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

1 Introduction 15

1.1 Alzheimer's Disease 15

1.1.1 Pathology 15

1.1.2 The two hallmarks: plaques and tangles 16

1.1.3 Genetic Component 16

1.1.4 Mouse Models 18

1.1.5 Currently available mouse models in AD 18

1.2 Gene expression and regulation 19

1.3 Transcriptional profiling 20

1.3.1 Short-read RNA-sequencing 22

1.3.2 Long-read sequencing approaches 25

1.3.3 Hybrid approach of short and long read sequencing 26

1.3.4 Isoform quantification 26

1.4 Aims and Objectives 28

1.5 Future Directions 29

2 General Methodology 30

2.1 RNA Extraction 30

2.2 Polymerase Chain Reaction (PCR) 30

2.3 Agarose Gel Electrophoresis 31

2.4 Bioanalyzer and Tapestation 31

2.5 Qubit 32

2.6 Target Capture using IDT Probes 33

3 Pacific Biosciences: Isoform Sequencing 34

3.1 Introduction 34

3.1.1 Single-molecule real time sequencing 34

3.2 Iso-Seq: Lab Pipeline . . . 38
3.2.1 CDNA synthesis . . . 38
3.2.2 PCR optimisation and DNA Amplification . . . 39
3.2.3 AMPure Bead Purification . . . 39
3.2.4 Target Capture using IDT Probes . . . 40
3.2.5 SMRT Bell Template Preparation . . . 41
3.2.6 Primer Annealing and Polymerase Binding . . . 42
3.3 Iso-Seq: Bioinformatics Pipeline . . . 43
3.3.1 ERCC . . . 43
3.3.2 Classify . . . 45
3.3.3 Cluster . . . 46
3.3.4 Genome/Transcriptome Alignment . . . 47
3.3.5 Genome Mapping . . . 48
3.3.6 Cupcake . . . 48
3.3.7 SQANTI2 . . . 49
3.3.8 Isoform expression from Iso-Seq . . . 50
3.3.9 Validation of isoforms with RNASeq . . . 51
3.3.10 SQANTI2 filtering . . . 51
3.3.11 Quantification of human transgene expression . . . 52
3.3.12 Classification of Alternative Splicing Events . . . 52
3.3.13 Limitations . . . 52
3.4 Iso-Seq: Optimisation . . . 53
3.4.1 Varying CCS parameters . . . 53
3.4.2 Additional parameters . . . 54
3.5 Iso-Seq Protocol . . . 57
3.5.1 Requirement of Sample quality . . . 58
3.5.2 General . . . 58
3.5.3 First Strand Synthesis . . . 62
3.5.4 PCR Cycle Optimisation . . . 62
3.5.5 Large-Scale PCR . . . 64
3.5.6 Bead Purification of Large-Scale PCR Products . . . 64
3.5.7 Pooling Fraction 1 (1X) and 2 (0.40X) . . . 65

3.5.8 Target Capture using IDT probes 65
3.5.9 SMRTbell Template Preparation 70

4 Oxford Nanopore: cDNA Sequencing 73
4.1 Oxford Nanopore 73
4.1.1 Mechanism 74
4.1.2 ONT Kits 74
4.2 Minion Sequencing: Lab Pipeline 74
4.2.1 cDNA synthesis 74
4.2.2 ONT MinION Library Preparation 75
4.3 Minion Sequencing: Bioinformatics Pipeline 76
4.3.1 Base-calling 76
4.3.2 Quality Control of Run and Base-called Reads 76
4.3.3 Filtering of Base-called Reads 77
4.3.4 Removing of Nanopore and cDNA sequencing adapters 77
4.3.5 Filtering and processing of trimmed reads 78
4.3.6 Genome Alignment 79
4.3.7 Transcript Collapse 79
4.3.8 Isoform Quantification 80
4.3.9 Limitations of Oxford Nanopore 80
4.4 ONT Protocol 80
4.4.1 cDNA Synthesis and Amplification 81
4.4.2 Bead Purification of Large Scale PCR Products 81
4.4.3 ONT MinION Library Preparation 82
4.4.4 Priming the Flow Cell 84
4.4.5 Library loading into the Flow Cell 85

5 Whole Transcriptome 87
5.1 Introduction 87
5.1.1 Mouse model of AD amyloidopathy: J20 87
5.1.2 Mouse model of AD tauopathy: rTg4510 87
5.2 Methods 89
5.3 Results 90

5.3.1 Bioinformatics output 94
5.3.2 Genome Mapping 94
5.3.3 Transcriptome annotation 95
5.3.4 Exploring the PacBio transcriptome of the mouse reference genome 95
5.3.5 Transcriptome abundance hierarchal between samples 96
5.3.6 Iso-Seq vs RNA-Seq 97
5.3.7 Change in endogeneous expression 97
5.3.8 LncRNA 98
5.3.9 Isoform diversity 98
5.3.10 Splicing Events 98
5.3.11 Validation of novel isoforms 98

6 Targeted Transcriptome **99**

7 ONT vs PacBio **100**

7.0.1 Comparisons of PacBio vs ONT for RNA-Sequencing 100

8 Proteomics (Eli-Lilly) **101**

9 Conclusion **102**

List of Figures

1.1 Challenges of using short-reads for transcript assembly 24

2.1 Evaluation of RNA integrity with Bioanalyzer and Tapestation 33

3.1 PacBio SMRT 36

3.2 Generation of Circular Consensus Sequence 37

3.3 PacBio Isoseq Bioinformatics Pipeline 44

3.4 Optimisation of CCS generation 54

3.5 No significant correlation between RIN and Whole Transcriptome Iso-Seq run output 56

3.6 No significant correlation between RIN and Whole Transcriptome Iso-Seq run output 57

5.1 Tg4510 WT and TG samples sequenced using Whole and Targeted Iso-Seq 89

5.2 Tg4510 WT and TG samples sequenced using Whole and Targeted ONT . 90

5.3 Whole Transcriptome Iso-Seq run output in transgenic and wild type Tg4510 mouse model 91

5.4 No significant correlation between RIN and Whole Transcriptome Iso-Seq run output 92

List of Tables

3.1 cDNA synthesis 62

3.2 PCR conditions for cDNA synthesis 63

3.3 Large Scale PCR 64

3.4 PCR conditions for Large Scale PCR 64

4.1 Repair DNA and Ends] 82

4.2 Loading Flow Cells 86

5.1 Run Yield Output from Whole Transcriptome Iso-Seq of Tg4510 91

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

1.1 Alzheimer's Disease

Alzheimer's disease (AD) is a devastating neurodegenerative disorder, clinically characterised by progressive memory loss, cognitive decline, and behavioural impairment. The most common form of dementia, it is estimated to affect XXX worldwide with numbers expecting to increase to X by 2050, ensuing both a heavy economic and social burden amounting to £XXX each year. Despite international efforts to better understand the disorder for drug discovery and development, there are currently no cure and existing medication only act to reduce symptoms.

1.1.1 Pathology

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

Neuroimaging scans and post-mortem brain analysis from patients reveal significant brain atrophy caused by neuronal and synaptic loss. Further microscopic examination reveal accumulation of beta-amyloid (Abeta) in amyloid plaques and aggregation of tau in neurofibrillary tangles, which are now believed to manifest years before presentation of clinical symptoms and diagnosis. In addition to these neuropathological changes, there is increasing evidence for the causative role of the innate immune system. Despite the well characterisation of these neuropathological hallmarks, the exact biological mechanisms driving AD onset and pathogenesis are still widely unknown.

1.1.2 The two hallmarks: plaques and tangles

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

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

1.1.3 Genetic Component

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

PSEN2 (presenilin 1 and 2, which are components of BACE) (Chai, 2007).

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

To date, the most recent GWAS meta-analysis of 74,000 AD individuals identified over XX significant LOAD risk loci, many of which were annotated to the non-coding cis regulatory regions of gene (Lambert et al., 2013). At least 42 genes/loci have been associated with LOAD at genome-wide significance in at least one GWAS (Verheijen and Sleegers, 2018). These genes included BIN1 (bridging integrator 1), CLU (clusterin), CR1 (complement receptor 1), PICALM (phosphatidylinositol binding clathrin assembly protein), with the most significant genetic locus annotated to APOE (apolipoprotein E); inheritance of both APOE allele increases the risk of AD development by X%. Common biological pathways emerging from these GWAS studies are immune response, lipid metabolism, endocytosis, and cell adhesion molecule (CAM) pathways ((Verheijen and Sleegers, 2018)).

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

1.1.4 Mouse Models

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

1.1.5 Currently available mouse models in AD

1.2 Gene expression and regulation

Common observation from gene expression analysis is that genes typically express multiple isoforms, and the greater the number of annotated isoforms, the greater the number of expressed isoforms (with a plateau of 12 isoforms) (Djebali et al. (2012)). However, as most studies are performed on bulk-tissues, it is unclear whether this is a consequence of multiple isoforms in one single cell or from multiple isoforms from multiple single cells. Perhaps assumed but the expression of alternative isoform is also not consistent, with usually a dominant isoform (Djebali et al. (2012))

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

1.3 Transcriptional profiling

Transcriptome profiling by the identification of full landscape of transcribed elements is critical to elucidate the functional relationship between the genomic loci and molecular mechanisms that drive development and diseases. Transcriptome profiling of disease-relevant tissue has enabled discovery of pathogenic coding and non-coding splicing variants in rare diseases, that would have otherwise been missed by exome and whole-genome sequencing in Mendialian disease diagnosis (Cummings et al. (2016), Kremer et al. (2017))

With Mendialian diseases such as Duchenne muscular dystrophy, pathogenic variants that result in aberrant splicing (exon inclusion, exon skipping, exon extension, intronic splice gain, exonic splice gain) can have significant downstream impacts (i.e. loss of function). A genetic variant can result in aberrant splicing in the following ways (Cummings et al. (2016)):

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

Transcriptome diversity is highly regulated by various mechanisms:

- Alternative transcription initiation (ATI))
- Alternative cleavage and alternative polyadenylation (APA))
- Alternative splicing (AS))

Alternative splicing and polyadenylation is a widespread phenomenon that facilitates generation of multiple distinct mRNA transcripts or isoforms from one gene, which are subsequently translated to different protein isoforms with unique, and potentially, antagonistic functions (E. T. Wang et al., 2008). AS further regulates gene expression through various mechanisms: non-sense mediated decay, miRNA-mediated mRNA degradation, altered translational efficiency of isoforms. In contrast, alternative polyadenylation regulates RNA transportation, localization, stability, and translation by generating splice isoforms with different cleavage sites.

Alternative splicing is essential in shaping transcriptome and proteome diversity - over 95% of 22,000 protein-coding multi-exonic human genes are estimated to undergo alternative splicing, with up to 70% containing multiple polyadenylation sites and 30% having multiple first exons due to alternative transcription start sites. Each gene is estimated to have on average six transcript isoforms Dunham et al. (2012), and this figure is likely to increase with more transcriptomic studies. It occurs most prevalently in the brain implicating its role in neuronal development and maintenance (Pan et al., 2008) (Mazin et al., 2014) (Raj, Blencowe, 2015). It is predicted that a single cell, with a transcription of 600,000 molecules, will have generated 5 - 15 conservative isoforms per gene, and 2-4 exon cassette isoforms ((Karlsson and Linnarsson, 2017)) (a single oligodendrocyte contained 2000 conservative transcripts associated with 700 genes, and 1000 unique isoforms).

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

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

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

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

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

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

Natural Antisense Transcripts

3'Polyadenylation Polyadenylation of 3'end of mRNA regulates mRNA stability and translation efficiency. Studies using long-read sequencing of human transcriptome have revealed differences in poly(A) length distribution between genes, and even between isoforms of the same gene with protein-coding isoforms having shorter poly-A tails than intron-retaining isoforms (Workman et al. (2019). This is line with studies showing that hyperadenylation targets intron-retaining transcripts for degradation (Bresson et al. (2015))

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

1.3.1 Short-read RNA-sequencing

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

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

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

Typically, several millions of 25-150bp (typically <700bp) reads are generated and aligned to genome to identify transcribed sequences. Major advances, including generation of reads that retain information on transcript orientation, allowing input of low yield or quality, have revolutionised the field. Also, now possible to sequence transcriptome *de novo*, allowing characterisation of novel organisms.

