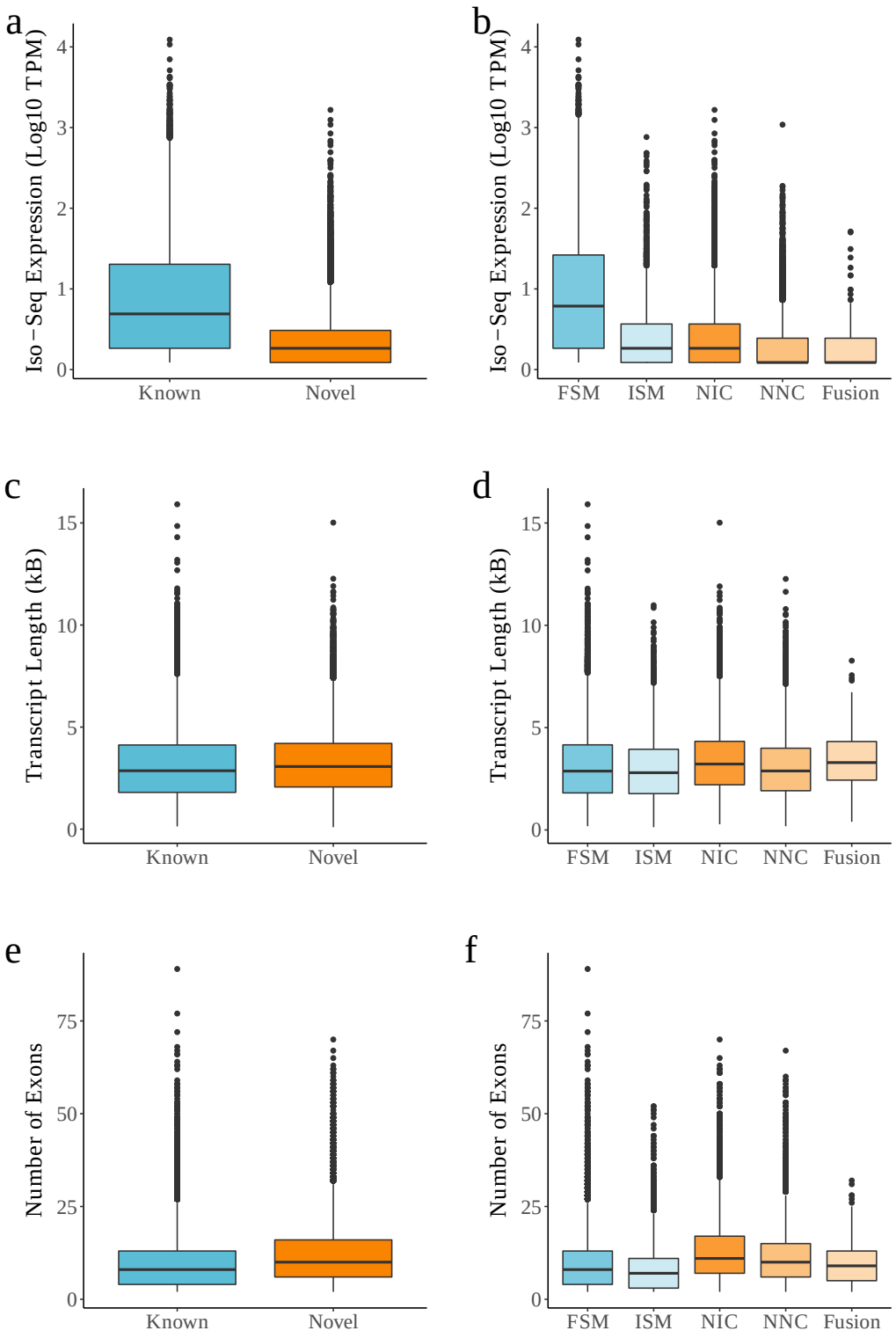


Figure 3.17: Novel isoforms were less expressed, longer and had more exons than known isoforms: Shown is the **a)** Iso-Seq transcript expression, the **c)** transcript length, and the **e)** the number of exons of novel and known isoforms. The known and novel isoforms can be further subdivided and classified, with the **b)** Iso-Seq expression **d)** transcript length and **f)** number of exons for each category. According to SQANTI, known isoforms are subdivided into FSM and ISM, and novel isoforms are subdivided into NIC, NNC, and fusion. FSM – Full Splice Match, ISM – Incomplete Splice Match, NIC – Novel In Catalogue, NNC – Novel Not in Catalogue.

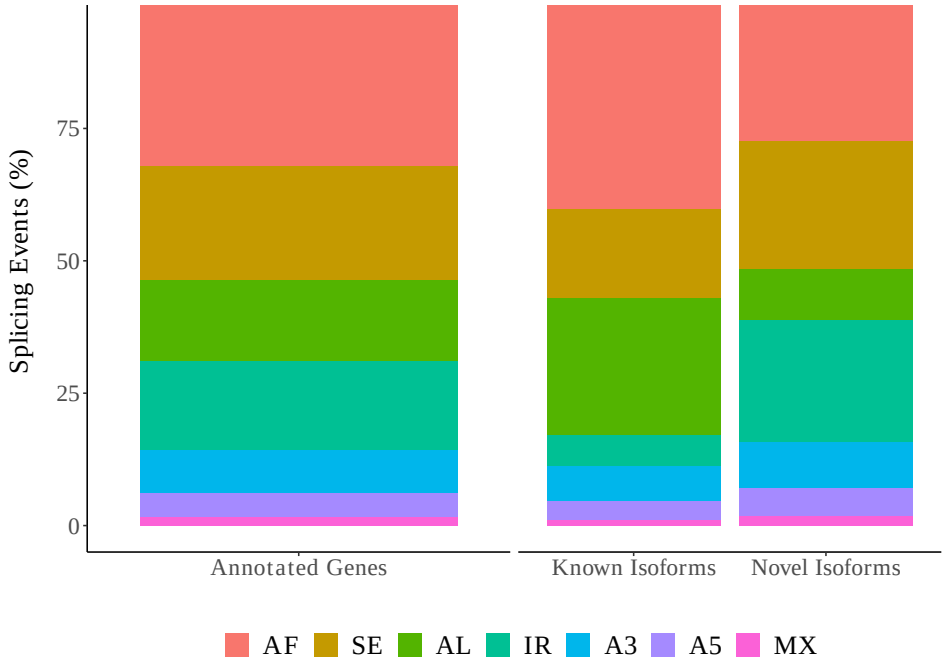


Figure 3.18: Alternative first is the most prevalent AS event, and novel isoforms are more likely to be characterised with complex AS events: Shown is the proportion of AS events in annotated genes, and further subdivided by known and novel isoforms. Novel isoforms were more likely to be characterised by all AS events, with the exception of AF and AL. MX and SE events were determined using SUPPA2, IR with SQANTI2 and A3', A5', AF and AL with custom scripts. AF – Alternative First Exon, AL – Alternative Last Exon, A5' – Alternative 5' prime, A3' – Alternative 3' prime, IR – Intron Retention, MX – Mutually Exclusive, SE – Skipped Exon

to be associated with NMD than known isoforms (Fisher's Test: $P < 2.23 \times 10^{-16}$, odds ratio = 4.16).

These isoforms with both IR and NMD were found to more lowly expressed than isoform only with NMD and no IR ($W = 7.50 \times 10^{6}$, $P = 1.67 \times 10^{-42}$, **Figure 3.19b**), those of which were also more lowly expressed than isoforms with no NMD. Furthermore, only a small number of genes were associated with isoforms where IR and NMD were mutually exclusive (n = 277, 1.91% of total genes, **Figure 3.19a**), providing additional support for the hypothesized relationship between these two transcriptional control mechanisms.

Number of Splicing Events	Frequency
1	7315 (55.89%)
2	3393 (25.92%)
3	1724 (13.17%)
4	548 (4.19%)
5	108 (0.83%)

Table 3.5: Shown is the number of splicing events observed in genes that are alternatively spliced. Majority of genes are detected with only one or two splicing events.

3.3.8 Fusion 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 cancer32. Although fusion events are thought to be rare, we found that 0.4% of transcripts included exons from two or more adjacent genes (mouse cortex: n = 297 fusion transcripts associated with 218 genes (1.51%)).

3.3.9 LncRNA

Although the majority of isoforms (93.6%, 43,450) mapping to known genes were classified as protein-coding by the presence of an ORF, a relatively large number of isoforms (n = 1,141) were mapped to genes annotated as encoding lncRNA (n = 734 genes). Compared to isoforms not defined as lncRNA (non-lncRNA) by reference genome, these lncRNA isoforms were found to be longer (Mann-Whitney-Wilcoxon test, $W = 3.52 \times 10^{7}$, $P = 8.24 \times 10^{-98}$, **Figure 3.20a**), despite containing fewer exons ($W = 4.56 \times 10^{7}$, $P < 2.23 \times 10^{-308}$, **Figure 3.20b**) and being enriched for mono-exonic molecules(23.9% vs 2.02%) - corroborating previous findings from

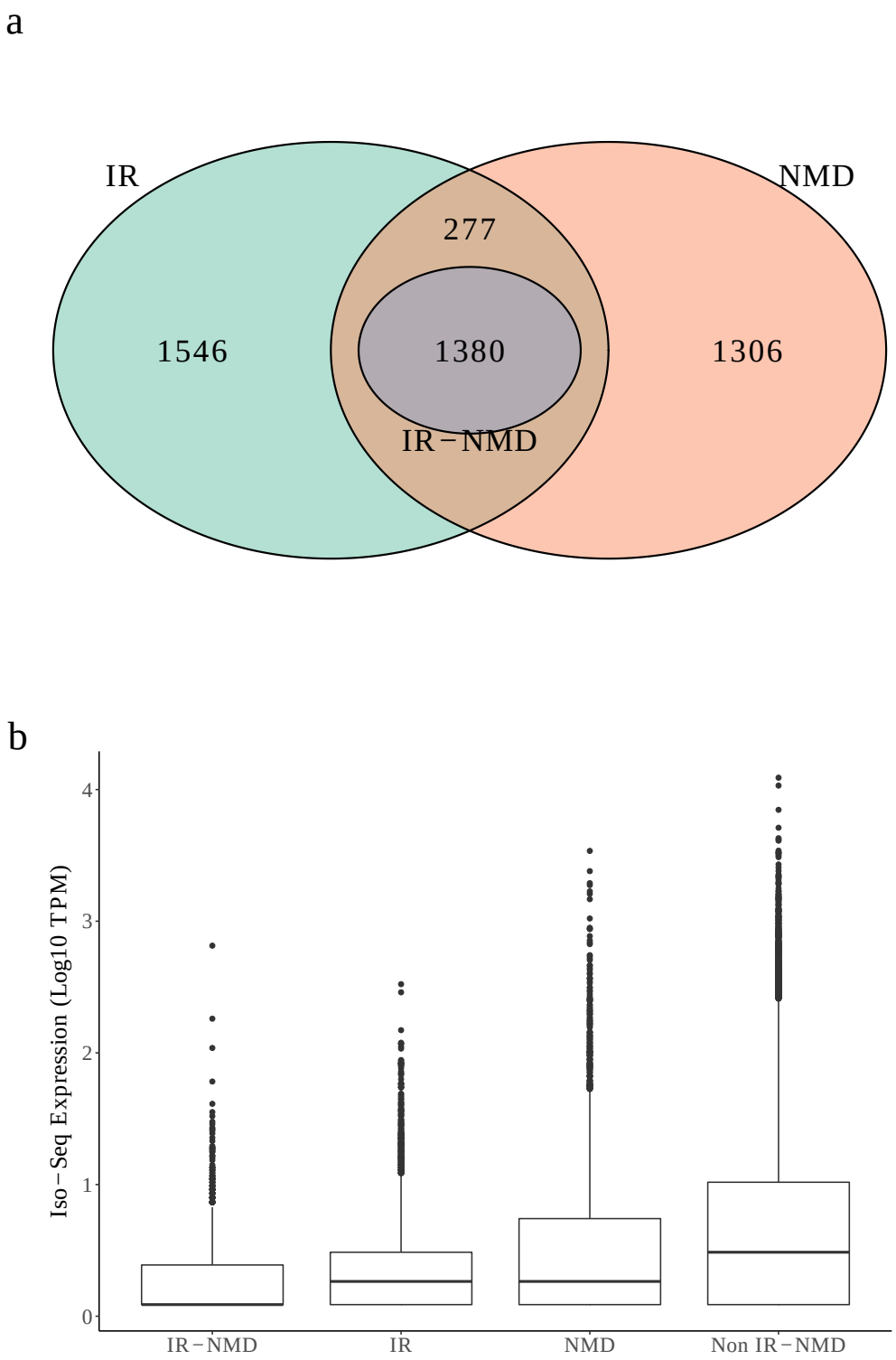


Figure 3.19: Intron retention is associated with nonsense-mediated mRNA decay (NMD) and reduced expression: Shown is the overlap of genes associated with isoforms characterised with intron retention (IR), nonsense-mediated mRNA decay (NMD), and transcripts with both IR and NMD (IR-NMD). Of note, genes with isoforms characterised by both IR and NMD were further classified into genes that contain isoforms where both events are observed together (purple) and where they are mutually exclusive (dark orange). As such, 13800 genes were associated with IR-isoforms that were predicted for NMD, and 168 genes that contained IR-isoforms and NMD-isoforms. Isoforms that were characterised with both IR and NMD were particularly lowly expressed compared to isoforms with either IR, NMD or neither events. IR – Intron Retention, NMD – Nonsense-mediated mRNA decay.

other long-read studies(,[7125]). These lncRNA isoforms were found to be more lowly expressed than non-lncRNA isoforms ($W = 3.16 \times 10^7$, $P = 5.67 \times 10^{-40}$), with fewer RNA isoforms identified per lncRNA gene (mean n = 1.55, range = 1 - 34 vs mean n = 3.29, range = 1 - 86; $W = 7.40 \times 10^6$, $P = 5.76 \times 10^{-107}$, **Figure 3.20e**).

Importantly, over a third (448, 39.3%) of these annotated lncRNA isoforms contained a putative ORF, supporting recent observations that lncRNA have potential protein coding capacity, with shorter ORFs than non-lncRNA isoforms (mean length = 139bp, s.d = 127bp vs mean length = 519bp, s.d = 393bp; $W = 1.75 \times 10^7$, $P = 8.33 \times 10^{-195}$).

3.3.10 Novel Genes

Although the vast majority of isoforms were annotated to known genes, 0.5% (n = 223 isoforms) did not and potentially represent "novel" genes (n = 189 genes). These novel genes were all multi-exonic (mean length = 1.75kb, s.d = 1.21kb, range = 0.098 - 6.86kb, mean number of exons = 2.5) and were identified uniformly across the genome/chromosome, with over half the identified transcripts from these genes predicted to be non-coding (n = 143 (64.1%) novel-gene transcripts), shorter and more lowly expressed than annotated genes (length: $W = 7.79 \times 10^6$, $P = 5.22 \times 10^{-45}$; expression: $W = 2.29 \times 10^6$, $P = 1.5 \times 10^{-73}$).

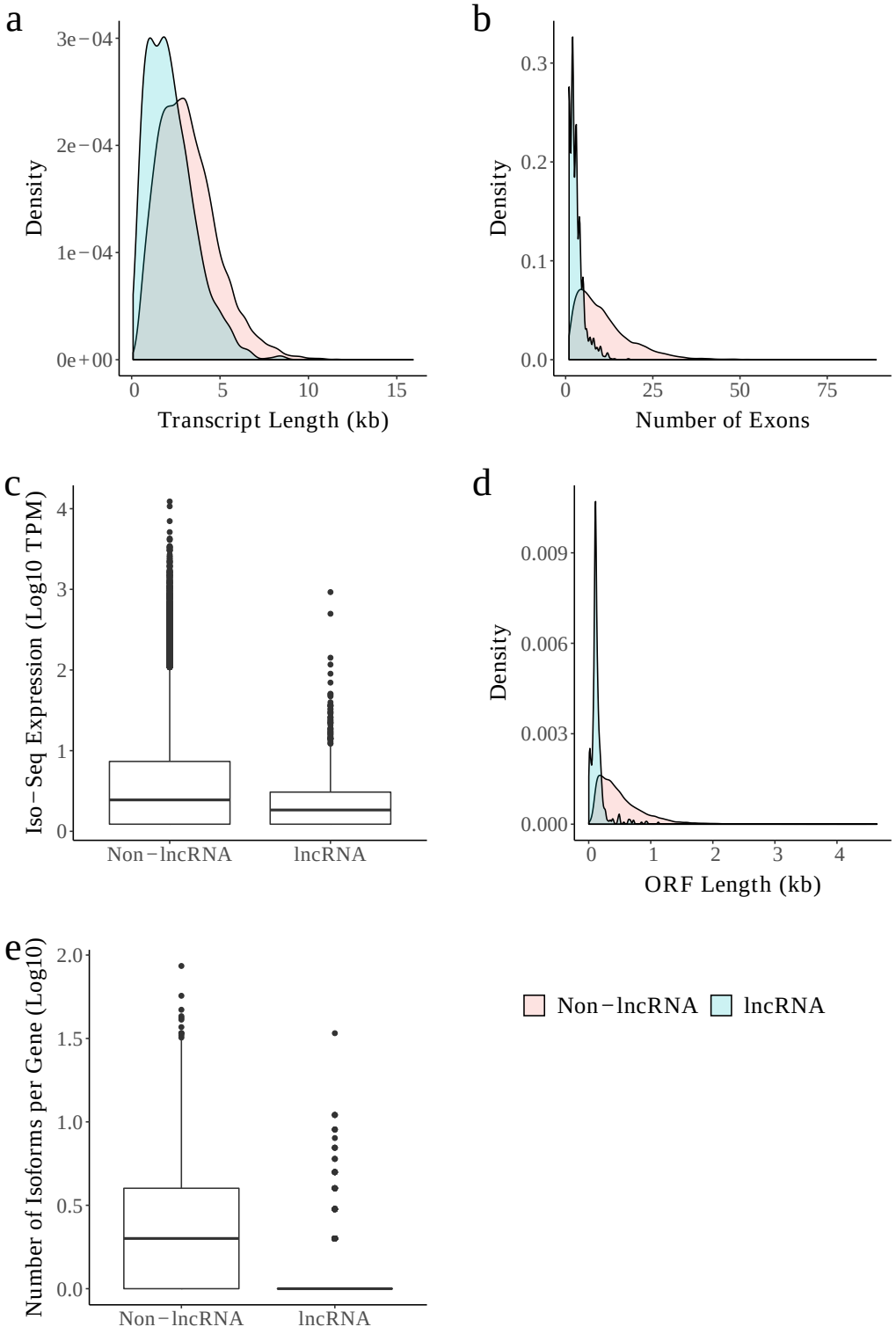


Figure 3.20: LncRNA isoforms were more lowly expressed and typically longer than non-lncRNA transcripts, despite containing fewer exons: Shown is the distribution of the **a)** transcript length, **b)** number of exons, **c)** transcript expression, **d)** ORF length and the **e)** diversity of isoforms annotated to lncRNA and non-lncRNA.lncRNA – long non-coding RNA

3.4 Discussion

"The apparent length limitation to 6kb is most likely a combined result of ineffective size selection and the limitation ofthe sequencing chemistry (P4-C2, Methods) used in this study"; what are the proportion of transcripts relative to genome in size? The length of clustered transcripts closely reflect size distribution of the input full-length reads. "Final transcripts include a large number of isoforms greater than 3 kb that are not accessible by simply using CCS reads."

Although skipped exons are known to be the most common AS events in mouse, our data conversely suggests that splice variants from a single gene are predominantly generated through alternative first exons

Chapter 4

Targeted Transcriptome

4.1 Introduction

One current limitation of whole transcriptome sequencing is the low coverage/sequencing depth achieved per gene due to the distribution of reads across the whole transcriptome. Consequently, while whole transcriptome sequencing allows identification of novel genes (genes not previously annotated to the genome) and novel isoforms, it may not detect isoforms particularly those of low expression resulting in many false negatives. This can be circumvented by the use of target capture, which enriches a selective panel of genes that are then only sequenced. Multiple samples can further be pooled and sequenced together by barcoding samples at cDNA synthesis, which simplifies laboratory workflow and minimises associated sequencing costs.

4.2 Methods

The extracted RNA from mouse rTg4510 samples were prepared for targeted transcriptome sequencing on the PacBio's Sequel (n = 24, Table 4.1), a subset of which were also sequenced on the Oxford Nanopore's MinION (n = 18, Table 4.1). Three biological replicates were selected at each age (2, 4, 6 and 8 months) across wildtype and transgenic mice, multiplexed using barcodes (listed in Table 2.1) and ran on the Sequel as three batches. Iso-Seq library preparation and SMRT sequencing is described in Chapter X. Following the Iso-Seq lab pipeline (Chapter

2.1.2), 200ng RNA from each sample was used for first strand cDNA synthesis (Chapter 2.1.2.1) and amplified using PCR with 14 cycles (Figure 3.1, Chapter 2.1.2.4). Purification with 0.4X and 1X AMPure PB beads selectively and successfully enriched cDNA with different molecular weights (Figure 3.2). The two fractions were then recombined at equimolar quantities and library preparation was successfully performed (Figure 3.2). Sequencing was performed for each sample on the PacBio Sequel using a 1M SMRT cell. Processing of raw reads were performed using the Iso-Seq bioinformatics pipeline outlined in Chapter X. RNA from the same samples (n = 24) was also prepared with TruSeq Stranded mRNA Sample Prep Kit (Illumina) and subjected to 125bp paired-end sequencing using a HiSeq2500 (Illumina), and used as junction support of the long reads.

