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

Abbreviations

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
CLU	Clusterin
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

GWAS	Genome-wide association studies
IR	Intron Retention
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
ONT	Oxford Nanopore Technologies
ORF	Open Reading Frame
PacBio	Pacific Biosciences
PICALM	Phosphatidylinositol Binding Clathrin Assembly Protein
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
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
ToFU	Transcript isOforms: Full-length and Unassembled

TPM	Transcripts per Million
TSS	Transcription Start Sites
TTS	Transcription Termination Sites

Chapter 1

Introduction

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.

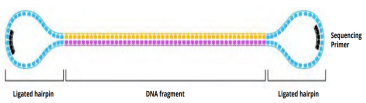
Unlike RNA-Sequencing, Pacific Bioscience's Single Molecule Real Time sequencing (SMRT) is able to generate long reads is due to its ability to mimic natural, uninterrupted, processive DNA synthesis, through three important innovations:

1. Creation of a circular template, SMRTbell, enclosed with hairpin adapters at end of the inserted target double-stranded DNA, allowing uninterrupted DNA polymerisation (Figure 2.1a).
2. Sequencing of each SMRTbell in a separate nanometre-wide well (zero-mode-waveguide - ZMW), and all wells contained within a single SMRT chip.[20] Due to the very nanoscale size of the ZMW and reduced detection volume, a single nucleotide incorporation can

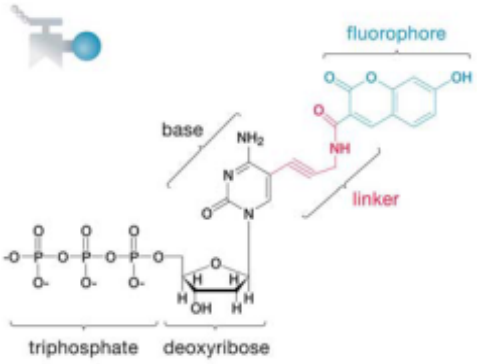
be sensitively detected against the high background of labelled nucleotides, achieving a high-signal-to-noise ratio (Figure 2.1c)).

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[21] (Figure 2.1b).

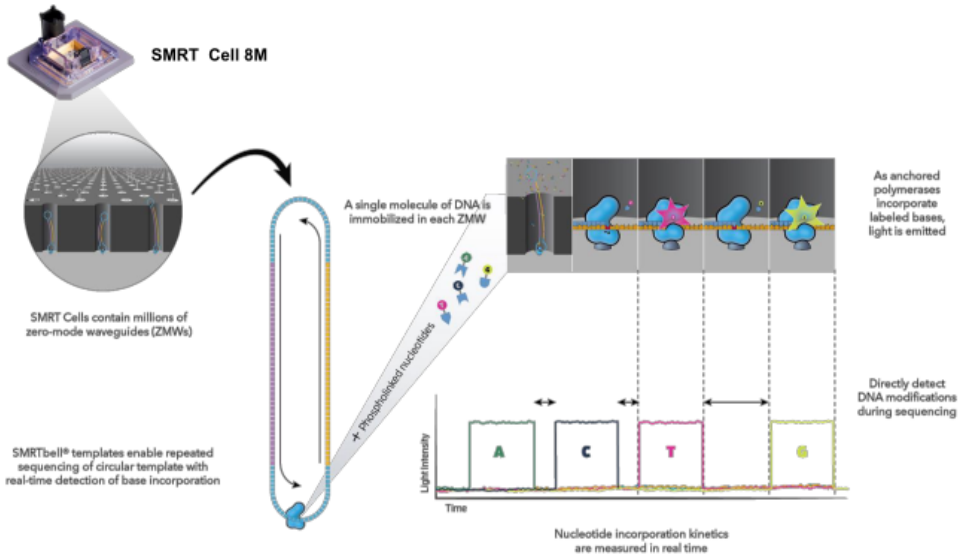
In summary, SMRT sequencing detect fluorescence events that correspond to addition of one specific nucleotide by a polymerase attached to the bottom of a tiny well.



(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, 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). Further due to the circular nature of the SMRT bell, while sequencing and subsequent base-calling error can occur randomly at a rate of XXX, generating a raw accuracy of only 80%, the generation of CCS from a coverage of 15 passes provides >99% accuracy per base rate from sequence overlaps (Eid et al. 2009). The number of passes and subsequent generation of the CCS, however, is hindered by the length of the insert, whereby a long target DNA >XXX kB would only generate one single subread.

Accuracy of SMRT sequencing dependent on the number of times the fragment is rad - depth of sequencing of the individual SMRT bell template. Randomness of sequencing errors in subreads, consisting of more indels than mismatches suggest that the final output from CCS assembly should be free from systematic biases.[22] Nontheless, CCS reads retain errors, with a bis for indels in homopolymers.

With rapidly-advancing technology and chemistry, PacBio released a faster polymerase with chemistry v3 in 2018, increasing read lengths to an average 30kb polymerase read length. Late 2018 v3 chemistry increases longevity of polymerase.

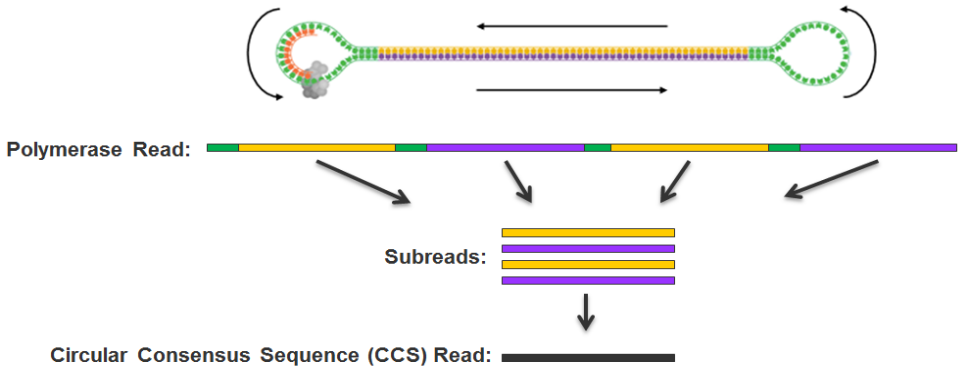


Figure 2.2: Generation of Circular Consensus Sequence: CCS is generated by the collapse of multiple subreads, which sequence correspond to the double-stranded cDNA of interest. The greater the number of "passes" sequenced by the polymerase, the longer the polymerase read, the more subreads generated,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

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: 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).

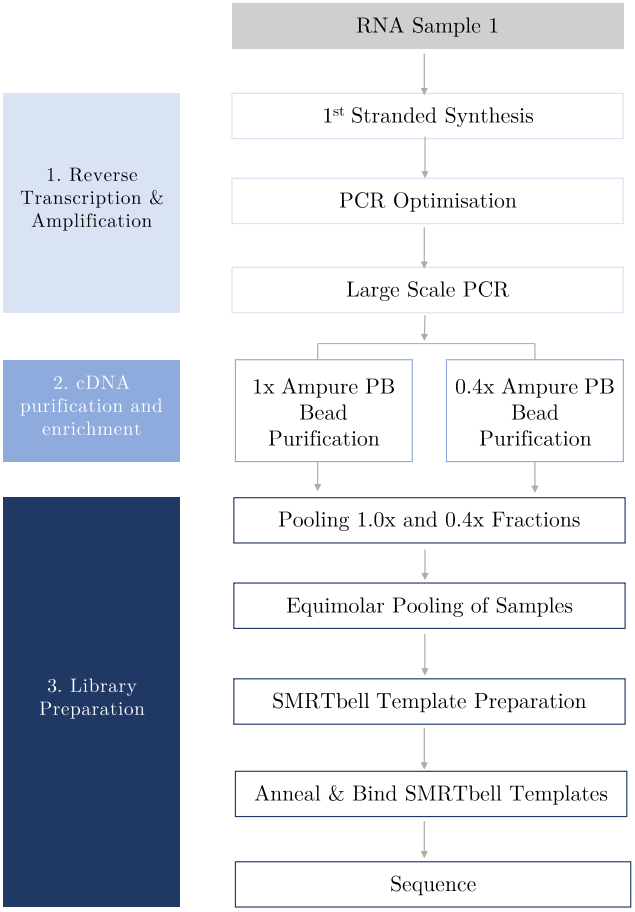


Figure 2.3: An overview of the lab Iso-Seq pipeline used for whole transcriptome profiling. The lab pipeline, as adapted from official Iso-Seq protocol, involves three main steps: 1) reverse transcription and amplification of cDNA (Section 2.1.2.1), 2) cDNA purification with ampure beads (Section 2.1.2.5) and 3) library preparation involving ligation of SMRT bell templates, and primer and polymerase binding (Section 2.1.2.9)

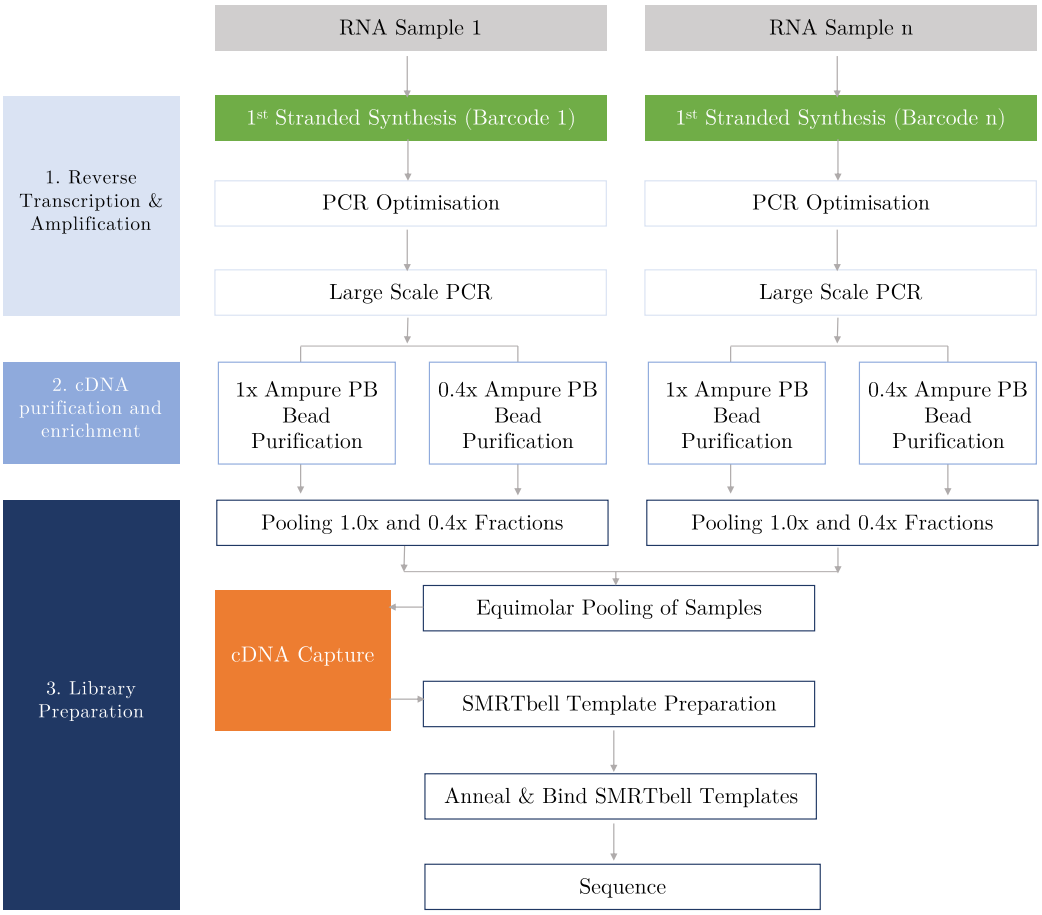


Figure 2.4: An overview of the lab Iso-Seq pipeline used for whole transcriptome profiling. The lab pipeline for targeted transcriptome profiling involves all the steps in the standard Iso-Seq lab pipeline (Figure 2.3), with the addition of the target capture step (Boxed orange, Section 2.1.2.8) and 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.[23] 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)([24]).