However, despite its power to identify and quantify gene expression (transcriptional profiling at a gene level), RNA-Seq is severely limited in assembling and reconstructing transcripts due to the reliance of short-reads that are only able to span a small part of the transcript rather than the full length (Figure X) Gordon et al. (2015)Wang et al. (2016); short reads have an average length of 100-500bp, whereas transcripts are on average 2-3kb - 50% of human transcripts are > 2.5Kb, with a range from 186bp to 109kb (Piovesan, Caracausi, Antonaros, Pelleri, & Vitale, 2016) Sharon et al. (2013). In particular, there are three transcriptional features that are difficult to characterise with short reads Kuo et al. (2017):

1. Transcript start sites (TSS) and Transcript termination sites (TTS) ,for which any interior multiple TSS and TTS sites within a transcribed locus would be undetected due to overlapping exons and splicing junctions, and low coverage
2. Exon chaining given that short-reads typically only span one splice junction. Thus, while short-reads may able to accurately identify the exons present, the exact sequence and linking of the exons are predicted by short-read assemblers with challenges (Figure 1.1).
3. Transcriptional Noise, particularly of reads in intronic regions that are falsely identified as intron retention, or of reads in intergenic regions that are erroneously

classified as fusion gene.

It is therefore unclear which combination of exons are spliced in, and whether alternative (distant) exons pairs are included in mutually exclusive or independent fashion (i.e. whether events are coordinated though some distant alternative exons have shown to be correlated included (Fagnani et al. (2007))Furthermore, short-read RNA-sequencing fails to capture the connectivity of exons and informs whether the alternative processive events are coordinated (coordination is defined by two or more alternative RNA processing events are dependent of each other and the probability of this occurrence is greater than the observation of the sole event). –> Molecular co-association of distant human alternative exons

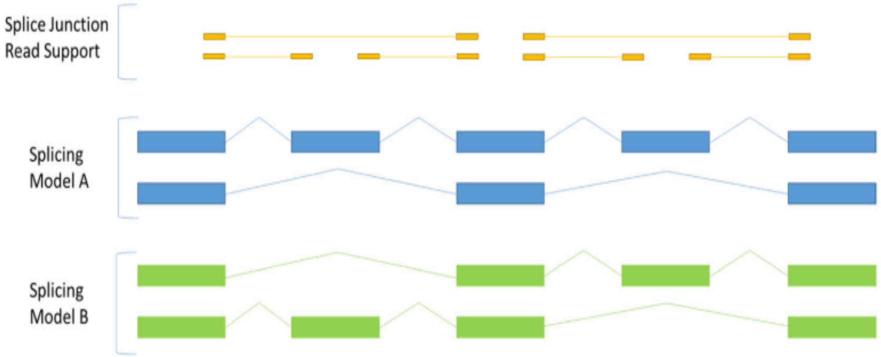


Figure 1.1: Challenges of using short-reads for transcript assembly: Example of a transcript model that is impossible to resolve using short-reads (yellow). Figure and caption taken from Kuo et al. (2017)

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

Attempts to overcome challenges with transcriptome assembly included generation of "synthetic long reads", by tagging full-length complementary DNAs with unique molecular identifiers (UMIs) before cluster amplification and sequencing on Illumina (Tilgner et al., 2015). With the presence of UMIs, transcript isoforms can be reconstructed for up to 4Kb for isoform discovery and expression analysis (Stark, Grzelak, & Hadfield, 2019). [However. . .] RNA-Seq is thus impaired to profile the transcriptome at an isoform-level, investigate cis-acting mechanisms with transcripts, and characterise the functional aspects of isoform diversity (Tardaguila et al., 2018)(Hayer et al., 2015).

1.3.2 Long-read sequencing approaches

The limitations with RNA-Seq were addressed with the emergence of long-read, third-generation sequencing approaches, which generated longer reads that were able to span the full-length transcript. Rather than massively-parallel sequencing of templates in "wash-and-scan" fashion that resulted in de-phasing and subsequently shorter reads, both platforms allowed real-time sequencing of templates in an uninterrupted and processive manner. Two technologies currently dominate this space: Single Molecule Real Time (SMRT) from Pacific Biosciences (PacBio) and protein nanopore sequencing technology from Oxford Nanopore Technologies (ONT). The performance and cost specifications of these two platforms are outlined in Table X. Other long read sequencing methods and protocols, synthetic long read (SLR) (Tilgner et al. (2015)) or sparse isoform sequencing (spISO-seq) (Tilgner et al. (2018)), however these require more complex workflows.

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

1.3.3 Hybrid approach of short and long read sequencing

Despite the ability of long-read sequencing (particularly, Iso-Seq) to discover large number of novel and longer transcripts and identify complex splicing events such as alternative adenylation, there are inherent biases to sequencing the more highly-expressed and relatively shorter transcripts. Consequently, while the new chemistry has improved the error rate and increased throughput, the coverage is still insufficient for accurate transcript quantification and sensitive differential transcript analysis based on long reads alone (Koren et al., 2012). Furthermore, there is currently no consensus to validate or functionally characterise these transcripts (B. Wang, Kumar, Olson, & Ware, 2019). The current standard for such application is thus a hybrid approach of aligning the short-reads to the long-reads to improve alignment and assemblage, and for downstream isoform quantification.

1.3.4 Isoform quantification

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

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

1. Inclusion level, calculated for a regulated exon by aligning reads either to candidate alternative exons and its junctions (inclusion reads), or to flanking exons and subsequently skipping the candidate alternative exon (skipping/exclusion reads) (Chen et al. 2012)

2. Percent-Spliced-In (PSI), calculated by proportion of isoforms that include the exon (Venables et al. 2008)(Katz et al. 2010). If the PSI value is calculated for a particular splicing event, it can be considered equivalent to the inclusion level.

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

Isoform abundance calculated by aligning short-reads to transcriptome is preferential to alignment with reference annotation library (RefSeq/GENCODE) in narrowing down the isoforms expressed and thus subsequently enabling more reliable abundance quantification. Reference annotation library is constructed on all data from the same species, and inclusion of annotated but not truly expressed isoforms can increase variability of abundance estimates. Finally, if the reference library is incomplete, then truly expressed isoforms would be completely missed and RNA-Seq reads would be incorrectly assigned to annotated isoform (**?**u2013)

Differential splicing analysis

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

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

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

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

1.4 Aims and Objectives

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

Gene expression and mRNA isoforms vary widely across tissues (Wang et al. (2008)), thus sequencing the disease-relevant tissue (in this case entorhinal cortex) is important for understanding the pathology of AD. However, it is consequently important to note that other tissues may have to be considered to fully grasp the whole picture of AD development.

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

1.5 Future Directions

At time of writing, there have been other major advances in the field that would unfortunately not be explored. This include, single cell transcriptomics and direct RNA-Sequencing: analysis of mRNA expression at the resolution of individual, "single", cells, allowing representation of cell-to-cell variation rather than taking the stochastic average from bulk measurements, and thereby resolving heterogeneity. This is currently achieved by the capture and analysis of single cells using a microfluidic or droplet-based technology. Importance of single cell approaches highlighted in Karlsson and Linnarsson (2017) with few isoforms shared between cells (7% of all detected isoforms shared between all cell-types, though this increased to 60% for exon-cassette isoforms).

Single-cell studies have highlighted the difference in transcriptome diversity at a single cell level, with small overlap of isoforms between cells (**?**arlsson2017). Previous methods on quantifying transcripts at a single cell level have relied on RNA-fluorescence in-situ Hybridisation (RNA-FISH), which is limited in terms of throughput and characterisation of complex splicing events (**?**yrne2017)

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

Chapter 2

General Methodology

2.1 RNA Extraction

Total RNA from mouse entorhinal cortex was extracted by Dr. Isabel Castanho using the AllPrep DNA/RNA Mini Kit (Qiagen). Samples were selected for long-read sequencing based on RNA quality and quantity, previously determined using RNA ScreenTape assay (Agilent) on 2200 TapeStation System or the RNA 6000 Nano kit (Agilent) on the 2100 Bioanalyzer (Section 2.4).

2.2 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.

Single-stranded DNA generated from SMARTer PCR synthesis kit in the official Iso-Seq protocol was amplified by PCR using PrimeSTAR GXL DNA Polymerase (ClonTech). Also performed for targeted sequencing.

2.3 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.

For this thesis, visualisation of DNA through gel electrophoresis was required primarily for optimising the number of PCR cycles for amplification, and for validating transcripts identified from Iso-Seq in Chapters X. A 1.5% agarose gel was made, with the separated DNA visualised using ethidium bromide on XXXX, as detailed in Section 3.5.4.1.

2.4 Bioanalyzer and Tapestation

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

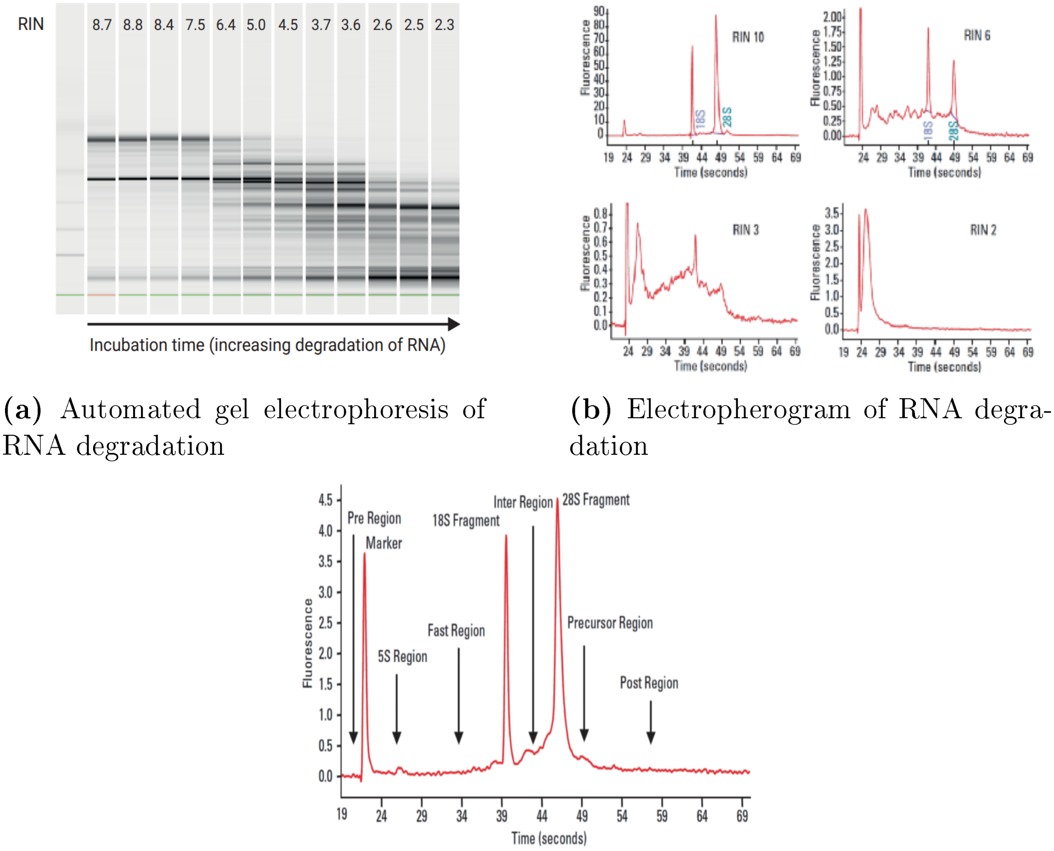
ScreenTape.

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 assay for accurate determination of library molarity (Section X). However, the D5000 ScreenTape (Agilent) was used 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 2.1). 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.

2.5 Qubit

Qubit assays 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 3.5.2.2.



(a) Automated gel electrophoresis of RNA degradation

(b) Electropherogram of RNA degradation

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

Figure 2.1: 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.

2.6 Target Capture using IDT Probes

For gene enrichment in targeted sequencing, the official PacBio protocol "cDNA Capture Using IDT xGen® Lockdown Probes" was used, with minor changes adapted from the official IDT protocol "xGen hybridisation capture of DNA libraries", as outlined in Figure X. In brief, the samples with unique non-overlapping barcodes were pooled in equal molarity post ampure bead purification and hybridised with target-specific probes; the captured cDNA was then washed with multiple wash buffers, amplified and purified with AMPure beads. Pooling 6 – 9 libraries prior to target capture simplifies laboratory workflow and minimises associated sequencing costs.

Chapter 3

Pacific Biosciences: Isoform Sequencing

3.1 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.