		Sample demographics				Sequencing Platform			
						PacBio Iso-Seq		Oxford Nanopore	
Sample	Phenotype	Age (Months)	RIN	Concentration (ng/ul)	Batch (Barcodes)	Whole Transcriptome	Targeted Transcriptome	Whole Transcriptome	Targeted Transcriptome
K19	WT	4	8.8	236	1 (PB_BC_1)		X		
K23	WT	8	9.1	143	1 (PB_BC_2)	X	X		
K21	WT	6	9	138	1 (PB_BC_3)		X		
K18	TG	2	8.8	136	1 (PB_BC_4)	X	X	X	
K20	TG	4	9.1	80.4	1 (PB_BC_5)		X		
K17	WT	2	9.2	77.1	1 (PB_BC_6)	X	X		
S19	WT	4	9.1	84.9	2 (PB_BC_1)		X		X
K24	TG	8	9.2	65.4	2 (PB_BC_2)	X	X		X
L22	TG	8	8.7	68.6	2 (PB_BC_3)	X	X		X
M21	WT	2	9.2	72.3	2 (PB_BC_4)	X	X	X	X
O18	TG	2	8.9	115	2 (PB_BC_5)	X	X		X
O23	WT	8	9	91.8	2 (PB_BC_6)	X	X		X
O22	TG	6	9.1	83.5	2 (PB_BC_7)		X		X
P19	WT	6	8.9	92.2	2 (PB_BC_8)		X		X
T20	TG	6	9	68.7	2 (PB_BC_9)		X		X
Q20	TG	8	8.6	99.7	3 (PB_BC_1)	X	X		X
Q21	WT	2	9.2	83.3	3 (PB_BC_2)	X	X		X
S18	TG	2	8.9	115	3 (PB_BC_3)	X	X		X
S23	WT	8	9.1	95.5	3 (PB_BC_4)	X	X		X
Q18	TG	6	8.8	87.2	3 (PB_BC_5)		X		X
Q17	WT	6	8.7	85.8	3 (PB_BC_6)		X		X
L18	TG	4	8.8	145	3 (PB_BC_7)		X		X
Q23	WT	4	9	70.8	3 (PB_BC_8)		X		X
T18	TG	4	9	85	3 (PB_BC_9)		X		X

Table 4.1: Mouse rTg4510 samples sequenced using whole and targeted transcriptome approach with PacBio Iso-Seq and ONT nanopore sequencing

Target	Number of Probes	Genome Co-ordinates	Strand	Full Region (bp)	Exons inc UTR (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				

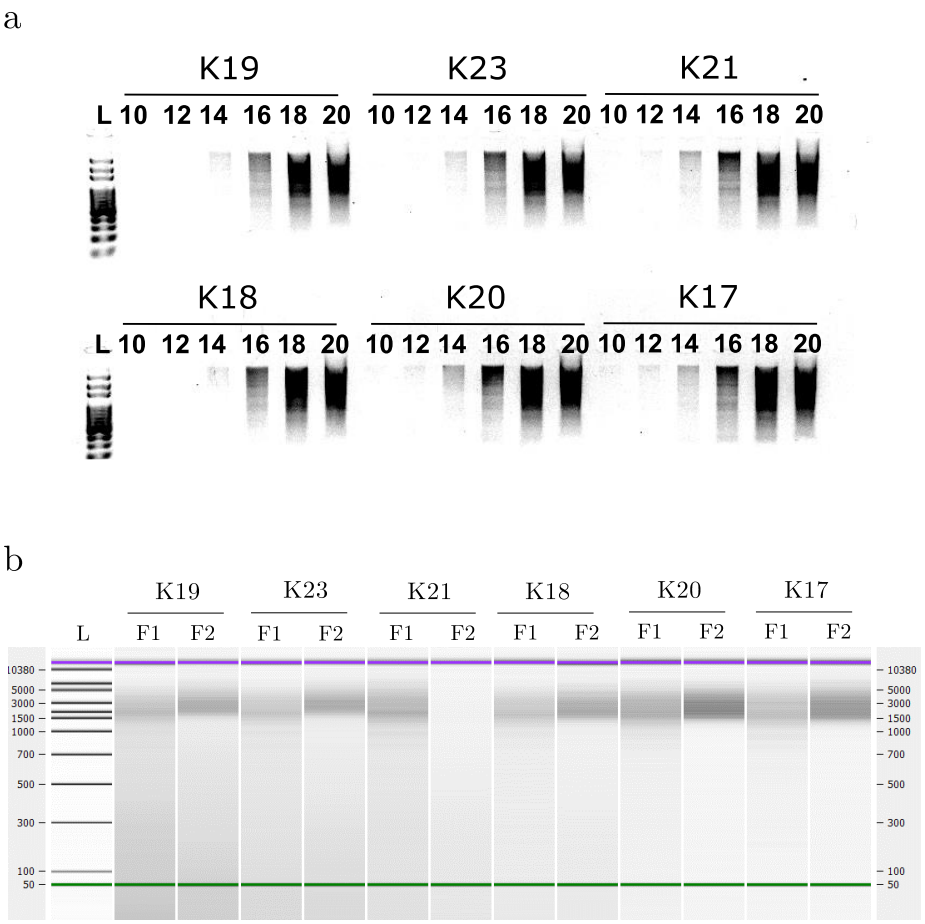


Figure 4.1: The first stage between the targeted and whole transcriptome sequencing is the same with samples typically amplified using 14 cycles followed by enrichment of high molecular weight cDNA in Fraction 2: a) Like whole transcriptome sequencing, samples were amplified using 14 cycles (Figure 3.1) whereby cycles below generated insufficent cDNA and cycles above showed signs of over-amplification. The samples shown here (K19, K23, K21, K18, K20, K17) were multiplexed and sequenced in Batch 1 (see Table 4.1. Ladder (L) shown is 100bp DNA ladder. **b)** Similar to whole transcriptome sequencing, ampilfied cDNA was further divided into two fractions (denoted here as F1 and F2) and purified with 1X (F1) and 0.4X (F2) Ampure beads. As shown in the bioanalyzer gel, there was an enrichment of higher-molecular weight cDNA in Fraction 2 compared to Fraction 1 across all the samples (with the exception of Sample K21 with loss of Fraction 2). Green and purple line represent the lower marker at 50bp and the upper marker at 17kb respectively. F1 - Fraction 1 containing cDNA purified with 1X Ampure beads; F2 - Fraction 2 containing cDNA purified with 0.4X Ampure beads.

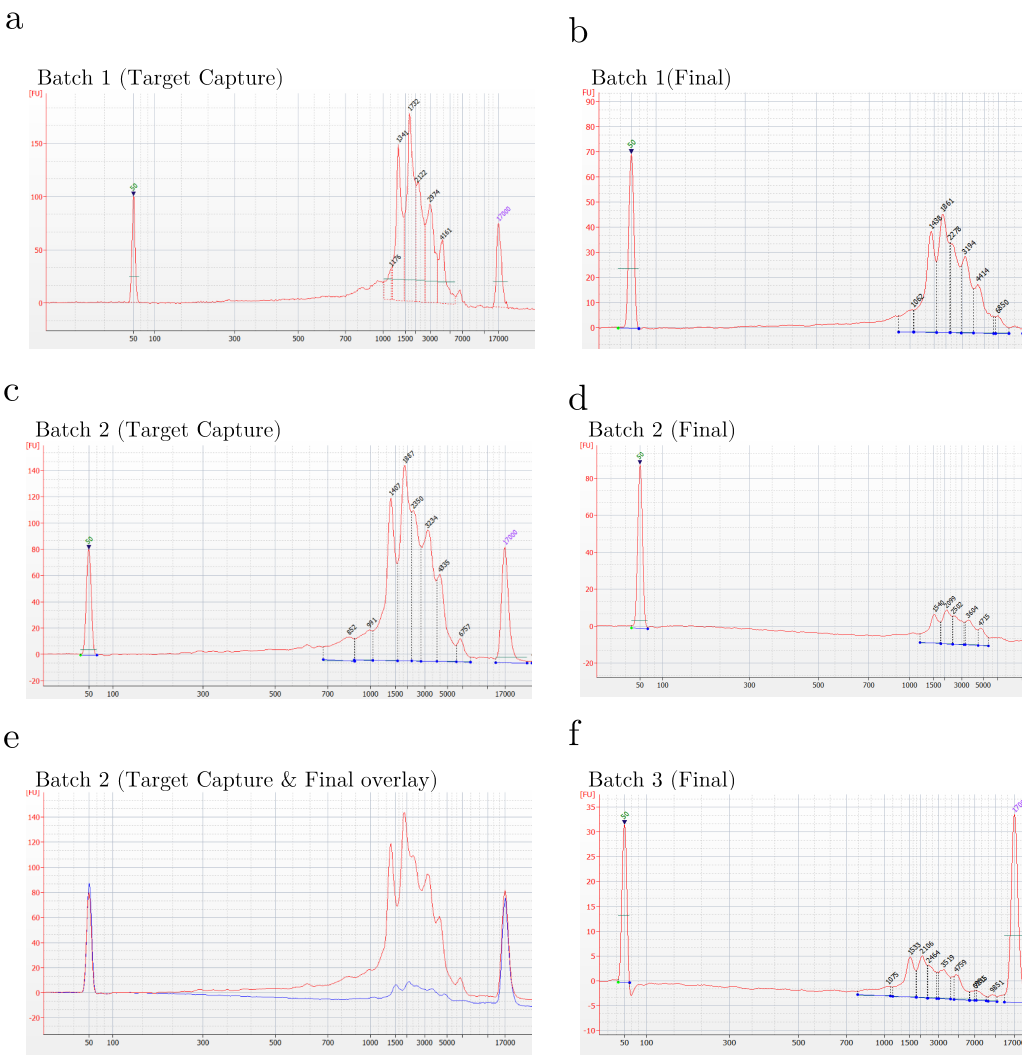


Figure 4.2: Successful target capture and library preparation across all batches, as shown by enrichment of transcripts with specific lengths: a) and **c)** are bioanalyzer electropherogram traces of Batch 1 (n = 6) and Batch 2 (n = 9) respectively after enrichment of cDNA with selective IDT probes (Section 2.1.2.9). **b)**, **d)** and **f)** are bioanalyzer electropherogram traces of Batch 1, 2 and 3 respectively after library preparation (denoted here as "Final", Section 2.1.2.10. **e)** An overlay of Batch 2 after target capture and library preparation.

As can be seen across all figures, target capture appears to be successful with detected peaks, reflecting enrichment of target transcripts with specific lengths, which differs from the broad peaks that are evident in whole transcriptome sequencing (Figure 3.2). Library preparation with ligation of SMRT bell templates retained these targeted transcripts with good peak overlay, as seen in figure e). The difference in peak height (i.e. cDNA quantity) between target capture and library preparation is due to a difference in input cDNA concentration when running Bioanalyzer - input cDNA after library preparation was diluted with a 1:5 dilution factor to maximise amount of cDNA available for sequencing, whereas input cDNA after target capture was not diluted.

4.3 Results

4.3.1 Run performance and sequencing metrics

Following library preparation and SMRT sequencing, a total of XXGb (s.d = XXGb) were obtained (Table 4.2). Of note, 6 samples were first trialled and multiplexed in Batch 1 to determine the yield output and coverage depth - PacBio recommends starting with 4 - 8 samples for multiplexing. Having noticed that an average yield output (24Gb) with a high off-target sequencing, implicating saturation of target genes with 6 samples, the number was increased to 9 samples in Batch 2 and Batch 3. Despite more samples, the sequencing run for Batch 2 and 3, performed by Exeter's Seqeuncing Service, had a poor loading rate (38.1% P1 of Batch 3 vs 71% of Batch 1) and low subsequent yield. The samples were also potentially degraded after having been stored in -20°C for over 6 months due to Covid-19 lockdown.

The yield difference between the first and last two batches was evident in the number of CCS reads (total = 996K; Batch 1 = 469K, Batch 2 = 306K, Batch 3 = 2221K Figure 4.3a) and FLNC reads (total = 930K; Batch 1 = 399K, Batch 2 = 275K, Batch 3 = 256K, Figure 4.3a) generated, after applying the bioinformatics Iso-Seq pipeline (same as the whole transcriptome approach with the exception of removing barcodes rather than general primers). However, calculation of the on-target rate suggested that while Batch 2 and 3 had lower output yield, the coverage of target genes was significantly greater than Batch 1 due to the increased sample size (mean rate in Batch 1 = 34.5%; mean rate in Batch 2 = 46.2%; mean rate in Batch 3: 42.9%, Figure 4.4). The on-target rate is defined as the proportion of mapped transcripts with sequences overlapping at least one target probe.

In addition to batch variability, the number of full-length transcripts obtained per sample varied within each batch (Figure 4.3b). This variability was not associated with RIN (corr = 0.147, P = 0.492, Spearman's rank) and is unlikely to be due to library preparation, given that samples were pooled in equal molarity during target capture. However, there was no significant difference in the number of full-length transcripts between WT and TG across the batched runs (Wilcoxon rank sum test, W = 73, P = 0.977, Figure 4.3c).

Sample	Total Bases (GB)	Polymerase Reads	Read Length						Productivity			Control				Local Base Rate	Template		Notes
			Polymerase		Subread		Insert		P0	P1	P2	Total Reads	Read Length Mean	Concordance			Adapter Dimer (0-10bp)	Short Insert (11-100bp)	
			Mean	N50	Mean	N50	Mean	N50						Mean	Mode				
Batch 1	24.2	712250	34016	70473	1402	1852	3024	3808	4.62% (46613)	71.58% (722026)	24.76% (249707)	9,690	31,505	0.84	0.87	2.31	0	0	Sequenced in November 2019
Batch 2																			Sequenced in July 2020
Batch 3	19.3	383292	50472	100255	1557	2017	3158	3898	18.68% (189549)	38.11% (386743)	43.56% (442054)	3,440	52,533	0.85	0.87	2.86	0	0	Samples were kept at -20 for over 9months. Sequel broke down mid-run. Sequencing was prepared by Exeter's Sequencing Services

Table 4.2: Iso-Seq run yield for each batch of Tg4510 mouse samples sequenced using targeted transcriptome approach

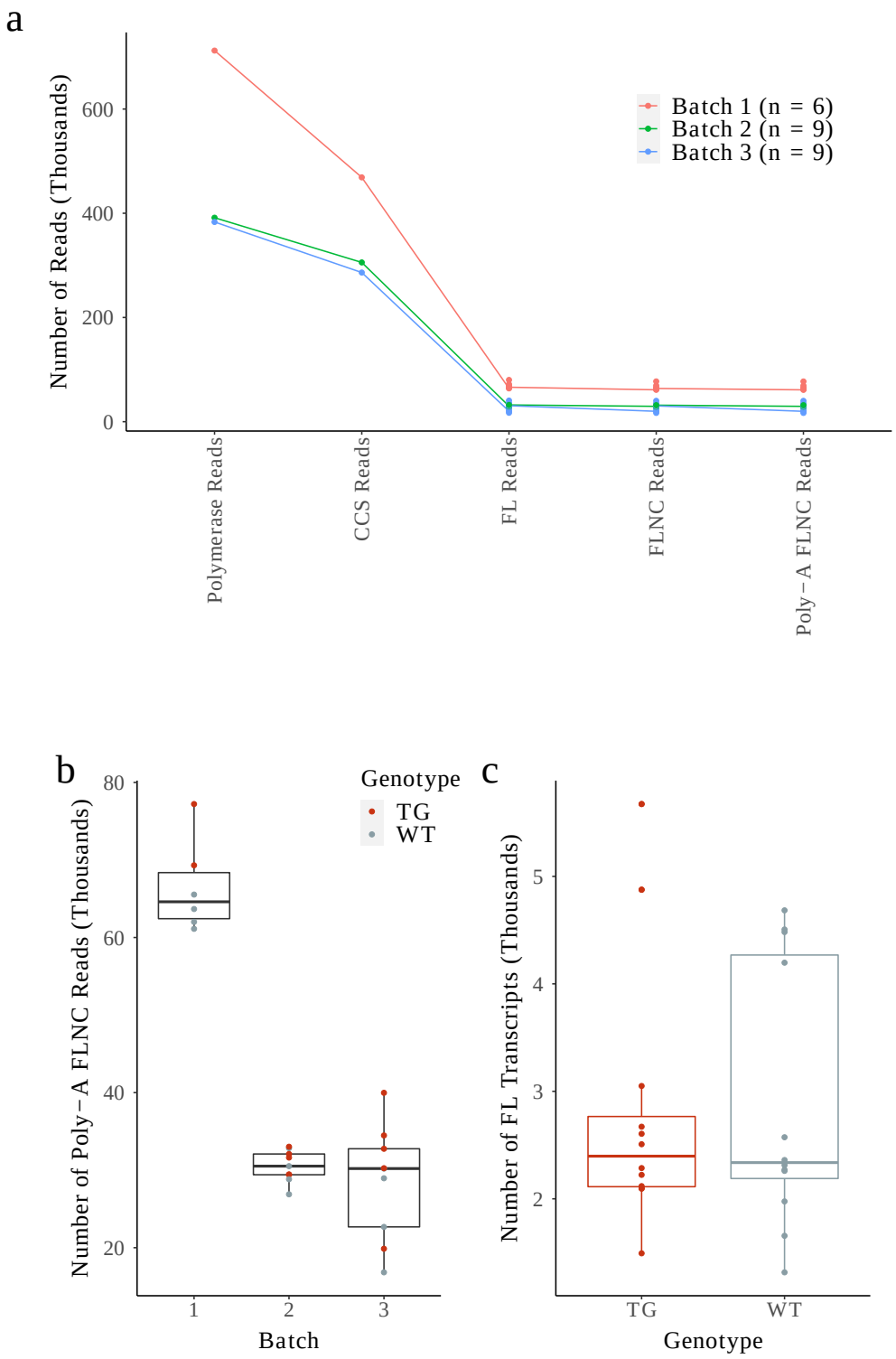


Figure 4.3: Despite batch variability in targeted transcriptome sequencing, no difference in the number of full-length transcripts was observed between wildtype and transgenic mice. **a)** Samples (n = 24) were multiplexed and sequenced in three runs (Batch 1, 2 and 3) with varied performance, as indicated by the number of polymerase reads through to poly-A FLNC reads. In the bioinformatics pipeline, the samples were demultiplexed and individually processed after generation of CCS reads from each run. **b)** Sample variability within each batch was observed from the number of poly-A FLNC reads generated. However, **c)** no statistical difference was observed in the overall number of full-length transcripts detected between wildtype and transgenic. Full-length transcripts were collapsed from poly-A FLNC reads in Iso-Seq Cluster. CCS - Circular Consensus Sequence, FLNC - Full-Length Non-Concatemer, FL - Full-Length, WT - Wildtype, TG - Transgenic

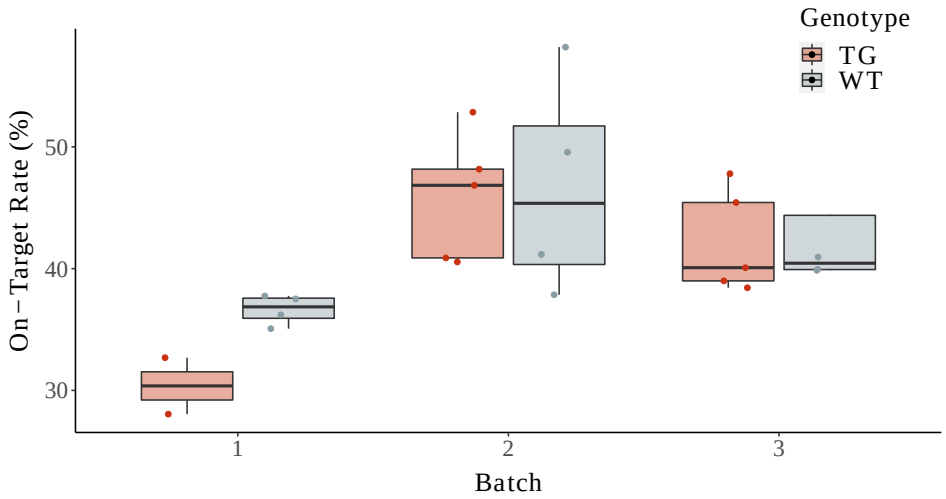
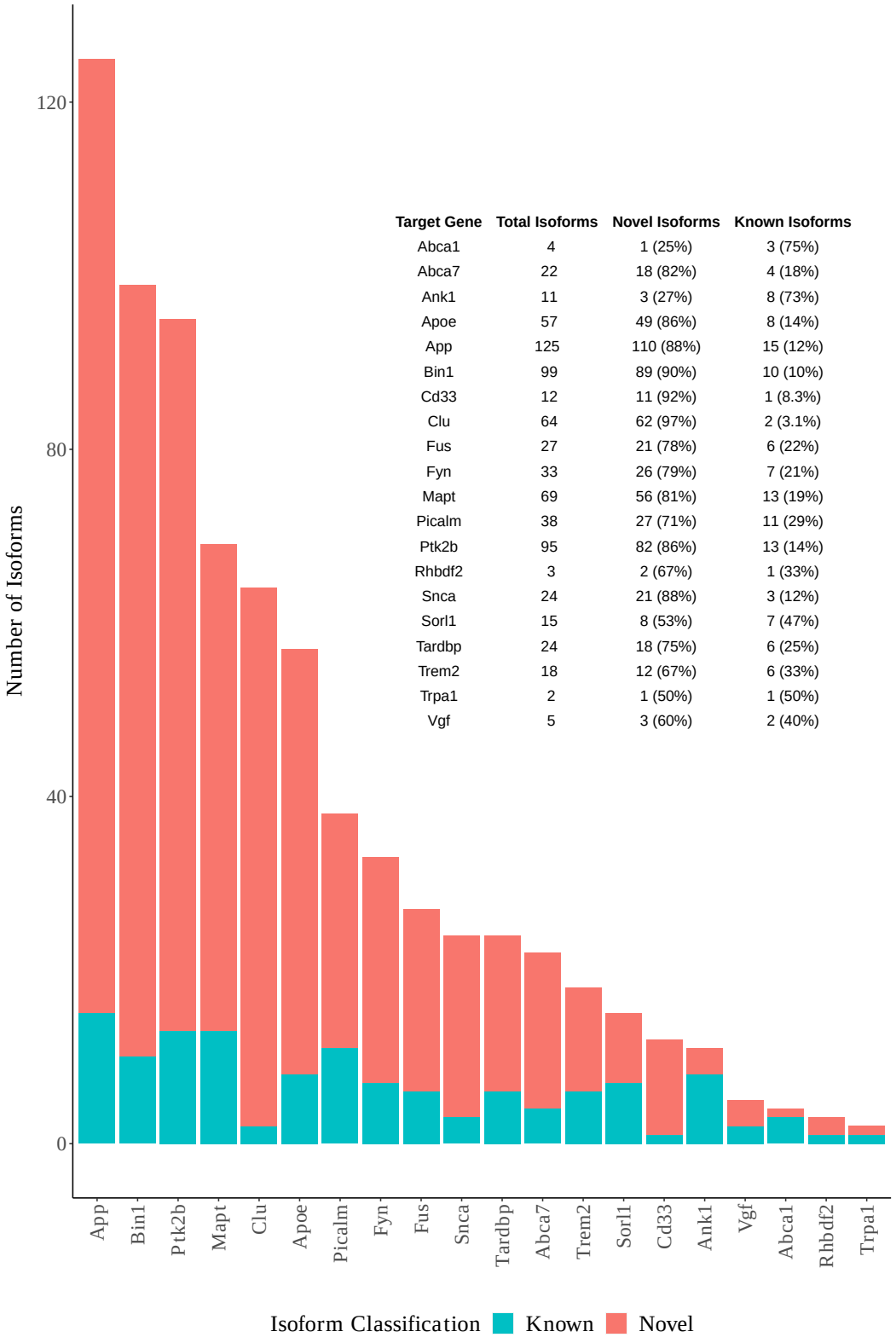


Figure 4.4: Coverage of target genes was greater in Batch 2 and 3 than Batch 1 due to more samples multiplexed and sequenced. Samples (n = 24) were multiplexed and sequenced in three runs (Batch 1 = 6 samples, Batch 2 = 9 samples, Batch 3 = 9 samples). Despite lower run yield output (4.2), Batch 2 and Batch 3 had a higher on-target rate, which refers to the proportion full-length transcripts associated with target genes. A difference in the on-target rate between wildtype and transgenic samples was observed in Batch 1, which is a likely reflection of the sample variability in sequencing (Figure 4.3b). WT - Wildtype, TG - Transgenic

4.3.2 Transcriptome annotation

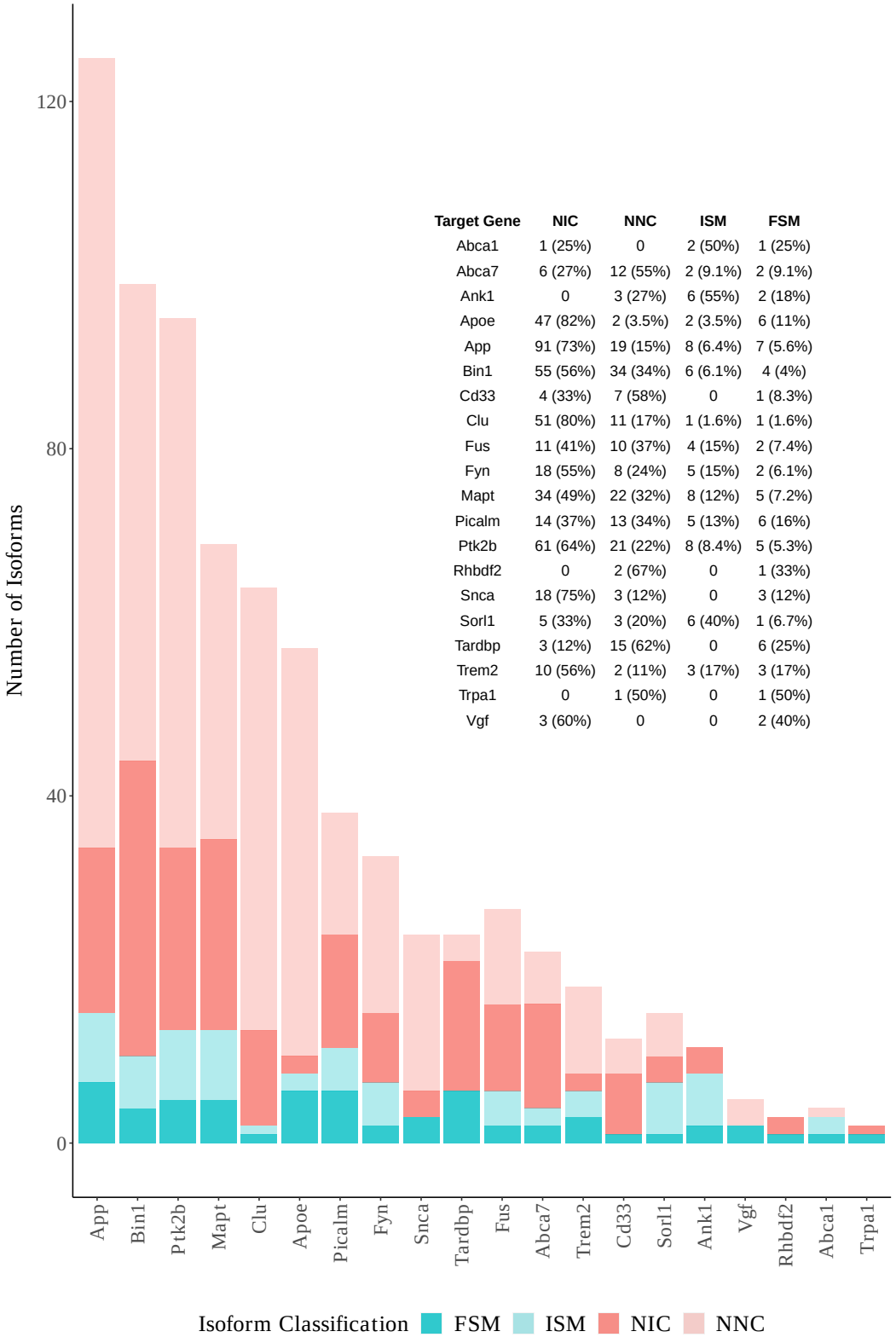
After filtering for technical artefacts (563 (1.69%) isoforms were removed due to intrapriming, 314 (0.94%) isoforms were removed due to RT switching, 1,267 (3.80%) were removed due to likely partial degradation), a total of 4,780 isoforms were detected across 20 AD-associated target genes across all the samples (n = 24). Of these isoforms, an overwhelming majority were novel (n = 4601, 96.2%) with no RNA-Seq support (n = 24 samples, total number of uniquely mapped reads = 360 million) at the junction (n = 4,033, 84.4%). This is likely to be reflection of the low coverage of RNA-Seq reads per sample (mean number of uniquely mapped reads = 15 million) to comprehensively span these novel junctions, rather than an indication of the invalidity of these isoforms given the stringent processing of the Iso-Seq bioinformatics pipeline. Nevertheless, following the recommendations from *SQANTI2* and to ease comparison with the whole transcriptome approach (Chapter X), we took the more stringent approach to only include novel isoforms with RNA-Seq support. All downstream analyses and statistics reported in this chapter thereon were based on the subset of *SQANTI2*-filtered isoforms (n = 747 isoforms, Figure 4.5).