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 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.

2.1.2.3 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.

2.1.2.4 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.

2.1.2.5 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:

$$\frac{concentration(\frac{ng}{ul}) \times 10^6}{660(\frac{g}{mol}) \times average\ library\ size\ in\ bp^*} = concentration\ in\ nM \tag{2.1}$$

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

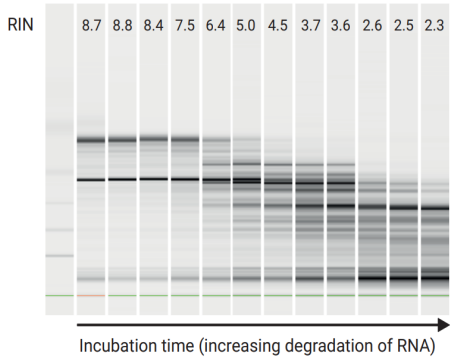
2.1.2.6 Bioanalyzer

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.

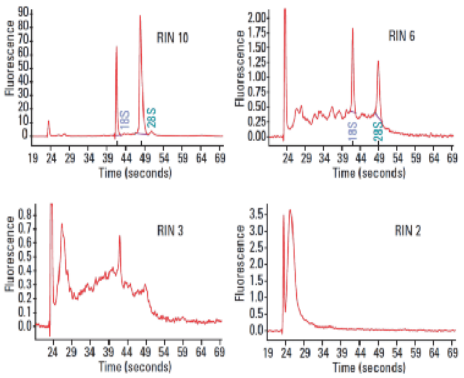
For this thesis, most of the assessments of DNA quality in the Iso-Seq and ONT protocol were performed on the DNA 12000 Kit (Agilent) on the 2100 Bioanalyzer (Agilent)for accurate

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).

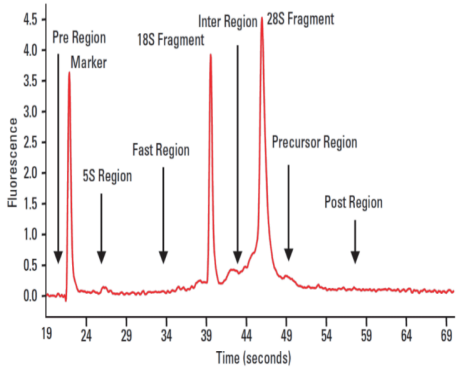
RNA extracted by Dr Isabel Castanho was also run on RNA ScreenTape assay and the Bioanalyzer RNA analysis to provide accurate evaluation of RNA degradation; this is represented by a RNA Integrity Number (RIN) between 1 and 10, where 1 is indicative of high degradation, and 10 of low degradation and thus high integrity (Figure 2.5). 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 2.5: 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.

2.1.2.7 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.

2.1.2.8 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.

2.1.2.9 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).

2.1.2.10 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.

2.1.2.11 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

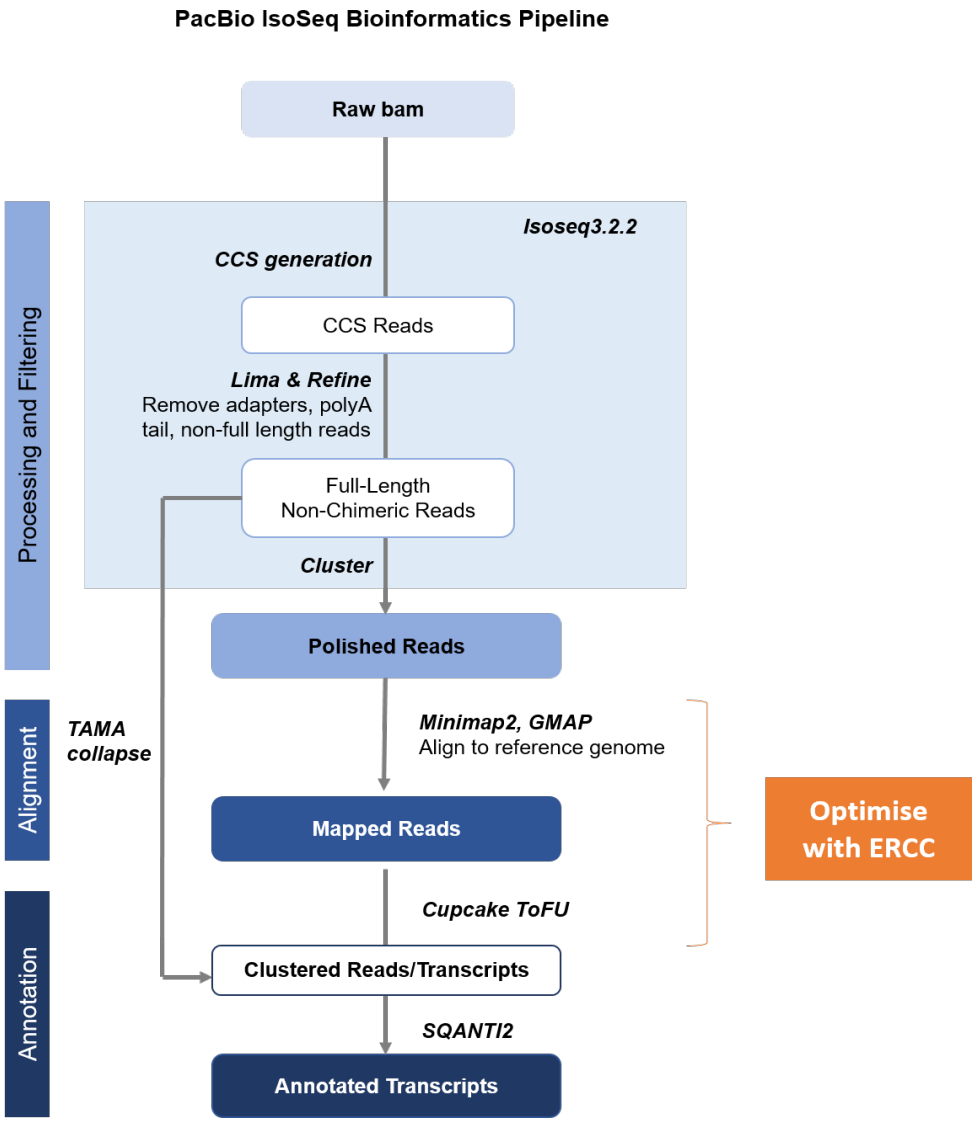


Figure 2.6: PacBio Isoseq Bioinformatics Pipeline: Pipeline is adapted from ToFU [10]

reads). Currently there were three methods of correcting long reads:[25]

- Hybrid error correction strategy using short-reads: LSC[26] which maps short reads, and LoRDEC which build De Brujin graph of short reads[27]
- Self-correction using long reads only: Long-read multiple aligner (LoRMA)[28]
- 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 ERCC

One source of error from long-read sequencing can occur at reverse transcription, 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,[7] most ERCC reads fell within +/- 5bp at both 5' and 3' ends, with 3' end slightly more accurate than 5' end. From,[12] 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 ([7]) 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.

2.1.3.3 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

2.1.3.4 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. 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 represents 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

2.1.3.5 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%).

2.1.3.6 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)

2.1.3.7 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 (ref) 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[29] : 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 2.7):

- 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

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

Figure 2.7: 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

Further filtering of isoforms from technical artifacts

This was developed to remove artifacts from library preparation: i.e. intrapriming of polyA that usually happens in antisense strands and also lack of junction support in NNC; increase % of FSM transcripts, and removes NIC.

SQANTI2 further filters isoforms, based on the following rules:

1. FSM with a reliable 3' end by:
    - >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.

Reverse Transcription switch is determined on a junction level on both plus and minus strands by aligning each splice junction to reference file (splice junction defined as 3'/end of the exon to the 3'/end of the intron). The transcript is considered to be an artifact of reverse transcription

if any of the junctions were labelled as RT switch.

all junctions were either canonical or has short read coverage (> 3 reads)

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:

$$FL\ \ TPM(x_{sample}, y_{sample}) = \frac{Raw\ \ FL\ \ count(x_{isoform}, y_{sample})}{Total\ \ FL\ \ count(y_{sample})} * 10^6 \tag{2.2}$$

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.[30] 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 Quantification of human transgene expression

As reported in,[31] human-specific MAPT sequence was selected from a 2kb region present in the 3'UTR after using BLAT to identify divergent sequence in human and mouse MAPT. Counts of this human-specific MAPT sequence in the CCS and polished reads from WT and TG were then plotted as a ratio of unique reads to the total number of input reads.

2.1.3.12 Classification of Alternative Splicing Events

SUPPA2 was used to classify alternative splicing events with the parameter –f ioe in isoforms retained from SQANTI2 filter. Splicing events included Alternative 5' Splice Sites (A5), Alternative 3' Splice Sites (A3), Alternative First Exons (AF), Alternative Last Exons (AL), Mutually Exclusive Exons (MX), Retained Intron (RI), Skipping Exon (SE).

2.1.3.13 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:[13]

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

Chapter 3

Whole Transcriptome

Lowly expressed gene and minor isoform quantification

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 fron-

totemporal 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) (Figure **??**). As a technological comparison and validation of the IsoSeq approach, a subset of samples were also sequenced on ONT (Figure **??**). 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.[31] 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.[31] 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 (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 2.1.2.3). 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.