3.1.1 Single-molecule real time sequencing

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 3.1a).
2. Sequencing of each SMRTbell in a separate nanometre-wide well (zero-mode-waveguide

- ZMW), and all wells contained within a single SMRT chip (Levene et al. (2003)). 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 3.1c)).

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 (Figure 3.1b (Mccarthy (2010).

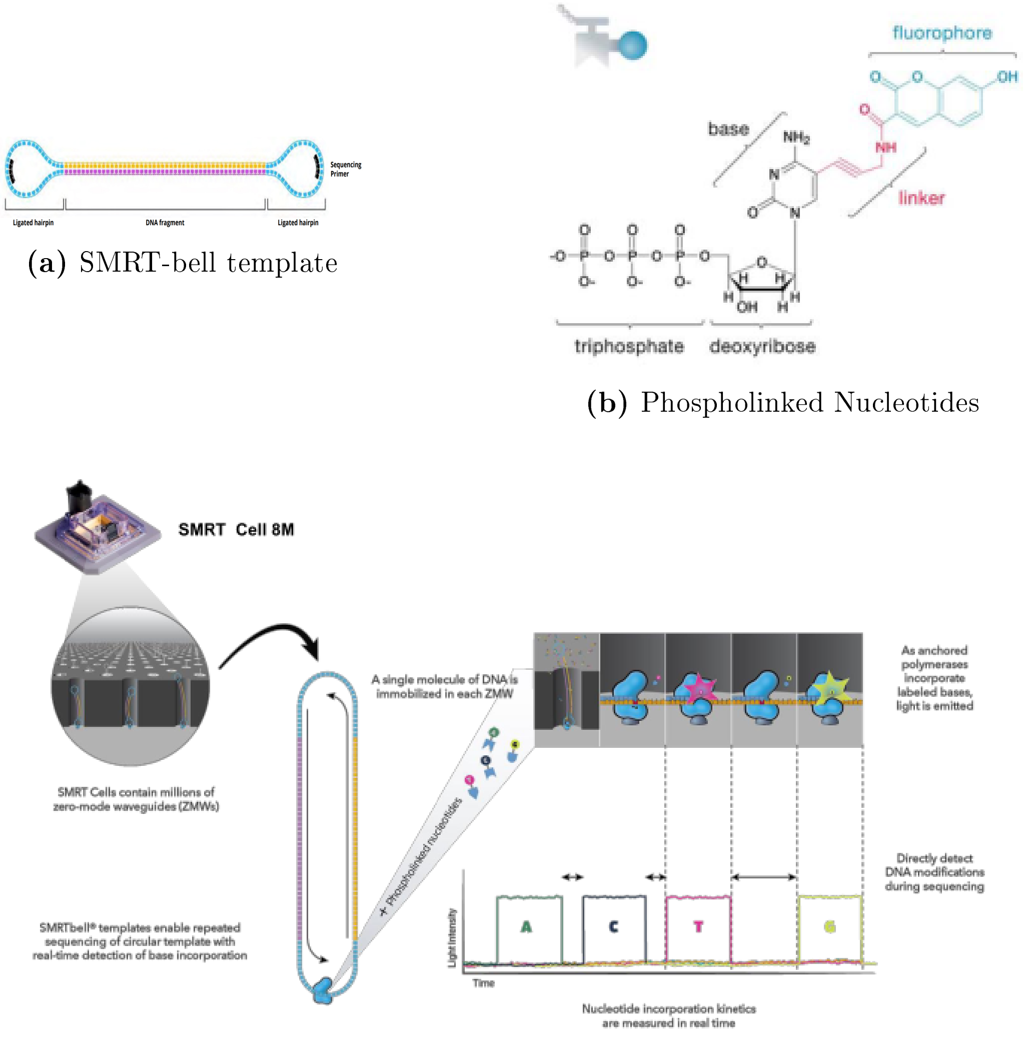
In summary, SMRT sequencing detect fluorescence events that correspond to addition of one specific nucleotide by a polymerase attached to the bottom of a tiny well.

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.

3.1.1.1 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)(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 -



Sequencing Primer
Ligated hairpin
DNA fragment
Ligated hairpin
fluorophore
OH
base
NH2
NH
linker
triphosphate
deoxyribose
SMRT Cell 8M
SMRT Cells contain millions of zero-mode waveguides (ZMWs)
SMRTbell® templates enable repeated sequencing of circular template with real-time detection of base incorporation
A single molecule of DNA is immobilized in each ZMW
+ Phospholinked nucleotides
As anchored polymerases incorporate labeled bases, light is emitted
Directly detect DNA modifications during sequencing
Light intensity
Time
A
C
T
G
Nucleotide incorporation kinetics are measured in real time

(a) SMRT-bell template

(b) Phospholinked Nucleotides

(c) PacBio Single-molecule real time sequencing

Figure 3.1: 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

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 Weirather et al. (2017). 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.

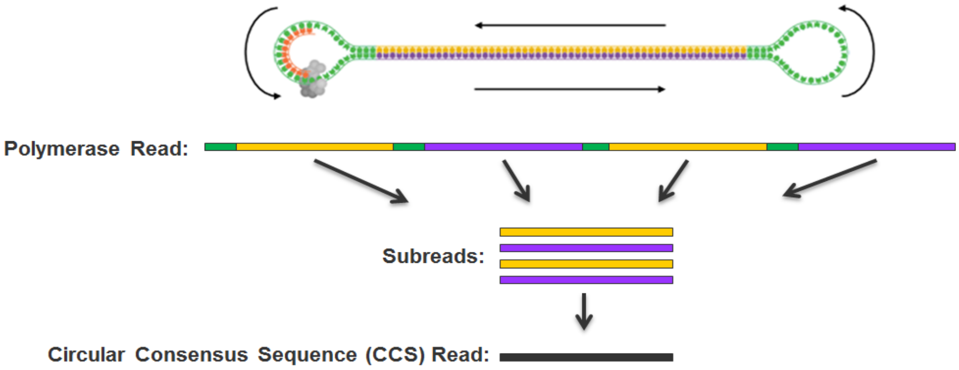


Figure 3.2: Generation of Circular Consensus Sequence: CCS is generated by the collapse of multiple subreads, which sequence correspond to the double-stranded cDNA of interest. The greater the number of "passes" sequenced by the polymerase, the longer the polymerase read, the more subreads generated, and subsequently the higher the quality of CCS. Picture adapted from PacBio

3.1.1.2 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.

3.2 Iso-Seq: Lab Pipeline

In brief, the Iso-Seq protocol involved 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. The cDNA 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.

3.2.1 CDNA synthesis

As part of the official Iso-Seq protocol, SMARTer PCR cDNA Synthesis Kit (Clontech) was used to convert 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 Ramsköld et al. (2012). 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. 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)(Cartolano et al. (2016)).

For this thesis, 200ng of total RNA was used for each sample for consistency and to ease downstream analyses.

3.2.2 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 was determined (Figure X). As described in Section X (Chapter 2), 5uL PCR aliquots were collected every two cycles (cycle 10, 12, 14, 16, 18) and run on a 1.5% Agarose gel electrophoresis. With 200ng total input of total RNA, cycles 14 – 15 were selected for large scale amplification across all the mouse samples to generate sufficient amount of double-stranded cDNA product for SMRTbell library construction (Figure X).

3.2.3 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), as described in Section X (Chapter 2). In brief, ds-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 2100 (Agilent), as described in Section X, Chapter X. Two fractions per sample were then recombined at equimolar quantities and library preparation performed using SMRTbell Template Prep Kit v1.0 (PacBio) (Figure X).

The molarity was calculated by the following equation:

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

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

3.2.4 Target Capture using IDT Probes

As an additional step to the standard protocol for whole transcriptome sequencing for the targeted approach, amplified and purified cDNA was further enriched using IDT probes (Section X). In brief, probes and blocking oligonucleotides were first added to cDNA, to allow hybridisation of probes to the target sequences/genes and oligonucleotides to the poly-T tract and cDNA primer to reduce non-specific binding. The hybridised library fragments were then incubated with washed magnetic streptavidin beads, and amplified using Takara Hot-Start polymerase (rather than KAPA HiFi from the official IDT protocol). The amplified library, still containing streptavidin beads, then underwent AMPure bead purification (Section X) for the elution of target cDNA (Figure X). SMRT Bell template preparation, primer and polymerase annealing was then proceeded.

*Modifications to the protocol: waiting times at room temperature during hybridisation, lid heat temperatures, method of washing beads at room temperature; all modifications are incorporated from official IDT protocol, post amplification clean-up for consistency

Targeted Sequencing

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), 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.

[Other methods of Targeted Sequencing i.e. CRISPR]

Target genes is enriched from dsDNA using complementary IDT xGen Lockdown probes, which are individually synthesised, 5' biotinylated DNA oligonucleotides. The hybrid capture is carried out using IDT xGEN hybridisation and wash kit protocol, using streptavidin-coated magnetic beads to bind and extract biotinylated probe-hybridised target DNA.

120nt-long probes were designed to a panel of 20 AD-associated genes: Bin1, Trem2, Cd33, Vgf, FynMapt, 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). To ensure full coverage of all transcripts, a list of probes was manually curated on the following criteria (Figure X): • 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 A full list of the number of probes per gene can be found in Supplementary X. The probes for all the target genes were delivered and resuspended in one pooled tube, in equimolar amounts. Note: no spike-in was added as control i.e. extra probe

3.2.5 SMRT Bell Template Preparation

As described in Section X (Chapter 2), DNA Damage and End Repair was 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).

3.2.6 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.

3.3 Iso-Seq: Bioinformatics Pipeline

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

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

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

Analysing long-read sequencing data requires a different approach to short-read, as the initial processing focuses on reducing the high error rate (due to low read coverage relative to short reads). Currently there are three methods of correcting long reads Zhao et al. (2019):

- Hybrid error correction strategy using short-reads: LSC Au et al. (2012) which maps short reads, and LoRDEC which build De Brujin graph of short reads Salmela and Rivals (2014)
- Self-correction using long reads only: Long-read multiple aligner (LoRMA) Salmela et al. (2017)
- Reference-based correction by alignment of reads to reference genome by spliced-aware 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.

3.3.1 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,

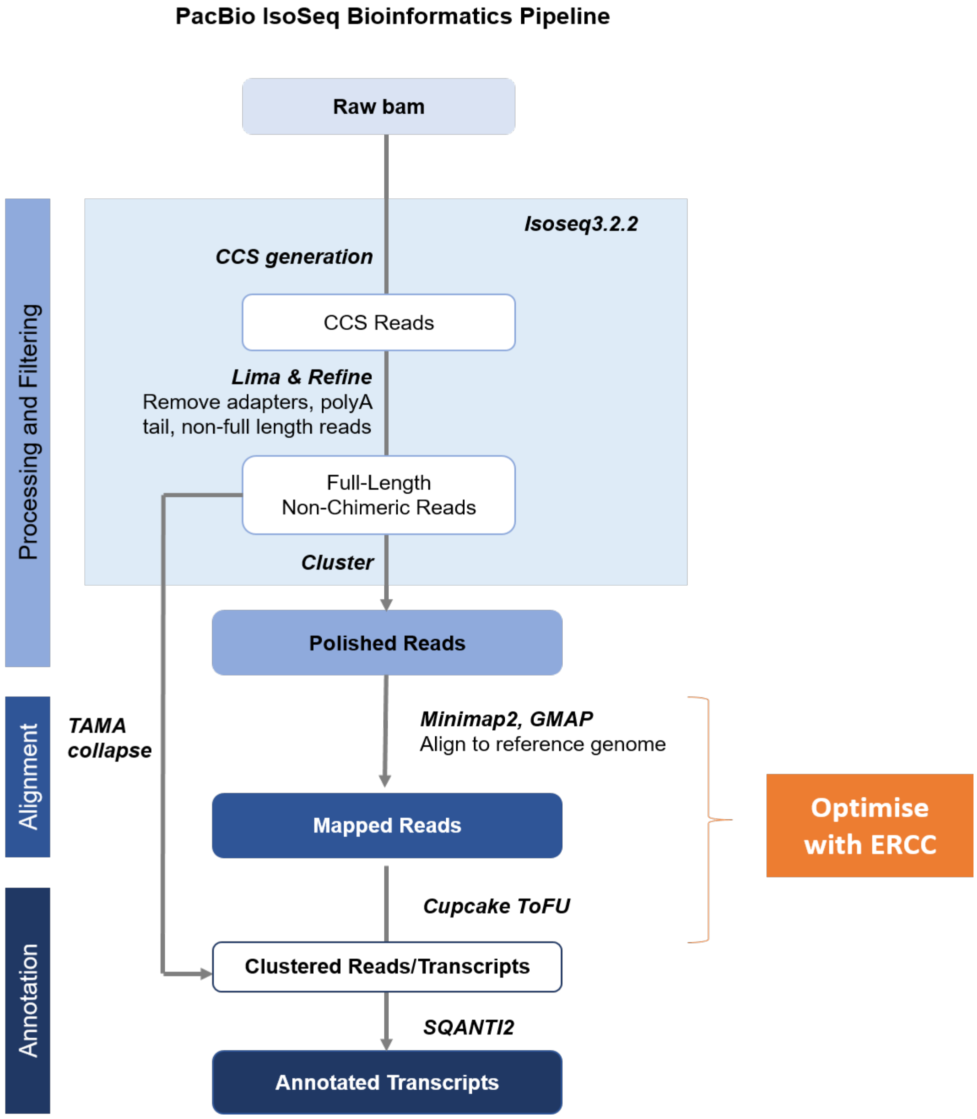


Figure 3.3: PacBio Isoseq Bioinformatics Pipeline: Pipeline is adapted from ToFU Gordon et al. (2015)

with known start and end positions can be used as benchmark. As detailed in Karlsson and Linnarsson (2017), most ERCC reads fell within +/- 5bp at both 5' and 3' ends, with 3' end slightly more accurate than 5' end. From Sharon et al. (2013), 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 ((Karlsson and Linnarsson, 2017)) 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 are >2kB, there is a need for longer control sequences.