4.3.3 Comparison with whole transcriptome



Target Gene	Total Isoforms	Novel Isoforms	Known Isoforms
Abca1	4	1 (25%)	3 (75%)
Abca7	22	18 (82%)	4 (18%)
Ank1	11	3 (27%)	8 (73%)
Apoe	57	49 (86%)	8 (14%)
App	125	110 (88%)	15 (12%)
Bin1	99	89 (90%)	10 (10%)
Cd33	12	11 (92%)	1 (8.3%)
Clu	64	62 (97%)	2 (3.1%)
Fus	27	21 (78%)	6 (22%)
Fyn	33	26 (79%)	7 (21%)
Mapt	69	56 (81%)	13 (19%)
Picalm	38	27 (71%)	11 (29%)
Ptk2b	95	82 (86%)	13 (14%)
Rhbdf2	3	2 (67%)	1 (33%)
Snca	24	21 (88%)	3 (12%)
Sorl1	15	8 (53%)	7 (47%)
Tardbp	24	18 (75%)	6 (25%)
Trem2	18	12 (67%)	6 (33%)
Trpa1	2	1 (50%)	1 (50%)
Vgf	5	3 (60%)	2 (40%)

Figure 4.5: Wide isoform diversity observed in AD-associated genes with many novel isoforms detected. Shown is the number of isoforms detected per target gene, classified by novel and known, after sequential processing and filtering in the bioinformatics Iso-Seq pipeline. Novel isoforms refer to isoforms that are not known in current existing annotations.



Target Gene	NIC	NNC	ISM	FSM
Abca1	1 (25%)	0	2 (50%)	1 (25%)
Abca7	6 (27%)	12 (55%)	2 (9.1%)	2 (9.1%)
Ank1	0	3 (27%)	6 (55%)	2 (18%)
Apoe	47 (82%)	2 (3.5%)	2 (3.5%)	6 (11%)
App	91 (73%)	19 (15%)	8 (6.4%)	7 (5.6%)
Bin1	55 (56%)	34 (34%)	6 (6.1%)	4 (4%)
Cd33	4 (33%)	7 (58%)	0	1 (8.3%)
Clu	51 (80%)	11 (17%)	1 (1.6%)	1 (1.6%)
Fus	11 (41%)	10 (37%)	4 (15%)	2 (7.4%)
Fyn	18 (55%)	8 (24%)	5 (15%)	2 (6.1%)
Mapt	34 (49%)	22 (32%)	8 (12%)	5 (7.2%)
Picalm	14 (37%)	13 (34%)	5 (13%)	6 (16%)
Ptk2b	61 (64%)	21 (22%)	8 (8.4%)	5 (5.3%)
Rhbdf2	0	2 (67%)	0	1 (33%)
Snca	18 (75%)	3 (12%)	0	3 (12%)
Sorl1	5 (33%)	3 (20%)	6 (40%)	1 (6.7%)
Tardbp	3 (12%)	15 (62%)	0	6 (25%)
Trem2	10 (56%)	2 (11%)	3 (17%)	3 (17%)
Trpa1	0	1 (50%)	0	1 (50%)
Vgf	3 (60%)	0	0	2 (40%)

Figure 4.6: Majority of the novel isoforms detected of the target genes has at least one novel donor or acceptor splice sites. Shown is the number of isoforms detected per target gene, further classified into FSM (Full Splice Match), ISM (Incomplete Splice Match), NIC (Novel In Catalogue) and NNC (Novel Not in Catalogue).

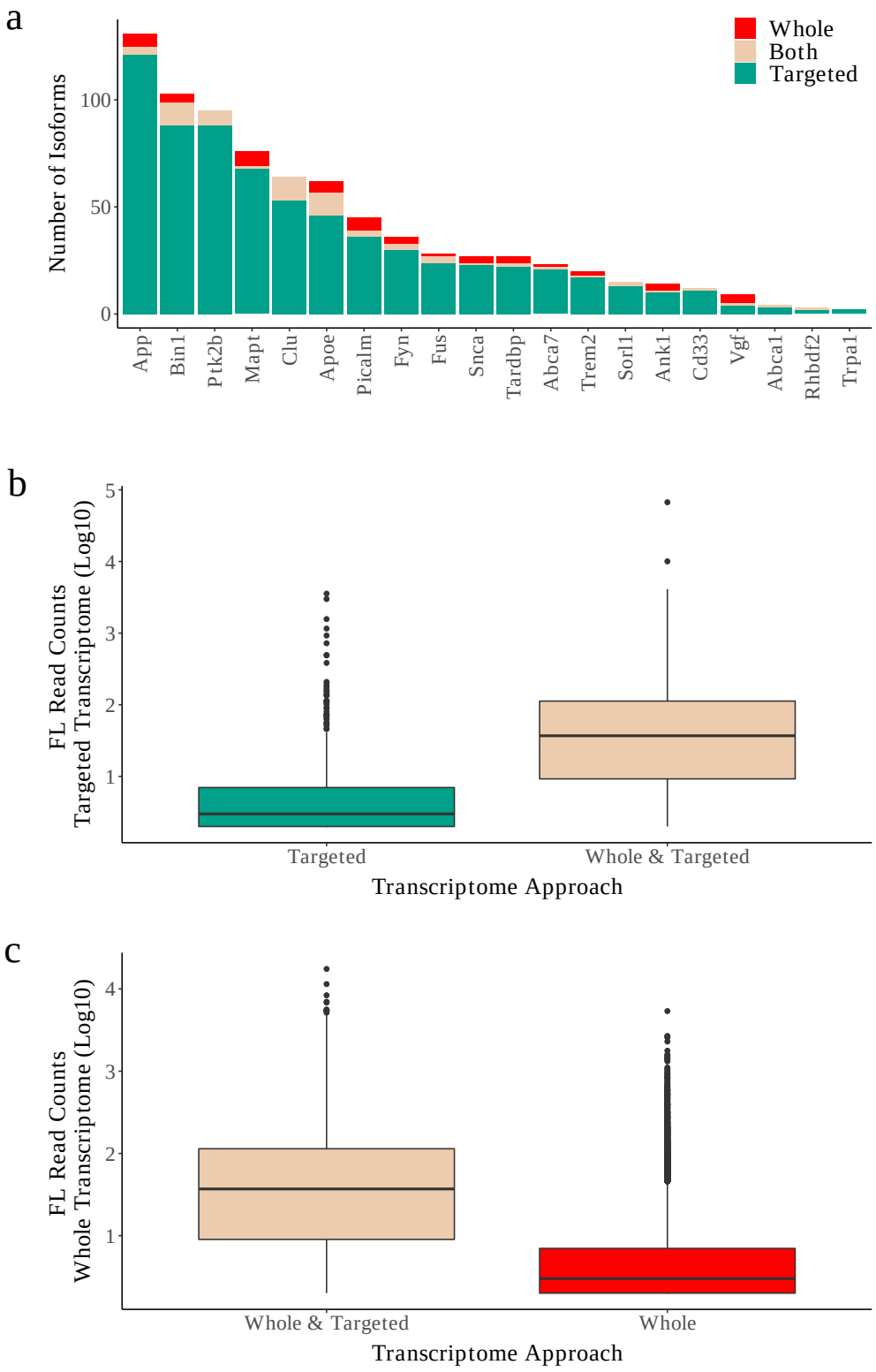


Figure 4.7: Majority of the novel isoforms detected of the target genes has at least one novel donor or acceptor splice sites. Shown is the number of isoforms detected per target gene, further classified into FSM (Full Splice Match), ISM (Incomplete Splice Match), NIC (Novel In Catalogue) and NNC (Novel Not in Catalogue).

Chapter 5

Transcriptional differences between WT and TG mice

5.1 Introduction

Following the accurate characterisation of the mouse transcriptome using long-read sequencing from a global (whole transcriptome approach, **Chapter 3**) and targeted perspective (Chapter **Chapter 4**), this chapter aims to exploit these datasets to investigate the transcriptional changes in the mouse entorhinal cortex associated with tau pathology.

There have been multiple studies recently that explore the transcriptional differences in transgenic mice harbouring different mutations associated with AD. However, all of these studies to date have been undertaken with short-read RNA-Seq - which, while it offers obvious advantage compared to microarrays in accurately quantifying gene expression, is severely limited in detecting and characterising transcripts (as discussed in detail in **Section 1.3.2**).

While there have been significant advances to process long-read sequencing data for transcriptome annotations, methods to harness long-read data for differential analyses have been limited. Currently, all the new tools developed to process long-read sequencing data (such as Oxford Nanopore's recommended cDNA transcriptome tutorial,[72] *FLAIR*[73]) integrate old tools, which were initially designed to analyse short-read sequencing data, for differential gene and

isoform analyses. Systematically assessed and benchmarked for detecting differential splicing and expression at isoform level in RNA-Seq studies, *DESeq2*, *DexSeq* and *NOISeq* have been most widely used.

5.2 Methods

5.2.1 Iso-Seq Processing and Isoform Quantification

All analyses pertaining to this chapter follows on from **Chapter 3** and **Chapter 4**: raw Iso-Seq reads from the individual samples were processed using *IsoSeq3*, which were them merged to one complete transcriptome at the global and targeted level, before transcript collapse with *Cupcake*, alignment with *Minimap2*, annotation with *SQANTI3* (v3.3) and finally, additional filtering with *TAMA*. Of note, transcriptome was reannotated with *SQANTI3* due to SQANTI2 being no longer maintained and the addition of novel features in *SQANTI3*, including the generation of a functionally-labelled annotation from the long-reads.

The full-length long-read counts (abundance) for each sample, required for downstream analyses, were obtained from one of *cupcake's* output files (read_stat.txt), which documented the source of all the full-length transcripts that were used for isoform collapse. Given that samples were sequenced individually under the whole transcriptome approach, we were thus able to differentiate and count the transcripts using the Run ID. For the targeted transcriptome approach, whereby the samples were barcoded and thus could not be differentiated by sequencing run, we used the ID (original CCS read) documented in the output file (flnc.report.csv) from *Iso-Seq3 Refine* after sample demultiplexing.

5.2.2 Quantification of human MAPT transgene expression

As a quality check of sample identity, the presence of human- and mouse-specific Mapt/MAPT sequences was determined in full-length transcripts across all the samples. Species-specific MAPT sequence, located in a 2kb region present in the 3'UTR, was identified after using BLAT[74] to compare human and mouse MAPT/Mapt sequence for divergent transcript sequences[70]).

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 , to assess whether there was a change in splicing patterns associated with rTg4510 pathology and across age.

5.2.4 Differential expression analysis

After trialling various ad-hoc methods, I chose to explore the transcriptomic changes between wildtype and transgenic mice with *tappAS* (v1.0.0),[75] which was also developed by the same authors as *SQANTI* (A.Conesa's group) and was recommended as an extension to the Iso-Seq pipeline for the functional annotations of isoforms. Accessible as a user-friendly Java application, *tappAS* was chosen as the framework for differential expression analysis due to the flexibility to explore both genotype effects and progressive changes across age, and to optimise parameters as background scripts were fully accessible and clearly written.

tappAS requires three inputs:[75]

1. An experimental design file, which allows comparisons to be made 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* (another tool developed by A.Conesa's group, https://isoannot.tappas.org), as a "scaffold" for transcript-level annotations. For the purpose of this study, the annotation file would be the conglomerate, long-read defined transcriptome of all the samples merged.
3. A transcript level expression matrix, which can either be derived directly from the full-length long-read transcript counts, or from mapping and transcript quantification of short-reads to the long-read defined transcriptome using *Kallisto*(v0.46.0). Raw transcript counts were tabulated per sample.

As a method comparison, the expression from both short- and long-read was used as quantification at the gene and transcript level, such that four differential analyses were performed using the whole and targeted transcriptome datasets (**Table 5.1**). To explore the utility and power of long reads for transcriptome annotation, results from the differential gene analyses were also compared to that generated from I.Castanho's analyses.[70]

	Datasets	Annotation	Quantification
1	Whole Transcriptome (n = 12)	Iso-Seq reads processed using bioinformatics pipeline	Iso-Seq FL reads
2			RNA-Seq reads
3	Targeted Transcriptome (n = 24)		Iso-Seq FL reads
4			RNA-Seq reads

Table 5.1: Summary of the differential gene and transcript analyses for mouse transcriptome using whole and targeted Iso-Seq transcriptome datasets. Using the Iso-Seq defined transcriptome as the "scaffold" rather than mouse reference genome, the analyses primarily differed on the quantification input. FL - Full length.

Count Normalisation

Very lowly-expressed transcripts with a sum of expression value less than 1 CPM (counts per million) 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[76] (Trimmed Mean of M-values) to account for differences in library size (sequencing depth) and sample RNA library composition, which is particularly important when comparing samples from different genotypes. The difference in RNA composition is determined by calculating a "scaling factor" for each sample relative to a reference sample (sample with the least varying read counts), which is the weighted average of all the log2 ratio of transcript counts between the two samples (M-values). The weighted average does not consider log2 ratio of transcripts with significant differences ("biased" transcripts that are widely present in one sample and not the other, and vice versa) and of transcripts with highest or lowest expression, hence "trimmed mean", to avoid effect of outliers. Of note, TMM assumes that the majority of the transcripts are not differentially expressed. Gene abundance was deduced from the sum of normalised counts of associated isoforms, after removing transcripts with low or highly-varied expression values.

Differential Gene and Isoform Expression Analysis

To elucidate transcriptional changes for both genotype and longitudinal effects between two groups and over time, *maSigPro*[77–79] was used for both differential gene and transcript expression analysis, implemented as part of *tappAS*. Briefly, maSigPro performs a two-step regression strategy to first define a negative binomial general linearised models[78] for each gene or transcript, accounting for both genotype and age (Equation Eq. (5.1)), 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.

Adapting the model[77] to our scenario, let *I* denote the genotype groups (wildtype - WT, transgenic - TG) and $\mathcal{J}$ as the age (2, 4, 6, 8 months) for each particular group, and assuming that gene or transcript expression in measured in replicated samples (*R*).

$$\begin{aligned} y_{ijr} = \beta_0 + \beta_1 D_{ijr} \\ + \delta_0 T_{ijr} + \delta_1 T_{ijr} D_{ijr} \\ + \gamma_0 T_{1ijr}^2 + \gamma_1 T_{ijr}^2 D_{ijr} \\ + \lambda_0 T_{ijr}^3 + \lambda_1 T_{ijr}^3 D_{ijr} + \varepsilon_{ijr} \end{aligned}$$

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 refers to reference group (WT) and 1 refers to experimental group (TG)
T = age at 2, 4, 6, 8 months described using a polynomial model with a degree of 3
$\beta_0, \delta_0, \gamma_0, \lambda_0$ = regression coefficients for reference group (WT) relating to the age
$\beta_1, \delta_1, \gamma_1, \lambda_1$ = regression coefficients for the difference between the experimental group (TG) and reference group (WT) at each age

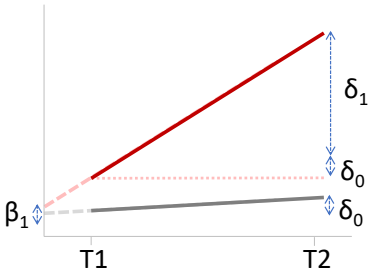
therefore, if:

$FDR(\beta_1) < 0.05$ = significant expression difference between WT and TG at 2 months
$FDR(\delta_0) < 0.05$ = significant expression difference in WT across 2 and 4 months
$FDR(\delta_1) < 0.05$ = significant expression difference between WT and TG across 2 and 4 months

...

Equation 5.1: Linear regression model to determine differential gene and transcript expression. The model, adapted from *MaSigPro* and implemented as part of *tappAS*, describes gene or transcript expression between two groups (WT - wildtype, TG - transgenic) at four different time points (age in months). FDR - False discovery rate

A gene or transcript with different profiles between WT and TG mice would have a different corresponding regression model with a statistically significant coefficient (**Figure 5.1**), and was considered differentially expressed if P-value adjusted for multiple testing, by controlling the false discovery rate (FDR) with the Benjamin and Hochberg correction, was < 0.05 and R^2 > 0.5. 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



Inference from regression coefficients
β_1 : WT vs TG at T1
δ_0 : WT over time
δ_1 : WT vs TG over time

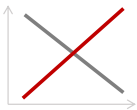
Model	β_1 CasevsControl	δ_0 Time	δ_1 TimexCase	Condition	Effects
1	✓	x	x		Genotype
2	✓	✓	x		Genotype + Age
3	x	✓	x		Age
4	x	✓	✓		Interaction
5	x	x	✓		Interaction
6	✓	x	✓		Interaction
7	✓	✓	✓		Interaction

Figure 5.1: Different conditions modelled for exploring rTg4510 genotype across age. An example of six different models generated with *maSigPro* using Equation Eq. (5.1) for 2 experimental groups (WT - Wildtype/Control, TG - Transgenic/Case) across two time points/age (T1, T2). The regression coefficients from Equation Eq. (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-value (FDR) < 0.05 and > 0.05 respectively. A significance 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).

maSigPro labels the coefficients in the results table as "CaseVsControl", "Time", "TimexCase" for β_1, δ_0, δ_1. In the case where there is more time points/ages (as experimented in the Targeted Transcriptome datasets), the significance for the additional regression coefficients relating to the additional time variables are reported (Time2, Time2xCase, Time3, Time3xCase)

differentially expressed genes with meaningful biological implications.[77]

Following the identification of statistically significant gene models (**Figure 5.1**), the specific conditions (phenotype or age-associated changes) for which the genes show statistically significant profile changes (the significant variables) were identified by using an iterative backward stepwise approach.[79] The procedure therefore first started with all the variables imputed (different phenotype x age at different time points). At each iteration step, the P-value associated to each variable was determined and only the variables with a P-value < 0.05 were retained.

Differential Isoform Usage

In addition to assessing expression changes across conditions through differential isoform expression analysis, the relative expression, and as such the usage, of these isoforms can also change (see **Section 1.3.5**). A gene is therefore identified as exhibiting differential isoform usage (DIU) if the fraction of the associated isoforms (Isoform Fraction) is significantly altered between conditions, which could result in switching of the major isoform.

In accounting for biological replicates, the isoform fraction (IF) for each isoform was defined as:

$$IF_{cig} = \frac{\bar{E}_{cig}}{\sum_{i=1}^{n} \bar{E}_{cig}} \tag{5.1}$$

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

Equation 5.2: Calculation of isoform fraction for differential isoform usage analysis. Equation is adopted from *tappAS*

Despite abundant evidence of widespread isoform diversity,[5] most protein-coding genes have been reported to typically express a few dominant isoforms,[80,81] while the remaining are very lowly expressed and unlikely to be main contributors to the proteome.[80]As such, minor isoforms were filtered to avoid finding genes associated with differential isoform usage due to "flat" behaviour of these minor isoforms[75] (relatively small non-negligible expression changes

of minor isoforms in the opposing direction of the predominant isoforms). *tappAS* provides two strategies to filter lowly-expressed isoforms: 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 alternatively if its proportion relative to the major isoform is below a pre-specified threshold (default: FC = 2). A major isoform is defined as the isoform with the highest expression across all the conditions, with the remaining isoforms annotated as minor.

Implemented as an additional filtering step after *tappAS* and recommended in other bioinformatic tools,[82] lowly expressed genes were also filtered as there would be less confidence in isoform fraction used for determining genes with significant differential isoform usage.

5.3 Results

5.3.1 Change in endogeneous expression

As expected, human-specific *MAPT* sequences were only detected in reads from TG mice, confirming stable activation of human *MAPT* transgene (**Figure 5.2a**) and supporting findings from Castanho et al.[70] Alignment of these human-specific transcripts to the mouse genome were mapped either to the mouse prion protein gene (*Prnp*) with high identity but short overlap (**Figure 5.2b,c**), given that the transgene contains exons 2-3 of mouse *Prnp*,[83] and to the mouse *Mapt* gene with low identity but long overlap (**Figure 5.2b,d**). Applying filter thresholds for downstream analysis removed these human-specific *MAPT* transcripts (**Figure 5.2b**).

5.3.2 Transcriptome Annotation

While identifying widespread RNA isoform diversity amongst genes expressed in the mouse entorhinal cortex (described in **Chapter 3**), no difference was observed in the number of genes (mean number = 13,302 genes) or isoforms (mean number = 35,157 isoforms) detected between wildtype and trangenic mouse at the two ages (WT: n = 6, TG: n = 6). Further characterisation of the transcriptome revealed similar profile of isoform diversity between the two phenotypes and ages (**Table 5.2**), with half of the isoforms annotated as known (mean number = 18,567 isoforms (52.8%), as also shown in **Section 3.3.6**) and with a similar distribution in isoform length and number of exons (median: 9, range = 1-89)

5.3.3 Alternative Splicing

No significant difference was observed XX of known transcripts were identified to have intron retention; XX of known transcripts were identified to be fusion genes. XX of know transcripts identified to have non-sense-mediated decay.