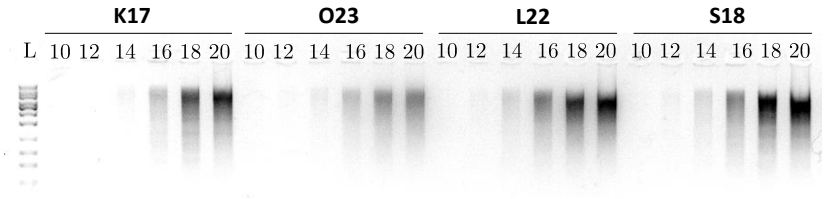


Figure 3.1: Samples were typically amplified using 14 cycles after performing PCR cycle optimisation: An example of gel image from PCR cycle optimisation of Samples K17, O23, L22 and S18. PCR aliquots were collected every two cycles (10, 12, 14, 16, 18, 20) and then run on gel electrophoresis. 14 cycles 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 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 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

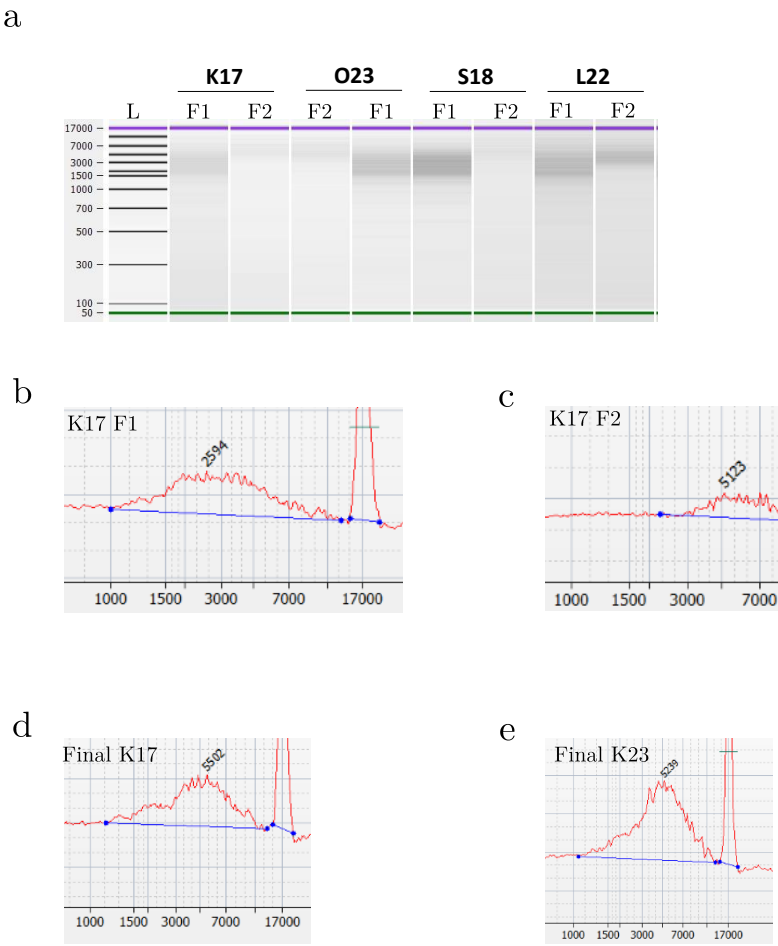


Figure 3.2: Library preparation was performed for each sample with successful cDNA purification and ligation with SMRT bell templates: Following large scale amplification using the optimum cycle number (as determined from Figure 3.1), the resulting cDNA was divided into two fractions (denoted here as 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

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.

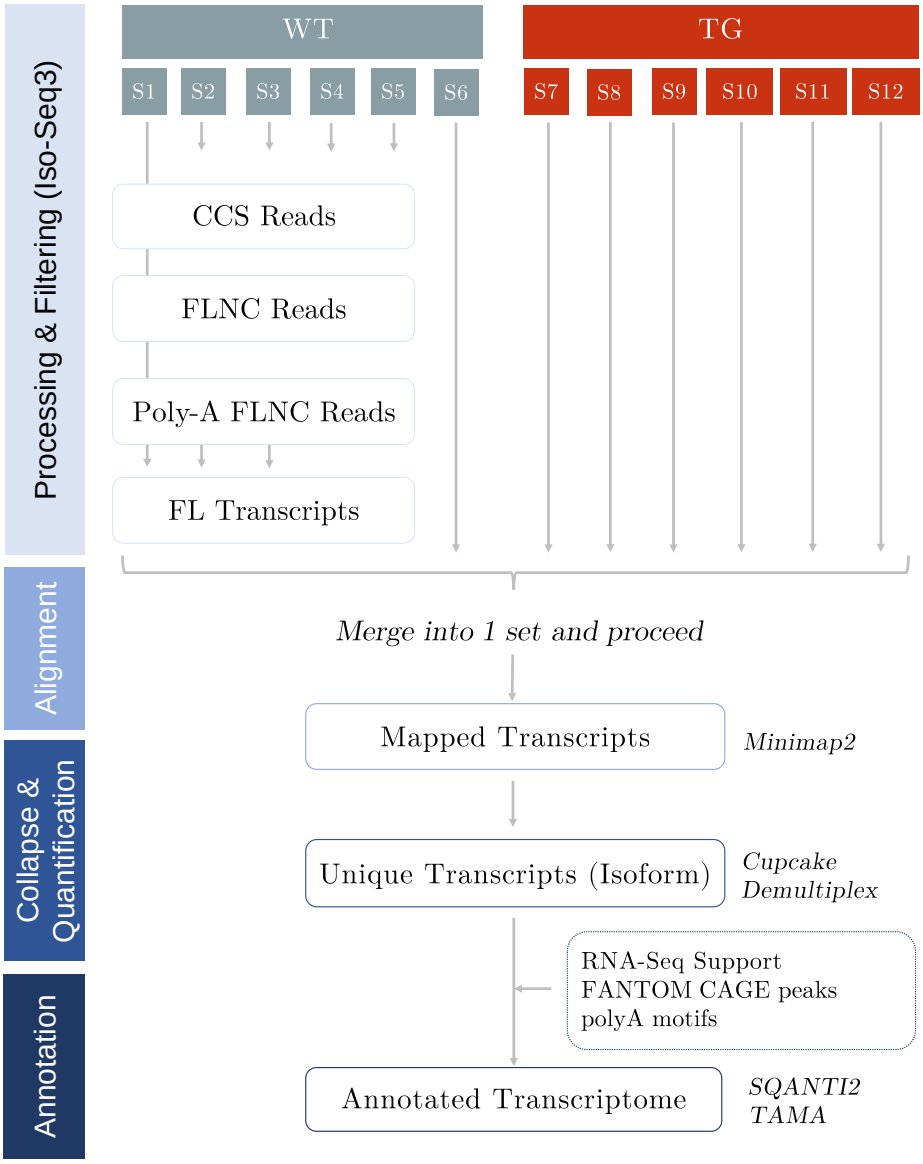


Figure 3.3: PacBio Isoseq Bioinformatics Pipeline: Pipeline is adapted from ToFU [10]

Alternative splicing events were assessed using a range of packages and custom scripts: mutu-

ally 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.

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 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 are 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

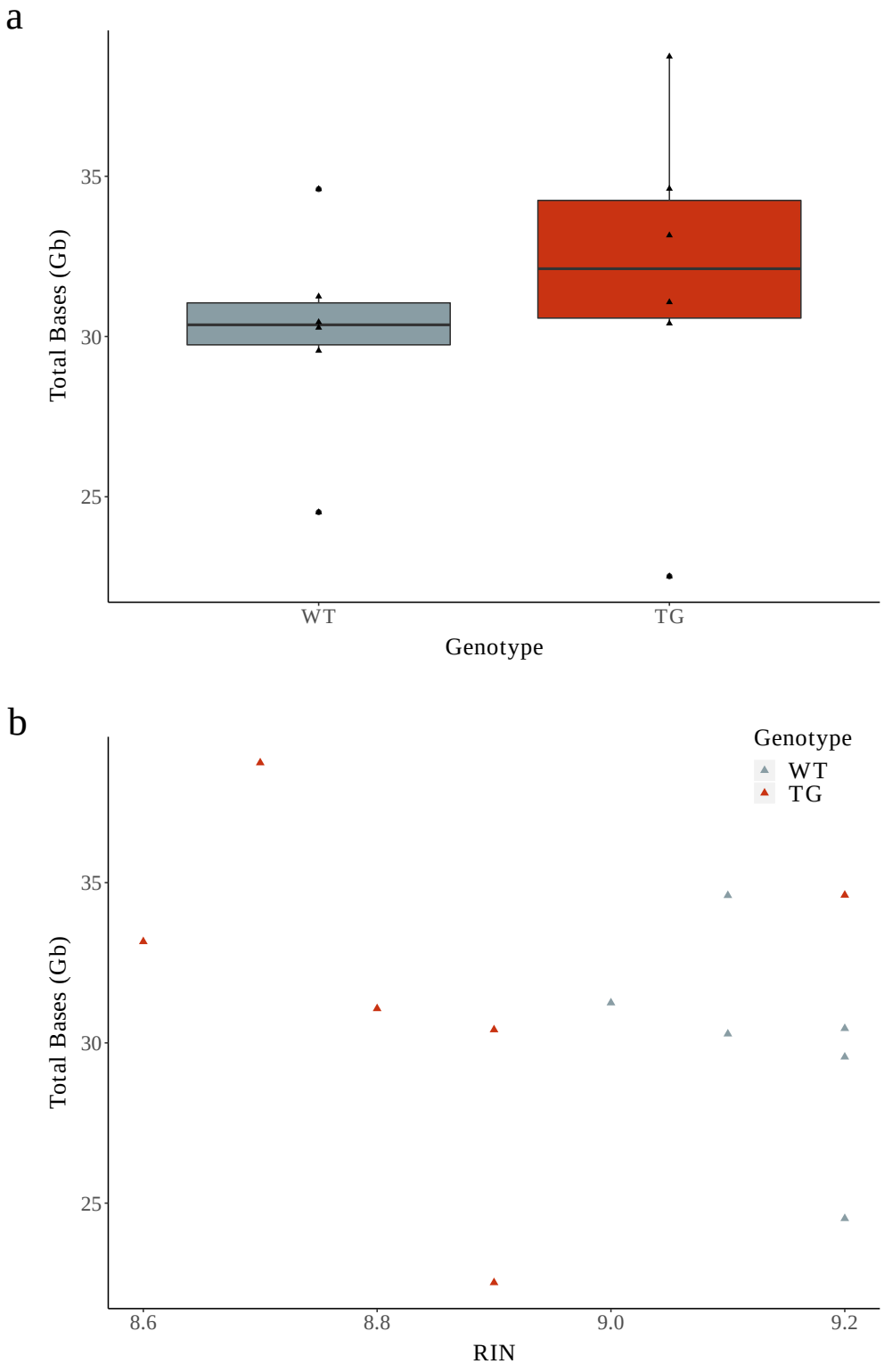


Figure 3.4: Whole Transcriptome Iso-Seq runs generated ~30Gb per sample, independent of RIN score: Sequential Iso-Seq run generated **a)** a range of 30-35Gb per sample of the whole transcriptome, with no significant difference observed between WT and TG Tg4510 mice. Of note, two samples with <25Gb in 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 3.5a). 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 3.5b), 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 3.5c), 266K transcripts were retained. No difference was observed in the number of transcripts generated between WT and TG (n = 12, two-tailed unpaired t-test, t = -0.005, df = 10, P = 0.996) or by age (n = 12, t = -1.58, df = 10, P = 0.15).

3.3.2 Transcriptome annotation

After further collapsing and filtering of transcripts, 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 3.6a). 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.