3.3.2 Classify

CCS Generation: In the first stage, the raw subreads (stored as a BAM file, unaligned.bam) from each "productive" ZMW are processed individually and collapsed to generate a CCS (Figure 3.2), 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 are 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 are identified and then removed using lima. CCS with unwanted orientations are removed and are 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 are 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 are 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 rare ($<$0.5%). Conversely, it is challenging to differentiate and remove PCR-induced artificial chimera from true biological chimera. PCR-induced artefacts are defined as cDNA sequences that appear to be fusion transcripts, but are 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

3.3.3 Cluster

In the second stage, Iso-Seq uses an iterative isoform-clustering algorithm (ICE – iterative clustering for error, called Quiver for PacBio RSII data and Arrow for PacBio Sequel data) to group all FLNC reads that are 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 are 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 are further required for a cluster. Two possible issues: reads belong to incorrect clusters, and reads that belong together are 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 are met. CCS are then trimmed and processed for primer and poly-A sequence removal to generate full-length non-chimeric (FLNC) reads, which are clustered if they are 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 compared 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 are 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 compared 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

3.3.4 Genome/Transcriptome Alignment

High quality isoforms are then aligned to the reference genome (as opposed to transcriptome as otherwise miss novel isoforms using BLASR) using splice-aware 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-aware 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%).

3.3.5 Genome Mapping

HQ-isoforms from the pooled dataset were aligned to mouse genome using Minimap2, and a total of XXX reads (XX%) were mapped. Errors for substitution, insertion and deletion are 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)

3.3.6 Cupcake

To avoid redundancy of transcripts, aligned and filtered HQ transcripts are 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 are 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 are filtered by 5'degradation due to the lack of a cap protection employed in the cDNA synthesis step (Clontech SMARTer cDNA kit).

3.3.7 SQANTI2

High-quality, clustered, filtered isoforms from Cupcake are characterised using SQANTI2, a pipeline initially developed by Conesa et al. [ref] and refined by Elizabeth Tseng (Pacific Bioscience's specialist) [ref]. In addition to providing the descriptive statistics for each transcript (such as the chromosome, the strand, the length and the number of exons), it further allows input of public datasets and RNA-Seq data to further characterise and validate isoforms from Iso-Seq:

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

SQANTI2 pipeline involves:

1. Performs a reference-based correction of sequences [check??]
2. Generates gene models and classifies transcripts based on splice junctions (See Section X)
3. Predicts Open Reading Frames (ORF) for each transcript
4. Remove isoforms potential to be artefacts

3.3.7.1 SQANTI2 classification of isoforms

Using SQANTI classifications based o splice junctions, the transcriptome can be segregated into:

- Well-known annotated genes with known transcripts, further 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 transcripts
  - 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/ac-

ceptor 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

- Unannotated, novel genes with novel transcripts
- Others
  - Antisense: the query isoform does not have overlap a same-strand reference gene but is anti-sense to an annotated gene.
  - Genic Intron: the query isoform is completely contained within an annotated intron.
  - Genic Genomic: the query isoform overlaps with introns and exons.
  - 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

3.3.8 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:

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

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 Abrams et al. (2019). 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.

3.3.9 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

3.3.10 SQANTI2 filtering

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 are 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 are labelled as RT switch.

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

3.3.11 Quantification of human transgene expression

As reported in Castanho et al. (2020), 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.

3.3.12 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).

3.3.13 Limitations

While PacBio's Iso-Seq have major potential for transcriptome annotation, there are currently several major limitations that need to be addressed with further development of library preparation and bioinformatic data analyses Kuo et al. (2017):

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

3.4 Iso-Seq: Optimisation

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

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

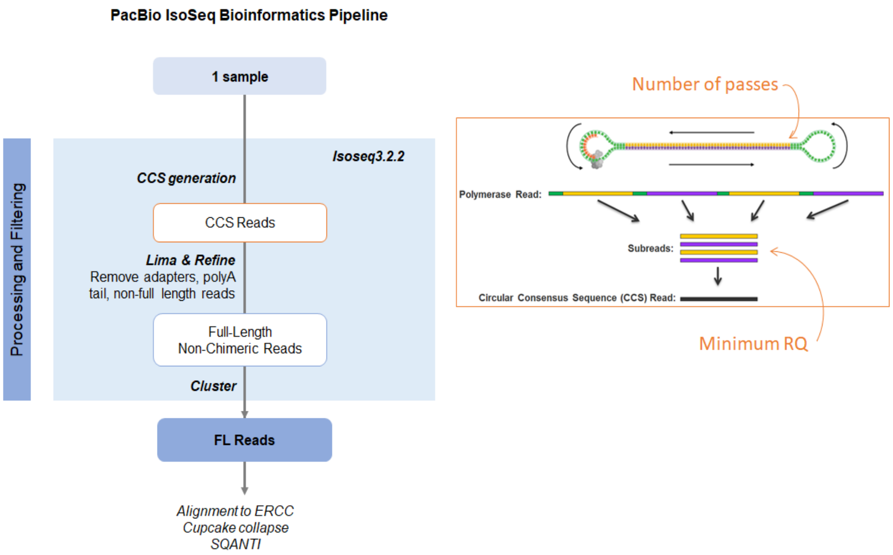
3.4.1 Varying CCS parameters

As described in Section 3.3.2, the proportion of raw subreads that can be successfully collapsed to generate a CCS is widely influenced by the number of passes (default: 3 passes) and minimum base accuracy (default: 99%), which settings are widely varied in scientific community.

To determine the most optimum parameters for CCS generation, CCS was generated on a subset of 1 sample using a combination of parameters, and then further validated with 2 whole samples (Figure 3.4):

Conclusion: 0.9 and 1 pass

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



(a) Factors influencing successful CCS generation of raw subreads

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

(b) 1st CCS parameter optimisation

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

(c) 2nd CCS parameter optimisation

Figure 3.4: Optimisation of CCS generation: Successful CCS generation of raw subreads is dependent on the number of polymerase passes and minimum RQ of subreads. A two-step approach was taken to determine the optimum parameters, using (b) whole range of parameters on 10% of one sample, and (c) extending analyses to two samples but with a more refined combination (as determined from the results of first step))

3.4.2 Additional parameters

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

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

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

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

3.4.2.1 Application of additional parameters to Whole Transcriptome

A merged analysis of whole transcriptome samples (n = 12, WT = 6, TG = 6) was performed with alignment to the mouse genome (mm10), with 276,035 reads (99.3%) mapped to mouse genome, 365 reads (0.13%) mapped to ERCC and 1,568 reads (0.56%) unmapped.

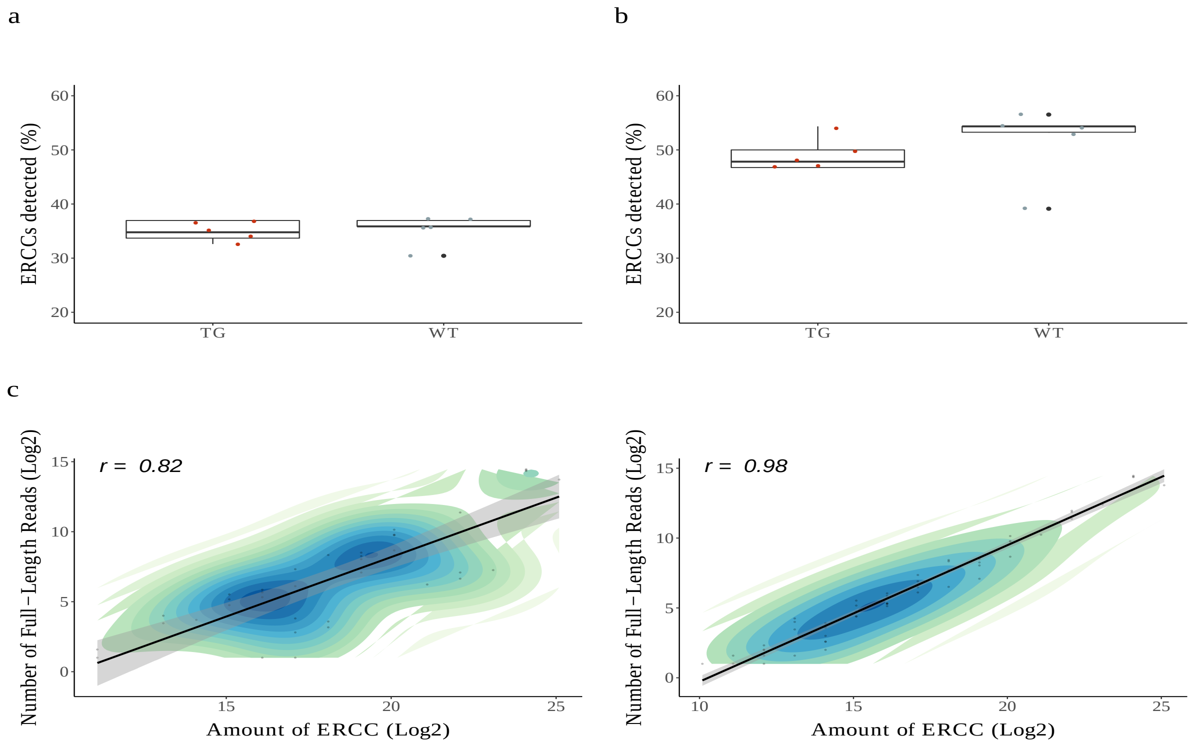


Figure 3.5: No significant correlation between RIN and Whole Transcriptome Iso-Seq run output: Samples with RIN >8 were selected for Whole Transcriptome Iso-Seq, with TG samples having distinctly lower RIN values than WT samples. However, no significant difference was observed for run output between WT and TG (Figure 5.1)

using cupcake's collapse default (99%) and reduced threshold (85%) for alignment identity.