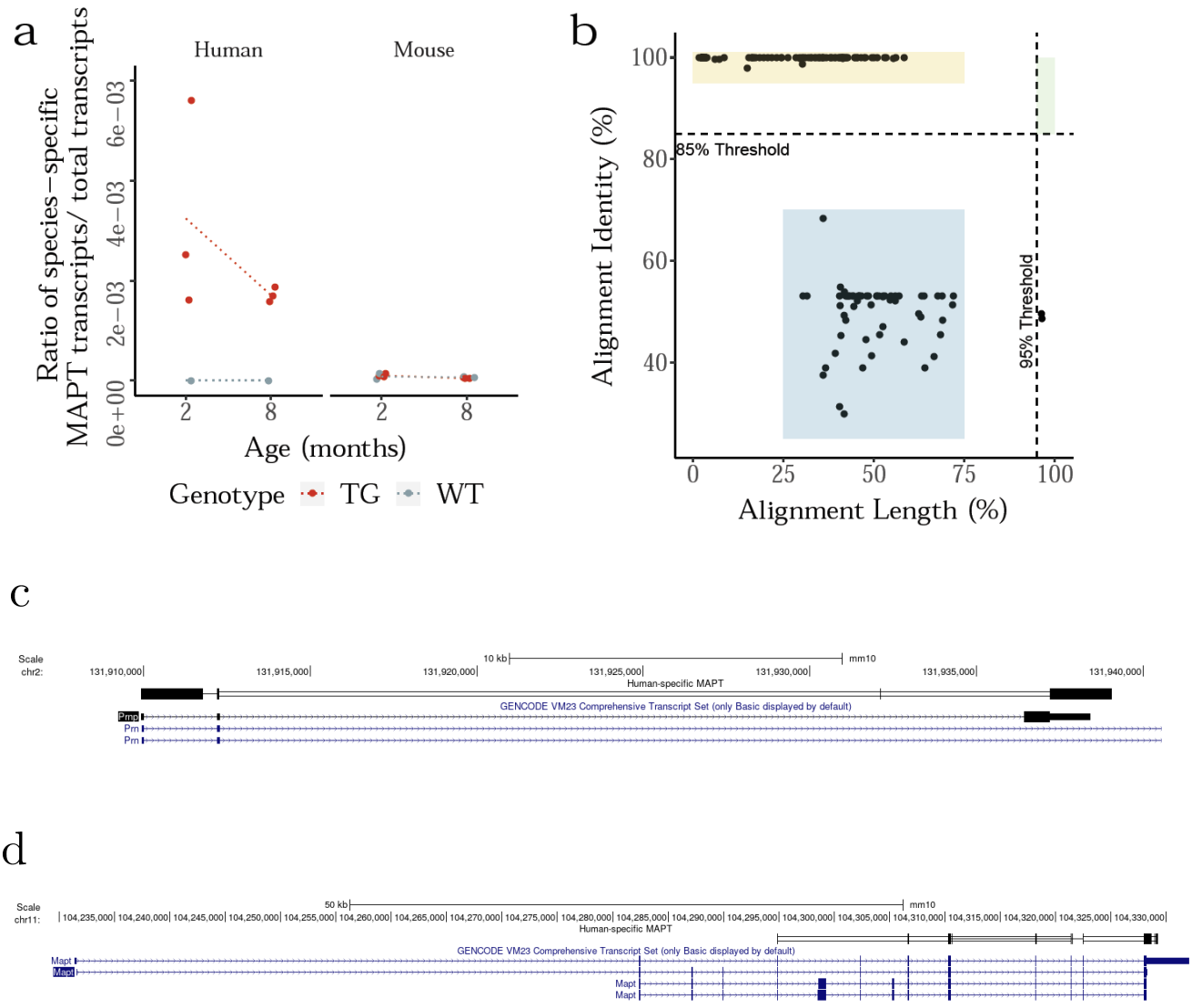


Figure 5.2: Human-specific *MAPT* sequences only present in transgenic mice with poor alignment to mouse *Prnp* and *Mapt* gene: Presence of human- and mouse-specific *MAPT*/*Mapt* sequences was determined in full-length transcripts generated from Iso-Seq merged dataset. **a)** Ratio of full-length transcripts that were mapped to human-specific *MAPT* and mouse-specific *Mapt* sequences. Dotted lines represent the mean paths across ages. **b)** As expected, human-specific *MAPT* transcripts were poorly aligned to mouse genome. Transcripts were either aligned to mouse *Prnp* gene (boxed yellow) or mouse *Mapt* gene (boxed blue). Alignment to mouse *Prnp* gene was near 100% within a short region, given that the transgene contains exon 2 and 3 of mouse *Prnp* gene.[83] Conversely, while human-specific *MAPT* gene was sufficiently divergent from mouse *Mapt* gene for transgene quantification, it still mapped to the mouse *Mapt* gene 3'UTR albeit poorly. **c)** USCS 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 reference mouse genome (mm10). Tracks were cropped and modified to remove irrelevant genes within the same locus. UTR - Untranslated region

	Wildtype (n = 6)	Transgenic (n = 6)	Wildtype, 2 months (n = 3)	Wildtype, 8 months (n = 3)	Transgenic, 2 months (n = 3)	Transgenic, 8 months (n = 3)
Total Number of Genes	14083	14183	12798	12947	12384	13418
Annotated Genes	13911 (98.78%)	14018 (98.84%)	12696 (99.2%)	12816 (98.99%)	12286 (99.21%)	13289 (99.04%)
Novel Genes	172 (1.22%)	165 (1.16%)	102 (0.8%)	131 (1.01%)	98 (0.79%)	129 (0.96%)
Total Number of Isoforms	41081	41671	31407	32630	28982	35169
FSM	18287 (44.51%)	18574 (44.57%)	15503 (49.36%)	15870 (48.64%)	14675 (50.63%)	16892 (48.03%)
ISM	3164 (7.7%)	3242 (7.78%)	2243 (7.14%)	2368 (7.26%)	2066 (7.13%)	2590 (7.36%)
NIC	11781 (28.68%)	12033 (28.88%)	8356 (26.61%)	8856 (27.14%)	7518 (25.94%)	9805 (27.88%)
NNC	7354 (17.9%)	7343 (17.62%)	4980 (15.86%)	5175 (15.86%)	4427 (15.27%)	5520 (15.7%)
Genic Genomic	55 (0.13%)	54 (0.13%)	37 (0.12%)	41 (0.13%)	28 (0.1%)	43 (0.12%)
Antisense	93 (0.23%)	100 (0.24%)	49 (0.16%)	74 (0.23%)	65 (0.22%)	74 (0.21%)
Fusion	253 (0.62%)	242 (0.58%)	180 (0.57%)	176 (0.54%)	155 (0.53%)	177 (0.5%)
Intergenic	94 (0.23%)	83 (0.2%)	59 (0.19%)	70 (0.21%)	48 (0.17%)	68 (0.19%)
Genic Intron	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Isoform Length (bp)	Median: 2946, Range: 83-15016	Median: 2955, Range: 83-15913	Median: 2987, Range: 88-15016	Median: 2890, Range: 83-14850	Median: 2798, Range: 88-14302	Median: 3013, Range: 83-15913
Number of Exons	Median: 9, Range: 1-89	Median: 9, Range: 1-89	Median: 9, Range: 1-89	Median: 9, Range: 1-89	Median: 9, Range: 1-77	Median: 9, Range: 1-89
Number of Isoforms within 50bp CAGE	34574 (84.16%)	35097 (84.22%)	26539 (84.5%)	27689 (84.86%)	24398 (84.18%)	29911 (85.05%)

Table 5.2: Overview of the whole transcriptome Iso-Seq datasets generated from mouse rTg4510, subsected by phenotype and age. Annotations from wildtype (n = 6) and transgenic mouse (n = 6) 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). Isoform can be further classified as known (FSM, ISM) or novel (ISM, NIC, NNC, Genic Genomic, Antisense, Fusion, Intergenic, Genic Intron), as described in **Section 2.1.3.9**. FSM – Full Splice Match, ISM – Incomplete Splice Match, NIC – Novel In Catalogue, NNC – Novel Not in Catalogue.

5.3.4 Differential Gene Expression Analysis

Usage of Iso-Seq reads alone detected robust changes in gene expression

Strikingly, despite the lower long-read sequencing coverage and smaller sample size (n = 6 TG, n = 6 WT) sequenced on the long-read platform, a significant number of genes (n = 446 genes) were identified as differentially expressed when we used the Iso-Seq long reads for both annotation and expression. Using *EnrichR*, differentially expressed genes (n = 466) were found to be highly enriched in lysosome (KEGG 2021 Human: odds ratio = 5.99, adjusted P = 4.41 x 10^{-6}) and in particularly the TGF-βpathway (WikiPathway 2021 Human: odds ratio = 58.98, adjusted P = 2.69 x 10^{-3}). Classification of the differentially expressed genes by effects (depicted in **Figure 5.1** and showcased in **Figure 5.4**), further identified the majority (n = 340 genes, **Figure 5.3**) to be associated with tau pathology (n = 23, **Figure 5.4a**) and progressive with age (n = 340, **Figure 5.4d,e,f,g**).

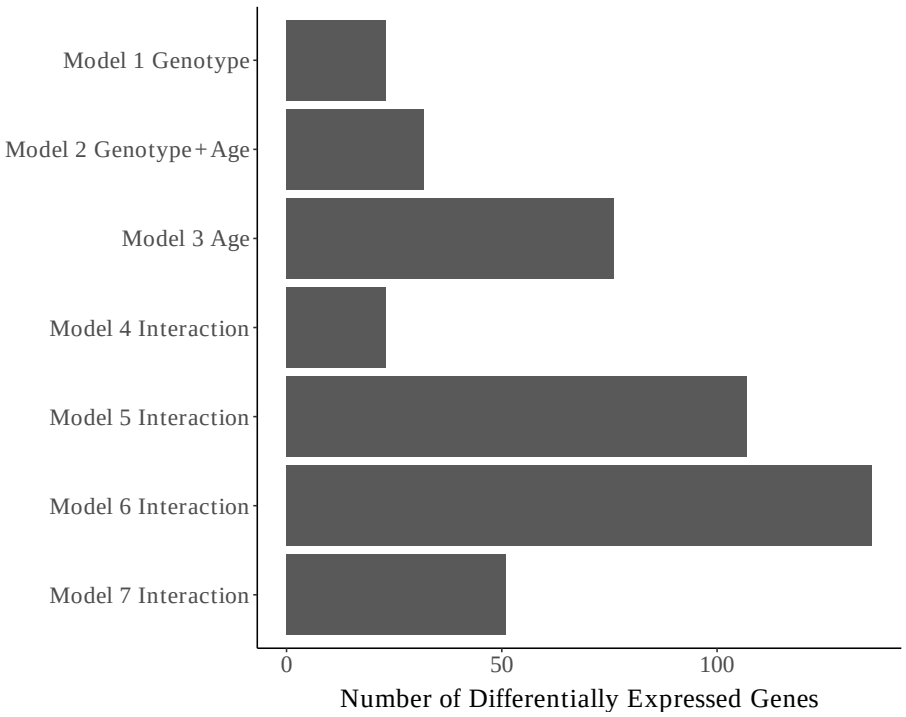


Figure 5.3: Differentially expressed genes were identified across all the different conditions, with most genes exhibiting an interaction effect of rTg4510 genotype and age Shown is a bar plot of the number of differentially expressed genes (n = 446) classified by the different conditions, as modelled in **Figure 5.1**. The differentially expressed genes were identified from the whole transcriptome datasets (WT = 6, TG = 6, across age 2 and 8 months) using Iso-Seq read counts as abundance. WT - Wildtype, TG - Transgenic.

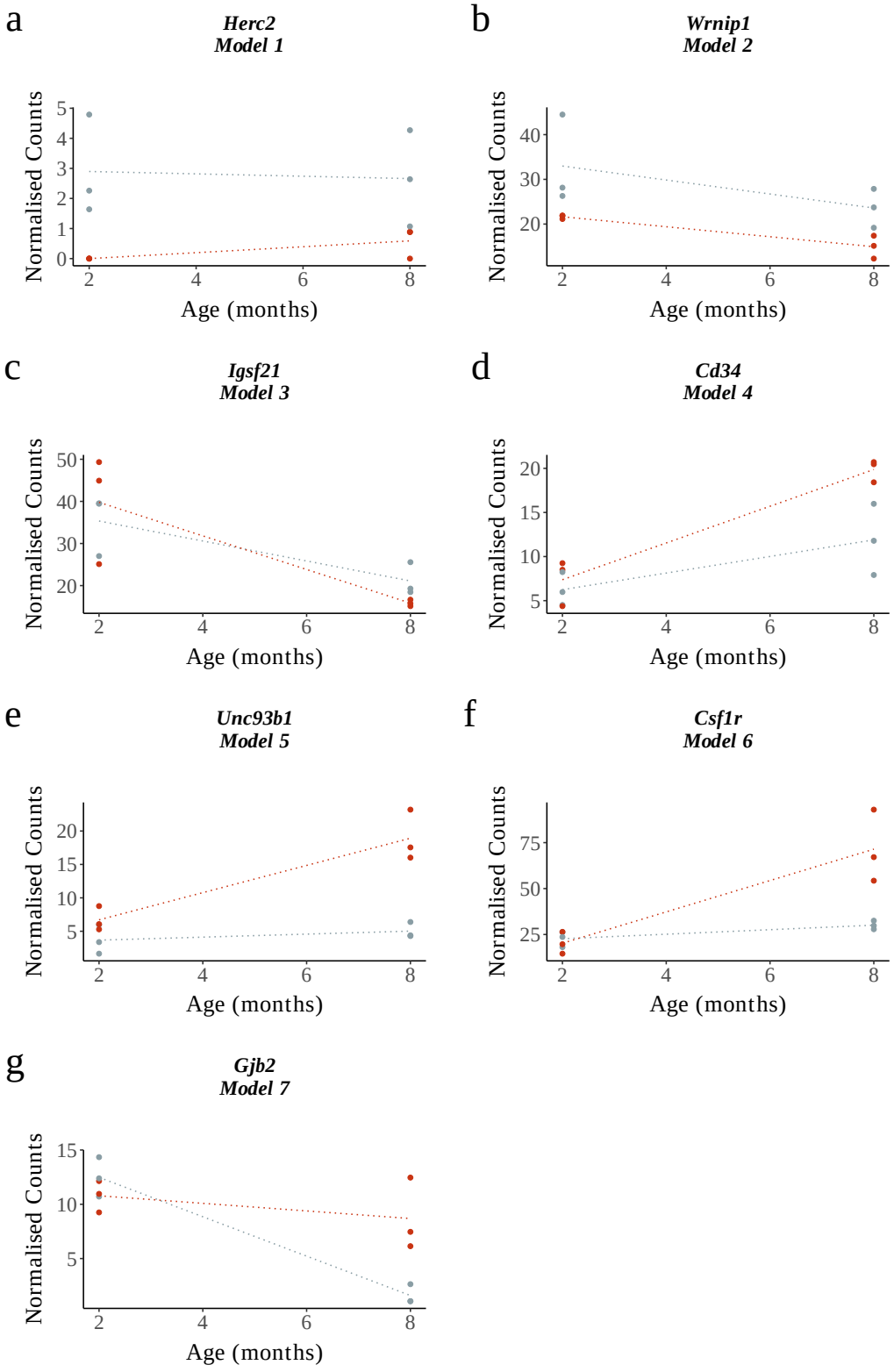


Figure 5.4: Differential expressed genes exhibiting genotype, age and interaction effects Shown are examples of differentially expressed genes classified under the different models, using the whole transcriptome dataset (WT = 6, TG = 6, across age 2 and 8 months) using Iso-Seq read counts as abundance: **a** *Herc2* with a genotype effect, **b** *Kcng1* with a genotype and age effect, **c** *Clnc3* with an age effect, and **d** *Cd34*, **e** *Unc93b1*, **f** *Csf1r* and **g** *Gjb2* with an interaction effect. Dotted lines represent the mean paths across ages.

Iso-Seq-identified-differential gene expression changes, associated with rTg4510 phenotype and age, were validated with RNA-Seq reads

With RNA-Seq reads (n = 29 TG, n = 30 WT) used as expression abundance, we identified 832 genes as differentially expressed, a significant proportion of which were also identified when using the reference genome instead as annotation (n = 483 genes, 58.8%.)

Usage of Iso-Seq reads alone identified differentially-expressed genes previously reported to be highly associated with AD

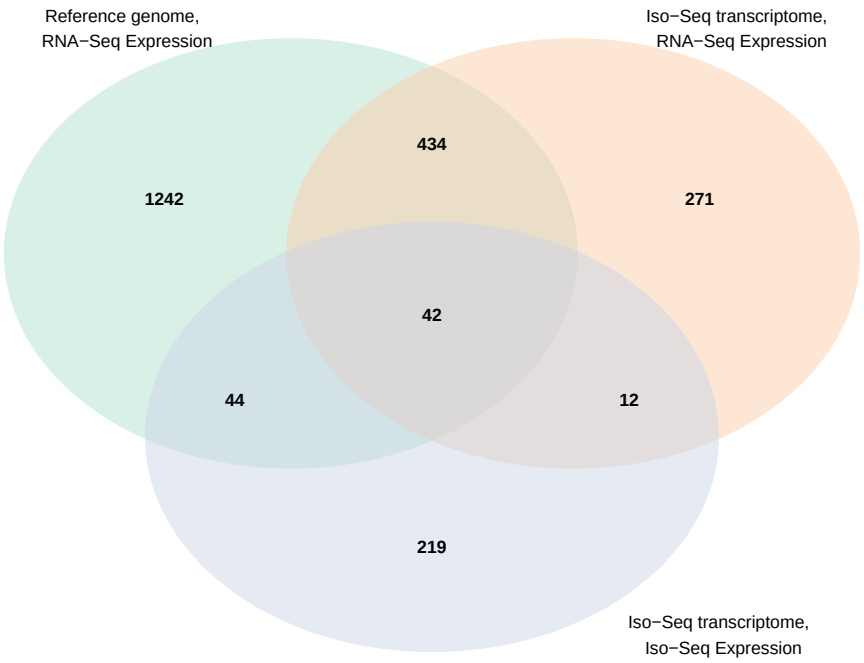
Of the differentially-expressed genes identified using Iso-Seq reads alone, the top ranked tau-associated differentially-expressed genes that was found to be progressive with age in our previous analyse[70] was also recapitulated (**Figure 5.6a,b**): *Gfap*, which encodes for glial fibrillary acidic protein (GFAP), a cytoskeletal protein that acts as a marker for astrocyte activation and proliferation - its increased expression was reported to correlate with increased neurofibrillary tangles density in Alzheimer's disease entorhinal cortex tissues.[84] Higher GFAP levels have also be documented in cerebrospinal fluid (CSF) in patients with AD compared to healthy control,[85] and more recently, in plasma of cognitively intact older adults at risk of AD.[86]

Other top-ranked differentially-expressed genes have been reported to be involved in AD development and pathology, notably *C4b* - a member of the complement immune system (**Figure 5.6c,d**), *Slc14a1* encoding the urea transporter 1, *Tgfbr1* encoding the TGF-βreceptor protein (**Figure 5.6e,f**) and *Unc93b1*, a transmemebrane protein required for toll (**Table 5.3**). Up-regulated expression of these genes have been observed previously in AD patients and other AD mouse models.[87–89]

Iso-Seq defined transcriptome enable identification of tau-pathology associated changes in genes not previously known in reference

Using the Iso-Seq reads as annotation and RNA-Seq reads for expression, we identified robust expression changes in a few genes (n = 3) that we would have otherwise not detected if we used the reference genome for annotation. These "novel" genes, not present in existing genome annotations (mm10), have been previously characterised as more lowly-expressed and often were antisense to known genes, with a large proportion sharing exonic regions either at the 5'UTR, 3'UTR or within the gene body (**Section 3.3.10**). The most significant differentially-

a



b

Figure 5.5: While usage of RNA-Seq reads as expression identified more differentially expressed genes associated with rTg4510 genotype and across age, a significant number of these were detected when using Iso-Seq read counts as expression. Shown are venn diagrams of the number of differentially expressed genes associated with **a)** rTg4510 genotype and progressive with age (interaction effect) **b)** and genotype alone, that were identified when using RNA-Seq reads as expression and the reference genome (mm10) as annotation (green circle), RNA-Seq reads as expression but Iso-Seq defined transcriptome as reference annotation (orange circle), and Iso-Seq reads as both expression and annotation (purple circle)

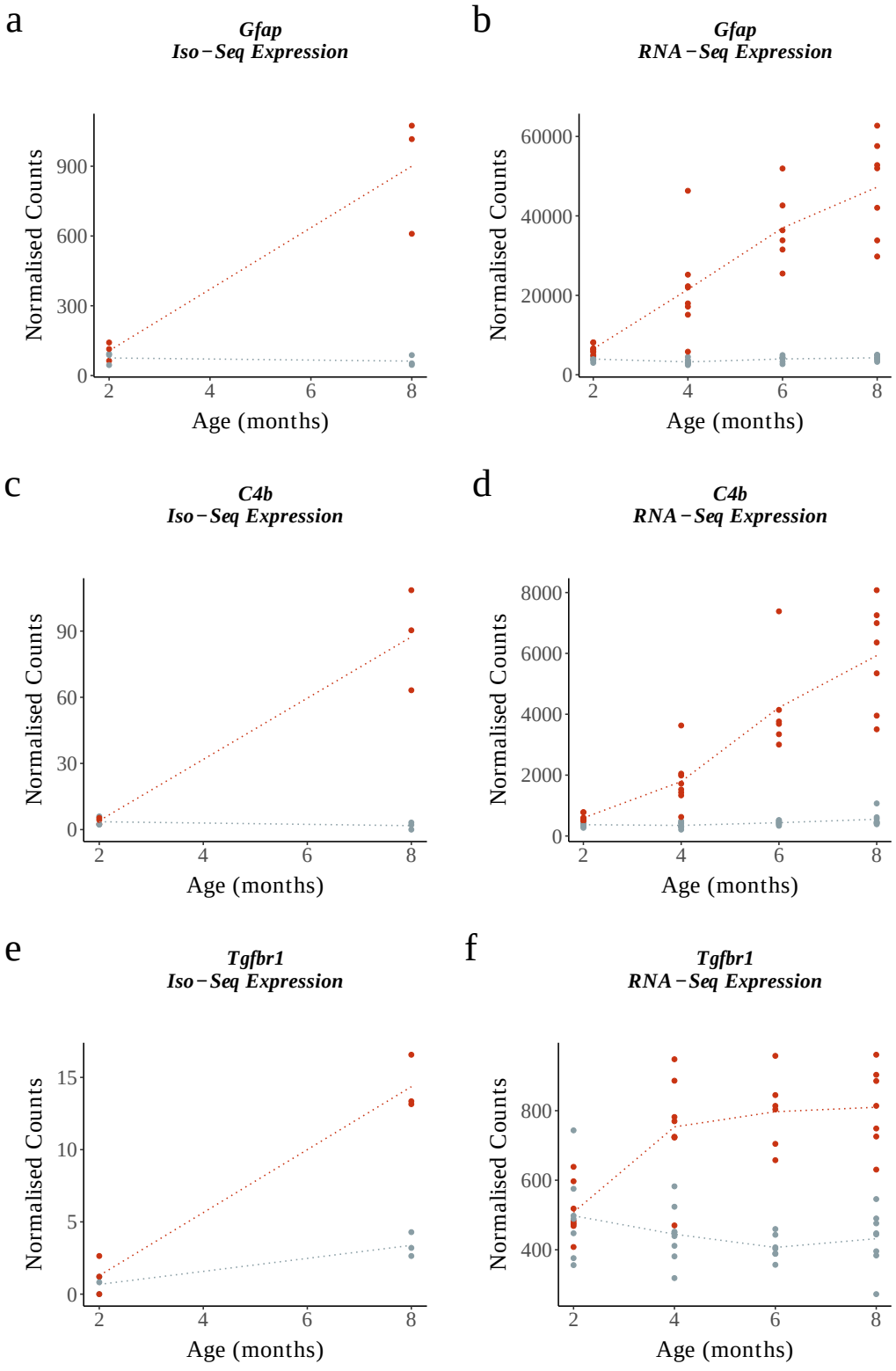


Figure 5.6: Novel isoforms were less expressed, longer and had more exons than known isoforms: Shown is the **a)** Iso-Seq transcript expression, the **c)** transcript length, and the **e)** the number of exons of novel and known isoforms. The known and novel isoforms can be further subdivided and classified, with the **b)** Iso-Seq expression **d)** transcript length and **f)** number of exons for each category. According to SQANTI, known isoforms are subdivided into FSM and ISM, and novel isoforms are subdivided into NIC, NNC, and fusion. FSM – Full Splice Match, ISM – Incomplete Splice Match, NIC – Novel In Catalogue, NNC – Novel Not in Catalogue.