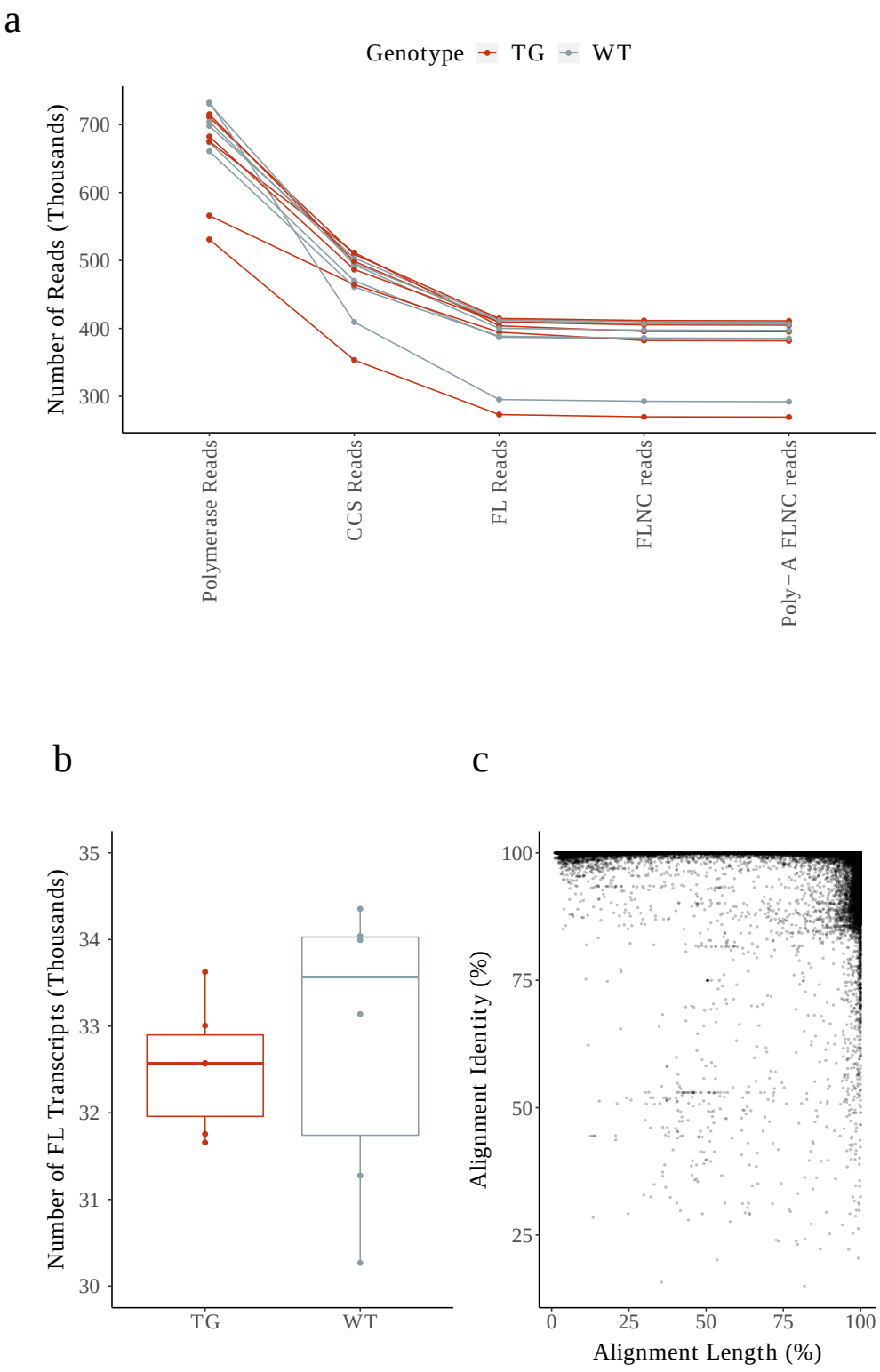


Figure 3.5: Sequential processing of Iso-Seq Reads generated around 32K transcripts per sample with good alignment to reference genome: **a)** Processing of Iso-Seq reads generated a similar number of reads across all sample throughout Iso-Seq3 bioinformatis pipeline, with the exception of 2 earlier 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

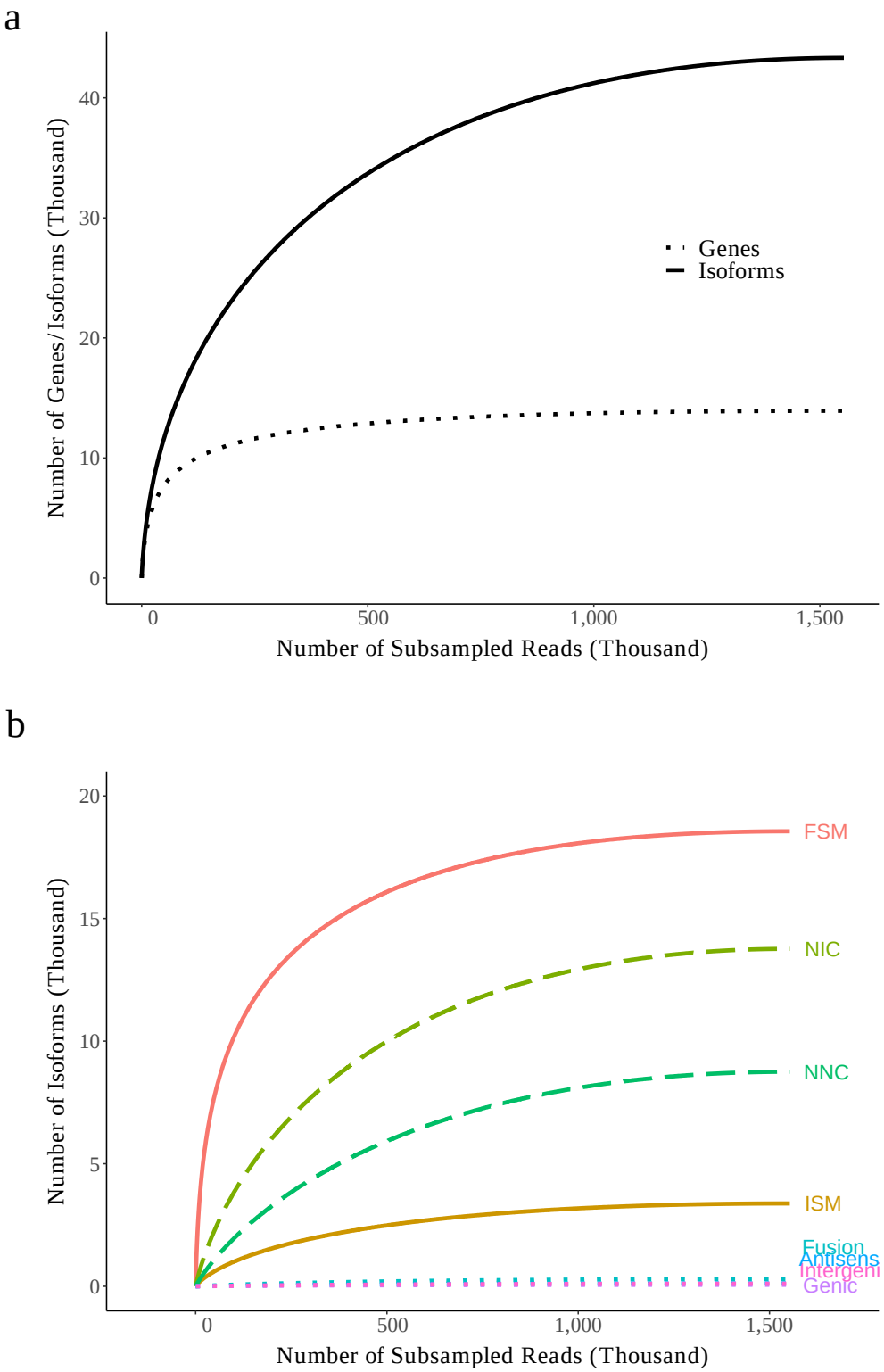


Figure 3.6: Rarefaction curve of Iso-Seq merged dataset indicated saturation and good coverage of genes and isoforms:

3.3.3 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 3.7a), with about half (47.8%) detected in 2 - 3 samples with very low transcript expression (Figure 3.7b). 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 3.8a). 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 3.8b), highlighting the power of Iso-Seq to quantify highly-expressed transcripts.

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 3.9a), 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 3.9c), and exon number (corr = 0.24, P = 7.97 x 10^{-155}, Figure 3.9d).

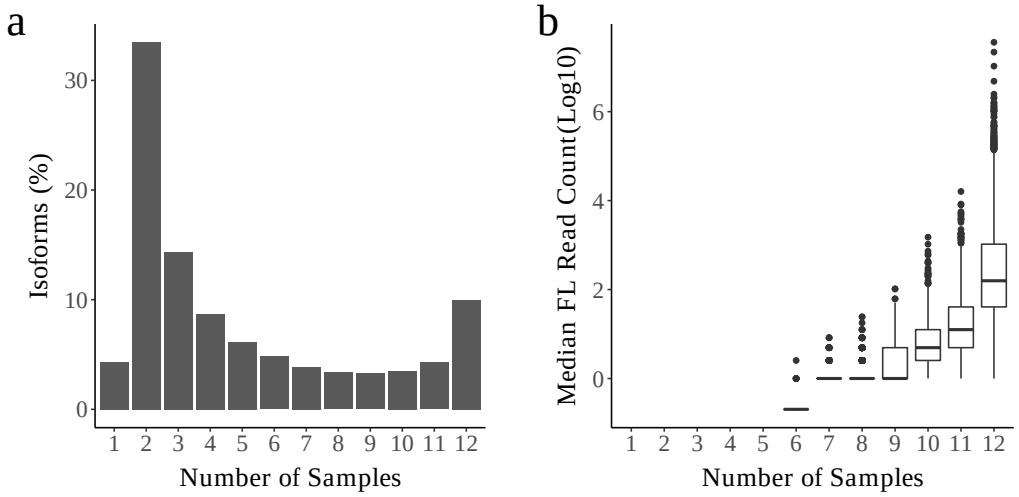


Figure 3.7: Highly-expressed isoforms are more likely to be sequenced across sampples and accurately quantified: Shown is **a)** the distribution of isoforms detected in the number of mouse samples, with a third detected in any two of the total 12 samples. However, **b)** quantification of these isoforms 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

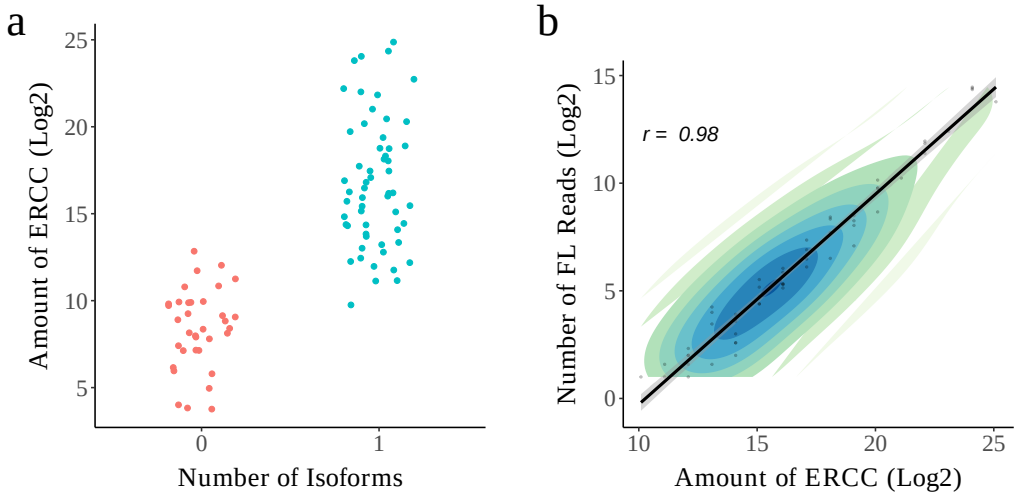


Figure 3.8: Over 60% of ERCCs were detected with highly accurate quantification **a** Highly-concentrated ERCCs were detected as single molecules, as expected, and **b** the number of full-length reads associated for each detected ERCC was highly correlated to known amount. FL - Full Length