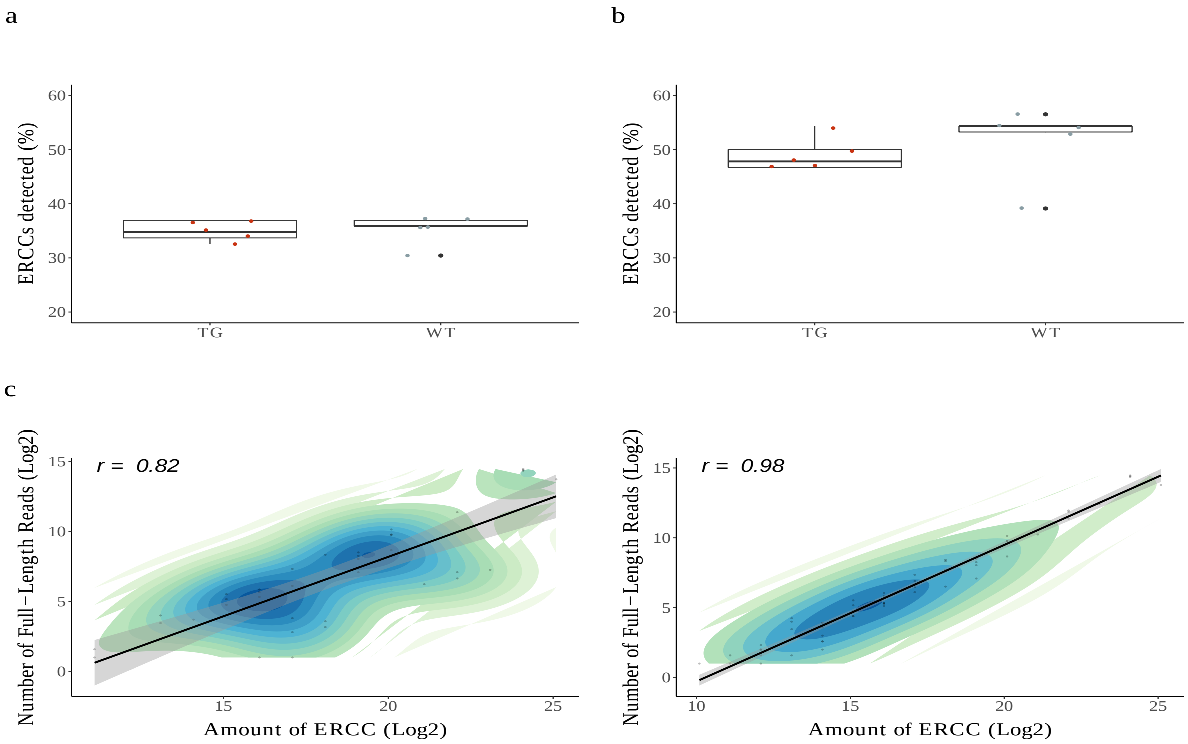


Figure 3.6: No significant correlation between RIN and Whole Transcriptome Iso-Seq run output: Samples with RIN >8 were selected for Whole Transcriptome Iso-Seq, with TG samples having distinctly lower RIN values than WT samples. However, no significant difference was observed for run output between WT and TG (Figure 5.1)

3.5 Iso-Seq Protocol

3.5.1 Requirement of Sample quality 58

3.5.2 General 58

3.5.2.1 Ampure Bead Purification 59

3.5.2.2 Assessment of DNA quantity using Qubit 60

3.5.2.3 Assessment of DNA library size using Tapestation or Bioanalyzer 61

3.5.3 First Strand Synthesis 62

3.5.4 PCR Cycle Optimisation 62

3.5.4.1 Running an agarose gel 63

3.5.5 Large-Scale PCR 64

3.5.6 Bead Purification of Large-Scale PCR Products 64

3.5.6.1 Fraction 1: 2nd purification 65

3.5.7 Pooling Fraction 1 (1X) and 2 (0.40X) 65
3.5.8 Target Capture using IDT probes 65
3.5.8.1 Prepare beads for Capture 66
3.5.8.2 Binding cDNA to beads 67
3.5.8.3 Perform heated washes 68
3.5.8.4 Perform room temperature washes 69
3.5.8.5 Amplification of Captured DNA Sample 69
3.5.9 SMRTbell Template Preparation 70
3.5.9.1 Repair DNA Damage and Ends 70
3.5.9.2 DNA Purification 70
3.5.9.3 Prepare Blunt Ligation Reaction 71
3.5.9.4 Adding Exonuclease to remove failed ligation products . . 71
3.5.9.5 First Purification of SMRTbell Templates 72
3.5.9.6 Second Purification of SMRTbell Templates 72

3.5.1 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

3.5.2 General

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

3.5.2.1 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.

3.5.2.2 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.

3.5.2.3 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.

3.5.3 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 3.1)

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 3.1: Long Description

3.5.4 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

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

Table 3.2: PCR conditions for cDNA synthesis

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)

3.5.4.1 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

3.5.5 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 3.3)
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 3.4

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 3.3: 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 3.4: PCR conditions for Large Scale PCR

3.5.6 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

3.5.6.1 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 2.4)

3.5.7 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".

3.5.8 Target Capture using IDT probes

Prepare hybridisation

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 30minutes)
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

3.5.8.1 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.

3.5.8.2 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

3.5.8.3 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.

3.5.8.4 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

3.5.8.5 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

3.5.9 SMRTbell Template Preparation

3.5.9.1 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

3.5.9.2 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)

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

3.5.9.3 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

3.5.9.4 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.

3.5.9.5 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

3.5.9.6 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 2.5)
3. Determine library size using the BioAnalyzer with DNA 12000 Kit (Section 2.4)

Chapter 4

Oxford Nanopore: cDNA Sequencing

4.1 Oxford Nanopore

In 2014, Oxford Nanopore Technology (ONT) introduced another long-read sequencing technology akin to PacBio's SMRT, in the ability to also generate long reads that are able to resolve the exon structure of mRNA transcripts. However, rather than mimicking the natural DNA synthesis as is the focus in all major sequencing applications including the PacBio's SMRT, ONT's nanopore-based sequencing adopts an entirely different approach; the DNA sequence is inferred from fluctuations in a current applied across a membrane as it passes through a protein pore.

Inhibited mostly by the ability to deliver very high-molecular weight DNA to the pore, nanopore sequencing is able to generate much longer reads than SMRT sequencing (from 500bp to currently 2.3Mb (Payne et al. 2019)).

Basecalling accuracy of reads have dramatically increased, with raw base-called error rate reduced from <1% from SMRT sequencing and <5% for nanopore sequencing.

In comparison to SMRT sequencing, accuracy of nanopore reads is independent of DNA length, but reliant on achieving optimal translocation speed of DNA through pore, which often decreases in the late stages of sequencing, thereby negatively impacting quality.

Contrary to SMRT sequencing, each DNA fragment is read only once:

- 1D: each strand of dsDNA fragment is read independently, and single pass accuracy is final accuracy of fragment
- 1D^2: sequence complementary strand of dsDNA fragment immediately after, allowing determination of more accurate consensus strand, achieving a median consensus accuracy of 98%.

Low complexity stretches, including homopolymers, are difficult to resolve with current pores (R9) and basecallers, as translocation of homopolymers do not change the sequence of nucleotides within pore, thereby resulting in a constant signal. Significant major advances in technology have been made over the past 3 years, with 4 pore version released in 2019 (R9.4, R9.4.1, R9.5.1 and R10.0) Amarasinghe et al. (2020).

4.1.1 Mechanism

In contrast to Pacbio's XXX by XXX Sequel, ONT's nanopore sequencing can be performed in a handheld MINion device (10 × 3 × 2 cm, 90 g), housing a flow cell at its centre where the DNA sample is loaded. (Also on mobile device: Samarakoon et al. (2020)) Each flow cell contains a sensor array, consisting of 512 channels, each with 4 cells that can in turn house one nanopore. However, while the current MinION contains a total of 2048 nanopores, only one of the four wells in each channel can be active at any time, as controlled by Application Specific Integrated Circuit (ASIC), allowing up to 512 independent DNA molecules to be sequenced simultaneously.

1D vs 2D vs 1D2 *6-mer* *flow cell and the different nanopore?*

4.1.2 ONT Kits

4.2 Minion Sequencing: Lab Pipeline

4.2.1 cDNA synthesis

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

4.2.2 ONT MinION Library Preparation

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

4.2.2.1 Repair DNA and Ends

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

4.2.2.2 Adapter Ligation

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

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

4.2.2.3 Priming the Flow Cell

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

4.3 Minion Sequencing: Bioinformatics Pipeline

4.3.1 Base-calling

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

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

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

4.3.2 Quality Control of Run and Base-called Reads

There are a number of developed bioinformatic tools that provide a quality assessment of base-called reads, which provide information on:

1. Performance of the sequencing run for each flow cell
2. Distribution of base called read lengths
3. Distribution of quality scores: over base pair per read, over time across all reads across the flow cell
4. No of reads generated over time

4.3.3 Filtering of Base-called Reads

Base-called reads are filtered using NanoFilt (part of Nanopack) based on the following parameters:

1. Filter on a minimum average read quality score of 7
2. No Filter on a minimum read length
3. No Filter on a maximum read length

4.3.4 Removing of Nanopore and cDNA sequencing adapters

Similarly to the Iso-Seq protocol, nanpore ligation adapters and cDNA sequencing primer sequences are removed to prevent spurious alignment, using PoreChop. This tool finds and trims adapters at the end of reads, and remove any "chimeric" reads with adapters in the middle. As detailed in WTAC's course, the nanopore.read.py script was edited to generate the output file for downstream analyses, and the adapters.py was amended to include the specific adapter sequence in accordance to developer's instructions.

```
python porechop-runner.py
     --format fastq \ # input format
     -pass_reads.fq \ # input reads
     -o pass_reads_choped.fq \ # output trimmed reads

     # Additional parameters to run porechop
     # Discard reads with adapters in the middle, i.e. "chimeric reads"
     # An adapter would only be trimmed if it has 90% identity (default)
     # Report for presence and location of adaptors for every read as output_adaptor_alignments_stats
     --discard_middle \
     --adapter_threshold 90
     --verbosity 4 \

     # Parameters to control trimming of adaptors from read ends
     # 100 bases at each end of reads are searched for adaptor
     # Adapter alignments smaller than 15bases will be ignored
     # 1 base removed next to adapters found at the ends of reads
     # Adapters at ends of reads must have >75% identity before removed

     --end_size 100 \
     --min_trim_size 15 \
     --extra_end_trim 1 \
     --end_threshold 75 \
     > output_adaptor_alignment_stats
```

Listing 4.1: PoreChop command

4.3.4.1 Definition of sequence adapters for removal

After cDNA synthesis with SMARTer PCR cDNA synthesis kit (ClonTech), and ligation of ONT adapters, cDNA have the structure depicted in Figure X. Due to the nature of cDNA sythesis primers and ONT adapter sequences, it is able to:

1. identity the beginning (Transcription Start Site - TSS) and the end (Transcription End Site - TES) of the cDNA sequence
2. differentiate between the plus strand of the cDNA (strand corresponding to the original mRNA sequence with a poly-A tail) and the negative strand (complementary strand to original mRNA sequence with a poly-T tail)

The motor protein binds to either 5' end of the cDNA strand for translocation through the pore. The sequence of the ONT adapters used in the SQK-LSK109 kit are recorded in Table X.

4.3.5 Filtering and processing of trimmed reads

The trimmed reads from Porechop contains both plus and negative strands, and only reads containing at least one adapter at either end are retained to maximise read usage. However, it is important to be more stringent for isoform identification and quantification, and retain reads with both adapters at the end. It is further necessary to reverse complement the reads that correspond to the negative strands; all the reads would then start with the TSS and end with the polyA sequence. Using cutadapt, the polyA sequence are then trimmed with cutadapt with a window gap of 40 bases from the end (as polyT primer in cDNA synthesis introduces 30As and to avoid spurious base calling).

4.3.6 Genome Alignment

Similarly to Iso-Seq, trimmed reads are then aligned to the reference genome (as opposed to transcriptome as otherwise miss novel isoforms) using splice-aware aligner Minimap. However, unlike the Iso-Seq pipeline which generates high-quality transcripts, ONT reads are more prone to errors and having a lower mapping coverage (check). The reads from Minimap2 are thus further filtered based on alignment quality (identity and percentage of alignment length) using htsbox.

4.3.7 Transcript Collapse

Similarly to Iso-Seq, aligned and filtered reads need to be collapsed to generate unique transcripts. This is done with Cupcake in Iso-Seq, and TAMA collapse for ONT with the below parameters 4.2. The unique transcripts are then merged with the reference genome using TAMA merge for annotation and further characterisation. TAMA collapse and merge can also be performed for Iso-Seq reads, however, this is not part of the official pipeline.

```
python tama_collapse.py \
  -s aligned.sam \  # input file
  -f mm10.fa \    # genome fasta sequence
  -p tama_collapse_output \

  # Parameters to control collapse of transcripts by:
  # common exon start/end site rather than the longest feature
  # using only reads where 95% are mapped to the genome
  # 80% identity
  # Difference of 50 bases at 5' and 3' end for reads to be collapsed
  # Difference of 20 bases at splice junctions for reads to be collapsed
  # Merge duplicate transcript groups

  -e common_ends -c 95 -i 80 -x capped -a 50 -z 50 -m 20 -d  merge_dup

python tama_merge.py -f file.list -a 50 -z 50 -m 20 -p tama_merge_output
```

Listing 4.2: Tama collapse and merge command

FLAIR: Full-Length Alternative Isoform analysis of RNA (FLAIR) Three steps are involved: Correct splice sites with short reads if incorrect splice site is within 10base pairs away from correct splice site, collapse reads to generate consensus sequences. This involves first grouping reads with identical splice junctions - "first pass nanopore isoform transcriptome"; the representative isoform within each group is determined by the most supported transcription and end site. All the reads, including reads that were aligned but not able to be fully corrected, are re-aligned to the "first-pass isoform" with the best alignment. First-pass isoforms that have fewer than three supporting reads are filtered out; three supporting reads selected as threshold as this gave the highest base sensitivity without compromising on precision.