Gene	FDR	R^2	log2-fold change (8 months)	Mean Gene Expression			
				Wildtype		Transgenic	
				2 months	8 months	2 months	8 months
C4b	3.9E-39	0.941	4.44	3.57	1.79	4.03	87.4
Gfap	8.88E-36	0.935	3.09	75.6	62.3	106	900
Tgfbr1	1.06E-20	0.880	3.48	0.673	3.39	1.28	14.4
Slc14a1	1.94E-16	0.872	2.92	8.56	12.3	5.2	39.4
Unc93b1	1.69E-15	0.853	1.5	3.68	5.04	6.68	18.9

Table 5.3: Tabulated are the top-ranked genes identified as differentially expressed in rTg4510 using *maSigPro* with Iso-Seq defined transcriptome for annotation and Iso-Seq FL read count as expression. Gene expression is determined from the sum of normalised expression of associated transcripts. FDR - False Discovery Rate.

expressed novel gene was identified in Chromosome 10 (**Figure 5.7a**), with progressive down-regulation in TG over time (**Figure 5.8a**). The other differentially-expressed novel genes were found antisense to *Fgfr1op* (**Figure 5.7b**), within the gene-body, and to *Htra1* (**Figure 5.7c**), at the 5'UTR. Both genes were associated with increased expression in TG compared to the WT over time (**Figure 5.8b,d**). Notably, while *Fgfr1op* was not identified as differentially-expressed (**Figure 5.8c**), *Htra1* was also found to have higher expression in TG compared to WT (**Figure 5.8e**).

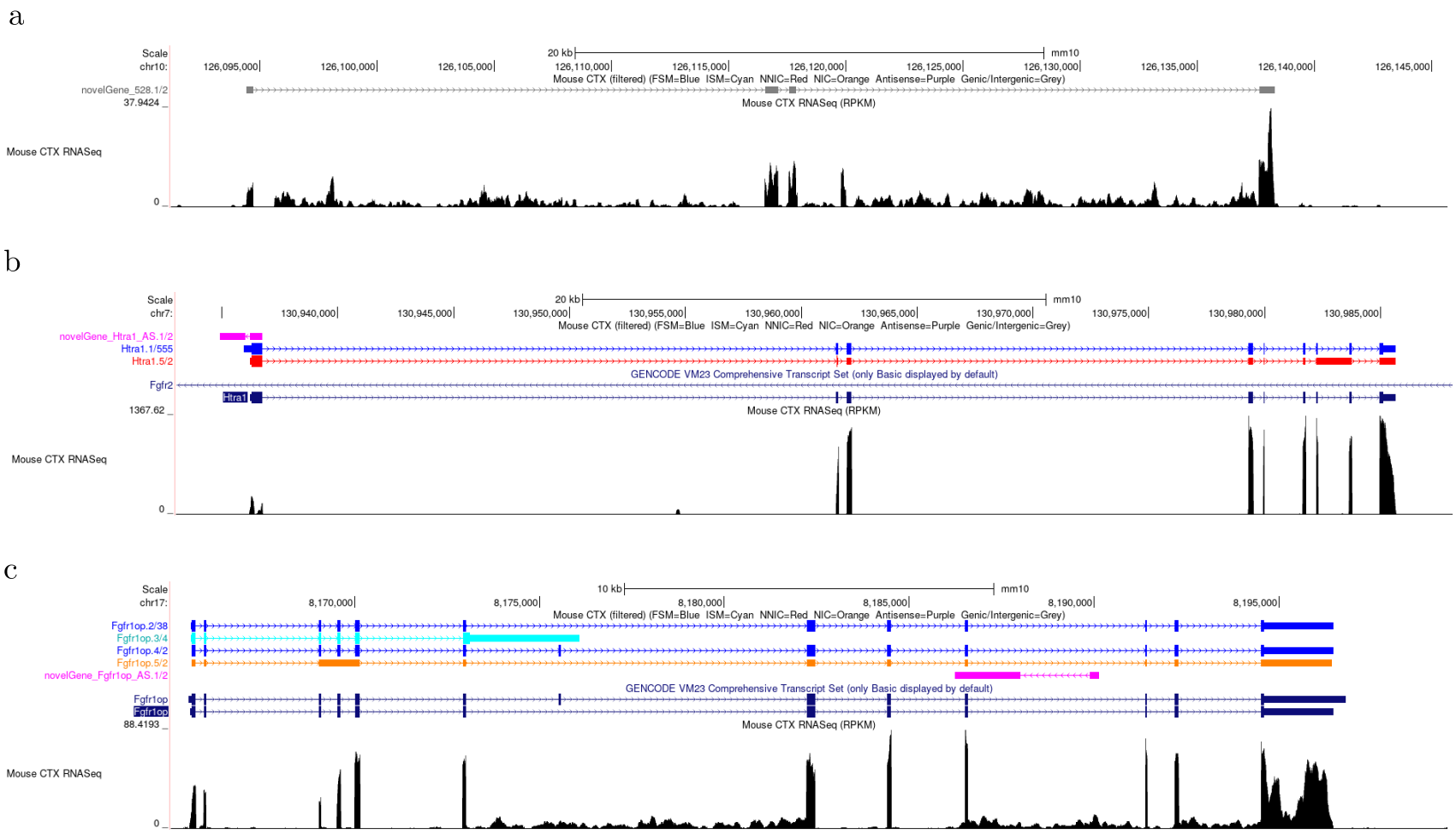


Figure 5.7: Tracks of novel genes that were differentially expressed in rTg4510 mice: Shown are UCSC genome browser Iso-Seq tracks of three novel genes - **a)**novel gene in Chromosome 10, **b)** novel gene antisense to *Fgfr1op*, and **c)** novel gene antisense to *Htra1* - that were identified as differentially expressed with Iso-Seq reads (n = 12 samples) as annotation and RNA-Seq reads (n = 59 samples) as expression. The isoforms were coloured based on *SQANTI* classification, with the novel gene coloured grey for genic/intergenic in panel a) and pink for antisense in panels b) and c). Shown are also the reference genome annotations (mm10) and RNA-Seq reads.

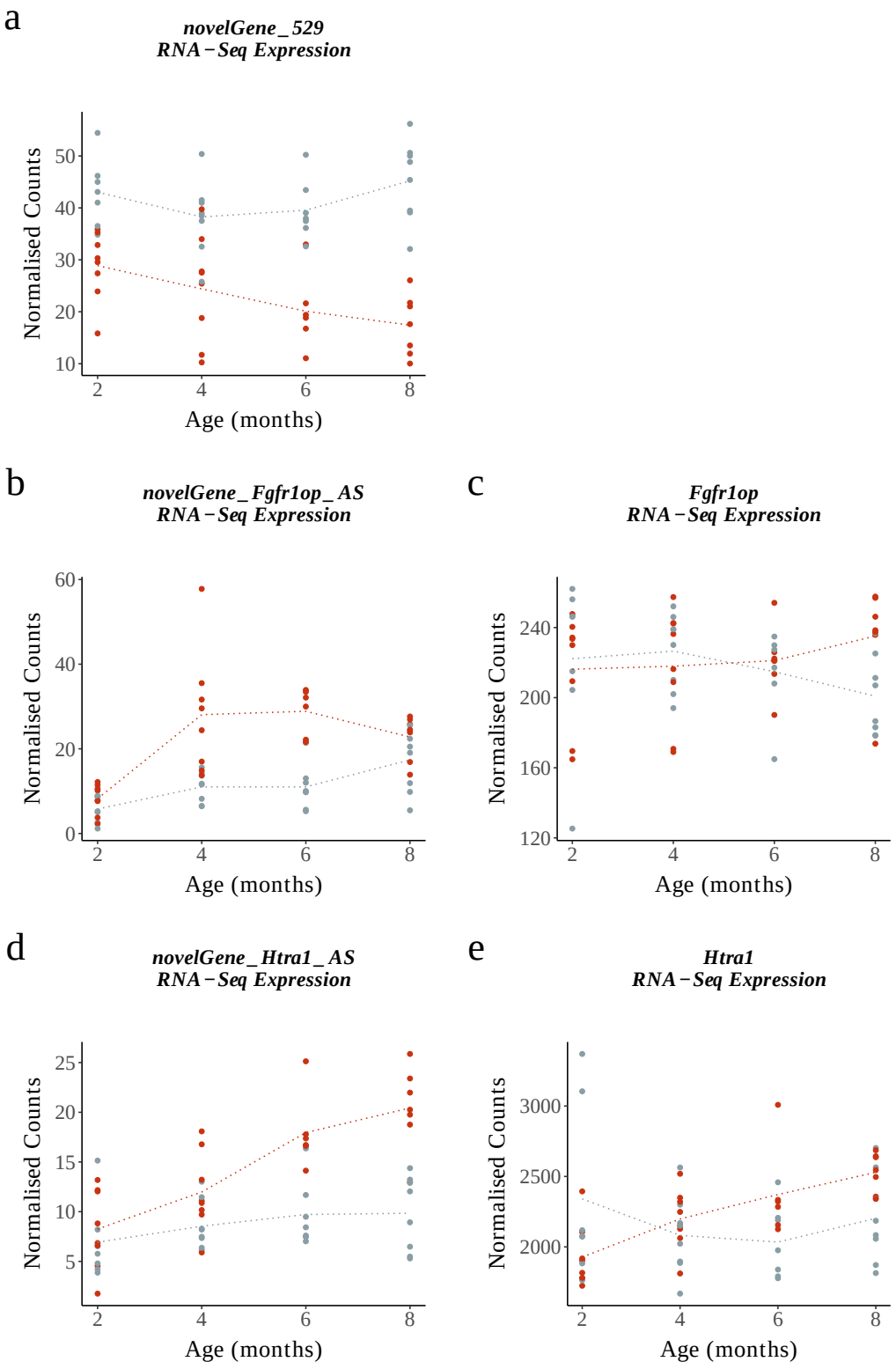


Figure 5.8: Novel isoforms were less expressed, longer and had more exons than known isoforms: Shown is the **a)**

5.3.5 Differential Isoform Expression Analysis

One of the added advantages of long reads over short-reads is the power to accurately identify isoforms and to explore isoform expression changes across conditions and over time. Using *MaSigPro* with Iso-Seq reads as both reference annotation and expression, we identified hundreds of differentially expressed isoforms (n = 582 isoforms), of which 36 isoforms (6.19%) were associated with rTg4510 genotype, and 378 isoforms (64.9%) associated with progressive tau pathology.

Robust changes in isoform expression associated with progressive tau pathology were identified with Iso-Seq annotation and expression

Strikingly, the two most significant progressive-tau-associated differentially expressed isoforms 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, which was significantly up-regulated with progressive tau pathology (**Figure 5.9a**, **Figure 5.10a**), indicating that increased *Gfap* (**Figure 5.6a**) and *C4b* (**Figure 5.6c**) gene expression in rTg4510 mice were primarily driven by one associated isoform. Notably, expression of other minor novel isoforms was also higher in rTg4510 mice aged 8 months (**Figure 5.9b**, **Figure 5.10b**). A similar finding was also observed when using RNA-Seq reads as abundance, though tau-pathology associated expression changes in minor isoforms were significantly more pronounced - a reflection of the comparably higher sequencing coverage of RNA-Seq reads to Iso-Seq reads.

Other isoforms that were significantly unregulated with progressive tau pathology were annotated to genes that have been previously strongly implicated in AD pathology and development (**Figure 5.11**). This included ENSMUST00000151120.8 associated with *Ctsd*, which encodes for Cathepsin D, a lysosomal protease that is involved in degradation of Aβ,[90] tau,[91] and has recently been identified as a key regulator of Aβ42/40 ratio.[92] Other isoforms include ENSMUST00000172785.7 associated with *H2-D1* - that encodes for major histocompatability complex (MHC) class 1, an immune-related gene that has been found to be upregulated in microglial cells of a different mouse model of neurodegeneration with AD-like phenotypes[93] - ENSMUST00000028624.8 associated with *Gatm*, a mitochondrial protein that has been recently revealed as a key protein signature of AD from a large proteomic analysis of human

cortex and CSF,[94] - and ENSMUST00000030765.6 associated with *Padi2*/*Pad2*, an enzyme that has been found abnormally activated in astrocytes of patients with AD.[95]

Iso-Seq reads can accurately quantify highly-expressed isoforms and identify significant expression changes in isoforms annotated to highly-expressed genes

Observations of differential isoform expression using Iso-Seq reads as expression were recapitulated with RNA-Seq reads, highlighting the power to accurately quantify isoforms at high expression (**Figure 5.12**). However, on closer examination, the majority of isoforms identified as differentially expressed with progressive tau pathology (n = 321, 84.9%) were not similarly identified as differentially expressed when RNA-Seq reads were used as expression. Given that the wide majority of Iso-Seq-identified-differentially-expressed isoforms were very-lowly expressed (295 isoforms,91.9%, with mean full-length counts < 24 FL, n = 12 samples) with low read count, a seemingly significant expression change may result in calling the isoform differentially expressed (**Figure 5.13**). However, even for more highly-expressed isoforms, while a change in mean expression was observed, the variance was substantial due to a small sample size (n = 3 replicates). Conversely, with a greater sample size (n = 6 replicates) and higher sequencing coverage of RNA-Seq reads, the usage of RNA-Seq reads as expression reduced the probability of calling an isoform as differentially expressed due to chance (**Figure 5.14**).

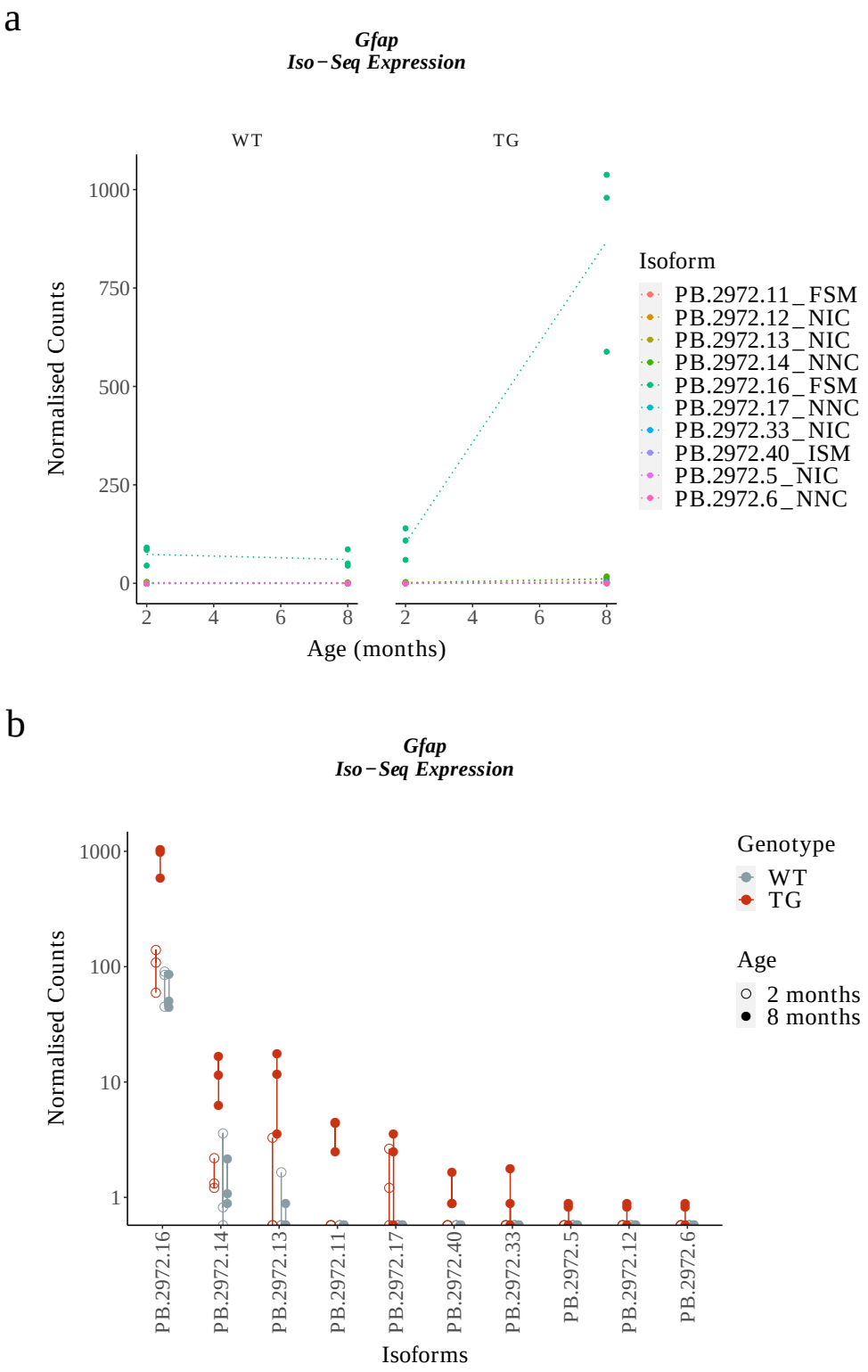


Figure 5.9: Significant upregulation of known isoform of *Gfap* with progressive tau pathology. Shown are **a)** normalised counts of *Gfap* with associated isoforms in WT and TG across 2 time points (aged 2 and 8 months) for emphasis of dominant isoform (PB.2972.16, ENSMUST00000067444.9), and **b)** alternative plot of *Gfap* with y-axis log-transformed for better visualisation of minor isoform expression. Novel minor isoforms, PB.2972.13 and PB.2972.14, were also identified as differentially expressed. Normalised counts of isoforms are derived directly from Iso-Seq full-length reads. FSM - Full Splice Match, ISM - Incomplete Splice Match, NIC - Novel In Catalogue, NNC - Novel Not in Catalogue. WT - Wildtype, TG - Transgenic. Dotted lines represent the mean paths across ages.

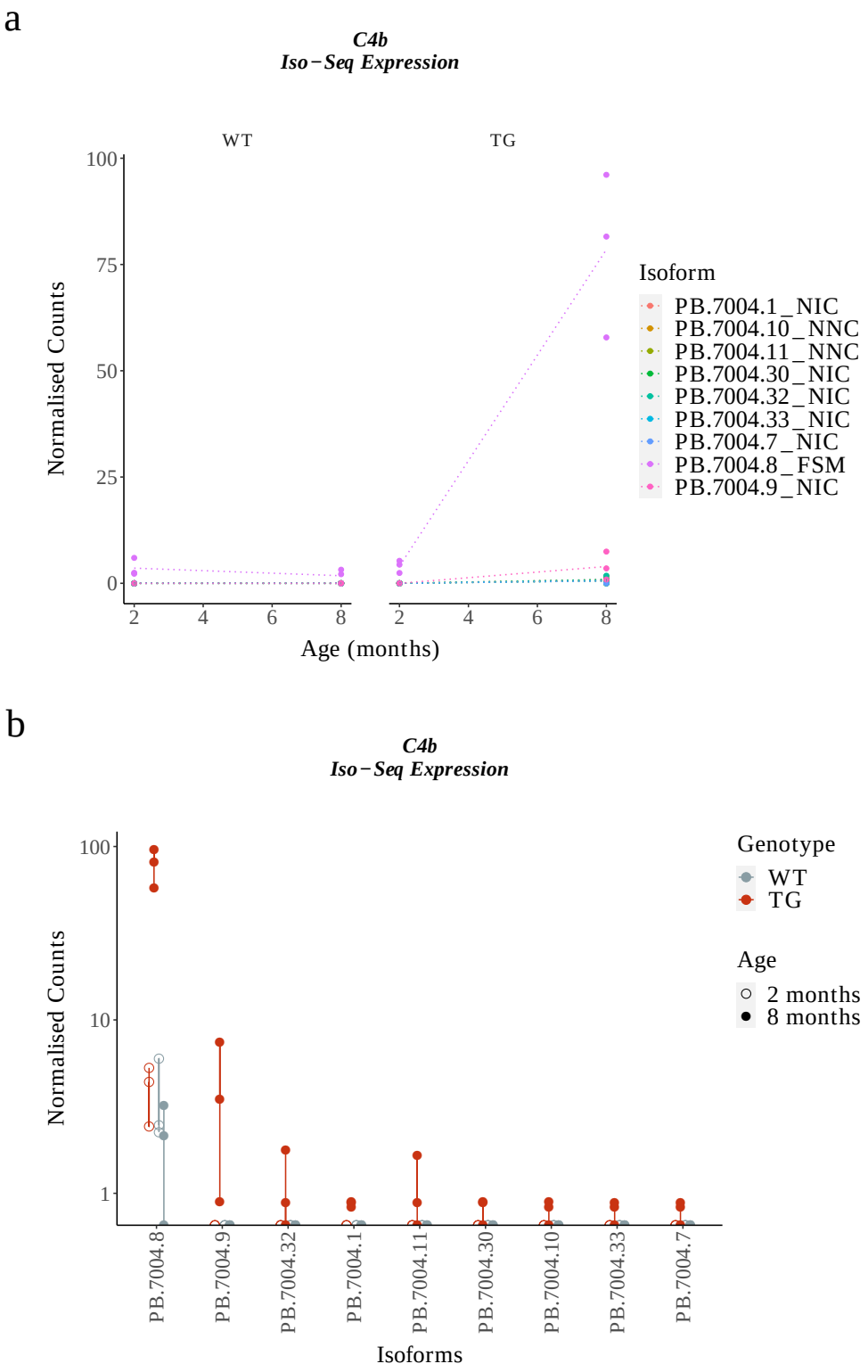


Figure 5.10: Significant upregulation of known isoform of *C4b* with progressive tau pathology. Shown are **a)** normalised counts of *C4b* with associated isoforms in WT and TG across 2 time points (aged 2 and 8 months) for emphasis of dominant isoform (PB.7004.8, ENSMUST00000069507.8), and **b)** alternative plot of *C4b* with y-axis log-transformed for better visualisation of minor isoform expression. Normalised counts of isoforms are derived directly from Iso-Seq full-length reads. FSM - Full Splice Match, ISM - Incomplete Splice Match, NIC - Novel In Catalogue, NNC - Novel Not in Catalogue. WT - Wildtype, TG - Transgenic. Dotted lines represent the mean paths across ages.