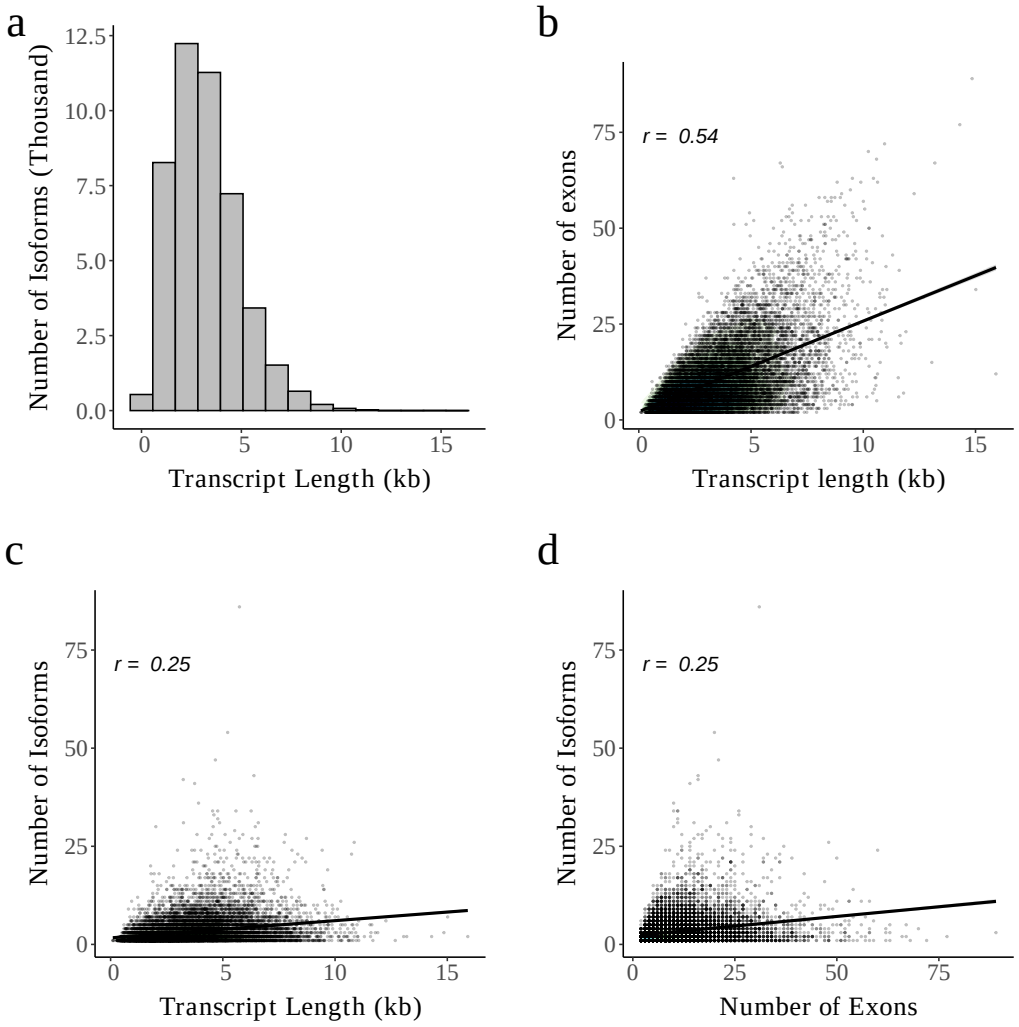


Figure 3.9: Longer genes with more exons were associated with more isoforms: **a** The majority of isoforms have a length between 1 - 5kb. **b)** The number of exons was correlated with the transcript length, and the **c)** the number of isoforms was correlated with the length and **d)** and the number of exons per gene. Gene length and exon number is represented by the longest transcript. kb - kilobases

3.3.4 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. Using SQANTI to characterise isoforms similarly to the Iso-Seq analysis, RNA-Seq defined transcriptome revealed significantly more isoforms (XX). 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 (XX vs XX, two-tailed unpaired t-test,), fewer exons (XX vs XX, two-tailed unpaired ,) and less supported by CAGE peaks (XX vs XX, two-tailed unpaired t-test,). Considering only isoforms that had a complete exact match as defined by gffcompare, more than XX% of isoforms detected from Iso-Seq dataset could not be readily recapitulated, the majority of which were novel isoforms and genes.

3.3.5 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 3.2). 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 x 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 x 108, P < 2.23 x 10^{-308} 3.10a,b) and longer (Mann-Whitney-Wilcoxon test, W = 2.37 x10^{8}, P = 2.13 x 10^{-42}, Figure 3.10c,d) with more exons (Mann-Whitney-Wilcoxon test, W = 1.94 x 10^{8}, P < 2.23 x 10^{-308}, Figure 3.10e,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

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 3.2: 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

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 3.11). 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

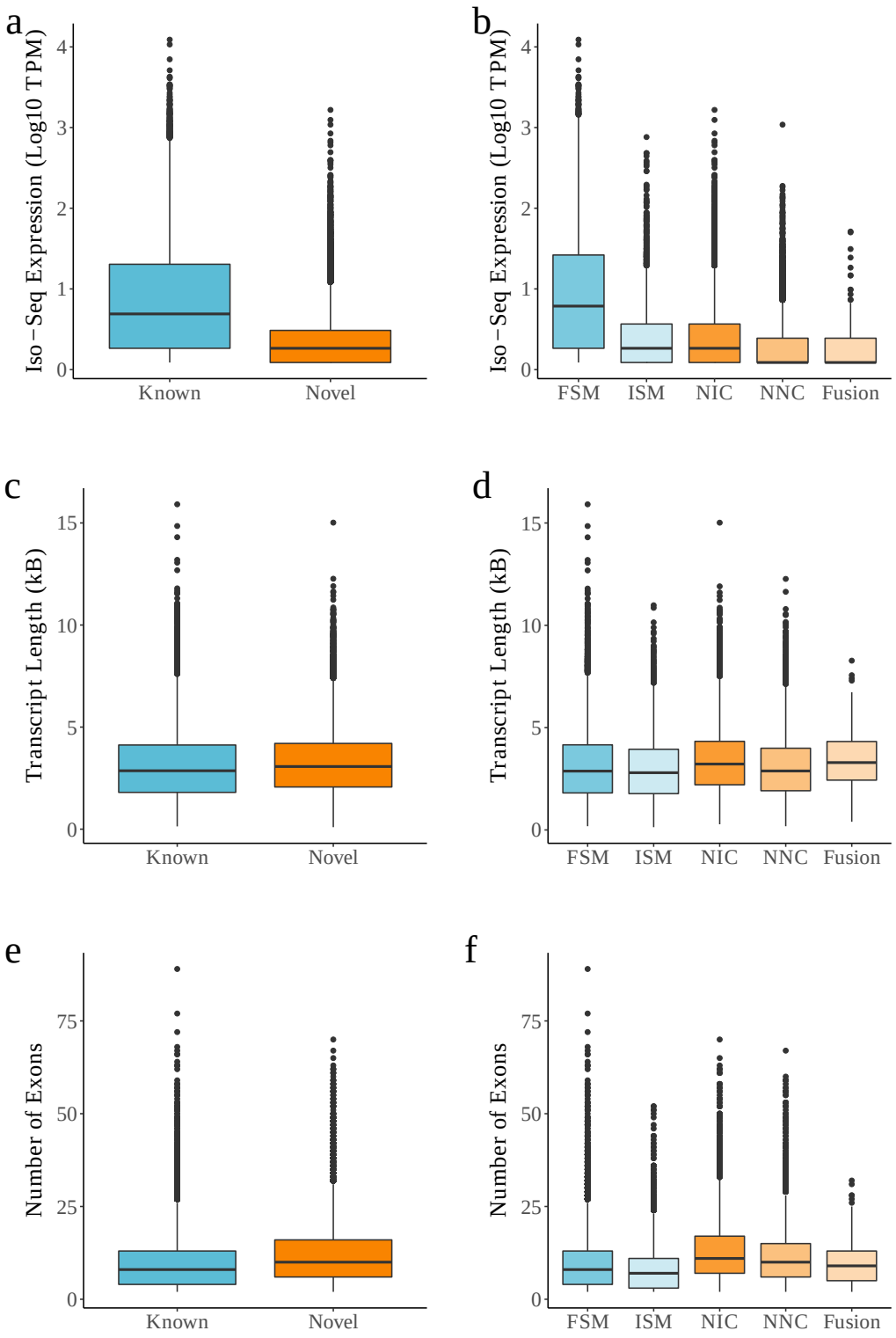


Figure 3.10: Novel isoforms were less expressed, longer and had more exons than known isoforms: Shown is the **a)** Iso-Seq transcript expression, the **c)** transcript length, and the **e)** the number of exons of novel and known isoforms. The known and novel isoforms can be further subdivided and classified, 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.

= 1.81; SE: P < 2.23 x 10^{-16}, odds ratio = 1.57, Figure 3.11).

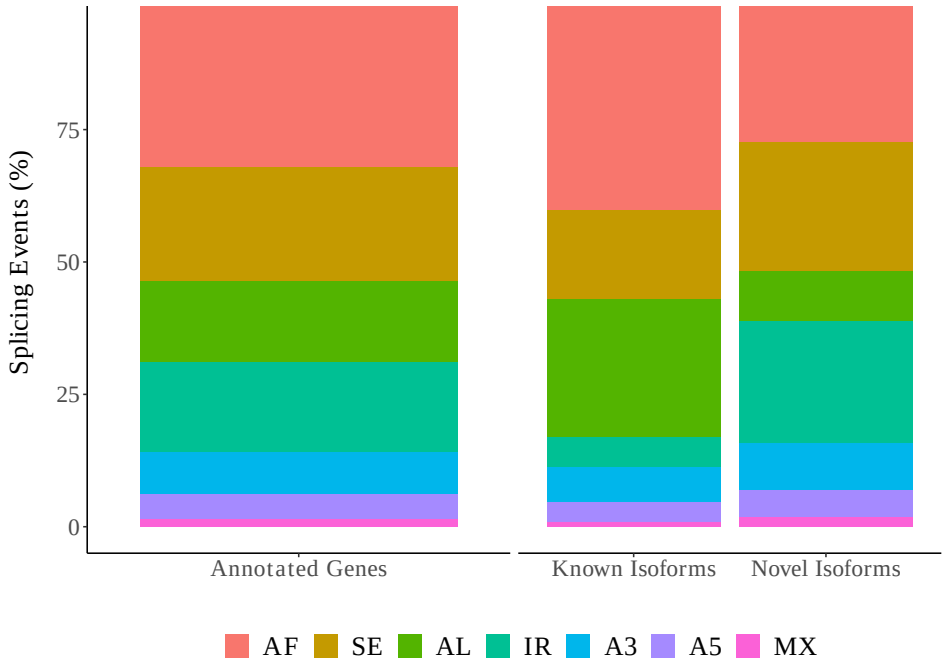


Figure 3.11: Alternative first is the most prevalent AS event, and novel isoforms are more likely to be characterised with complex AS events: Shown is the proportion of AS events in annotated genes, and further subdivided by known and novel isoforms. Novel isoforms were more likely to be characterised 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

3.3.6 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 3.3), 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 to be associated with NMD than known isoforms (Fisher's Test: P < 2.23 x 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 x 10^6, P = 1.67 x 10^{-42}, Figure 3.12b), 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 3.12a), 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 3.3: 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.