4.3.8 Isoform Quantification

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

4.3.9 Limitations of Oxford Nanopore

Refer to Workman et al. (2019) for information on strand break etc

4.4 ONT Protocol

This protocol was adapted from Wellcome Trust Advanced Course: RNA Transcriptomics (2018), provided by J.Ragoussis (referred as WTAC), the official ONT protocol "1D amplicon/cDNA by Ligation (SQK-LSK109)", and directed under the guidance of K.Moore, Exeter's sequencing services. In brief, this protocol aimed to complement the Iso-Seq Protocol (Section **??**) as a direct comparison of the two sequencing technologies. It was therefore important to ensure that all other steps, bar library preparation, were consistent (Figure X). Consequently, cDNA synthesis and amplification 3.2.1 was performed twice in parallel for the sample of interest, and the pipeline branched upon the respective library preparation.

4.4.1 cDNA Synthesis and Amplification 81
4.4.2 Bead Purification of Large Scale PCR Products 81
4.4.3 ONT MinION Library Preparation 82
4.4.3.1 Repair DNA and Ends 82
4.4.3.2 Bead Purification of cDNA end-repaired products 83
4.4.3.3 Prepare Ligation Reaction 83
4.4.3.4 Bead Purification of ligated cDNA 83
4.4.4 Priming the Flow Cell 84
4.4.5 Library loading into the Flow Cell 85

4.4.1 cDNA Synthesis and Amplification

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

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

4.4.2 Bead Purification of Large Scale PCR Products

1. Pool 800μL PCR reactions (16 x 50μL PCR reactions) and add 0.90X volume of AMPure PB (200μL) magnetic beads.

2. 20-30μL loss is expected from evaporation,therefore would not be able to recover 800μL of cDNA . Note: only prepare 1 Fraction for downstream library preparation rather than 2 Fractions in Iso-Seq
3. Proceed with AMPure PB Bead Purification (Section X), with 51μL of TE Buffer
4. Quantify DNA amount and concentration of using a high-sensitivity Qubit (Section X)
5. Determine library size using the BioAnalyzer with DNA 12000 Kit (Section 2.4)

4.4.3 ONT MinION Library Preparation

4.4.3.1 Repair DNA and Ends

1. Thaw DNA CS (DCS) at room temperature, spin down, mix by pipetting, and place on ice
2. Prepare the NEBNext FFPE DNA Repair Mix and NEBNext End repair / dA-tailing Module reagents in accordance with manufacturer's instructions, and place on ice
3. Prepare a PCR reaction mix for each sample in microcentrifuge tube (Table 4.1)
4. Mix gently by flicking tube and spin down
5. Incubate in thermal cycle at 20° C for 5 minutes and 65° C for 5 mins

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

Table 4.1: Repair DNA and Ends]

4.4.3.2 Bead Purification of cDNA end-repaired products

1. Proceed with AMPure PB Bead Purification (Section 3.5.2.1), with 1X of AMPure Beads and elute with 61μL of nuclease-free water
   - Incubate on a Hula mixer (rotator mixer) for 5 minutes at room temperature, rather than shaking in a VWR vortex mixer at 2000rpm for 10 minutes at room temperature

4.4.3.3 Prepare Ligation Reaction

1. Prepare the following reagents:
   - Spin down Adapter Mix (AMX) and T4 Ligase from the NEBNext Quick Ligation Module, and place on ice.
   - Thaw Ligation Buffer (LNB) at room temperature, spin down and mix by pipetting. Due to viscosity, vortexing this buffer is ineffective. Place on ice immediately after thawing and mixing.
   - Thaw Elution Buffer (EB) and S Fragment Buffer (SFB) at room temperature, mix by vortexing, spin down and place on ice.
2. Prepare PCR reaction mix in a 1.5 ml Eppendorf DNA LoBind tube
3. Mix gently by flicking the tube, and spin down
4. Incubate the reaction for 10 minutes at room temperature (up to 4hrs)

4.4.3.4 Bead Purification of ligated cDNA

1. Prepare the AMPure beads for use by allowing to equilibrate to room temperature for a minimum of 15minutes. Resuspend by vortexing.
2. Add 40μl of resuspended AMPure XP beads to the reaction and mix the bead/DNA solution thoroughly.
3. Incubate on a Hula mixer (rotator mixer) for 5 minutes at room temperature.
4. Spin down both tubes (for 1 second) to collect beads
5. Place tubes in a magnetic bead rack, and wait until the beads collect to the side of the tubes and the solution appears clear (2 minutes).
6. With the tubes still on the magnetic bead rack, slowly pipette off cleared supernatant

and save in other tubes. Avoid disturbing the bead pellet.

7. With the tubes still on the magnetic bead rack, wash the beads by adding either 250μl S Fragment Buffer (SFB). Flick the beads to resuspend, then return the tube to magnetic rack and allow the beads to pellet. Remove the supernatant using a pipette and discard.
8. Repeat the previous step.
9. Remove residual supernatant by taking tubes from magnetic bead rack and spin to pellet beads. Place the tubes back on magnetic bead rack and pipette off any remaining supernatant
10. Remove tubes from magnetic bead rack and allow beads to air-dry (with tube caps open) for 30 seconds
11. Elute with 15μl Elution Buffer (EB). Tap tubes until beads are uniformly resuspended. Do not pipette to mix
12. Elute DNA by letting the mix stand at room temperature for 10 minutes
13. Spin the tube down to pellet beads, then place the tube back on the magnetic bead rack. Let beads separate fully and transfer supernatant to a new 1.5ml Lo-Bind tube. Avoid disturbing beads.
14. Quantify DNA amount and concentration of Fraction 1 and Fraction 2 using a high-sensitivity Qubit (Section X). Determine library size using the BioAnalyzer with DNA 12000 Kit (Section 2.4)

4.4.4 Priming the Flow Cell

1. Prepare the following reagents:
   - Thaw the Sequencing Buffer (SQB), Loading Beads (LB), Flush Tether (FLT) and one tube of Flush Buffer (FLB) at room temperature before placing the tubes on ice as soon as thawing is complete.
   - Mix the Sequencing Buffer (SQB) and Flush Buffer (FLB) tubes by vortexing, spin down and return to ice.
   - Spin down the Flush Tether (FLT) tube, mix by pipetting, and return to ice.
2. Open the lid of the nanopore sequencing device and slide the flow cell's priming port cover clockwise so that the priming port is visible.

3. Priming and loading the SpotON Flow Cell
    - Take care to avoid introducing any air during pipetting
    - Care must be taken when drawing back buffer from the flow cell. The array of pores must be covered by buffer at all times. Removing more than 20-30μl risks damaging the pores in the array.
4. After opening the priming port, check for small bubble under the cover. Draw back a small volume to remove any bubble (a few μls):
    - Set a P1000 pipette to 200 μl
    - Insert the tip into the priming port
    - Turn the wheel until the dial shows 220-230μl, or until you can see a small volume of buffer entering the pipette tip
    - Visually check that there is continuous buffer from the priming port across the sensor array.
5. Prepare the flow cell priming mix: add 30μl of thawed and mixed Flush Tether (FLT) directly to the tube of thawed and mixed Flush Buffer (FLB), and mix by pipetting up and down.
6. Load 800μl of the priming mix into the flow cell via the priming port, avoiding the introduction of air bubbles. Wait for 5 minutes.
7. Thoroughly mix the contents of the LB tube by pipetting. The Loading Beads (LB) tube contains a suspension of beads. These beads settle very quickly. It is vital that they are mixed immediately before use.

4.4.5 Library loading into the Flow Cell

1. Prepare sample with for library as in Table
2. Gently lift the SpotON sample port cover to make the SpotON sample port accessible.
3. Load 200 μl of the priming mix into the flow cell via the priming port (not the SpotON sample port), avoiding the introduction of air bubbles.
4. Mix the prepared library gently by pipetting up and down just prior to loading.
5. Add 75 μl of sample to the flow cell via the SpotON sample port in a dropwise fashion. Ensure each drop flows into the port before adding the next.

6. Gently replace the SpotON sample port cover, making sure the bung enters the SpotON port, close the priming port and replace the MinION lid.

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

Table 4.2: Loading Flow Cells

Chapter 5

Whole Transcriptome

5.1 Introduction

5.1.1 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.

5.1.2 Mouse model of AD tauopathy: rTg4510

Unlike with APP, there are currently no known mutations in MAPT linked to AD. Mouse models, such as rTG4510, that recapitulate AD tauopathy are therefore developed through harbouring missense mutations in MAPT that are associated with tauopathy in familial frontotemporal dementia (FTD). In the case with rTg4510, the human tau transgene carrying the P301L mutation is over-expressed under the calcium calmodulin kinase II promotor (CaMK2a) and is largely restricted to the forebrain (such as hippocampus and cortex). These mice also exhibit cognitive and behavioural impairments, with neurofibrillary tangles developing as early as 2 months, and associated neuronal and synaptic loss evident by 9 months. Starting from the neocortex and progressing rapidly to the hippocampus, the age-dependent spread of neuropathology in rTG4510 mouse closely reflects the spread of NFTs in human AD, as classified into Braak stages. However, it is important to note that the genomic integration of CAMK2a and MAPT transgene has been to have off-target effects with disruption in the endogenous mouse genes, including XXX. [Figure X: rTg4510 with image of why it is called regulatable due to the mouse line]

5.2 Methods

As detailed in Chapter X, Pacific Biosciences Iso-Seq dataset was generated with whole transcriptome approach using high quality RNA from i) 12 mouse cortex (WT = 6, TG = 6) (Figure 5.1). As a technological comparison and validation of the IsoSeq approach, a subset of samples were also sequenced on ONT (Figure 5.2). 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 [Figure X].

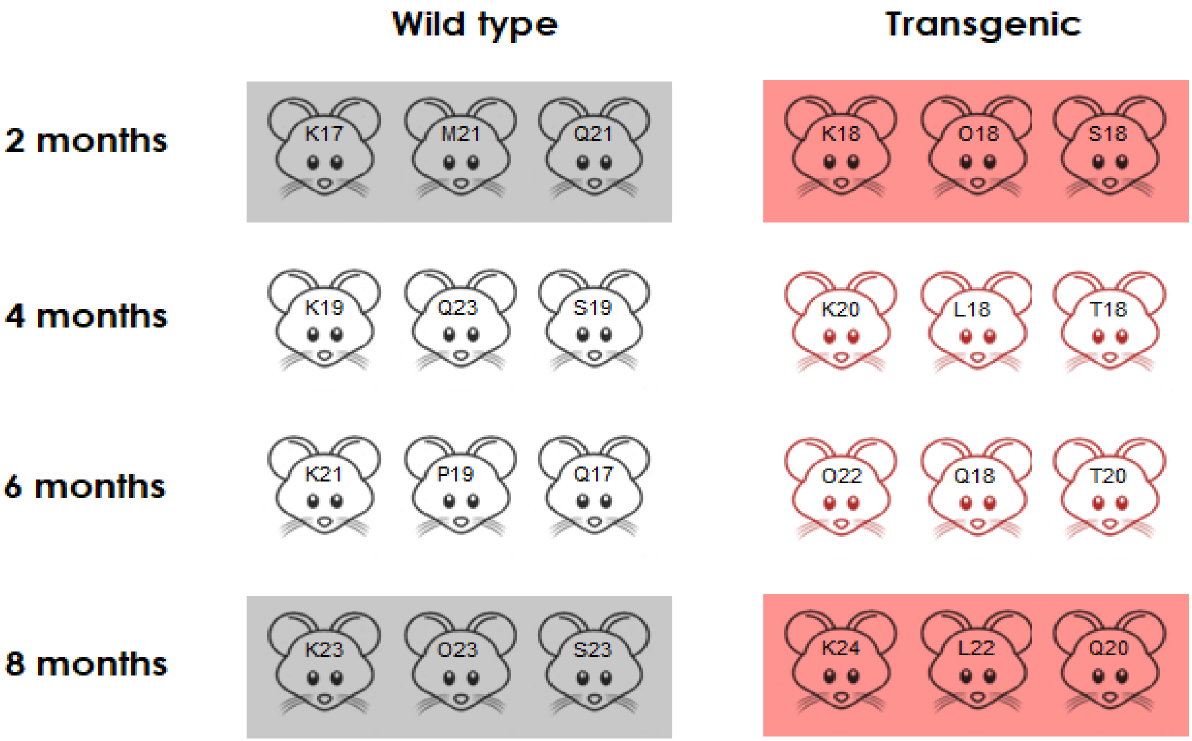


Figure 5.1: Tg4510 WT and TG samples sequenced using Whole and Targeted Iso-Seq: 12 samples at baseline and final age timepoint (WT = 6, TG = 6, ages = 2 months, 8 months) were sequenced using Whole Iso-Seq (highlighted boxes) and an additional 12 samples (WT = 6, TG = 6, ages = 4 months and 6 months) were sequenced using Targeted Iso-Seq (outlined). Text on each mouse figure refer to sample names