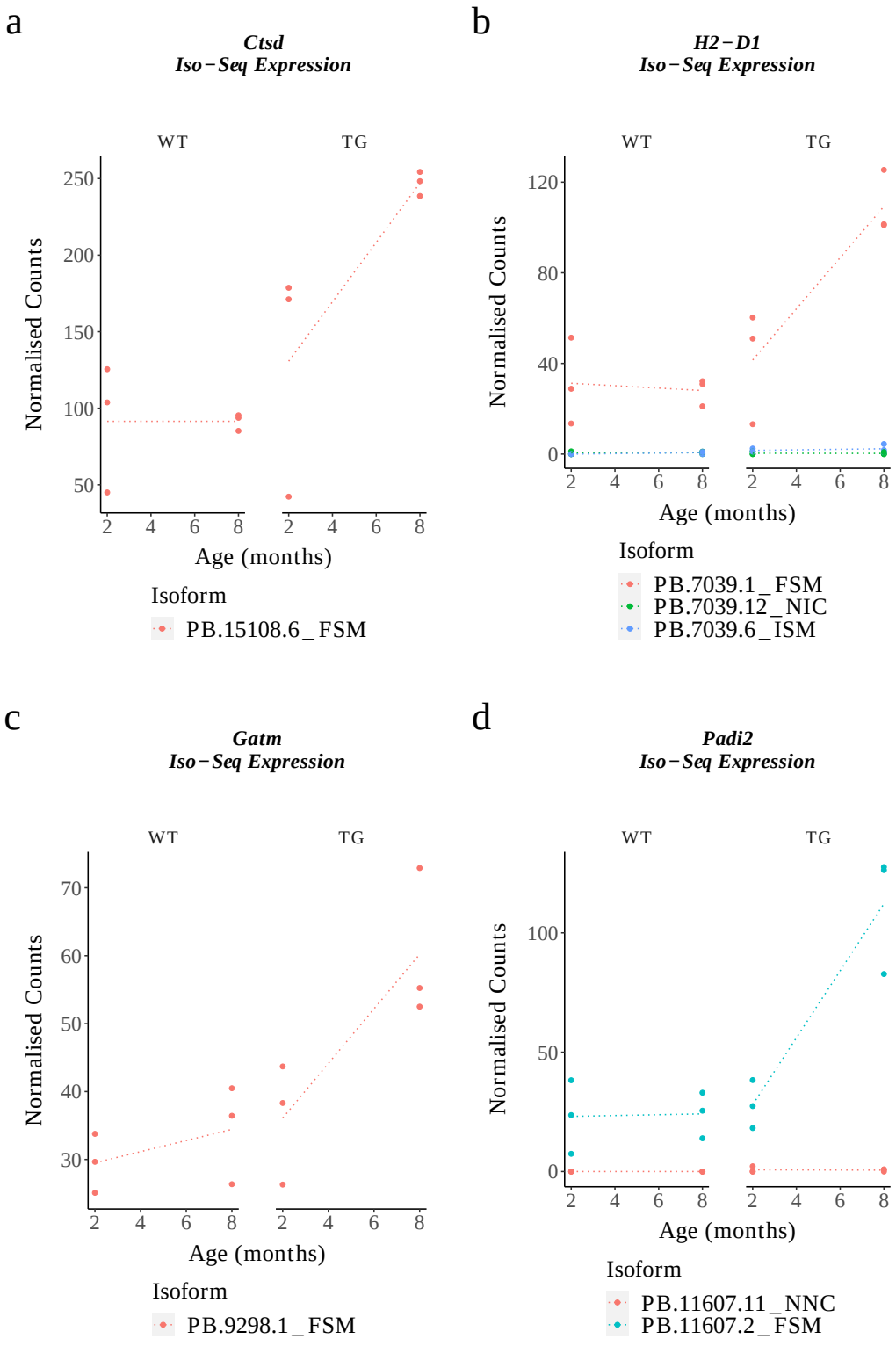


Figure 5.11: Robust changes in transcript expression of isoforms annotated to genes that are strongly implicated in AD. Using Iso-Seq reads (n = 6 WT, n = 6 TG, across 2 and 8 months) as annotation and expression, differential isoform expression was observed for **a)** ENSMUST00000151120.8 (PB.15108.6) annotated to *Ctsd*, **b)** ENSMUST00000172785.7 (PB.7039.1) annotated to *H2-D1*, **c)** ENSMUST00000028624.8 (PB.9298.1) annotated to *Gatm*, and **d)** ENSMUST00000030765.6 (PB.11607.2) associated with *Padi2*/*Pad2*. WT - Wildtype, TG - Transgenic. Dotted lines represent the mean paths across ages.

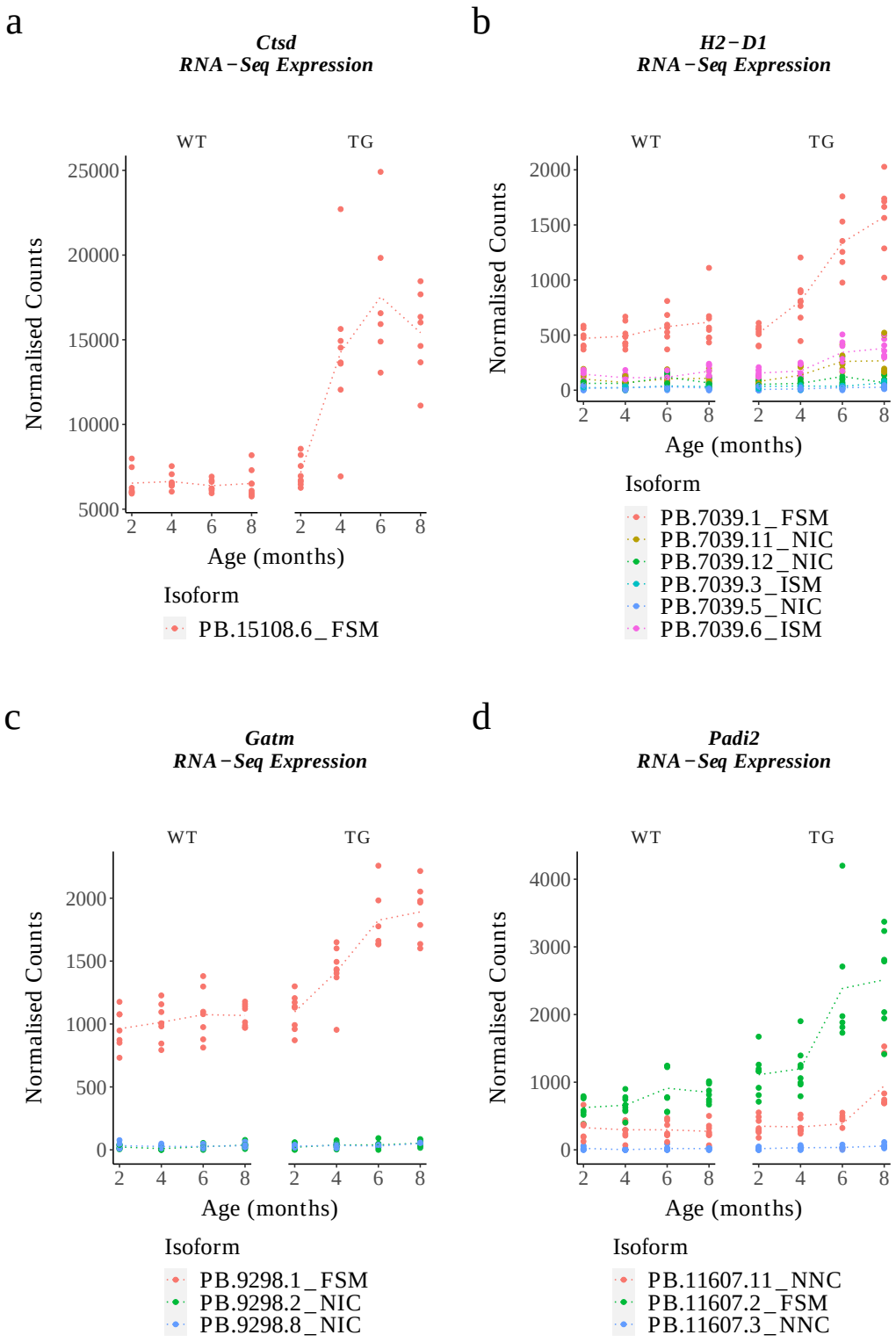


Figure 5.12: Changes in transcript expression of genes strongly implicated in AD were similarly detected using RNA-Seq reads. Usage of RNA-Seq reads (n = 30 WT, n = 29 TG, across 2, 4, 6 and 8 months) as expression similarly identified isoforms as differentially expressed (**Figure 5.11**). **a)** ENSMUST00000151120.8 (PB.15108.6) annotated to *Ctsd*, **b)** ENSMUST00000172785.7 (PB.7039.1) annotated to *H2-D1*, **c)** ENSMUST00000028624.8 (PB.9298.1) annotated to *Gatm*, and **d)** ENSMUST00000030765.6 (PB.11607.2) associated with *Padi2*/*Pad2*. WT - Wildtype, TG - Transgenic. Dotted lines represent the mean paths across ages.

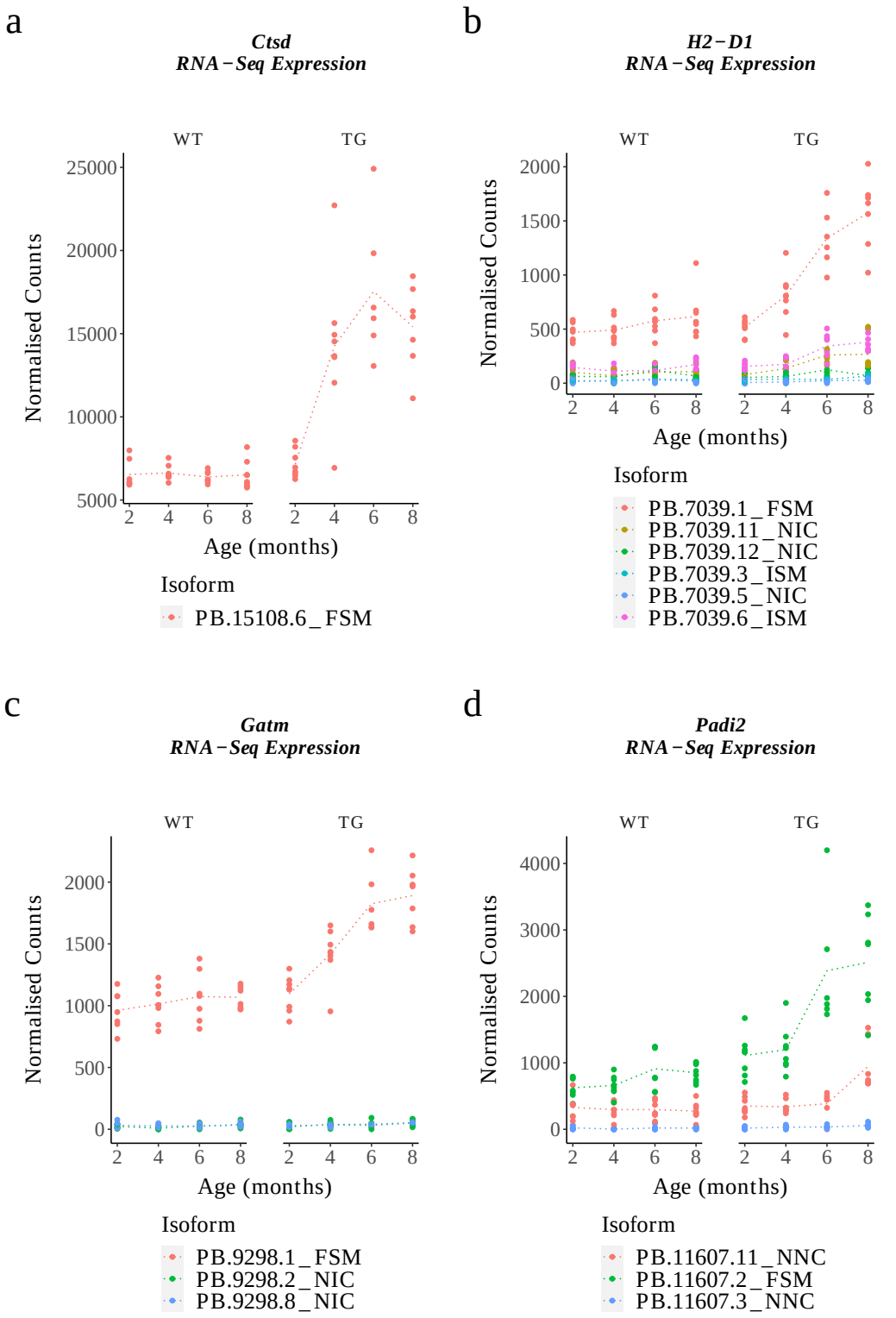


Figure 5.13: Differential isoform expressed observed with Iso-Seq reads as expression were not recapitulated using RNA-Seq reads. Shown are normalised counts **a)** *Cd34* and **c)** *Ubqln1* with associated isoforms identified from using Iso-Seq reads as expression, and of the same two genes **b)** *Cd34* and **d)** *Ubqln1* identified from using RNA-Seq reads as expression. Significant changes in isoform expression identified using Iso-Seq reads as expression - PB.1063.2 associated with *Cd34* and PB.4255.13 and PB.4255.4 associated with *Ubqln1* - was not recapitulated when using RNA-Seq reads as expression, due to lower sequencing coverage. Notably, minor isoforms had a higher expression with RNA-Seq alignment than direct detection with Iso-Seq reads. FSM - Full Splice Match, ISM - Incomplete Splice Match, NIC - Novel In Catalogue, NNC - Novel Not in Catalogue. Dotted lines represent the mean paths across ages.

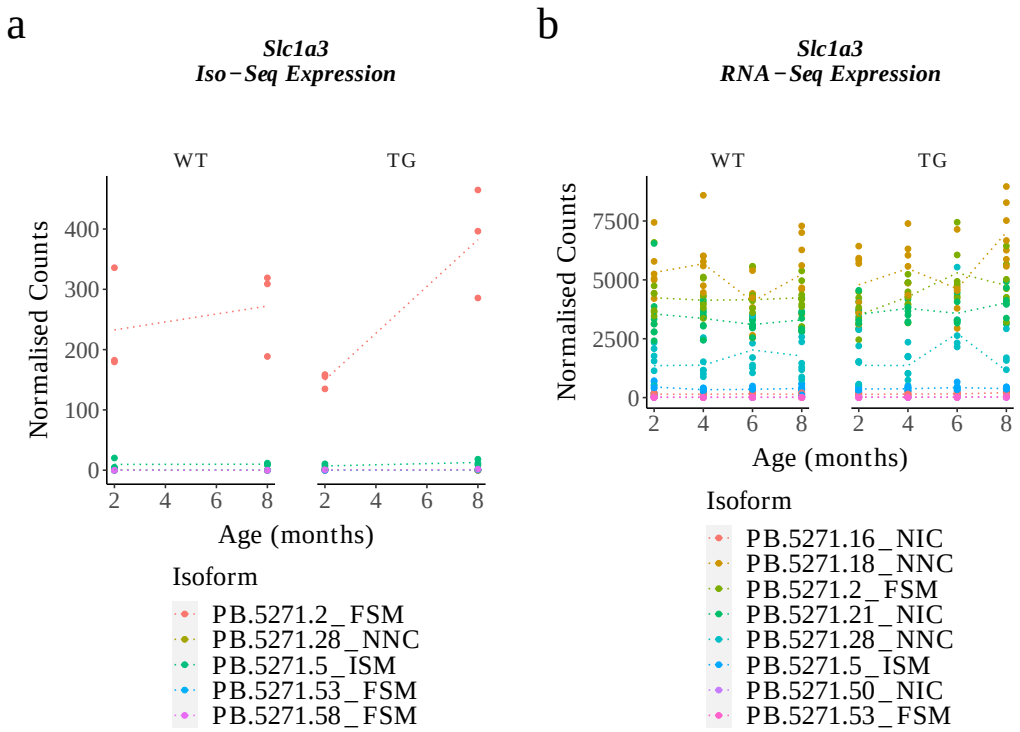


Figure 5.14: Usage of RNA-Seq reads as expression, with larger sample size, reduce probability of calling isoforms differentially expressed due to chance Shown are **a)** *Slc1a3* with associated isoforms identified as differentially expressed when Iso-Seq reads were used as expression (n = 6 WT, n = 6 TG, across 2 time points), and **b)** *Slc1a3* when RNA-Seq reads were used as expression (n = 30 WT, n = 29 TG, across 4 time points).

Using Iso-Seq reads as expression, the more-highly expressed isoforms PB.527.1 (red) in Figure a) was identified as differentially expressed. The expression change, however, of the same isoform was less prominent when RNA-Seq reads were used due to greater sample size and higher sequencing coverage. Notably, all the minor isoforms had a higher expression when RNA-Seq reads were used, resulting in some novel isoforms not being pre-filtered (as in the case with Iso-Seq reads as expression due to low full-length read count)

FSM - Full Splice Match, ISM - Incomplete Splice Match, NIC - Novel In Catalogue, NNC - Novel Not in Catalogue. Dotted lines represent the mean paths across ages.

5.3.6 Differential Isoform Usage Analysis

Contributing to the complexity of transcript regulation through alternative splicing, while the expression of a gene may be constant between conditions, the relative expression of the isoforms (and thus isoform proportion) can change. This phenomenon is known as differential transcript/isoform usage (DIU), and was also assessed using *tappAS* for genes with more than one isoform.

Usage of Iso-Seq reads results in many false positives due to insufficient sequencing depth

Using Iso-Seq reads for annotation and expression and after filtering lowly-expressed isoforms, we identified 400 genes that were characterised with differential isoform usage. However upon further examination, the majority of these genes were lowly expressed and the associated isoforms that were observed to undergo differential usage had only 1-2 full-length long-read counts. We were therefore not confident in any of observed DIU changes detected with Iso-Seq abundance from the whole transcriptome sequencing due to low sequencing depth. Indeed, none of the DTU genes were significant after applying a gene expression threshold (described in **Section 5.2.4**).

In further support of this conclusion, there was a low overlap of DIU genes that were identified when we used RNA-Seq reads as expression. This was likely to be a reflection of lower sequencing depth of Iso-Seq reads, resulting in i) a different pool of isoforms after filtering lowly-expressed isoforms - a minor isoform was more likely to be filtered when using RNA-Seq reads as abundance rather than Iso-Seq reads, particularly if the overall gene expression was low, due to relatively greater expression difference, and ii) smaller differences in isoform expression using Iso-Seq reads are translated to misleading significant changes in isoform proportions, resulting in false detection of genes as DIU. These implications can be seen in the gene *Esyt2*, whereby the low Iso-Seq isoform counts resulted in misleading representation of isoform fraction, which was not recapitulated when RNA-Seq reads were used as expression (**Figure 5.16**).

Alignment of RNA-Seq reads to Iso-Seq defined transcriptome identified multiple

genes with differential transcript usage with major switching events

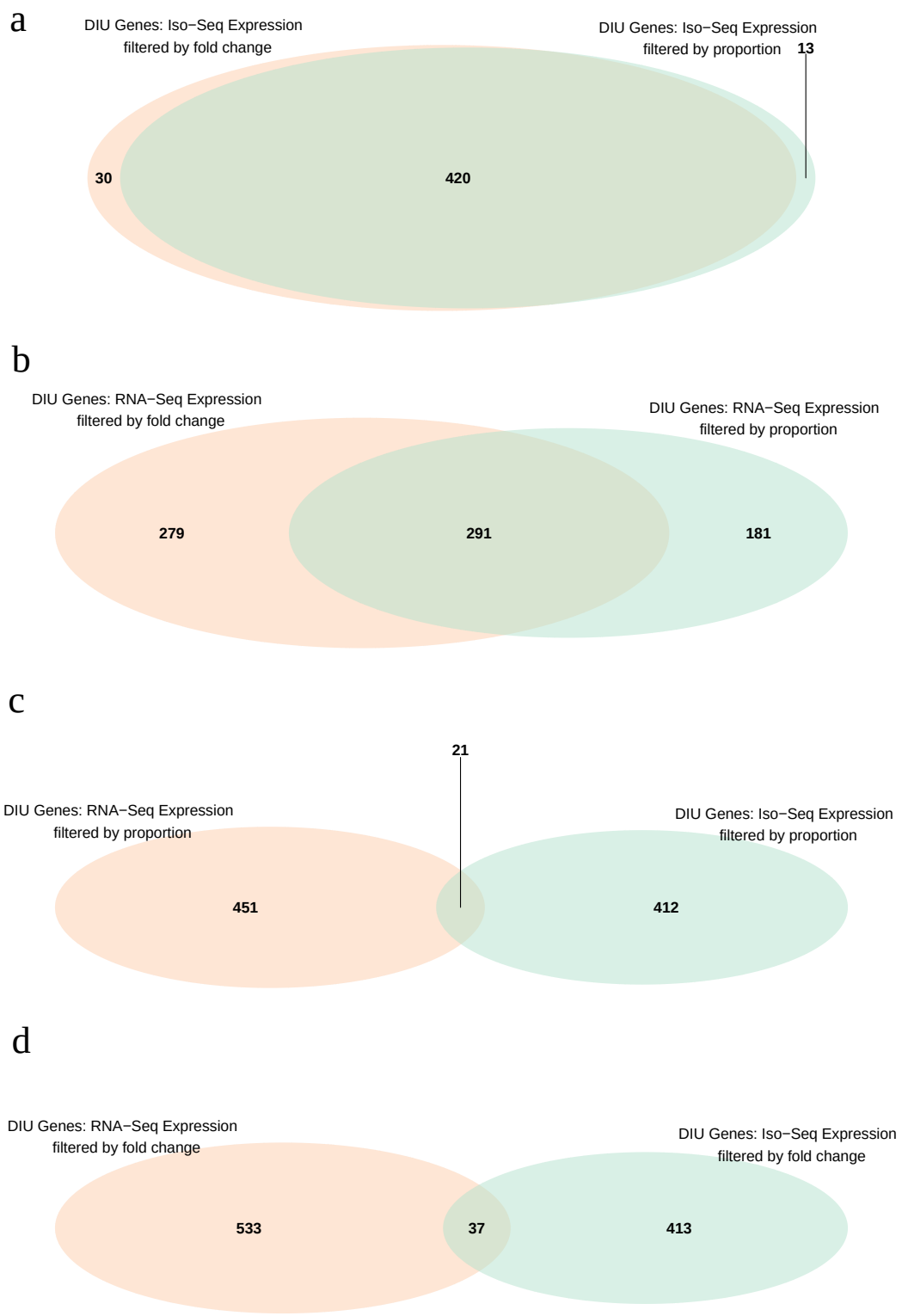


Figure 5.15: Comparison of number of differentially expressed genes identified from Whole Transcriptome datasets after using RNA-Seq or Iso-Seq reads as abundance and various strategies to filter lowly-expressed isoforms: Shown are venn diagrams that encapsulate the number of genes observed with significant differentially isoform usage, if **a)** Iso-Seq reads were used as abundance, and lowly expressed isoforms were pre-filtered either by relative proportion or fold change to major isoform, **b)** RNA-Seq reads were used as abundance, **c)** if relative proportion or **d)** relative fold change to major isoform, was chosen as strategy for filtering lowly-expressed isoforms

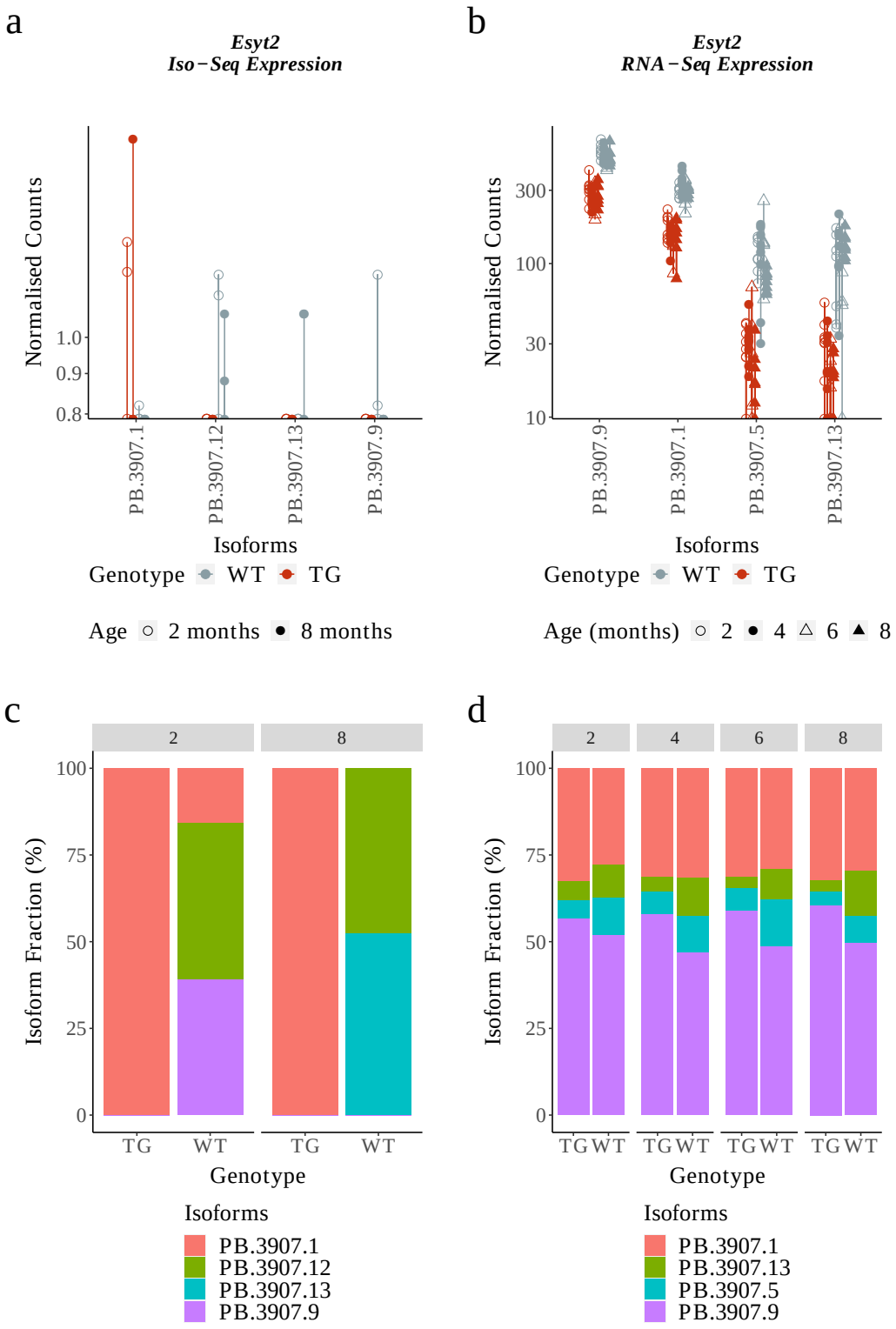


Figure 5.16: *Esyt2* was misidentified with differential isoform usage due to low Iso-Seq read counts: **a)** Isoform expression (normalised counts) and **c)** subsequent deduction of isoform fraction of *Esyt2* using Iso-Seq reads as expression, and the equivalent **b)** isoform expression and **d)** isoform fraction of the same gene, *Esyt2* with RNA-Seq reads as expression. Iso-Seq reads were used as annotation in both analyses.
As can be observed, the Iso-Seq normalised counts are significantly lower than RNA-Seq normalised counts for the associated isoforms, resulting in a misleading representation of isoform fraction with major isoform switching events (Figure c, PB.3907.1 becomes the dominant isoform in transgenic mice), which is not recapitulated with RNA-Seq reads as expression (Figure d)