3.3.7 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%)).

3.3.8 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 x 10^7, P = 8.24 x 10^{-98}, Figure 3.13a), despite containing fewer exons (W = 4.56 x 10^7, P < 2.23 x 10^{-308}, Figure 3.13b) and being enriched for mono-exonic molecules(23.9% vs 2.02%) - corroborating previous findings from other long-read studies(,[3215]). These lncRNA isoforms were found to be more lowly expressed than non-lncRNA isoforms (W = 3.16 x 10^7, P = 5.67 x 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 x

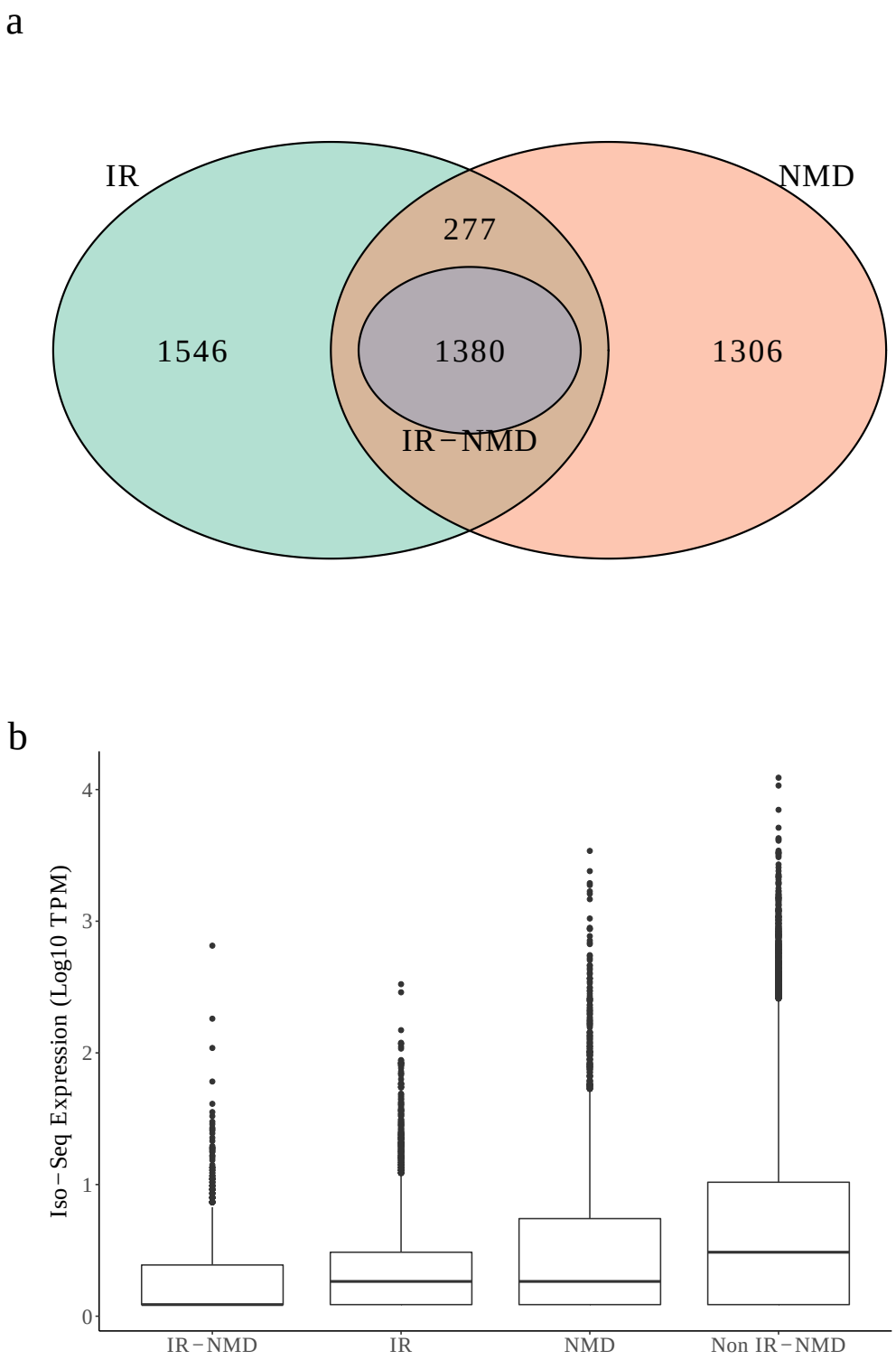


Figure 3.12: Intron retention is associated with nonsense-mediated mRNA decay (NMD) and reduced expression: Shown is the overlap of genes associated with isoforms characterised with intron retention (IR), nonsense-mediated mRNA decay (NMD), and transcripts with both IR and NMD (IR-NMD). Of note, genes 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.

10^6, P = 5.76 x 10^{-107}, Figure 3.13e).

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 x 10^7, P = 8.33 x 10^{-195}).

3.3.9 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 x 10^6, P = 5.22 x 10^{-45}; expression: W = 2.29x 10^6, P = 1.5 x 10^{-73}).

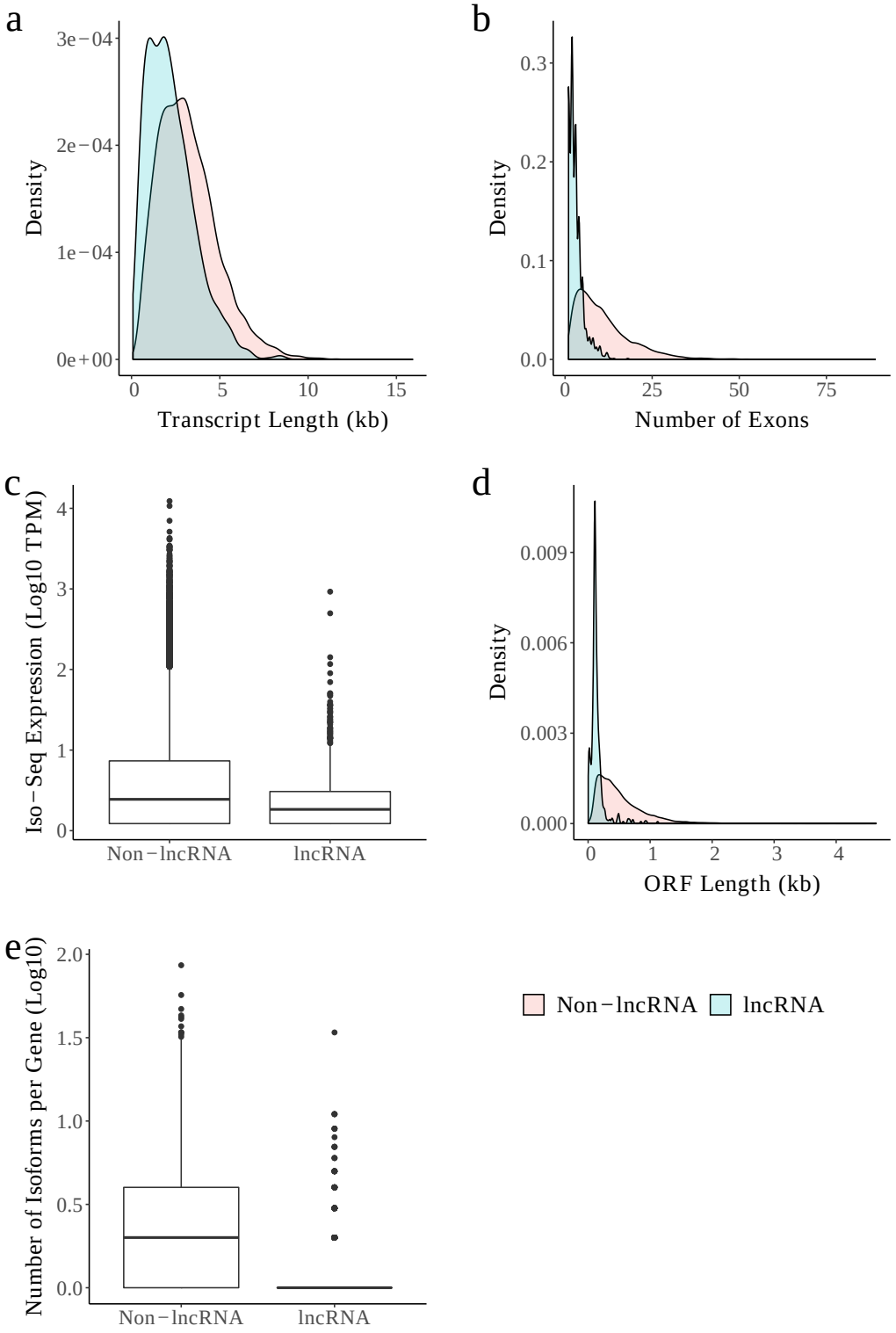


Figure 3.13: LncRNA isoforms were more lowly expressed and typically longer than non-lncRNA transcripts, despite containing fewer exons: Shown is the distribution of the **a)** transcript length, **b)** number of exons, **c)** transcript expression, **d)** ORF length and the **e)** diversity of isoforms 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 2.1.2.3). 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.

		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				

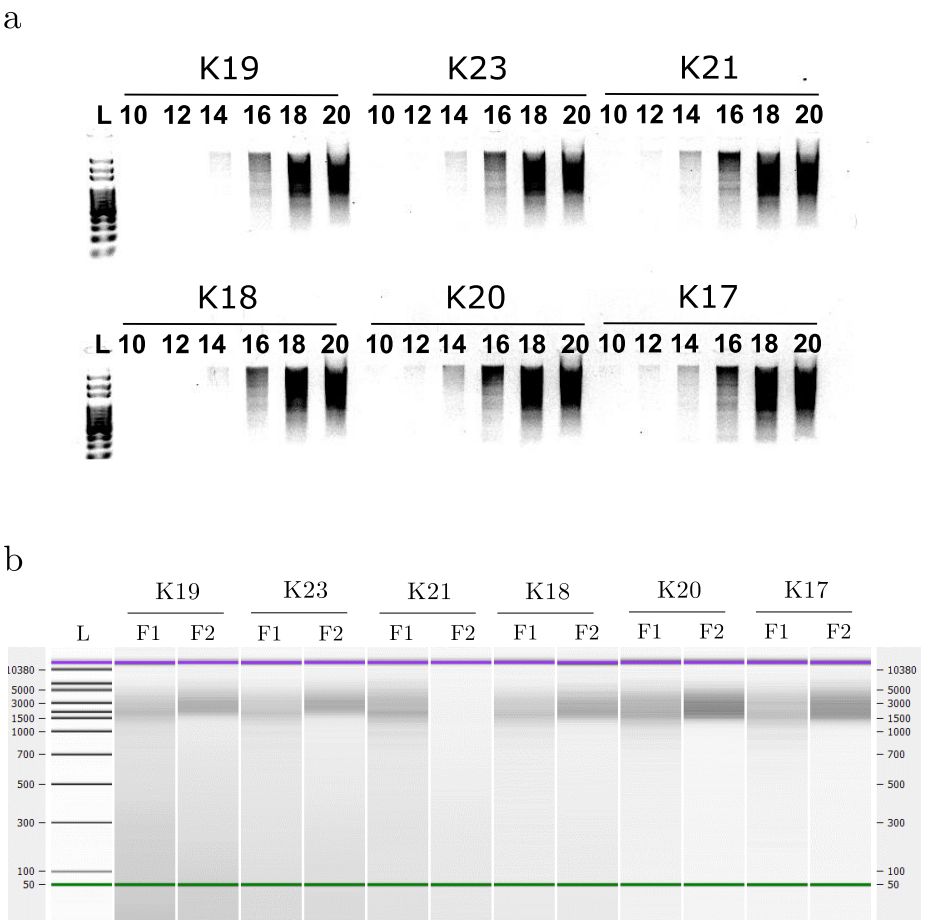


Figure 4.1: The first stage between the targeted and whole transcriptome sequencing is the same with samples typically amplified using 14 cycles followed by enrichment of high molecular weight cDNA in Fraction 2: a) Like whole transcriptome sequencing, samples were amplified using 14 cycles (Figure 3.1) 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.