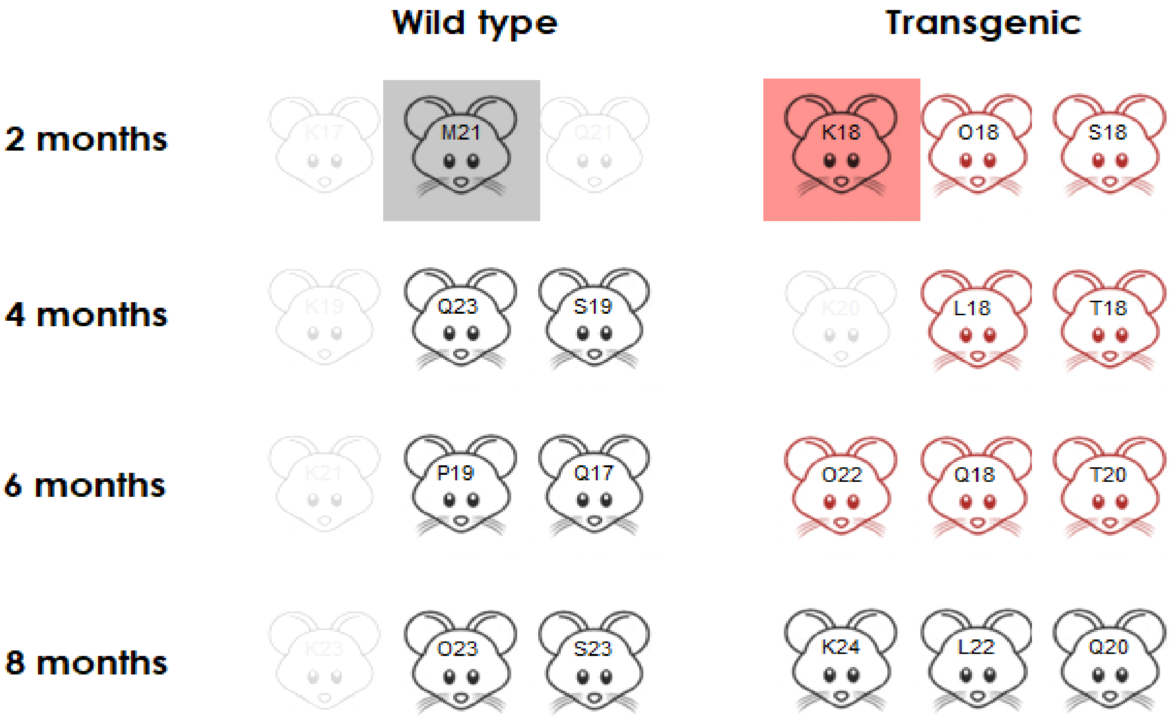


Figure 5.2: Tg4510 WT and TG samples sequenced using Whole and Targeted ONT: A subset of samples were also sequenced as whole transcriptome on ONT (WT = 1, TG = 1, age = 2 months, highlighted boxes) and targeted on ONT (WT = 7, TG = 11, outlined). Text on each mouse figure refer to sample names

5.3 Results

The mouse transcriptome of 12 pooled samples (WT and TG) was sequenced and analysed with the PacBio Sequel 1 platform for deep characterisation of full-length splice variants and identification of novel transcripts.

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 5.1). 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 5.4), 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 5.3). 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 5.1: Phenotypic information and Iso-seq run yield for each sample of Tg4510 sequenced using Whole Transcriptome approach

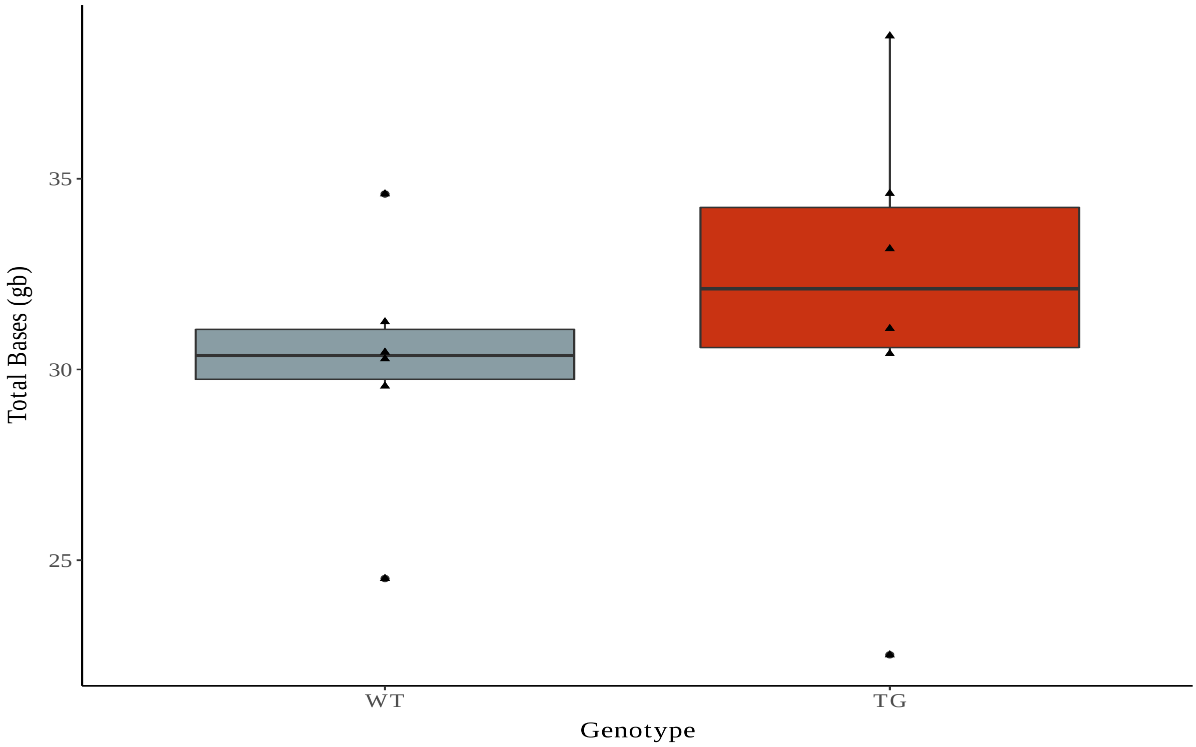


Figure 5.3: Whole Transcriptome Iso-Seq run output in transgenic and wild type Tg4510 mouse model: No significant difference in run output was observed between WT and TG Tg4510 mice. Of note, two samples with <25Gb in WT and TG refer to earlier samples sequenced with a lower chemistry (Sequencing Primer v3; Sequel Binding Kit 2.1) and diffusion loading

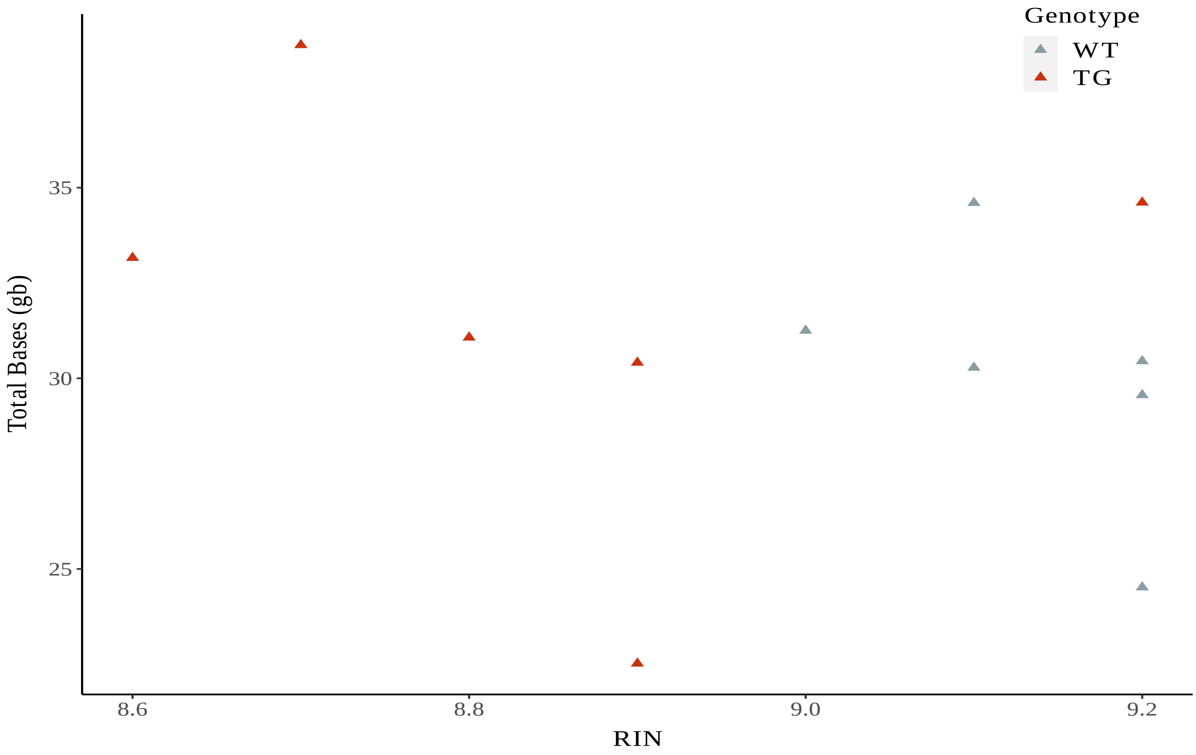


Figure 5.4: No significant correlation between RIN and Whole Transcriptome Iso-Seq run output: Samples with RIN >8 were selected for Whole Transcriptome Iso-Seq, with TG samples having distinctly lower RIN values than WT samples. However, no significant difference was observed for run output between WT and TG (Figure 5.1)

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

5.3.1 Bioinformatics output

Following the established bioinformatics pipeline [Figure X], a total of XXX successful CCS reads were generated (n = 12 samples, s.d = , range = XX - XX) and a total of XXX FLNC reads obtained, post trimming of barcodes with LIMA and clustering of transcripts with Iso-Seq3. As described in Section X, the FLNC reads from the same isoform were then clustered to generate a total XXX high-quality transcripts (n = 12 samples, s.d = , range = XX - XX) with accuracy >99% and XX low-quality transcripts (n = 12 samples, s.d = , range = XX - XX). XX isoforms (XX%) were longer than 500 bp, and XXX transcripts (XX%) were longer than 1-kb. No difference was observed in number of FLNC reads and number of transcripts between WT and TG.

The clustered transcripts have an average length of XXX nt, with the longest being XXX nt. "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."

Further stringent clustering and SQANTI filtering, as detailed in Section X, produced an average XX of unique transcripts and XX of genes with no differences identified between WT and TG in both rTg4510 (n = 12, two-tailed unpaired t-test) and J20 (n = 4, two-tailed unpaired t-test). Supplementary Table X records the number of transcripts prior and post SQANTI filtering.