5.3.7 Differential Feature Inclusion Analysis

Chapter 6

BDR

Chapter 7

Conclusion

Appendix

Appendix A

Iso-Seq Targeted and Whole Transcriptome Protocol

A.0.1 Requirement of Sample quality

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

- Double stranded DNA sample (dsDNA) generated from cDNA synthesis of extracted RNA
- Minimum freeze thaw cycles
- No exposure to high temperature (>65) or pH extremes (<6, >9),
- 1.8 - 2 OD260/280, and 2.0 - 2.2 OD260/230
- No insoluble material
- No RNA contamination or carryover contamination (e.g polysacharides)
- No exposure to UV or interacalating fluorescent dyes
- No chelating agents, divalent metal cations, denaturants or detergents

A.0.2 General

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

A.0.2.1 Ampure Bead Purification

Throughout the protocol, DNA is purified using ampure 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 the AMPure beads for use by allowing to equilibrate to room temperature for a minimum of 15minutes. 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 particularly for 0.4X ampure beads - too high concentration would result in retainment of undesired short inserts, too low 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 (for 1 second) to collect beads
6. 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).
    - 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, as otherwise 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 otherwise result in low yield due to difficulties during sample elution
13. Elute with specified amount of PacBio Elution Buffer (differs pending on the 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.0.2.2 Assessment of DNA quantity using Qubit

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

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

1. Set up and label the required number of Qubit assay tubes (0.5mL) for samples and 2 samples.
    - 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.0.2.3 Assessment of DNA library size using Tapestation or Bioanalyzer

Also as part of quality control across the various stages of library preparation in conjunction to performing Qubit assay, run DNA using D5000 ScreenTape or DNA 12000 Assay (Agilent), following manufacterer'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.0.3 First Strand Synthesis

1. For each sample, add 200ng of RNA with 1μL of barcoded/non-barcoded polyT primer in a micro centrifuge on ice (Table X), 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 X 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 90minutes, followed by 70°C for 10minutes
7. Add 90μL of PacBio Elution Buffer (EB) to each RNA reaction tubes: diluted first-strand cDNA (Table A.1)

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 Pol (1.25U/μL)	1
Total Volume per sample	45

Table A.1: Long Description

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

Table A.2: PCR conditions for cDNA synthesis

A.0.4 PCR Cycle Optimisation

1. Prepare a PCR reaction mix (Table X), scaled 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 X using 105°C heated lid
    - At cycles 10, 12, 14, 16 and 18, take 5μL from reaction tubes and transfer to new micro centrifuge tube
    - Flick and spin down reaction tubes, before returning them back to thermo cycler to continue for incubation
5. Run 5μL of cDNA from each sample and cycle on a 1% agarose gel (Section X) at 110V for 20minutes with 1μL 100bp ladder
    - Note: input of 5uL of cDNA rather than 10uL, 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 X)

A.0.4.1 Running an agarose gel

1. 1.5mg of agarose was weighed and placed into a beaker containing 100ml 1X TBE buffer
2. Beaker was microwaved for 10-20 seconds until the solution appears clear, and allowed to cool for 2-3 minutes
3. 1.75uL of ethidium bromide was added to beaker, and mix was poured into a casket
4. Gel was cooled for 20minutes

A.0.5 Large-Scale PCR

1. Set up and label 16 micro centrifuge tubes for each sample
2. Prepare a PCR reaction mix for each sample in 1.5mL LoBind eppendorf (Table A.3)
3. Add 50μL of respective diluted cDNA to each PCR reaction mix
    - Note: input of 50μL of cDNA rather than 100uL, 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.4

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 Pol (1.25U/μL)	16
Total Volume per sample for 16 PCR reactions	750

Table A.3: 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

Table A.4: PCR conditions for Large Scale PCR

A.0.6 Bead Purification of Large-Scale PCR Products

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.
2. Important to pipette exactly 500μL of PCR reactions and 200μL of AMPure PB magnetic beads as otherwise risk of significant DNA loss
3. Pool remaining PCR reactions and add 1X volume of AMPure PB magnetic beads. This is Fraction 1. Note: Inevitable sample loss through evaporation (20μL), therefore would not be able to recover 800μL of cDNA
4. Proceed with AMPure PB Bead Purification (Section X), with 100μL of EB to Fraction 1 and 22μL EB to Fraction 2
5. 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.0.6.1 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 **??**), with 22μL of EB to Fraction 1
3. Quantify DNA amount and concentration of Fraction 1 and Fraction 2 using a high-sensitivity Qubit (Section X)
4. Determine library size using the BioAnalyzer with DNA 12000 Kit (Section 2.1.2.7)

A.0.7 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:

A minimum 200ng of pooled cDNA is necessary for library construction, despite the minimum recommended 1ug in protocol. If performing target capture, proceed to "Target Capture with IDT Probes" below, otherwise skip to "SMRTbell Template Preparation".

A.0.8 Target Capture using IDT probes

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

1. Add 1 – 1.5 µg cDNA to a 0.2mL PCR tube
2. Add 1 μL of SMARTer PCR oligo and 1 µL PolyT 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/PolyT blocker completely in a LoBind tube using a DNA vacuum concentrator (speed vac)
    - Place the 0.2mL PCR Tube in a 1.5mL Eppendorf. Do not leave tubes in the speed vac 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 30min-

utes)

5. To the dried-down sample, add reagents listed in Table X
6. Cut off the punctured lid and replace with new PCR lid. Ensure fully sealed.
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 4 μL of xGen Lockdown Panel/Probe for a total volume of 17 μL. Mix and quick spin.
11. Leave the PCR tube at room temperature for 5minutes
12. Incubate in a thermo cycler at 65°C for 4 hours, lid set at 100°C

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.0.8.1 Prepare beads for 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 X
3. Aliquot 200μL of 1x Wash Buffer (Tube1) to new 1.5ml Eppendorf
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 1.5 mL 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 (2minutes). 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 the magnetic rackVortex/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 (2minutes). 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.

A.0.8.2 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 65C.

1. Transfer 17μL hybridized probe/sample mixture prepared in the "Preparing hybridization section" 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 45minutes, lid set at 70°C
    - Every 10-12minutes, 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 low-bind Eppendorf at 65°C for later transfer of sample
5. Preheat the following wash buffers to +65 degrees in water bath: 200μL of 1x Wash

Buffer (Tube 1), 500μL of 1x Stringent Wash Buffer (Tube S)

6. Proceed immediately to Heated Washes

A.0.8.3 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 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 (1minute)
    - 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 (TubeS) to +65°C
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.

A.0.8.4 Perform room temperature washes

1. Wash beads with 200μL of room temperature 1X Wash Buffer I (Tube1)
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 30secs and resting for 30secs, 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 (1minute). When clear, remove and discard supernatant
5. Wash beads with 200μL of room temperature 1X Wash Buffer II (Tube2)
6. Repeat steps 2 - 4
7. Wash beads with 200μL of room temperature 1X Wash Buffer III (Tube3)
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. If forgot, place tube back on magnetic rack, remove supernatant and re-elute with elution buffer.
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 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.0.8.5 Amplification of Captured DNA Sample

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

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

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.0.9 SMRTbell Template Preparation

A.0.9.1 Repair DNA Damage and Ends

1. Preparation a PCR reaction mix in a 1.5mL LoBind eppendorf (Table X)
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

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.0.9.2 DNA Purification

1. Proceed with AMPure PB Bead Purification (Section **??**), with 1X volume of AMPure Beads (50μL) and eluting with 32μL of EB
2. The End-Repaired DNA can be stored overnight at 4°C (or -20°C for longer)

A.0.9.3 Prepare Blunt Ligation Reaction

1. Add the following reagents in Table X 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 24hours)
4. Incubate at 65°C for 10minutes to inactivate the ligase, returning reaction to 4°C. Proceed with adding exonuclease.

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.0.9.4 Adding Exonuclease to remove failed ligation products

1. Add 1μL of Exonuclease III to pooled cDNA (ligated)
2. Add 1μL of Exonuclease VII to pooled cDNA (ligated)
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.0.9.5 First Purification of SMRTbell Templates

1. Proceed with AMPure PB Bead Purification (Section **??**), with 1X volume of AMPure Beads (42μL) and eluting with 50μL of EB

A.0.9.6 Second Purification of SMRTbell Templates

1. Proceed with AMPure PB Bead Purification (Section **??**), with 1X volume of AMPure Beads (50μL) and eluting with 10μL of EB

2. Quantify DNA amount and concentration of Fraction 1 and Fraction 2 using a high-sensitivity Qubit (Section 2.1.2.8)
3. Determine library size using the BioAnalyzer with DNA 12000 Kit (Section 2.1.2.7)

Appendix B

Oxford Nanopore Transcriptome Protocol

This protocol was adapted from Wellcome Trust Advanced Course: RNA Transcriptomics (2018), provided by J.Ragoussis (referred as WTAC), the official ONT protocol "1D amplicon/cDNA by Ligation (SQK-LSK109)", and directed under the guidance of K.Moore, Exeter's sequencing services. In brief, this protocol aimed to complement the Iso-Seq Protocol (Section **??**) as a direct comparison of the two sequencing technologies. It was therefore important to ensure that all other steps, bar library preparation, were consistent (Figure X). Consequently, cDNA synthesis and amplification 2.1.2.1 was performed twice in parallel for the sample of interest, and the pipeline branched upon the respective library preparation.

B.0.1 cDNA Synthesis and Amplification

For a direct comparison of ONT's minion sequencing and PacBio's Iso-Seq approach, the same methods for cDNA synthesis and amplification in the Iso-Seq protocol were used (Section **??** - **??**). There were attempts to perform cDNA synthesis and amplification from WTAC's protocol, particularly as it used the capped-dependent Teloprime kit (Appendix X). However, there were difficulties in achieving sufficient yield for downstream library preparation, in addition to complicating downstream comparative analyses.

Rationale for repeating ClonTech 2x and then pooling: Ideal scenario would be to use 400ng of Total RNA and then dilute in 180ul (rather than 90ul as per protocol). Concentration of diluted cDNA would be same as in previous experiments (200ng diluted in 90ul), therefore expect similar number of PCR cycles; only difference is with more diluted cDNA, able to split cDNA products for PacBio and ONT protocols (90ul each); Unfortunately, due to low concentration of starting Total RNA, not able to reverse-transcribe 400ng of total RNA. One other possible solution is to dilute 200ng in 180ul before proceeding with PCR cycle optimisation and large scale amplification. However, this would result in more PCR cycles required, resulting in more PCR bias and errors etc.

B.0.2 Bead Purification of Large Scale PCR Products

1. Pool 800μL PCR reactions (16 x 50μL PCR reactions) and add 0.90X volume of AMPure PB (200μL) magnetic beads.
2. 20-30μL loss is expected from evaporation,therefore would not be able to recover 800μL of cDNA . Note: only prepare 1 Fraction for downstream library preparation rather than 2 Fractions in Iso-Seq
3. Proceed with AMPure PB Bead Purification (Section X), with 51μL of TE Buffer
4. Quantify DNA amount and concentration of using a high-sensitivity Qubit (Section X)
5. Determine library size using the BioAnalyzer with DNA 12000 Kit (Section 2.1.2.7)

B.0.3 ONT MinION Library Preparation

B.0.3.1 Repair DNA and Ends

1. Thaw DNA CS (DCS) at room temperature, spin down, mix by pipetting, 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 cycle at 20° C for 5 minutes and 65° C for 5 mins

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	60

Table B.1: Repair DNA and Ends]

B.0.3.2 Bead Purification of cDNA end-repaired products

1. Proceed with AMPure PB Bead Purification (Section A.0.2.1), with 1X of AMPure Beads and elute with 61μL of nuclease-free water
    - Incubate on a Hula mixer (rotator mixer) for 5 minutes at room temperature, rather than shaking in a VWR vortex mixer at 2000rpm for 10 minutes at room temperature

B.0.3.3 Prepare Ligation Reaction

1. Prepare the following reagents:

- 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 1.5 ml Eppendorf DNA LoBind tube
3. Mix gently by flicking the tube, and spin down
4. Incubate the reaction for 10 minutes at room temperature (up to 4hrs)

B.0.3.4 Bead Purification of ligated cDNA

1. Prepare the AMPure beads for use by allowing to equilibrate to room temperature for a minimum of 15minutes. 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 (for 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 Lo-Bind tube. Avoid disturbing beads.
14. Quantify DNA amount and concentration of Fraction 1 and Fraction 2 using a high-sensitivity Qubit (Section X). Determine library size using the BioAnalyzer with DNA 12000 Kit (Section 2.1.2.7)

B.0.4 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 200 μl
    - 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.0.5 Library loading into the Flow Cell

1. Prepare sample with for library as in Table
2. Gently lift the SpotON sample port cover to make the SpotON sample port accessible.
3. Load 200 μl 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 75 μl of sample to the flow cell via the SpotON sample port in a dropwise 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.

Reagents	Volume (μl)
Sequencing Buffer (SQB)	37.5
Loading Buffer (LB), mixed immediately before use	25.5
DNA library	12
Total	75

Table B.2: Loading Flow Cells

Appendix C

cDNA Synthesis alternative approach

Bibliography

[1] Jan Verheijen and Kristel Sleegers. Understanding Alzheimer Disease at the Interface between Genetics and Transcriptomics, 2018.

[2] Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Nadav S. Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J. Luo, Eddie Park, Kimberly Persaud, Jonathan B. Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guig, and Thomas R. Gingeras. Landscape of transcription in human cells. Nature, 489(7414):101–108, sep 2012.

[3] Beryl Cummings, Jamie Marshall, Taru Tukiainen, Monkol Lek, Sandra Donkervoort, A Reghan Foley, Veronique Bolduc, Leigh Waddell, Sarah Sandaradura, Gina O'Grady, Elicia Estrella, Hemakumar Reddy, Fengmei Zhao, Ben Weisburd, Konrad Karczewski, Anne

O'Donnell-Luria, Daniel Birnbaum, Anna Sarkozy, Ying Hu, Hernan Gonorazky, Kristl Claeys, Himanshu Joshi, Adam Bournazos, Emily Oates, Roula Ghaoui, Mark Davis, Nigel Laing, Ana Topf, Peter Kang, Alan Beggs, Kathryn North, Volker Straub, James Dowling, Francesco Muntoni, Nigel Clarke, Sandra Cooper, Carsten Bonnemann, and Daniel MacArthur. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing, page 074153, 2016.

[4] Laura S. Kremer, Daniel M. Bader, Christian Mertes, Robert Kopajtich, Garwin Pichler, Arcangela Iuso, Tobias B. Haack, Elisabeth Graf, Thomas Schwarzmayr, Caterina Terrile, Eliška Koňaříkova, Birgit Repp, Gabi Kastenmüller, Jerzy Adamski, Peter Lichtner, Christoph Leonhardt, Benoit Funalot, Alice Donati, Valeria Tiranti, Anne Lombes, Claude Jardel, Dieter Gläser, Robert W. Taylor, Daniele Ghezzi, Johannes A. Mayr, Agnes Rötig, Peter Freisinger, Felix Distelmaier, Tim M. Strom, Thomas Meitinger, Julien Gagneur, and Holger Prokisch. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nature Communications, 8(1):1–11, jun 2017.

[5] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. Nature, 456(7221):470–476, 2008.

[6] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics, 40(12):1413–1415, 2008.

[7] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Pouya Kheradpour, Timo Lassmann, Qunhua Li, Xinying Lin,

Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephen C.J. Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A.L. Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Peter J. Good, Elise A. Feingold, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Morgan C. Giddings, Thomas R. Gingeras, Roderic Guigó, Timothy J. Hubbard, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, John A. Stamatoyannopoulos, Scott A. Tenenbaum, Zhiping Weng, Kevin P. White, Barbara Wold, Yanbao Yu, John Wrobel, Brian A. Risk, Harsha P. Gunawardena, Heather C. Kuiper, Christopher W. Maier, Ling Xie, Xian Chen, Tarjei S. Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J. Coyne, Timothy Durham, Manching Ku, Thanh Truong, Matthew L. Eaton, Alex Dobin, Andrea Tanzer, Julien Lagarde, Wei Lin, Chenghai Xue, Brian A. Williams, Chris Zaleski, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J. Luo, Eddie Park, Jonathan B. Preall, Kimberly Presaud, Paolo Ribeca, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, Yoshihide Hayashizaki, Alexandre Reymond, Stylianos E. Antonarakis, Gregory J. Hannon, Yijun Ruan, Piero Carninci, Cricket A. Sloan, Katrina Learned, Venkat S. Malladi, Matthew C. Wong, Galt P. Barber, Melissa S. Cline, Timothy R. Dreszer, Steven G. Heitner, Donna Karolchik, Vanessa M. Kirkup, Laurence R. Meyer, Jeffrey C. Long, Morgan Maddren, Brian J. Raney, Linda L. Grasfeder, Paul G. Giresi, Anna Battenhouse, Nathan C. Sheffield, Kimberly A. Showers, Darin London, Akshay A. Bhinge, Christopher Shestak, Matthew R. Schaner, Seul Ki Kim, Zhuzhu Z. Zhang, Piotr A. Mieczkowski, Joanna O. Mieczkowska, Zheng Liu, Ryan M. McDaniell, Yunyun Ni, Naim U. Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Vishwanath R. Iyer, Meizhen Zheng, Ping

Wang, Jason Gertz, Jost Vielmetter, E. Christopher Partridge, Katherine E. Varley, Clarke Gasper, Anita Bansal, Shirley Pepke, Preti Jain, Henry Amrhein, Kevin M. Bowling, Michael Anaya, Marie K. Cross, Michael A. Muratet, Kimberly M. Newberry, Kenneth McCue, Amy S. Nesmith, Katherine I. Fisher-Aylor, Barbara Pusey, Gilberto DeSalvo, Stephanie L. Parker, Sreeram Balasubramanian, Nicholas S. Davis, Sarah K. Meadows, Tracy Eggleston, J. Scott Newberry, Shawn E. Levy, Devin M. Absher, Wing H. Wong, Matthew J. Blow, Axel Visel, Len A. Pennachio, Hanna M. Petrykowska, Alexej Abyzov, Bronwen Aken, Daniel Barrell, Gemma Barson, Andrew Berry, Alexandra Bignell, Veronika Boychenko, Giovanni Bussotti, Claire Davidson, Gloria Despacio-Reyes, Mark Diekhans, Iakes Ezkurdia, Adam Frankish, James Gilbert, Jose Manuel Gonzalez, Ed Griffiths, Rachel Harte, David A. Hendrix, Toby Hunt, Irwin Jungreis, Mike Kay, Ekta Khurana, Jing Leng, Michael F. Lin, Jane Loveland, Zhi Lu, Deepa Manthravadi, Marco Mariotti, Jonathan Mudge, Gaurab Mukherjee, Cedric Notredame, Baikang Pei, Jose Manuel Rodriguez, Gary Saunders, Andrea Sboner, Stephen Searle, Cristina Sisu, Catherine Snow, Charlie Steward, Electra Tapanari, Michael L. Tress, Marijke J. Van Baren, Stefan Washietl, Laurens Wilming, Amonida Zadissa, Zhengdong Zhang, Michael Brent, David Haussler, Alfonso Valencia, Nick Addleman, Roger P. Alexander, Raymond K. Auerbach, Suganthi Balasubramanian, Keith Bettinger, Nitin Bhardwaj, Alan P. Boyle, Alina R. Cao, Philip Cayting, Alexandra Charos, Yong Cheng, Catharine Eastman, Ghia Euskirchen, Joseph D. Fleming, Fabian Grubert, Lukas Habegger, Manoj Hariharan, Arif Harmanci, Sushma Iyengar, Victor X. Jin, Konrad J. Karczewski, Maya Kasowski, Phil Lacroute, Hugo Lam, Nathan Lamarre-Vincent, Jin Lian, Marianne Lindahl-Allen, Renqiang Min, Benoit Miotto, Hannah Monahan, Zarmik Moqtaderi, Xinmeng J. Mu, Henriette O'Geen, Zhengqing Ouyang, Dorrelyn Patacsil, Debasish Raha, Lucia Ramirez, Brian Reed, Minyi Shi, Teri Slifer, Heather Witt, Linfeng Wu, Xiaoqin Xu, Koon Kiu Yan, Xinqiong Yang, Kevin Struhl, Sherman M. Weissman, Luiz O. Penalva, Subhradip Karmakar, Raj R. Bhanvadia, Alina Choudhury, Marc Domanus, Lijia Ma, Jennifer Moran, Alec Victorsen, Thomas Auer, Lazaro Centanin, Michael Eichenlaub, Franziska Gruhl, Stephan Heermann, Burkhard Hoeckendorf, Daigo Inoue, Tanja Kellner, Stephan Kirchmaier, Claudia Mueller, Robert Reinhardt, Lea Schertel, Stephanie Schneider, Rebecca Sinn, Beate Wittbrodt, Jochen Wittbrodt, Gaurav Jain, Gayathri Balasundaram, Daniel L. Bates, Rachel Byron, Theresa K. Canfield, Morgan J. Diegel, Douglas Dunn, Abigail K. Ebersol, Tristan Frum, Kavita Garg, Erica Gist, R. Scott Hansen, Lisa Boatman, Eric Haugen, Richard Humbert, Audra K. John-

son, Ericka M. Johnson, Tattyana V. Kutyavin, Kristen Lee, Dimitra Lotakis, Matthew T. Maurano, Shane J. Neph, Fiedencio V. Neri, Eric D. Nguyen, Hongzhu Qu, Alex P. Reynolds, Vaughn Roach, Eric Rynes, Minerva E. Sanchez, Richard S. Sandstrom, Anthony O. Shafer, Andrew B. Stergachis, Sean Thomas, Benjamin Vernot, Jeff Vierstra, Shinny Vong, Hao Wang, Molly A. Weaver, Yongqi Yan, Miaohua Zhang, Joshua M. Akey, Michael Bender, Michael O. Dorschner, Mark Groudine, Michael J. MacCoss, Patrick Navas, George Stamatoyannopoulos, Kathryn Beal, Alvis Brazma, Paul Flicek, Nathan Johnson, Margus Lukk, Nicholas M. Luscombe, Daniel Sobral, Juan M. Vaquerizas, Serafim Batzoglou, Arend Sidow, Nadine Hussami, Sofia Kyriazopoulou-Panagiotopoulou, Max W. Libbrecht, Marc A. Schaub, Webb Miller, Peter J. Bickel, Balazs Banfai, Nathan P. Boley, Haiyan Huang, Jingyi Jessica Li, William Stafford Noble, Jeffrey A. Bilmes, Orion J. Buske, Avinash D. Sahu, Peter V. Kharchenko, Peter J. Park, Dannon Baker, James Taylor, and Lucas Lochovsky. An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414):57–74, sep 2012.