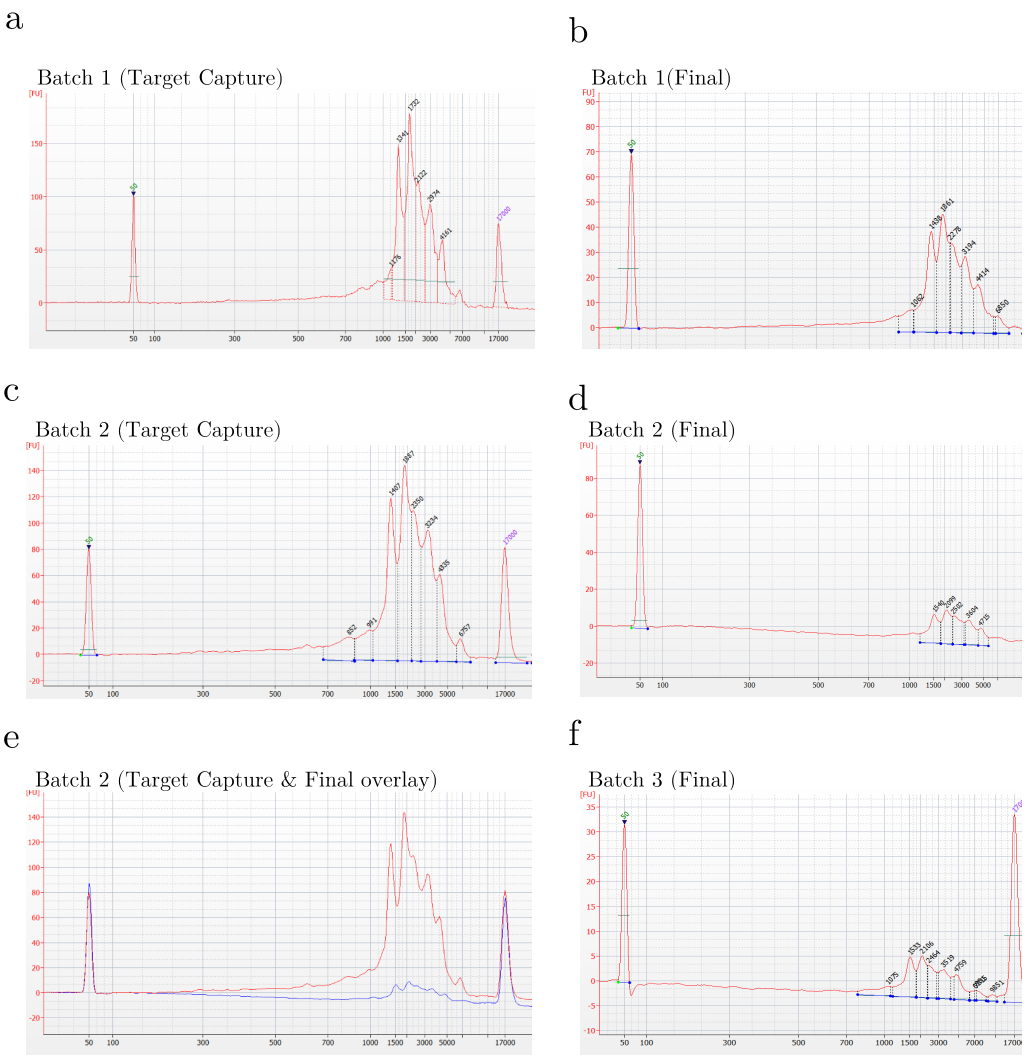


Figure 4.2: Successful target capture and library preparation across all batches, as shown by enrichment of transcripts with specific lengths: a) and **c)** are bioanalyzer electropherogram traces of Batch 1 (n = 6) and Batch 2 (n = 9) respectively after enrichment of cDNA with selective IDT probes (Section 2.1.2.8). **b)**, **d)** and **f)** are bioanalyzer electropherogram traces of Batch 1, 2 and 3 respectively after library preparation (denoted here as "Final", Section 2.1.2.9. **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

4.3.2 Transcriptome annotation

Across all the samples (n = 24), a total 757 isoforms were detected across 20 AD-associated target genes (Figure 4.5), of which *App* was detected with the highest number of isoforms (n = 121 isoforms) and *Trpa1* with the fewest (n = 2 isoforms).

4.3.3 Comparison with whole transcriptome

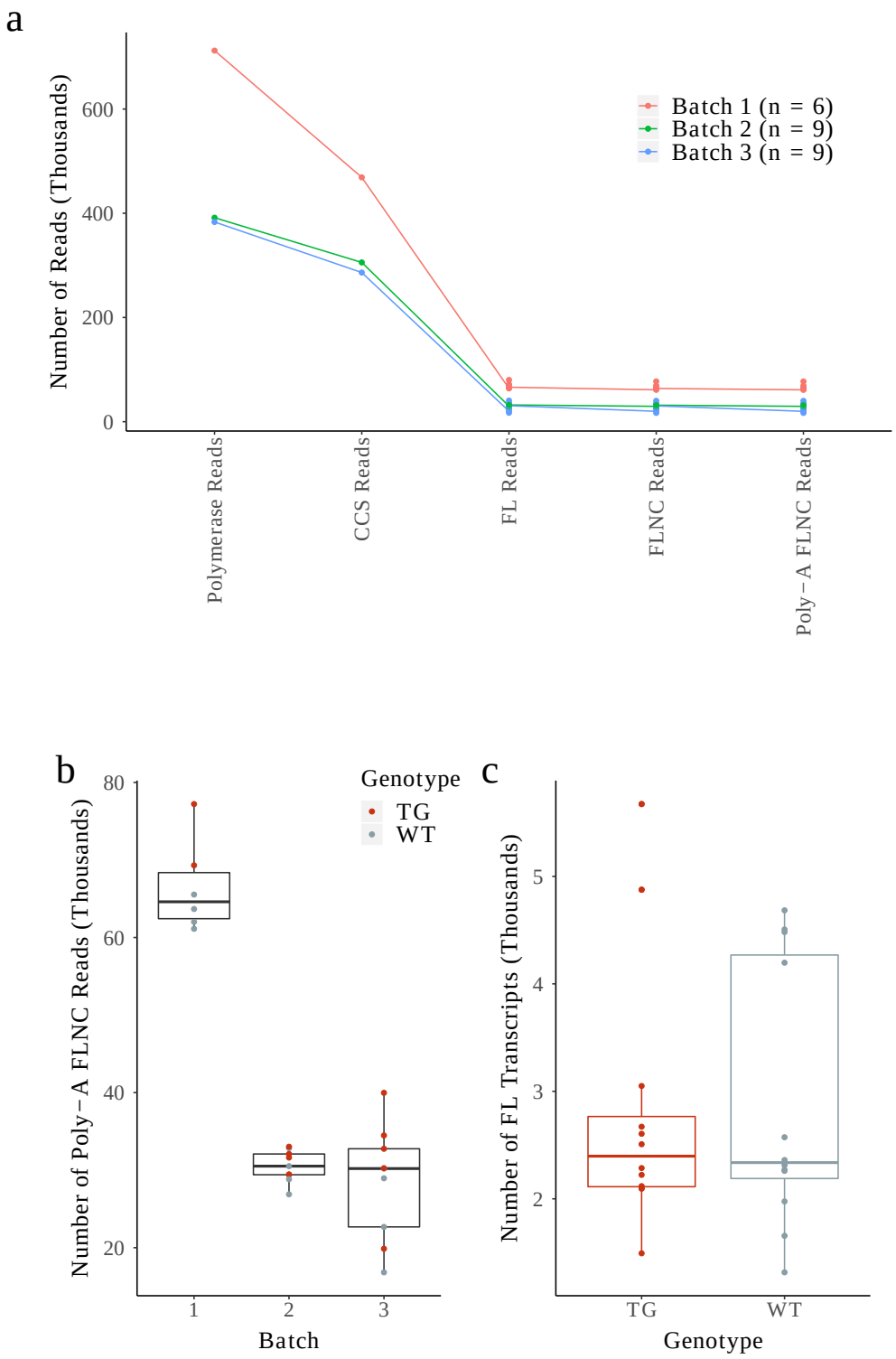


Figure 4.3: Despite batch variability in targeted transcriptome sequencing, no difference in the number of full-length transcripts was observed between wildtype and transgenic mice. **a)** Samples (n = 24) were multiplexed and sequenced in three runs (Batch 1, 2 and 3) with varied performance, as indicated 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