5.3.2 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 are 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)

1. Unmapped reads with no signficant mapping to the genome:

2. Multiple mapped reads showing multiple alignment
3. Reads mapped to positive strand of the genome
4. Reads mapped to negative strand of the genome

5.3.3 Transcriptome annotation

Post SQANTI filtering and removal of mono-exons, an average XX of unique transcripts were identified, with X% coding and x% non-coding. Corresponding with average transcript size in GENCODE and with bioanalyzer traces, the average length was XXX (range: XX, sd XX) with no difference reported between WT and TG in Tg4510 and J20 mouse models. A wide range in the number of isoforms was identified per gene (XX – XX), with majority of genes characterised by more than >1 detectable isoform, and a notable proportion of genes characterised by >10 isoforms – difference between WT and TG? [Saturation Curve]

Mapped isoforms were divided into:

1. known isoforms from known genes
2. novel isoforms from known genes
3. novel isoforms from novel genes

What percentage of FLNCs mapped?

5.3.4 Exploring the PacBio transcriptome of the mouse reference genome

5.3.4.1 Protein Coding and non coding RNA genes and transcripts

XX of protein coding transcripts from XX genes and XX non-coding transcripts from XX genes. Within the non-coding RNAs (ncRNAs), XX transcripts were longer than 200bp as classified as long noncoding RNAs. - Is there a difference in the number of exons between coding and non-coding transcripts i.e. single exon transcripts making up majority of non-coding RNAs. In Kuo et al. (2017), and Chen et al. (2017), observed a relatively higher proportion of mono-exon transcripts among non-coding RNAs than in human - likewse in mouse? rationale due to low sequencing depth and not able to detect many lowly expressed multiple exon RNA

5.3.4.2 Nonsense mediated decay products

5.3.4.3 Transcriptional Complexity

Ratio of transcripts to genes? Assessment of alternative TSS, remove all genes with only one representative transcript as default only one TSS - XX (XX%) had multiple TSS, high incidence likely due to a combination of library preparation error (resulting in wobble) and biological transcription start exons. Filtering transcripts with TSS caused by wobble, XX have multiple starting exons.

Assessment of alternative TTS, similarly remove all genes with only one representative transcript.

Assessment of retained introns and skipped exons in multi-transcript genes. Are there any alternative splicing differences between protein coding and lncRNA genes.

Comparison of the PacBio transcriptome with public annotation -

5.3.5 Transcriptome abundance hierarchal between samples

Annotated genes Average XX of annotated genes (XX out of XX; range: XX, sd: XX) with an average length of XX was identified. No difference in number of annotated genes or distribution of isoforms in annotated genes was identified between Tg4510 and J20 mouse. Average XX full splice match isoforms were identified with no difference between Tg4510 and J20 mouse [Table X for breakdown]

Different splicing events annotated by SUPPA2. Reported differences in terms of numbers between types of splicing events in WT vs TG in Tg4510 and J20?

XX of known transcripts were identified to have intron retention; XX of known transcripts were identified to be fusion genes. XX of know transcripts identified to have non-sense-mediated decay.

5.3.6 Iso-Seq vs RNA-Seq

- Correlation of IsoSeq TPM expression and RNASeq TPM expression - What is the threshold of gene expression observed in Iso-Seq data vs RNA-Seq data (genes that are observed in RNA-Seq but not in Iso-Seq) Isoform quantification using RNASeq - Comparison of PacBio isoforms with short-read assembly from Gordon et al. (2015) Wang et al. (2016) by using short-read assembler such as Cufflinks and de-novo for construction of short reads, and considering locus that are only in evident in PacBio dataset and short-reads, less than 20% of isoforms from PacBio dataset recapitulated

Unannotated novel genes with novel transcripts While majority of transcripts were annotated to known genes, X% [range: XX, sd: XX] of transcripts were mapped to unannotated "novel" genes. These novel transcripts with mean length XXX bps (range: XX, std: XX) were identified uniformly across the genome/chromosome, with XX% coding and XX% non-coding. Majority (XX%) of these novel genes had only 1 transcript, however, several novel genes were identified with 3 or more novel transcripts, highlighting the validity of these novel genes. The validity of these novel genes is further supported by other resources. Using publicly available FANTOM5 Cap Analysis of Gene Expression (CAGE) that maps transcript, transcription factors, transcription promoters and enhancers (ref), majority of these novel genes (XX out of XX) in the mouse samples lie within 5kB of a cage peak. Junctions/Exons of these novel genes are also supported by aligned short-read RNA-Seq reads (XX out of XX). [Novel transcript abundance] No reported differences in number of unannotated, novel transcripts were identified between Tg4510 and J20 mouse [XXX].

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6885035/ - Lorna's paper on example of isoforms in their selective panel of genes that have differential quantification associated with AD. Good examples to check with transcriptome data to see if this is also observed in mouse (I.e. TAU3 increase)

5.3.7 Change in endogeneous expression

Supporting findings from Castanho et al. (2020), human-specific MAPT sequences were only detected in CCS and FL reads from TG mice, confirming stable insertion and activation of MAPT transgene in TG mice.

5.3.8 LncRNA

lncRNA expression has been known to be more lowly expressed compared to protein-coding genes (Derrien et al. (2012)), and has been previously shown in other long-read studies (Tilgner et al. (2015))

5.3.9 Isoform diversity

Generally, higher gene expression, more isoforms observed (with linear increase); (Karlsson and Linnarsson, 2017)

5.3.10 Splicing Events

Single cell analysis (Karlsson and Linnarsson (2017)) noted that alternative TSS and TTS variation in the first exon representing more than 70% of splicing events, with only around 30% of 5' end transcripts located near annotated 5' (partially due to ongoing mRNA degradation, presence of unannotated alternative TSS, or strand invasion during reverse transcription resulting in template switching artefacts) and only around 70% of 3'ends located at the annotated 3'ends (attributed to alternative polyadenylation sites in 3'UTR).

5.3.11 Validation of novel isoforms

- Profile of H3K4me3 enrichment, a chromatin signature of active promoters (Qiao et al., 2020)

Chapter 6

Targeted Transcriptome

Chapter 7

ONT vs PacBio

7.0.1 Comparisons of PacBio vs ONT for RNA-Sequencing

- Comparing Read lengths - Mappability - Chimeric and gapped alignments - Error patterns - Isoform identification - Isoform abundance estimation

Sequencing quality (fraction of reads aligned) on read lengths for single pass reads (subreads for PacBio) and multi-pass consensus reads (CCS for PacBio and 2D reads for ONT) Fraction of Read aligned in bins

Context specific errors

Pacbio non-size selection and Oxford Nanopore non-size selection

Lowly expressed gene and minor isoform quantification

Chapter 8

Proteomics (Eli-Lilly)

Chapter 9

Conclusion

Bibliography

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. ISSN 14712105. doi: 10.1186/s12859-019-3247-x. URL https://doi.org/10.1186/s12859-019-3247-x.

Shanika L. Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis, feb 2020. ISSN 1474760X. URL https://doi.org/10.1186/s13059-020-1935-5.

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. ISSN 19326203. doi: 10.1371/journal.pone.0046679. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3464235/pdf/pone.0046679.pdf.

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. ISSN 15537404. doi: 10.1371/journal.pgen.1005610. URL https://dx.plos.org/10.1371/journal.pgen.1005610.

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. ISSN 19326203. doi: 10.1371/journal.pone.0157779. URL http://dx.plos.org/10.1371/journal.pone.0157779.

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. ISSN 22111247. doi: 10.1016/j.celrep.2020.01.063. URL `https://doi.org/10.1016/j.celrep.2020.01.063`.

Shi-Yi Chen, Feilong Deng, Xianbo Jia, Cao Li, and Song-Jia Lai. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. Scientific Reports, 7(1):7648, 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-08138-z.

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. ISSN 1946-6242. doi: 10.1101/074153. URL `http://stm.sciencemag.org/`.

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. ISSN 10889051. doi: 10.1101/gr.132159.111.

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. ISSN 00280836. doi: 10.1038/nature11233. URL `http://genome.crg.cat/encode{_}RNA{_}dashboard/`.

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. ISSN 14764687. doi: 10.1038/nature11247. URL http://encodeproject.org/ENCODE/.

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. ISSN 14747596. doi: 10.1186/gb-2007-8-6-r108. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-6-r108.

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. ISSN 15487105. doi: 10.1038/nmeth.4577. URL https://www.nature.com/articles/nmeth.4577.

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. ISSN 19326203. doi: 10.1371/journal.pone.0132628. URL https://www.ncbi.

nlm.nih.gov/pmc/articles/PMC4503453/pdf/pone.0132628.pdf.

Kasper Karlsson and Sten Linnarsson. Single-cell mRNA isoform diversity in the mouse brain. BMC Genomics, 18(1), 2017. ISSN 14712164. doi: 10.1186/s12864-017-3528-6. URL https://bmcgenomics.biomedcentral.com/track/pdf/10.1186/s12864-017-3528-6?site=bmcgenomics.biomedcentral.com.

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. ISSN 20411723. doi: 10.1038/ncomms15824. URL www.nature.com/naturecommunications.

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. ISSN 1471-2164. doi: 10.1186/s12864-017-3691-9. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3691-9.

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. ISSN 00368075. doi: 10.1126/science.1079700. URL http://www.ncbi.nlm.nih.gov/pubmed/12560545http://www.sciencemag.org/cgi/doi/10.1126/science.1079700.

Alice Mccarthy. Third generation DNA sequencing: Pacific biosciences' single molecule real time technology, jul 2010. ISSN 10745521.

Yunbo Qiao, Chao Ren, Shisheng Huang, Jie Yuan, Xingchen Liu, Jiao Fan, Jianxiang Lin, Susu Wu, Qiuzhen Chen, Xiaochen Bo, Xiangyang Li, Xingxu Huang, Zhen Liu, and Wenjie Shu. High-resolution annotation of the mouse preimplantation embryo transcriptome using long-read sequencing. Nature Communications, 11(1), 2020.

ISSN 20411723. doi: 10.1038/s41467-020-16444-w. URL https://doi.org/10.1038/s41467-020-16444-w.

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. ISSN 15461696. doi: 10.1038/nbt.2282.

Leena Salmela and Eric Rivals. LoRDEC: Accurate and efficient long read error correction. Bioinformatics, 30(24):3506–3514, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btu538. URL http://gatb-core.gforge.inria.fr.

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. ISSN 14602059. doi: 10.1093/bioinformatics/btw321. URL http://www.cs.helsinki.fi/u/lmsalmel/LoRMA/.

Hiruna Samarakoon, Sanoj Punchihewa, Anjana Senanayake, Jillian M. Hammond, Igor Stevanovski, James M. Ferguson, Roshan Ragel, Hasindu Gamaarachchi, and Ira W. Deveson. Genopo: a nanopore sequencing analysis toolkit for portable Android devices. Communications Biology, 3(1):1–5, dec 2020. ISSN 23993642. doi: 10.1038/s42003-020-01270-z. URL https://www.nature.com/articles/s42003-020-01270-z.

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. ISSN 1087-0156. doi: 10.1038/nbt.2705. URL http://www.nature.com/doifinder/10.1038/nbt.2705.

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. ISSN 15461696. doi: 10.1038/nbt.3242. URL http://www.nature.com/doifinder/10.1038/nbt.3242.

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. ISSN 15495469. doi: 10.1101/gr.230516.117. URL http://www.ncbi.nlm.nih.gov/pubmed/29196558http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5793787.

Jan Verheijen and Kristel Sleegers. Understanding Alzheimer Disease at the Interface between Genetics and Transcriptomics, 2018. ISSN 13624555. URL https://www.cell.com/trends/genetics/pdf/S0168-9525(18)30042-8.pdf.

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. ISSN 2041-1723. doi: 10.1038/ncomms11708. URL https://www.nature.com/articles/ncomms11708.pdfhttp://www.nature.com/doifinder/10.1038/ncomms11708.

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. ISSN 00280836. doi: 10.1038/nature07509. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2593745/pdf/nihms-72491.pdf.

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. ISSN 2046-1402. doi: 10.12688/f1000research.10571.1. URL https://f1000researchdata.s3.amazonaws.com/manuscripts/11392/aeb5b027-2cdd-4888-8967-bc4812e075eb{_}10571{_}-{_}kin{_}fai{_}au.pdf?doi=10.12688/f1000research.10571.1https://f1000research.com/articles/6-100/v1.

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. ISSN 15487105. doi: 10.1038/s41592-019-0617-2. URL `https://doi.org/10.1038/s41592-019-0617-2`.

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. ISSN 16648021.