[8] Kasper Karlsson and Sten Linnarsson. Single-cell mRNA isoform diversity in the mouse brain. BMC Genomics, 18(1), 2017.

[9] Cindy L. Will and Reinhard Lührmann. Spliceosome structure and function. Cold Spring Harbor Perspectives in Biology, 3(7):1–2, jul 2011.

[10] Janne J. Turunen, Elina H. Niemelä, Bhupendra Verma, and Mikko J. Frilander. The significant other: Splicing by the minor spliceosome, jan 2013.

[11] M. Aebi, H. Hornig, R. A. Padgett, J. Reiser, and C. Weissmann. Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. Cell, 47(4):555–565, nov 1986.

[12] A. I. Lamond, M. M. Konarska, and P. A. Sharp. A mutational analysis of spliceosome assembly: evidence for splice site collaboration during spliceosome formation. Genes & development, 1(6):532–543, 1987.

[13] Nihar Sheth, Xavier Roca, Michelle L. Hastings, Ted Roeder, Adrian R. Krainer, and Ravi Sachidanandam. Comprehensive splice-site analysis using comparative genomics. Nucleic

Acids Research, 34(14):3955–3967, 2006.

[14] Guillermo E. Parada, Roberto Munita, Cledi A. Cerda, and Katia Gysling. A comprehensive survey of non-canonical splice sites in the human transcriptome, sep 2014.

[15] Lydia Herzel, Diana S.M. Ottoz, Tara Alpert, and Karla M Neugebauer. Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function, 2017.

[16] Rachael E. Workman, Alison D. Tang, Paul S. Tang, Miten Jain, John R. Tyson, Roham Razaghi, Philip C. Zuzarte, Timothy Gilpatrick, Alexander Payne, Joshua Quick, Norah Sadowski, Nadine Holmes, Jaqueline Goes de Jesus, Karen L. Jones, Cameron M. Soulette, Terrance P. Snutch, Nicholas Loman, Benedict Paten, Matthew Loose, Jared T. Simpson, Hugh E. Olsen, Angela N. Brooks, Mark Akeson, and Winston Timp. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nature Methods, 16(12):1297–1305, dec 2019.

[17] Stefan M. Bresson, Olga V. Hunter, Allyson C. Hunter, and Nicholas K. Conrad. Canonical Poly(A) Polymerase Activity Promotes the Decay of a Wide Variety of Mammalian Nuclear RNAs. PLoS Genetics, 11(10):e1005610, oct 2015.

[18] Sean P. Gordon, Elizabeth Tseng, Asaf Salamov, Jiwei Zhang, Xiandong Meng, Zhiying Zhao, Dongwan Kang, Jason Underwood, Igor V Grigoriev, Melania Figueroa, Jonathan S Schilling, Feng Chen, and Zhong Wang. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. PLoS ONE, 10(7), 2015.

[19] Bo Wang, Elizabeth Tseng, Michael Regulski, Tyson A Clark, Ting Hon, Yinping Jiao, Zhenyuan Lu, Andrew Olson, Joshua C Stein, and Doreen Ware. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nature Communications, 7:11708, 2016.

[20] Donald Sharon, Hagen Tilgner, Fabian Grubert, and Michael Snyder. A single-molecule long-read survey of the human transcriptome. Nature Biotechnology, 31(11):1009–1014, 2013.

[21] Allison Piovesan, Maria Caracausi, Francesca Antonaros, Maria Chiara Pelleri, and Lorenza Vitale. GeneBase 1.1: A tool to summarize data from NCBI gene datasets and its application

to an update of human gene statistics. Database, 2016:baw153, dec 2016.

[22] Marie Louise Bang, Thomas Centner, Friderike Fornoff, Adam J. Geach, Michael Gotthardt, Mark McNabb, Christian C. Witt, Dietmar Labeit, Carol C. Gregorio, Henk Granzier, and Siegfried 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.

[23] Richard I. Kuo, Elizabeth Tseng, Lel Eory, Ian R. Paton, Alan L. Archibald, and David W. Burt. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics, 18(1):323, dec 2017.

[24] Matthew Fagnani, Yoseph Barash, Joanna Y. Ip, Christine Misquitta, Qun Pan, Arneet L. Saltzman, Ofer Shai, Leo Lee, Aviad Rozenhek, Naveed Mohammad, Sandrine Willaime-Morawek, Tomas Babak, Wen Zhang, Timothy R. Hughes, Derek Van der Kooy, Brendan J. Frey, and Benjamin J. Blencowe. Functional coordination of alternative splicing in the mammalian central nervous system. Genome Biology, 8(6):R108, jun 2007.

[25] Hagen Tilgner, Fereshteh Jahanbani, Tim Blauwkamp, Ali Moshrefi, Erich Jaeger, Feng Chen, Itamar Harel, Carlos D Bustamante, Morten Rasmussen, and Michael P Snyder. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nature Biotechnology, 33(7):736–742, may 2015.

[26] Hagen Tilgner, Fereshteh Jahanbani, Ishaan Gupta, Paul Collier, Eric Wei, Morten Rasmussen, and Michael Snyder. Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. Genome Research, 28(2):231–242, feb 2018.

[27] Diego Marques-Coelho, Lukas da Cruz Carvalho Iohan, Ana Raquel Melo de Farias, Amandine Flaig, Franck Letournel, Marie Laure Martin-Négrier, Françoise Chapon, Maxime Faisant, Catherine Godfraind, Claude Alain Maurage, Vincent Deramecourt, Mathilde Duchesne, David Meyronnet, Nathalie Streichenberger, André Mauès de Paula, Valérie Rigau, Fanny Vandenbos-Burel, Charles Duyckaerts, Danielle Seilhean, Serge Milin, Dan Christian Chiforeanu, Annie Laquerrière, Florent Marguet, Béatrice Lannes, Jean Charles Lam-

bert, and Marcos Romualdo Costa. Differential transcript usage unravels gene expression alterations in Alzheimer's disease human brains. npj Aging and Mechanisms of Disease, 7(1):1–15, dec 2021.

[28] Ashley Byrne, Anna E Beaudin, Hugh E Olsen, Miten Jain, Charles Cole, Theron Palmer, Rebecca M. DuBois, E Camilla Forsberg, Mark Akeson, and Christopher Vollmers. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nature Communications, 8, 2017.

[29] Daniel R. Garalde, Elizabeth A. Snell, Daniel Jachimowicz, Botond Sipos, Joseph H. Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, Michael Jordan, Jonah Ciccone, Sabrina Serra, Jemma Keenan, Samuel Martin, Luke McNeill, E. Jayne Wallace, Lakmal Jayasinghe, Chris Wright, Javier Blasco, Stephen Young, Denise Brocklebank, Sissel Juul, James Clarke, Andrew J. Heron, and Daniel J. Turner. Highly parallel direct RN A sequencing on an array of nanopores. Nature Methods, 15(3):201–206, mar 2018.

[30] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex DeWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time DNA sequencing from single polymerase molecules. Science, 323(5910):133–138, jan 2009.

[31] Kevin J Travers, Chen Shan Chin, David R Rank, John S Eid, and Stephen W Turner. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic acids research, 38(15):e159, 2010.

[32] 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, 299(5607):682–686,

jan 2003.

[33] Alice Mccarthy. Third generation DNA sequencing: Pacific biosciences' single molecule real time technology, jul 2010.

[34] Simon Ardui, Adam Ameur, Joris R. Vermeesch, and Matthew S. Hestand. Single molecule real-time (SMRT) sequencing comes of age: Applications and utilities for medical diagnostics, mar 2018.

[35] Anthony Rhoads and Kin Fai Au. PacBio Sequencing and Its Applications. Genomics, Proteomics & Bioinformatics, 13(5):278–289, oct 2015.

[36] Erick W. Loomis, John S. Eid, Paul Peluso, Jun Yin, Luke Hickey, David Rank, Sarah McCalmon, Randi J. Hagerman, Flora Tassone, and Paul J. Hagerman. Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. Genome Research, 23(1):121–128, jan 2013.

[37] Kin Fai Au, Vittorio Sebastiano, Pegah Tootoonchi Afshar, J. D. Durruthy, Lawrence Lee, Brian A Williams, H. van Bakel, Eric E Schadt, Renee A Reijo-Pera, Jason G Underwood, and Wing Hung Wong. Characterization of the human ESC transcriptome by hybrid sequencing. Proceedings of the National Academy of Sciences, 110(50):E4821–E4830, 2013.

[38] Spyros Oikonomopoulos, Anthony Bayega, Somayyeh Fahiminiya, Haig Djambazian, Pierre Berube, and Jiannis Ragoussis. Methodologies for Transcript Profiling Using Long-Read Technologies, jul 2020.

[39] Daniel Ramsköld, Shujun Luo, Yu Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, Gary P Schroth, and Rickard Sandberg. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature Biotechnology, 30(8):777–782, 2012.

[40] Maria Cartolano, Bruno Huettel, Benjamin Hartwig, Richard Reinhardt, and Korbinian Schneeberger. cDNA library enrichment of full length transcripts for SMRT long read sequencing. PLoS ONE, 11(6):e0157779, jun 2016.

[41] Dr Ioannis Ragoussis and Dr Spyridon Oikonomopoulos. RNA Transcriptomics. Wellcome Genome Campus Advanced Course 2018, pages 1–111, 2018.

[42] Liangzhen Zhao, Hangxiao Zhang, Markus V. Kohnen, Kasavajhala V.S.K. Prasad, Lianfeng Gu, and Anireddy S.N. Reddy. Analysis of transcriptome and epitranscriptome in plants using pacbio iso-seq and nanopore-based direct RNA sequencing, mar 2019.

[43] Kin Fai Au, Jason G Underwood, Lawrence Lee, and Wing Hung Wong. Improving PacBio Long Read Accuracy by Short Read Alignment. PLoS ONE, 7(10), 2012.

[44] Leena Salmela and Eric Rivals. LoRDEC: Accurate and efficient long read error correction. Bioinformatics, 30(24):3506–3514, 2014.

[45] Leena Salmela, Riku Walve, Eric Rivals, Esko Ukkonen, and Cenk Sahinalp. Accurate self-correction of errors in long reads using de Bruijn graphs. Bioinformatics, 33(6):799–806, 2017.

[46] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology, 34(5):525–527, may 2016.

[47] Abhinav Nellore, Andrew E. Jaffe, Jean Philippe Fortin, José Alquicira-Hernández, Leonardo Collado-Torres, Siruo Wang, Robert A. Phillips, Nishika Karbhari, Kasper D. Hansen, Ben Langmead, and Jeffrey T. Leek. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biology, 17(1):266, dec 2016.

[48] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michal Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis, 2016.

[49] Julie Cocquet, Allen Chong, Guanglan Zhang, and Reiner A. Veitia. Reverse transcriptase template switching and false alternative transcripts. Genomics, 88(1):127–131, jul 2006.

[50] Jonathan Houseley and David Tollervey. Apparent non-canonical trans-splicing is generated

by reverse transcriptase in vitro. PLoS ONE, 5(8), 2010.

[51] Douglas Kyung Nam, Sanggyu Lee, Guolin Zhou, Xiaohong Cao, Clarence Wang, Terry Clark, Jianjun Chen, Janet D. Rowley, and San Ming Wang. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. Proceedings of the National Academy of Sciences of the United States of America, 99(9):6152–6156, apr 2002.

[52] Zachary B Abrams, Travis S Johnson, Kun Huang, Philip R.O. Payne, and Kevin Coombes. A protocol to evaluate RNA sequencing normalization methods. BMC Bioinformatics, 20, 2019.

[53] Hiruna Samarakoon, Sanoj Punchihewa, Anjana Senanayake, Jillian M. Hammond, Igor Stevanovski, James M. Ferguson, Roshan Ragel, Hasindu Gamaarachchi, and Ira W. Deveson. Genopo: a nanopore sequencing analysis toolkit for portable Android devices. Communications Biology, 3(1):1–5, dec 2020.

[54] Parveen Goyal, Petya V. Krasteva, Nani Van Gerven, Francesca Gubellini, Imke Van Den Broeck, Anastassia Troupiotis-Tsaïlaki, Wim Jonckheere, Gérard Péhau-Arnaudet, Jerome S. Pinkner, Matthew R. Chapman, Scott J. Hultgren, Stefan Howorka, Rémi Fronzes, and Han Remaut. Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. Nature, 516(7530):250–253, dec 2014.

[55] Nicholas J. Loman and Mick Watson. Successful test launch for nanopore sequencing, mar 2015.

[56] Spyros Oikonomopoulos, Yu Chang Wang, Haig Djambazian, Dunarel Badescu, and Jiannis Ragoussis. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations OPEN. Nature Publishing Group, 6, 2016.

[57] Jason L Weirather, Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research, 6:100, 2017.

[58] Miten Jain, Ian T. Fiddes, Karen H. Miga, Hugh E. Olsen, Benedict Paten, and Mark Akeson. Improved data analysis for the MinION nanopore sequencer. Nature Methods, 12(4):351–356, mar 2015.

[59] Franka J. Rang, Wigard P. Kloosterman, and Jeroen de Ridder. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy, jul 2018.

[60] Chenhao Li, Kern Rei Chng, Esther Jia Hui Boey, Amanda Hui Qi Ng, Andreas Wilm, and Niranjan Nagarajan. INC-Seq: Accurate single molecule reads using nanopore sequencing. GigaScience, 5(1):34, dec 2016.

[61] Roger Volden, Theron Palmer, Ashley Byrne, Charles Cole, Robert J Schmitz, Richard E Green, and Christopher Vollmers. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. Proceedings of the National Academy of Sciences of the United States of America, 115(39):9726–9731, 2018.

[62] Adrien Leger and Tommaso Leonardi. pycoQC, interactive quality control for Oxford Nanopore Sequencing. Journal of Open Source Software, 4(34):1236, feb 2019.

[63] Nanopore summary statistics and basic QC tutorial. 2019.

[64] Wouter De Coster, Svenn D'Hert, Darrin T. Schultz, Marc Cruts, and Christine Van Broeckhoven. NanoPack: Visualizing and processing long-read sequencing data. Bioinformatics, 34(15):2666–2669, aug 2018.

[65] R Wick. rrwick/Porechop: Adapter Trimmer for Oxford Nanopore reads, 2017.

[66] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1):10, may 2011.

[67] Pychopper: A tool to identify, orient, trim and rescue full length cDNA reads.

[68] Heng Li. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics, 34(18):3094–3100, sep 2018.

[69] A R Jeffries, S K Leung, I Castanho, K Moore, J P Davies, E L Dempster, N J Bray, P O'neill, E Tseng, Z Ahmed, D Collier, S Prabhakar, L Schalkwyk, M J Gandal, E Hannon, and J Mill. Full-length transcript sequencing of human and mouse identifies widespread isoform diversity and alternative splicing in the cerebral cortex. bioRxiv, page 2020.10.14.339200, oct 2020.

[70] Isabel Castanho, Tracey K Murray, Eilis Hannon, Aaron Jeffries, Emma Walker, Emma Laing, Hedley Baulf, Joshua Harvey, Lauren Bradshaw, Andrew Randall, Karen Moore, Paul O'Neill, Katie Lunnon, David A. Collier, Zeshan Ahmed, Michael J. O'Neill, and Jonathan Mill. Transcriptional Signatures of Tau and Amyloid Neuropathology. Cell Reports, 30(6):2040–2054.e5, 2020.

[71] Thomas Derrien, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, David Martin, Angelika Merkel, David G. Knowles, Julien Lagarde, Lavanya Veeravalli, Xiaoan Ruan, Yijun Ruan, Timo Lassmann, Piero Carninci, James B. Brown, Leonard Lipovich, Jose M. Gonzalez, Mark Thomas, Carrie A. Davis, Ramin Shiekhattar, Thomas R. Gingeras, Tim J. Hubbard, Cedric Notredame, Jennifer Harrow, and Roderic Guigó. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Research, 22(9):1775–1789, 2012.

[72] Oxford Nanopore: Pipeline for differential gene expression (DGE) and differential transcript usage (DTU) analysis using long reads.

[73] Alison D Tang, Cameron M Soulette, Marijke J. van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J Wu, and Angela N Brooks. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nature Communications, 11(1), 2020.

[74] W. J. Kent. BLAT—The BLAST-Like Alignment Tool. Genome Research, 12(4):656–664, mar 2002.

[75] Lorena De La Fuente, Ángeles Arzalluz-Luque, Manuel Tardáguila, Héctor Del Risco, Cristina Martí, Sonia Tarazona, Pedro Salguero, Raymond Scott, Alberto Lerma, Ana

Alastrue-Agudo, Pablo Bonilla, Jeremy R.B. Newman, Shunichi Kosugi, Lauren M. McIntyre, Victoria Moreno-Manzano, and Ana Conesa. TappAS: A comprehensive computational framework for the analysis of the functional impact of differential splicing. Genome Biology, 21(1):1–32, may 2020.

[76] Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 11(3):1–9, mar 2010.

[77] Ana Conesa, María Jośe Nueda, Alberto Ferrer, and Manuel Talón. maSigPro: A method to identify significantly differential expression profiles in time-course microarray experiments. Bioinformatics, 22(9):1096–1102, may 2006.

[78] María José Nueda, Sonia Tarazona, and Ana Conesa. Next maSigPro: Updating maSigPro bioconductor package for RNA-seq time series. Bioinformatics, 30(18):2598–2602, sep 2014.

[79] Ana Conesa and María J Nueda. maSigPro User's Guide. Technical report, 2017.

[80] Mar Gonzàlez-Porta, Adam Frankish, Johan Rung, Jennifer Harrow, and Alvis Brazma. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biology, 14(7):R70, 2013.

[81] Iakes Ezkurdia, Jose Manuel Rodriguez, Enrique Carrillo-De Santa Pau, Jesús Vázquez, Alfonso Valencia, and Michael L. Tress. Most highly expressed protein-coding genes have a single dominant isoform. Journal of Proteome Research, 14(4):1880–1887, apr 2015.

[82] Kristoffer Vitting-Seerup and Albin Sandelin. The landscape of isoform switches in human cancers. Molecular Cancer Research, 15(9):1206–1220, sep 2017.

[83] Martin Ramsden, Linda Kotilinek, Colleen Forster, Jennifer Paulson, Eileen McGowan, Karen SantaCruz, Aaron Guimaraes, Mei Yue, Jada Lewis, George Carlson, Michael Hutton, and Karen H. Ashe. Age-dependent neurofibrillary tangle formation, neuron loss, and memory impairment in a mouse model of human tauopathy (P301L). Journal of Neuroscience, 25(46):10637–10647, nov 2005.

[84] Fumihiko Muramori, Katsuji Kobayashi, and Ichirou 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 and Clinical Neurosciences, 52(6):593–599, 1998.

[85] Aiko Ishiki, Maki Kamada, Yuki Kawamura, Chiaki Terao, Fumiko Shimoda, Naoki Tomita, Hiroyuki Arai, and Katsutoshi Furukawa. Glial fibrillar acidic protein in the cerebrospinal fluid of Alzheimer's disease, dementia with Lewy bodies, and frontotemporal lobar degeneration. Journal of Neurochemistry, 136(2):258–261, jan 2016.

[86] Pratishtha Chatterjee, Steve Pedrini, Erik Stoops, Kathryn Goozee, Victor L. Villemagne, Prita R. Asih, Inge M.W. Verberk, Preeti Dave, Kevin Taddei, Hamid R. Sohrabi, Henrik Zetterberg, Kaj Blennow, Charlotte E. Teunissen, Hugo M. Vanderstichele, and Ralph N. Martins. Plasma glial fibrillary acidic protein is elevated in cognitively normal older adults at risk of Alzheimer's disease. Translational Psychiatry, 11(1):1–10, jun 2021.

[87] Michele Zorzetto, Francesca Datturi, Laura Divizia, Cristiana Pistono, Ilaria Campo, Annalisa De Silvestri, Mariaclara Cuccia, and Giovanni Ricevuti. Complement C4A and C4B gene copy number study in Alzheimer's disease patients. Current Alzheimer Research, 13(999):1–1, nov 2016.

[88] Erika Castillo, Julio Leon, Guianfranco Mazzei, Nona Abolhassani, Naoki Haruyama, Takashi Saito, Takaomi Saido, Masaaki Hokama, Toru Iwaki, Tomoyuki Ohara, Toshiharu Ninomiya, Yutaka Kiyohara, Kunihiko Sakumi, Frank M. Laferla, and Yusaku Nakabeppu. Comparative profiling of cortical gene expression in Alzheimer's disease patients and mouse models demonstrates a link between amyloidosis and neuroinflammation. Scientific Reports, 7(1):1–16, dec 2017.

[89] Kerstin T.S. Wirz, Koen Bossers, Anita Stargardt, Willem Kamphuis, Dick F. Swaab, Elly M. Hol, and Joost Verhaagen. Cortical beta amyloid protein triggers an immune response, but no synaptic changes in the APPswe/PS1dE9 Alzheimer's disease mouse model. Neurobiology of Aging, 34(5):1328–1342, may 2013.

[90] John R. McDermott and Alison M. Gibson. Degradation of Alzheimer's β-amyloid protein by human cathepsin D. NeuroReport, 7(13):2163–2166, 1996.

[91] Agnes Kenessey, Parimala Nacharaju, Li Wen Ko, and Shu Hui Yen. Degradation of tau by lysosomal enzyme cathepsin D: Implication for Alzheimer neurofibrillary degeneration. Journal of Neurochemistry, 69(5):2026–2038, 1997.

[92] Caitlin N. Suire, Samer O. Abdul-Hay, Tomoko Sahara, Dongcheul Kang, Monica K. Brizuela, Paul Saftig, Dennis W. Dickson, Terrone L. Rosenberry, and Malcolm A. Leissring. Cathepsin D regulates cerebral Aβ42/40 ratios via differential degradation of Aβ42 and Aβ40. Alzheimer's Research and Therapy, 12(1):1–13, jul 2020.

[93] Hansruedi Mathys, Chinnakkaruppan Adaikkan, Fan Gao, Jennie Z. Young, Elodie Manet, Martin Hemberg, Philip L. De Jager, Richard M Ransohoff, Aviv Regev, and Li Huei Tsai. Temporal Tracking of Microglia Activation in Neurodegeneration at Single-Cell Resolution. Cell Reports, 21(2):366–380, 2017.

[94] Hong Wang, Kaushik Kumar Dey, Ping Chung Chen, Yuxin Li, Mingming Niu, Ji Hoon Cho, Xusheng Wang, Bing Bai, Yun Jiao, Surendhar Reddy Chepyala, Vahram Haroutunian, Bin Zhang, Thomas G. Beach, and Junmin Peng. Integrated analysis of ultra-deep proteomes in cortex, cerebrospinal fluid and serum reveals a mitochondrial signature in Alzheimer's disease. Molecular Neurodegeneration, 15(1):1–20, jul 2020.

[95] Akihito Ishigami, Takako Ohsawa, Masaharu Hiratsuka, Hirom Taguchi, Saori Kobayashi, Yuko Saito, Shigeo Murayama, Hiroaki Asaga, Tosifusa Toda, Narimichi Kimura, and Naoki Maruyama. Abnormal accumulation of citrullinated proteins catalyzed by peptidylarginine deiminase in hippocampal extracts from patients with Alzheimer's disease. Journal of Neuroscience Research, 80(1):120–128, apr 2005.