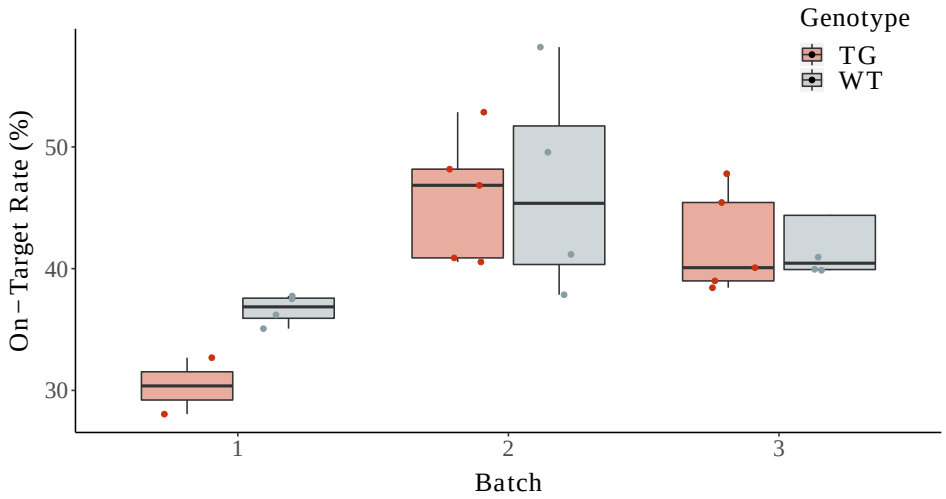
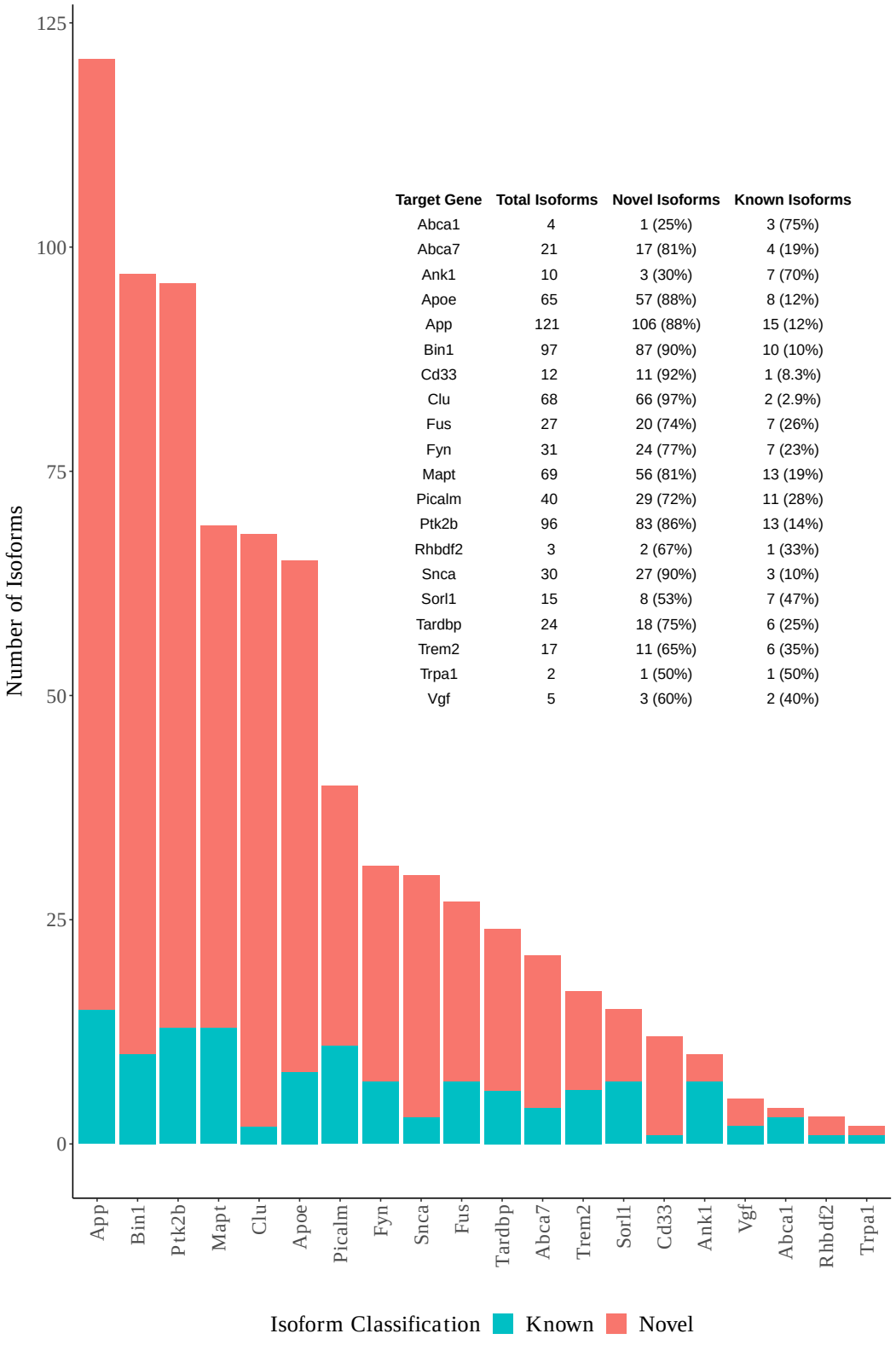
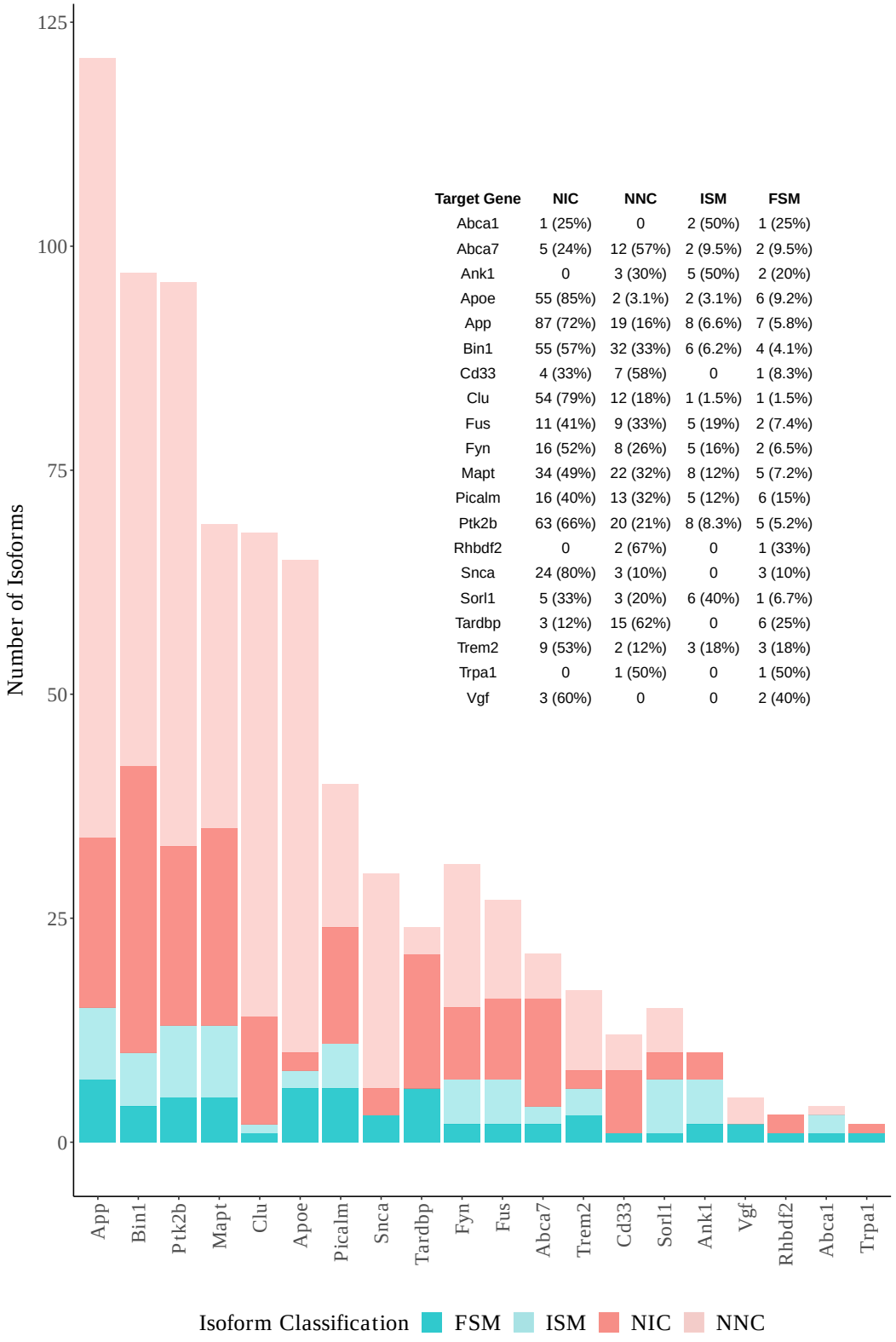


Figure 4.4: Coverage of target genes was greater in Batch 2 and 3 than Batch 1 due to more samples multiplexed and sequenced. Samples (n = 24) were multiplexed and sequenced in three runs (Batch 1 = 6 samples, Batch 2 = 9 samples, Batch 3 = 9 samples). Despite lower run yield output (4.2), Batch 2 and 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



Target Gene	Total Isoforms	Novel Isoforms	Known Isoforms
Abca1	4	1 (25%)	3 (75%)
Abca7	21	17 (81%)	4 (19%)
Ank1	10	3 (30%)	7 (70%)
Apoe	65	57 (88%)	8 (12%)
App	121	106 (88%)	15 (12%)
Bin1	97	87 (90%)	10 (10%)
Cd33	12	11 (92%)	1 (8.3%)
Clu	68	66 (97%)	2 (2.9%)
Fus	27	20 (74%)	7 (26%)
Fyn	31	24 (77%)	7 (23%)
Mapt	69	56 (81%)	13 (19%)
Picalm	40	29 (72%)	11 (28%)
Ptk2b	96	83 (86%)	13 (14%)
Rhbdf2	3	2 (67%)	1 (33%)
Snca	30	27 (90%)	3 (10%)
Sorl1	15	8 (53%)	7 (47%)
Tardbp	24	18 (75%)	6 (25%)
Trem2	17	11 (65%)	6 (35%)
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	5 (24%)	12 (57%)	2 (9.5%)	2 (9.5%)
Ank1	0	3 (30%)	5 (50%)	2 (20%)
Apoe	55 (85%)	2 (3.1%)	2 (3.1%)	6 (9.2%)
App	87 (72%)	19 (16%)	8 (6.6%)	7 (5.8%)
Bin1	55 (57%)	32 (33%)	6 (6.2%)	4 (4.1%)
Cd33	4 (33%)	7 (58%)	0	1 (8.3%)
Clu	54 (79%)	12 (18%)	1 (1.5%)	1 (1.5%)
Fus	11 (41%)	9 (33%)	5 (19%)	2 (7.4%)
Fyn	16 (52%)	8 (26%)	5 (16%)	2 (6.5%)
Mapt	34 (49%)	22 (32%)	8 (12%)	5 (7.2%)
Picalm	16 (40%)	13 (32%)	5 (12%)	6 (15%)
Ptk2b	63 (66%)	20 (21%)	8 (8.3%)	5 (5.2%)
Rhbdf2	0	2 (67%)	0	1 (33%)
Snca	24 (80%)	3 (10%)	0	3 (10%)
Sorl1	5 (33%)	3 (20%)	6 (40%)	1 (6.7%)
Tardbp	3 (12%)	15 (62%)	0	6 (25%)
Trem2	9 (53%)	2 (12%)	3 (18%)	3 (18%)
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).

Chapter 5

Transcriptional differences between WT and TG mice

Chapter 6

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 2.1.2.6)

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 2.1.2.7)
3. Determine library size using the BioAnalyzer with DNA 12000 Kit (Section 2.1.2.6)

Appendix B

Oxford Nanopore Transcriptome Protocol

Bibliography

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

[2] Karen E. Ocwieja, Scott Sherrill-Mix, Rithun Mukherjee, Rebecca Custers-Allen, Patricia David, Michael Brown, Susana Wang, Darren R. Link, Jeff Olson, Kevin Travers, Eric Schadt, and Frederic D. Bushman. Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. Nucleic Acids Research, 40(20):10345–10355, nov 2012.

[3] 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.

[4] 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.

[5] 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.

[6] 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. Johnson, 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.